SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Robotics, Cognition, Intelligence

Neural Radiance Fields for Ultrasound Imaging

Magdalena Wysocki

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Robotics, Cognition, Intelligence

Neural Radiance Fields for Ultrasound Imaging

Neuronale Strahlungsfelder für die Ultraschallbildgebung

Author:	Magdalena Wysocki
Supervisor:	Nassir Navab, Prof. Dr.
Advisors:	Mohammad Farid Azampour, M.Sc.
	Mehrdad Salehi, M.Sc.
	Benjamin Busam, Dr.
Submission Date:	15.08.2023

I confirm that this master's thesis in robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Munich, 15.08.2023 Magdalena Wysocki

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Mohammad Farid Azampour, for his inexhaustible guidance, valuable feedback, and constant encouragement throughout this thesis. His expertise and support were instrumental in shaping this work. Thanks to his mentorship, I have expanded my intellectual horizons. I am also thankful to Mehrdad Salehi, Benjamin Busam, Christine Eliers, Oliver Zettinig, and Walter Simson for their contributions and fruitful discussions that enriched the research process.

Furthermore, I extend my heartfelt gratitude to Prof. Nassir Navab and the Interdisciplinary Research Laboratory (IFL) members at Klinikum rechts der Isar for creating an environment that fosters collaboration and research excellence. I cannot express my gratitude enough for the support I received from everyone along this journey.

My appreciation also goes to my friend, Matan Atad, with whom I have enjoyed sharing the experience since our first day of Master's studies. I am truly grateful for your presence and camaraderie throughout this fulfilling adventure.

I am immensely grateful to my family for making it all possible. Especially to my sister, Edyta Zugaj, for always being my role model. Your belief in me kept me motivated.

Finally, my heartfelt gratitude goes to Olaf Wysocki, who provided academic guidance, tremendous editorial assistance, and the emotional comfort essential to complete this significant milestone. Your company and encouragement to pursue my passions allowed me to thrive. Your love has made all the difference.

Abstract

Ultrasound imaging plays a critical role in clinical practice as a valuable tool for visualizing patient anatomy. However, conventional 2D ultrasound scans present a significant limitation by lacking spatial context, which hinders the accurate identification and localization of structures within the scanned region. To improve spatial understanding, a popular solution for 3D reconstruction and visualization of anatomy is freehand 3D ultrasound. Yet, existing methods for 3D ultrasound reconstruction commonly average over multiple observations for a region of interest, leading to the loss of crucial directional information essential for understanding acoustic phenomena. After the volume is reconstructed, it becomes infeasible to render a B-mode image that represents view-dependent information in a way that simulates how an ultrasound system captures a frame.

This thesis addresses the challenge of view-dependent rendering and presents a new approach called neural radiance fields for ultrasound imaging (Ultra-NeRF). Through the application of deep learning techniques and incorporating a ray-based rendering method for ultrasound, a novel 3D ultrasound representation is introduced. This representation allows for the regression of anisotropic acoustic tissue properties, enabling the view-dependent rendering of B-mode images. However, estimating rendering parameters is a highly unconstrained task leading to ambiguity in representing a B-mode image in the parameter space. This thesis proposes regularization based on the relationship between tissue properties, which constrains the regression space of rendering parameters, enhancing the accuracy and reliability of the regressed parameters. The proposed method's ability to render view-dependent B-mode images is validated through experimental evaluations. These evaluations utilize simulated B-mode images of a liver and real B-mode images of a synthetic and exvivo phantom of a spine. It is experimentally demonstrated, that the method excels in rendering regions with ambiguous representation caused by view-dependent differences in B-mode images of the same region of interest.

To the best of my knowledge, this work is the first to tackle the challenge of view-dependent ultrasound image synthesis using neural radiance fields. By accurately preserving view-directional information, the approach offers several advantages, including the examination of missed cross-sectional planes, flexibility in scanning protocols, and potential automation of scanning protocols and image analysis.

Contents

1. Introduction

1.1. Motivation

Imaging of patient anatomy is an indispensable aspect of clinical practice, and ultrasound imaging plays a crucial role in providing valuable anatomical information invisible to a naked eye. Ultrasound has several advantages over other imaging modalities such as computed tomography (CT) or magnetic resonance imaging (MRI): (1) It does not use ionizing radiation which makes it a safer option for repeated or prolonged use for clinicians and patients alike; (2) It is more cost-effective compared to these other imaging modalities; (3) Ultrasound devices are typically more portable, making them more accessible in remote areas or in situations where mobility is a concern.

Proficiency in anatomy, physiology, and pathology is essential for operators to produce accurate and reliable ultrasound images, and acquiring such expertise demands considerable effort. The extensive training requirements for clinicians may lead to increased costs, potentially limiting access to ultrasound in regions with limited availability of skilled medical professionals. Despite the significant research focus on automatic scanning protocols in clinical practice, such as those utilizing robotic ultrasound acquisition [14, 20, 21, 22], the majority of ultrasound acquisitions are performed manually by clinicians using handheld probes. Consequently, the accuracy and reliability of ultrasound imaging heavily rely on the operator's expertise.

A fundamental challenge of ultrasound examination is that two-dimensional (2D) ultrasound scans lack spatial context presenting only a 2D cross-sectional view of the region of interest (ROI). This limited perspective makes it challenging for operators to accurately identify and locate structures within the ROI. Operators of ultrasound face two primary challenges in this regard. Firstly, they must rely on hand-eye coordination to navigate through the three-dimensional (3D) space using only 2D observations. This is a challenging task, as the limited perspective of 2D scans can make it difficult to accurately identify and locate structures within the ROI. Secondly, operators must mentally reconstruct the 3D volume of the scanned area to determine the location of a 2D plane within the structure, which requires a high degree of spatial visualization ability and cognitive effort, and is particularly challenging for less experienced operators.

3D ultrasound technology constitutes an effort to address and alleviate the previously mentioned challenges that arise during the application of conventional 2D ultrasound.

To enhance the clinician's comprehension of the spatial relationships between anatomical structures, 3D ultrasound provides an intuitive visualization of the patient's internal anatomy. This improvement in spatial understanding increases diagnostic accuracy and enhances the precision of imaging guidance. 3D ultrasound can be obtained using different acquisition techniques which could be divided into two categories: specialized probes directly providing volumetric image, such as 2D array transducer systems, 3D mechanical systems, and conventionally used freehand scanners along with 3D reconstruction techniques [17]. While dedicated hardware-based 3D acquisition methods provide a real-time 3D information and allow direct flow of visualization, the probes are expensive and the acquired ROI at each time step is limited to the array size. The adoption of 3D ultrasound technology addresses and mitigates the aforementioned challenges encountered when using conventional 2D ultrasound.

A typical approach to achieving freehand 3D ultrasound is commonly divided into two steps: data acquisition and 3D reconstruction. Data acquisition involves the simultaneous capture of both 2D frames and their respective positions by using specialized tracking technology [5]. This data is subsequently utilized to construct a 3D volume from multiple 2D slices. In recent years, several methods have been developed to improve the quality of 3D reconstruction by refining relative position of images. These methods involve the implementation of sensorless 3D ultrasound [35] and the use of advanced image formation techniques based on deep learning [42].

Reconstruction technique, commonly referred to as *volume compounding*, has become an established approach to generating high-quality 3D volumes of 3D ultrasound. A common framework involves processing multiple 2D frames to create a 3D description of the imaged anatomy and relies on a accurate fusion of multiple views of the same ROI [52]. However, ultrasound images acquired from different probe orientations exhibit varying tissue features due to speckles and artifacts. For example, acoustic shadows are heavily influenced by the probe orientation, with each view providing only a partial representation of the anatomy relative to the probe's orientation. When multiple observations of the same ROI are fused, an interpolation is performed to reconcile potentially conflicting information visible in ultrasound. Although this approach can produce a high-quality 3D representation of the imaged anatomy, it is infeasible to reconstruct a view-dependent ultrasound image from such a representation. This is because the interpolation process effectively blends information from multiple views, resulting in a loss of the unique details that are visible in each individual image.

Recently, computational sonography has introduced a novel 3D representation that preserves both the anatomical details of the imaged ROI and its directional information [13]. This paradigm offers a novel approach to retrieving a scalar intensity field dependent on the viewing direction and has been effectively applied to the direction-preserving 3D reconstruction [11]. Such a technique preserves direction-dependent

intensities, allowing for the recovery of a detailed ultrasound image from any viewing angle. By preserving this information, this approach offers a more comprehensive representation of the underlying anatomy compared to traditional 3D reconstruction methods.

Preserving directional information represents a key advantage, as it allows for the retrieval of any 2D cross-sectional plane, as if it had been directly acquired by an ultrasound system. Therefore, a view-dependent 3D representation permits the examination of cross-sectional planes that may have been missed during the initial acquisition phase, facilitates the implementation of more flexible scanning protocols, and has the potential to enable automated scanning protocols and image analysis.

1.2. Research Scope

The primary goal of this thesis is to introduce a novel 3D ultrasound representation that effectively retains view-directional information of the imaged anatomy. In pursuit of this objective, the study presented in this thesis addresses the following research questions:

- Can applying deep learning techniques facilitate the acquisition of a 3D ultrasound representation that accurately preserves viewing-direction-dependent intensities?
- What advantages can be derived from using a 3D ultrasound representation that accurately preserves viewing-direction-dependent information, compared to traditional techniques?
- To what extent do the regressed tissue characteristics accurately represent actual tissue properties?

1.3. Approach

In this thesis introducea a new paradigm of 3D ultrasound representation called neural radiance fields for ultrasound imaging (Ultra-NeRF). The proposed approach entails learning a scalar field over a volume to accurately represent the imaged anatomy derived from a set of 2D ultrasound frames while preserving directional information. This type of representation draws inspiration from computational sonography and the concept of *implicit neural representation*, as described by Sitzmann et al. [45]. To accomplish this objective, the method leverage a recent development in computer vision known as *neural radiance fields (NeRF)* [31], which divides the task into two distinct

components. The first component involves a network representing a 3D volume by a set of parameters. The second component entails a volumetric rendering function that maps regressed parameters to radiance. We propose an ultrasound-specific, ray-casting-based volume rendering approach that effectively accommodates anisotropic tissue properties regressed from a set of 2D B-modes and maps them to corresponding view-dependent intensities.

1.4. Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 presents essential concepts of medical ultrasound imaging, starting with the underlying ray-physics of ultrasound. Then, it introduces the characteristics of medical ultrasound and the imaging artifacts that can pose significant challenges for 3D ultrasound representation. Subsequently, it discusses volume compounding, a common approach to representing 3D ultrasound images. Lastly, it illustrates the concept of confidence maps, which provide valuable insights into the quality and reliability of ultrasound images.

Chapter 3, provides an in-depth analysis of NeRF, including its fundamental concepts. It introduces various 3D representations to demonstrate the difference between implicit neural representation, the cornerstone of NeRF, and other methods. Next, it explores the concept of ray tracing and its application in differentiable volume rendering, a crucial component of NeRF. Then, it discusses using NeRF in computer vision. It concludes by presenting the application of NeRF in ultrasound imaging.

Chapter 4 offers an overview of state-of the art of implict representation and neural radiance fields in medical imaging.

Chapter 5 presents the approach to addressing the challenges in view-dependent 3D ultrasound reconstruction. Before introducing the method, it discusses the setup and assumptions to provide context for the final approach. Next, it introduces Ultra-NeRF, a modified version of NeRF that incorporates specific features tailored for medical ultrasound imaging.

Chapter 6 presents validation of the proposed method by showing the results of experiments contacted on various tasks. First, it introduces the datasets used in the experiments and the evaluation methods employed, including qualitative and quantitative measures. Then, shows and discusses the results of experiments, including B-mode rendering, volume compounding, confidence maps, and evaluation of rendering parameters. Finally, it presents the limitations of the method.

Chapter 7 provides an outlook for feature research directions.

Chapter 8 concludes the thesis by summarizing the key findings and contributions.

2. Principles of Medical Ultrasound Imaging

A comprehensive understanding of the principles of ultrasound imaging and its underlying physics is crucial for successfully implementing a 3D ultrasound reconstruction method. This chapter introduces important aspects of medical ultrasound imaging and highlights the physical processes involved in the formation of ultrasound images. The primary sources for this chapter include the seminal book on diagnostic ultrasound imaging by Szabo [46], which provides a clear overview of the fundamental principles of medical ultrasound, as well as two important references by Allisy-Roberts and Williams [2], and Hoskins et al. [16] that offer in-depth knowledge of medical ultrasound physics.

2.1. Medical Ultrasound

Medical ultrasound imaging is a non-invasive technique that utilizes high-frequency sound waves to generate cross-sectional images of internal structures within the body. A conventional ultrasound system typically produces images in a brightness mode (B-mode), where the brightness at each point corresponds to the amplitude of the echo generated by the reflection of the ultrasound wave. The most prominent reflections attribute to tissue interfaces. Additionally, the texture of the image is influenced by the scattering of the sound waves, which arises from small inhomogeneities (scatterers) embedded in homogeneous tissues. Figure 2.1 illustrates a B-mode ultrasound image, which captures variations in brightness attributed to tissue interfaces and scatterers. B-mode ultrasound images are formed from multiple B-mode scanlines, each corresponding to a single ultrasound pulse transmitted from and then received by a transducer. The imaging geometry of B-mode ultrasound images varies depending on the type of transducer used. Therefore, transducers can be categorized based on the B-mode image type they produce. Two main types of transducers exist for external usage: linear and curvilinear probes. Linear probes consist of many transducer elements arranged in a straight line. Such a probe produces B-mode scan lines perpendicular to the transducer array and parallel to one another, resulting in a rectangular field. On the other hand, curvilinear probes have transducer elements arranged in a rounded array,

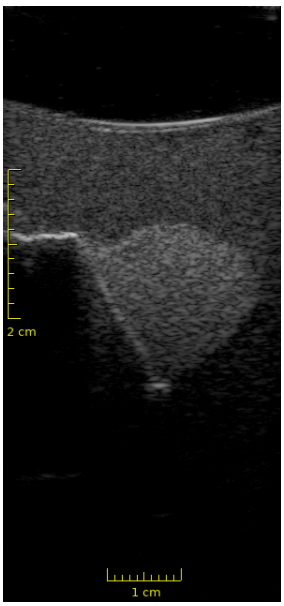

Figure 2.1.: A B-mode image of a thyroid phantom showcasing tissue interfaces appearing bright due to strong reflection of an acoustic wave. The presence of tissue inhomogeneities is depicted by a noisy pattern.

and the field of view in this type of transducer depends on the radius of the curvature. Figure 2.2 shows geometrical representations of B-mode images captured with linear and curvilinear transducer.

2.2. Ultrasound Physics

2.2.1. Basics of Ultrasound Ray Physics

Ultrasound waves are high-frequency longitudinal waves, which propagate through a physical medium, in the context of medical imaging the medium comprises different tissue types. As ultrasound waves move through the medium, they transport energy from the source into the tissue, and their propagation depends on the properties of the tissue itself. While the wave equation [19] can be used to model ultrasound waves, this thesis employs a geometrical approximation that treats the waves as rays.

The reflection of sound waves is a fundamental process in ultrasound imaging. Such reflections occur at the boundaries where there is a variation in the acoustic properties of the tissue. The primary property responsible for these reflections is the acoustic impedance (Z), which characterizes the medium's response to the wave and is a function

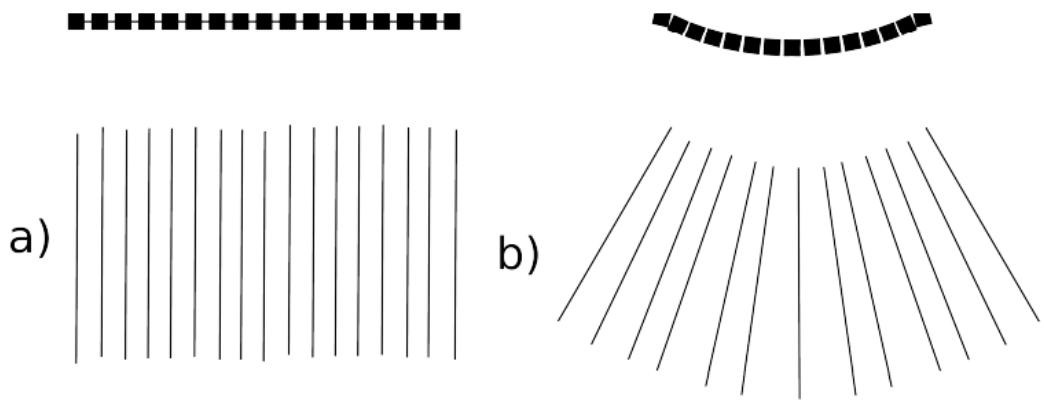

Figure 2.2.: Visualisation of the linear (a) and curvilinear (b) probe.

of the local pressure and particle velocity values in the medium. When acoustic waves encounter a boundary between two tissues with different impedance (Z_1 and Z_2), some of the energy from the sound wave reflects to its source, creating an echo that can be detected. This process also causes a loss of energy at the tissue boundary. The residual energy transmits to the subsequent medium as the wave propagates further into the tissue. The proportion of the reflected wave is represented by the intensity reflection coefficient (R), which denotes the ratio of the intensity of the reflected I_r and incident I_i waves. R is expressed in terms of acoustic impedance using the following formula:

$$R = \frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2} \tag{2.1}$$

Using the intensity reflection coefficient we can calculate the reflected intensity as follows:

$$I_r = RI_i \tag{2.2}$$

Equations 2.1 and 2.3 demonstrate that the observed intensity attributable to reflection becomes higher as the reflection coefficient increases with the difference in acoustic impedance between tissue types. As the total intensity remains constant, one can define the transmitted intensity I_t in terms of the intensity reflection coefficient as:

$$I_t = (1 - R)I_i \tag{2.3}$$

Figure 6.12 demonstrates how the incident wave behaves at the tissue boundary of tissue types admitting a different acoustic impedance.

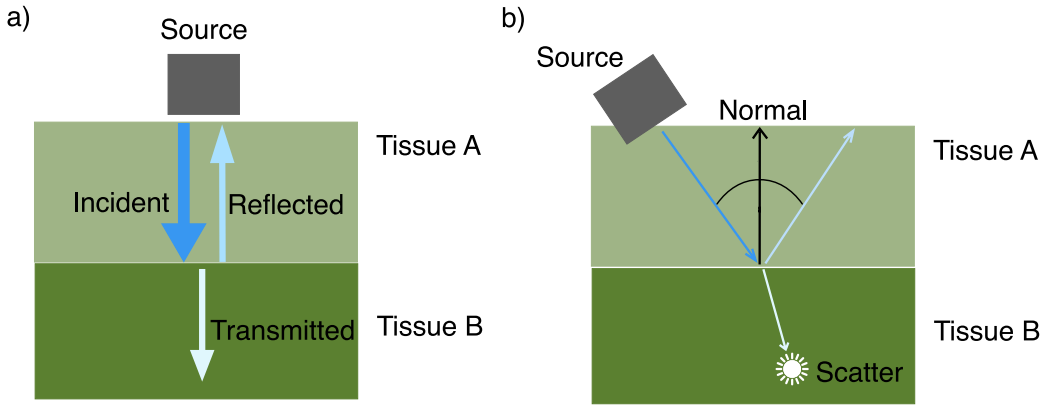

Figure 2.3.: (a) At an interface between two media a portion of the acoustic beam is reflected back towards the source while the remaining portion is transmitted further into the tissue. (b) The strength of the reflected signal captured by the source at a tissue interface depends on the incidence angle, which defines the amount of energy that is reflected back to the source, while the strength of the echo at a scattering point does not depend on the incidence angle due to diffuse scattering.

The presented model of reflection and transmission of ultrasound waves fails to consider the impact of the incidence angle on the amplitude of the reflected wave. According to the law of reflection, at a tissue interface, the strength of the reflected signal captured by the source depends on the incidence angle between the tissue interface and the acoustic beam, as shown in Figure 6.12. In order to accurately calculate the reflection coefficient, which incorporates the incidence angle, Fresnel's equations must be employed. Fresnel's equations describe the behavior of light at the boundary between two media with different refractive indices and is applicable to ultrasound waves. By taking into account the incidence angle, the reflection coefficient can be calculated more accurately as follows:

$$R = \left(\frac{Z_1 \cos\theta_i - Z_2 \cos\theta_t}{Z_1 \cos\theta_i + Z_2 \cos\theta_t} \right)^2 \tag{2.4}$$

Equation 2.1 is a special case where $\theta_i = \theta_t = 0$, which is in the case when the incident wave is perpendicular to the boundary between two media. In ultrasound imaging, Lambert's cosine law is commonly used to approximate the relation resulting from

Equation 2.4 giving the final formula for R

$$R = \cos\theta_i \cdot \frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2} \tag{2.5}$$

Note, that Equation 2.5 considers only specular reflection, while diffuse reflection is neglected.

Acoustic scattering is another source of observed acoustic echo, and it occurs due to small inhomogeneities in the tissue characteristics at a scale smaller than a wavelength. A unique echogenicity and speckle pattern characterize each tissue type. Therefore, how tissue scatter is distributed significantly impacts its appearance in the B-mode image [50]. A scattering point is characterized by a diffuse reflection, which means that the strength of the echo does not depend on the incidence angle. This is because the energy of the acoustic wave is scattered in many different directions. Therefore, its appearance does not change depending on the angle of incidence. The scattering pattern observed in B-mode images, known as speckle, can be modeled by a convolutional model of backscattering [29]. This model assumes that the scatterers within the tissue are distributed randomly, and the backscattered waves from each scatterer interfere randomly. Specifically, it can be expressed as a 2D convolution ($\otimes$) between a pulse function $P(x, y)$ realized by a 2D point spread function (PSF), and - a response function of the tissue $T(x, y)$, as shown below:

$$I(x, y) = T(x, y) \otimes P(x, y) \tag{2.6}$$

$$P(x, y) = \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right] \cos(2\pi f x)$$

In this model, a PSF is approximated with a cosine-modulated 2D Gaussian kernel, where f is the frequency of the ultrasound wave, and σ_x and σ_y modulate the shape of the Gaussian kernel along the lateral and axial directions, respectively [4]. Echoes, arising from scattered and reflected energy, are a fundamental component of ultrasound imaging. However, the intensity of the echo depends on the incident wave's strength. As an acoustic wave propagates through tissue, it experiences energy loss, mainly owing to the attenuation mechanism. Attenuation describes the reduction in ultrasound intensity with distance and is determined by the tissue's properties and the depth at which attenuation occurs. Attenuation declines as an acoustic wave propagates through tissue. In homogeneous tissue, this decline can be approximated using an exponential function, reflecting the attenuation rate as a function of depth. Furthermore, attenuation is frequency-dependent, with higher frequencies experiencing stronger attenuation. With this attenuation model, the intensity of an incident acoustic wave in a

homogeneous tissue at depth x, dependent on the wave frequency f, can be estimated using the attenuation coefficient α and initial intensity I_0 as:

$$I(x) = I_0 e^{-fx\alpha} \tag{2.7}$$

2.3. Imaging Artifacts

In medical ultrasound imaging, it is essential to distinguish between two types of imaging artifacts. The first type is contingent on the imaging system and is, therefore, considered a limitation of its performance. This includes imaging resolution and signal-to-noise ratio, key factors in determining image quality. The second type of artifact occurs due to properties inherent to the target tissue, such as strong shadows or reverberations. These artifacts may emerge when there is a significant difference in characteristics between two tissue types or when multiple reflections transpire between tissue types. This paragraph concentrates on the most prominent artifacts directly related to the propagation of the ultrasound in an inhomogeneous tissue.

2.3.1. Shadowing artifacts

Acoustic shadow is a phenomenon that can both provide informative features for diagnosis [9] and hinder various imaging tasks, such as segmentation, registration and 3D reconstruction [12, 25]. It appears at tissue interfaces with a significant difference in acoustic impedance, such as at a bone-tissue interface. Acoustic shadow is characterized by a rapid loss of brightness at the tissue interface caused by a strong reflection. At such interfaces, acoustic waves experience a high loss of energy due to reflectance, with the loss of brightness being stronger the higher the intensity reflection coefficient at the tissue interface. The reduction of wave transmission is anisotropic, meaning that the observation direction will impact the direction of the acoustic shadow, as presented in Figure 2.4. Different tissue regions may be occluded and impossible to observe depending on the observation direction due to acoustic shadow. As illustrated in Figure 2.4, the intersection of ROI from multiple observation directions will exhibit distinct appearances in acoustic shadow. This phenomenon is significant for algorithms that rely on averaging repeated observations from various viewpoints.

2.3.2. Attenuation artifacts

The phenomenon of attenuation artifacts in ultrasound imaging is similar to the shadowing artifacts, and occurs due to the presence of strongly attenuating tissues on the acoustic wave path. Such tissues absorb wave energy and cause a drop in

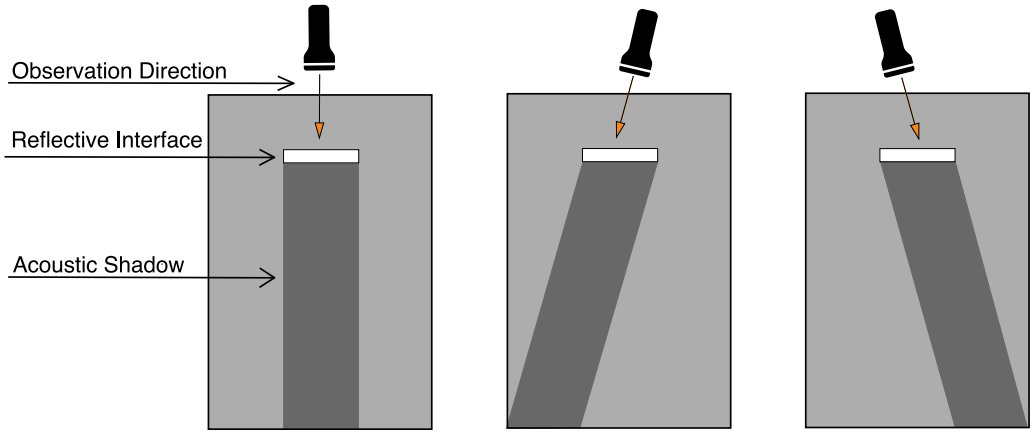

Figure 2.4.: The anisotropic nature of acoustic shadow. The extent of acoustic shadow varies depending on the observation direction, resulting in the occlusion of different tissue regions.

the transmitted energy towards the tissue, resulting in darker images of the tissues following a strongly attenuating structure compared to the image that would be obtained if the structure had been removed. As a result, the intensity of the same imaged ROI may differ when viewed from different viewpoints.

2.3.3. Reverberation artifacts

Reverberations are a common artifact observed in B-mode images during medical ultrasound imaging. These artifacts occur due to the assumption that "the ultrasound pulse travels only to targets that are on the beam axis and back to the transducer" [16]. However, the ultrasound pulse can become trapped between tissue at specific interfaces, leading to multiple reflections. The source detects these reflections after a delay and, as a result, appears in the image deeper than the tissue interface with a weaker amplitude. Reverberations are an anisotropic artifact highly dependent on the angle of incidence between the tissue interface and the acoustic wave, the number of reflections, and the strength of the signal. Consequently, even if the source of the reverberation is the same, it can appear at different ROIs.

2.3.4. Motion Artifacts

Motion artifacts are distortions in the imaged ROI in ultrasound imaging that result from the movement of either the probe or the patient. These artifacts can also arise due

to respiration or cardiac motion. The presence of motion artifacts in ultrasound imaging can lead to blurring and loss of image resolutions. Furthermore, motion artifacts can cause the ROI to appear at a different position, making it difficult to compare images taken at different times or from different viewpoints.

2.3.5. Deformation Artifacts

Deformation artifacts are a type of artifact in ultrasound imaging that results from physical deformations of a tissue. These deformations can occur due to internal or external forces applied to a tissue, such as pressure from the ultrasound probe or compression from surrounding tissues. As a result, the tissue can change its shape and position, causing the ROI to be misplaced with respect to B-mode images taken at different time steps.

2.4. Freehand 3D Ultrasound

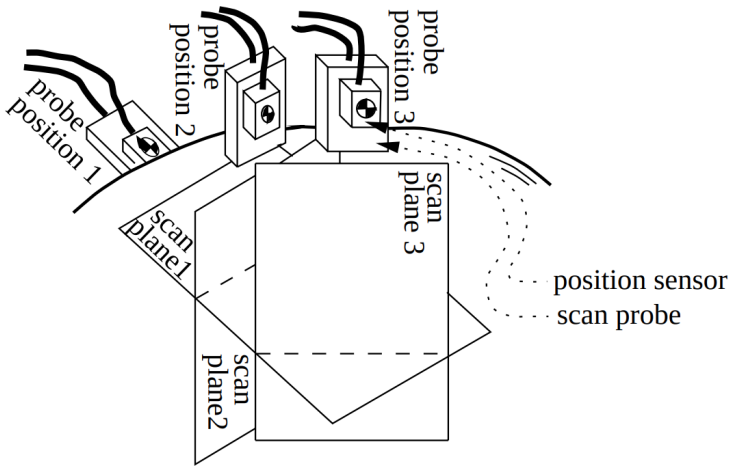

Figure 2.5.: Freehand 3D ultrasound: The ultrasound images are acquired using an ultrasound probe equipped with a position sensor that tracks its pose. This allows for freehand scanning, where the probe is manually moved by the operator over the region of interest. The position sensor provides real-time information about the position and orientation of the probe during the scanning process. Figure source: [39]

.

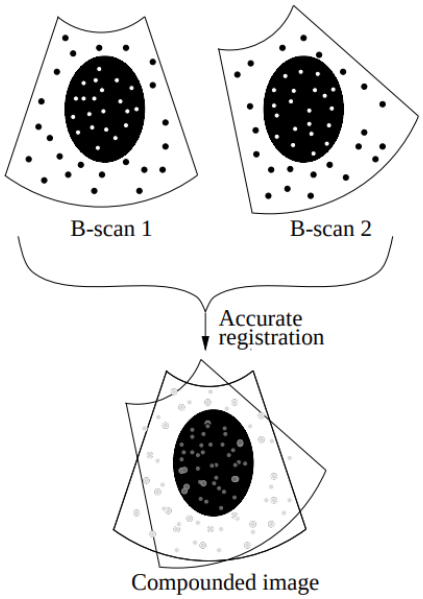

Figure 2.6.: Volume compounding: This illustration depicts the process of compounding B-mode images. After registration, the intensity of each pixel (or voxel) in the compounded volume is calculated by, for example, averaging the intensities from multiple observations. This method combines information from different images to create a final representation. Figure source: [39]

.

Freehand 3D ultrasound is a technique in ultrasound imaging that allows for the presentation of imaged tissue in three dimensions. Unlike 3D ultrasound systems that rely on specialized hardware solutions, freehand 3D ultrasound uses a conventional ultrasound transducer supported by a tracking device that measures its six degrees of freedom (6DoF) pose, including its rotation and translation in space. In this type of system, a set of 2D ultrasound frames is acquired along with tracking information, as presented in Figure 2.5. The resulting 3D dataset, also known as an ultrasound sweep, is then processed to form a 3D ultrasound volume through a process known as 3D ultrasound reconstruction. In medical ultrasound imaging, this reconstruction process is commonly referred to as volume compounding. This method combines multiple ultrasound sweeps or images into a 3D ultrasound volume. Each individual sweep provides a partial view of the target region, but by compounding or merging these sweeps, a more complete and comprehensive representation of the volume can be obtained. The compounding process involves spatially aligning and registering

the sweeps, accounting for any variations in probe position or patient movement. This alignment ensures that corresponding structures in different sweeps are correctly aligned in the compounded volume. Once the volume is compounded, it can be further processed or manipulated for various purposes, such as generating additional sweeps or slices at different orientations, enabling multi-planar visualization, performing measurements, or extracting quantitative information. A schematic illustration of this process is presented in Figure 2.6.

2.5. Confidence Maps

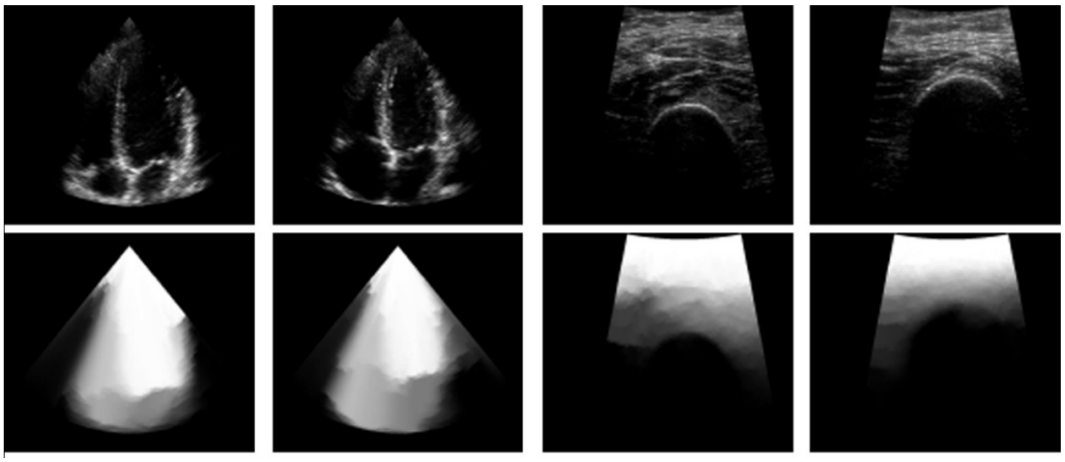

Figure 2.7.: Examples of confidence maps (bottom) generated from B-mode images (top) using a random walk algorithm; adapted from [23]

.

Medical ultrasound imaging acoustic artifacts ,presented in Section 2.3, significantly impact the acquired data's quality and reliability. Moreover, such artifacts can profoundly affect the accuracy and robustness of algorithms that rely on the fidelity of observed information. Consequently, numerous ultrasound techniques have been developed that provide uncertainty estimation to identify potentially unreliable regions. One widely-used method involves the calculation of confidence maps $C : \Omega - > [0, 1]$. These maps provide a per-pixel $p \in \Omega$ measure of uncertainty $C(p)$ in imaging information represented by the image intensity $I(p)$, where the value directly corresponds to the probability that the signal emanating from a particular pixel's location reaches the transducer [23]. Multiple studies have shown that confidence maps enhance the performance of various algorithms. For instance, in a study by Chatelain et al.[6], the

authors demonstrate that confidence maps can enhance the design of a control law and improve the quality of B-mode images in a robotic ultrasound acquisition. Similarly, another study by Jiang et al.[20] shows that uncertainty quantification improves the quality of robotic ultrasound imaging by enhancing the estimation of probe alignment with respect to the surface normal. In another study, Villa et al. [49] use confidence maps as an additional input channel to enhance the automatic segmentation of bone surfaces using a fully convolutional neural network (FCN).

3. An Introduction to Neural Radiance Fields

A core aspect of this thesis is the fundamental concept of NeRF. This chapter covers essential concepts from computer vision and computer graphics that have contributed to the development of NeRF, alongside an overview of the state-of-the-art. This chapter begins with an overview of 3D representations commonly used in neural rendering, focusing on implicit neural representations. It then proceeds to introduce the concept of neural rendering, which refers to a set of techniques that employ neural networks to generate realistic images or videos from a given 3D scene or object before delving into the details of NeRF.

3.1. 3D Representations

Representing the structure of a 3D scene is a crucial aspect of 3D computer vision. Over the years, various concepts have emerged, and the choice of a particular representation that is most suitable for a given task is determined by the specific application and the available data. This section presents an overview of 3D representations used in neural rendering and provide a rationale for employing a 3D implicit neural representation for representing a 3D scene.

3.1.1. 3D Representations in Neural Rendering

In the realm of 3D representations, explicit representations typically refer to discrete representations, whereas implicit representations pertain to continuous representations. Each of these groups can be further classified into surface and volumetric representations. Volumetric representations are capable of storing volumetric principles such as densities, and also facilitate the representation of multidimensional features such as colors. In contrast, surface representations are limited to the representation of surface properties alone, and are therefore unable to encapsulate volumetric features. Figure 3.1 shows examples of 3D representations with their classification as proposed by Tewari et al. [48].

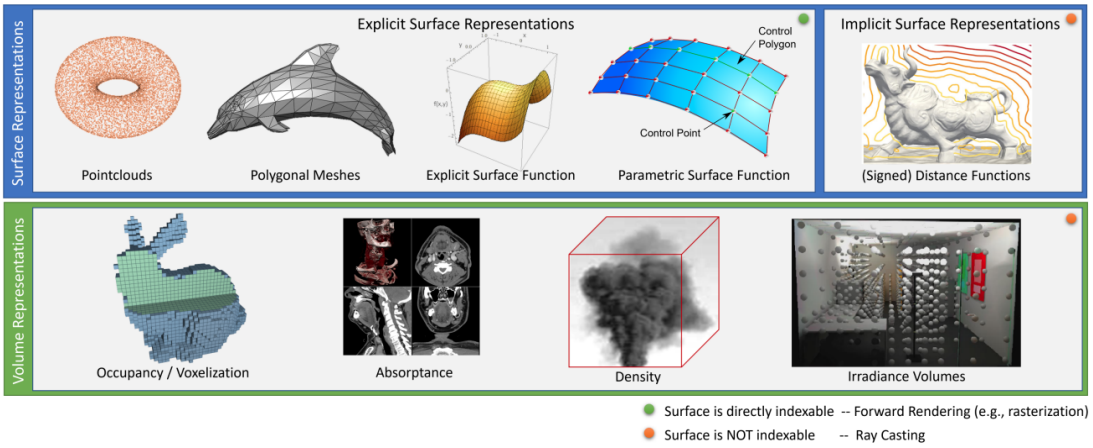

Figure 3.1.: Classification of surface and volume representations. Source [48].

Recent research has indicated that both explicit and implicit representations can be utilized in neural rendering. For example, Sitzmann et al.[44] employ voxel grids to encode a 3D scene, which allows for view-dependent appearance representation in their DeepVoxel approach. Park et al.[33] propose using a continuous signed distance function (SDF) as a means of shape representation in DeepSDF. In Free View Synthesis [38], the authors employ a mesh surface representation as a proxy for 3D scene representation from which they derive views. Furthermore, in Neural Point-Based Graphics [1], the authors utilize point clouds as a means of scene representation, which enables the rendering of new views.

3.1.2. Implicit Neural Representations

As presented in the previous section implicit representations can refer both to surface and volumetric representation. Following [47], the surface S_i using an implicit surface function $f_{\text{implicit}}(\cdot) \in \mathbb{R}$ is defined as the zero-level set:

$$S_i = \{\mathbf{x} \in \mathbb{R}^3 : f_{\text{implicit}}(\mathbf{x}) = 0\} \tag{3.1}$$

Whereas, the volume V_i is defined by a function $f_{\text{vol}}(\cdot) \in \mathbb{R}$:

$$V_i = \{f_{\text{vol}}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^3\} \tag{3.2}$$

The Equations 3.1 and 3.2 can be restricted to the domain in which a 3D scene is defined and extended to provide additional information. Moreover, both representations can be approximated by a variety of functions or a mixture thereof that can express the

respective representation. Notably, neural networks, particularly Multilayer Perceptron (MLP), which can be viewed as universal function approximators [15], can approximate both implicit representations. In this context, an implicit neural representation is a representation that is approximated by a neural network.

Implicit surface representations are widely adopted in many applications because they compactly encode complex shapes. Signed-distance functions (SDFs) are the most common form of implicit surface representation. In 1996, Curless and Levoy proposed a weighted SDF to represent surfaces. Following this idea, truncated SDFs (TSDFs) are used for real-time surface reconstruction [32]. More recently, Michalkiewicz et al. proposed incorporating an SDF as an individual layer in neural network architecture [30]. Other works have proposed alternative neural representations, such as continuous neural scene representations compatible with multiview geometry and represented as a MLP [43], or using periodic activation functions to represent natural signals[45].

3.2. Volume Rendering

Volume rendering is a set of techniques for visualizing volumetric data on a 2D plane. The most commonly used approaches to rendering are rasterization and ray-casting. In this thesis, we focus on ray-casting, a type of direct volume rendering, and then introduce neural rendering based on this technique.

3.2.1. Ray Casting

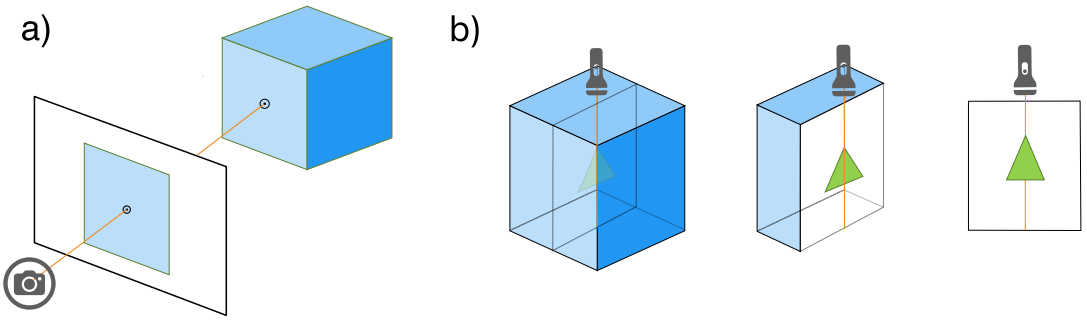

Figure 3.2.: Comparison between optical ray casting and ultrasound ray casting: (a) depicts a ray cast from a camera into a volume, passing through an image plane. (b) illustrates a ray cast from an ultrasound probe, passing through a volume and an image plane.

Ray casting is a widely employed technique in volume rendering, whereby the properties of the scene are integrated along a ray to create a projected image of volumetric data [26]. This technique involves casting rays from the image plane into the scene geometry. An optical alongside a pinhole camera model is commonly utilized in computer vision to define ray casting, whereby the intersection of each ray with the scene geometry is determined to obtain a color value for a particular pixel in the image plane. This fundamental principle is illustrated in Figure 3.2(a). To compute the color value for a specific pixel, we compute the volume rendering integral along a ray, defined as:

$$C(\mathbf{x}) = \int_0^{s_{\max}} \sigma(\mathbf{x} + s\mathbf{d}) C'(\mathbf{x} + s\mathbf{d}), ds, \tag{3.3}$$

where $C(\mathbf{x})$ is the final color value at position $\mathbf{x}$, $\sigma(\mathbf{x})$ is the volume's opacity at position $\mathbf{x}$, $C'(\mathbf{x})$ is the volume's color at position $\mathbf{x}$, s is the distance along the ray, $\mathbf{d}$ is the ray direction, and $s_{\max}$ is the maximum distance to traverse along the ray. In practical applications, this integral is evaluated solely for the subspace existing in the volume. It is essential to acknowledge that more than the optical model of ray casting is needed to fully capture the intricacies of ultrasound ray casting. As illustrated in Figure 3.2, an ultrasound ray cast from an ultrasound probe into a volume traverses through both the image plane and the volume itself. This distinction requires adjusting the volume rendering integral to incorporate this difference and accurately capture volumetric data from ultrasound scans.

3.2.2. Neural Rendering

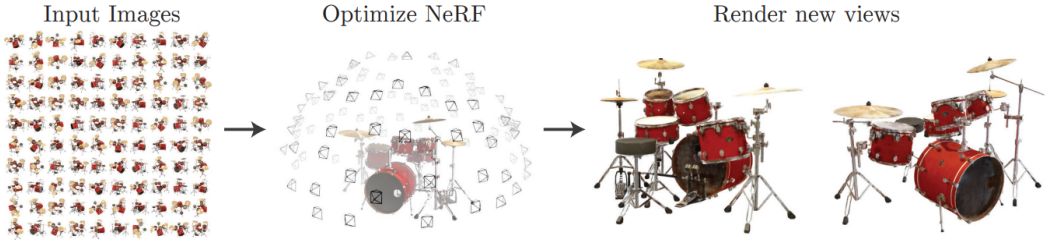

Figure 3.3.: Illustration of rendering process (from right to left): an unstructured set of pose annotated images of a scene; optimized neural network representing the scene; rendering of a 2D projection based as seen from a new viewing point. Source: [31]

Volume rendering techniques, such as ray-casting, rely on having complete knowledge of all physical parameters of the scene, including the scene geometry, to facilitate

rendering and provide inputs to the rendering process. These parameters can be estimated from existing observations in situations where they are not known. Recently, the field of neural rendering has emerged, aiming to learn the rendering process of a scene from a set of observations using neural networks. The state-of-the-art research in neural rendering defines it as *"deep image or video generation approaches that enable explicit or implicit control of scene properties such as illumination, camera parameters, pose, geometry, appearance, and semantic structure."* [48]. Typically, a neural rendering process follows a specific set of steps:

1. A set of images representing a 3D scene is provided as input to the process.
2. The process builds a neural representation of the scene.
3. The scene is rendered, and a new image is generated based on the specified scene properties.

These steps are illustrated in Figure 3.3 based on an example of NeRF.

3.3. Neural Radiance Fields

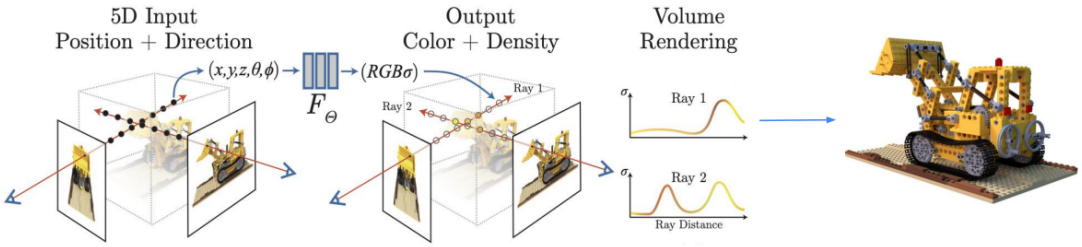

Figure 3.4.: Illustration of the NeRF scene representation and rendering procedure. Images are synthesized by sampling 5D coordinates, encompassing location and viewing direction, along camera rays. These coordinates are then inputted into a MLP to generate color and volume density information. Finally, volume rendering techniques are employed to combine these values into a cohesive image. Adapted from: [31]

The application of neural networks and neural rendering techniques to achieve photorealistic novel view synthesis has sparked considerable interest in continuous shape representation. Notably, a significant breakthrough in this area emerged with the introduction of NeRF [31]. NeRF introduces a framework that combines a neural

representation of a scene with fully differentiable volumetric rendering, allowing for the generation of 2D projections of a 3D scene from any conceivable viewing point. NeRF exhibits several key features that contribute to its success in photorealistic novel view synthesis:

- **Dataset Collection**: To learn a scene representation, NeRF requires a dataset of unstructured, densely located images capturing a scene from various viewing points, as depicted in Figure 3.3.
- **Pose Estimation**: Accurate pose estimation is necessary for NeRF to understand the viewing positions of the input images.
- **Physically Meaningful Representation**: NeRF learns physically meaningful color and density values in 3D space.
- **Ray Casting and Volume Integral**: NeRF utilizes ray casting and volume integration techniques to render novel views consistent with the 3D scene observed in the training data.

In addition to these key features, the fundamental concept of NeRF involves expressing the representation of a scene through a fully-connected neural network. This network maps a 5D vector, comprising both spatial location $\mathbf{x}$ and viewing direction $\mathbf{d}$, to volume density σ and radiance $\mathbf{c}$. To achieve this, NeRF defines a per-pixel camera ray denoted as $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$ represents the camera origin situated at the center of the pixel defining the near plane. This formulation enables NeRF to capture intricate scene details and accurately model the interplay between spatial location, viewing direction, volume density, and radiance. Ultimately, the computation of the color value for each pixel involves applying the volume rendering integral, which derives from the Equation 3.3:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(t)c(t)dt \tag{3.4}$$

$$\text{where } T(t) = \exp - \int_{t_n}^{t_f} \sigma(\mathbf{r}(s))ds \tag{3.5}$$

The volume rendering integral, as depicted in Equation 3.4, plays a crucial role in accumulating the ray-traced radiance field along the camera ray, ranging from the near plane (t_n) to the far plane (t_f). At each position along the ray, the contribution adds up to determine the final pixel color. The transmittance factor $T(t)$, outlined in Equation 3.5, governs the influence of each sample in the overall rendering process. Figure 3.4 provides a comprehensive visualization of the entire process, encompassing coordinate space sampling, neural representation construction, and culminating in

volume rendering. This figure effectively captures the key steps involved in the NeRF framework, offering a clear overview of the entire pipeline. While the presented formulation of NeRF demonstrates impressive capabilities in photorealistic novel view synthesis, it is important to acknowledge that it also has certain limitations and challenges:

- **Data Requirements**: NeRF relies on a large dataset of densely captured, unstructured images, making data collection time-consuming and resource-intensive.
- **Computational Complexity**: Training NeRF can be computationally expensive, especially for high-resolution images and complex scenes.
- **Limited Viewpoints**: NeRF may struggle to generalize to novel or unseen viewpoints if there are limited viewpoints in the training data.
- **Opaque Scenes**: NeRF's design assumes scenes with opaque objects, posing challenges for scenes with transparency or complex occlusions.
- **Scalability**: Scaling NeRF to larger and more complex scenes requires significant computational resources and memory.
- **Noise and Artifacts**: NeRF is sensitive to noise, and noisy images can introduce artifacts, affecting the quality of rendered results.
- **Limited Generalization**: NeRF may struggle to generalize to viewpoints significantly different from the training data.

Many of these challenges have been addressed by subsequent works, as outlined in several comprehensive reviews [10, 47, 48]. These works have made significant progress in mitigating the limitations and improving the performance of NeRF. However, it is essential to acknowledge the specific challenges when employing NeRF in the medical field, as the imaging technology used in this domain significantly differs from conventional cameras. The original NeRF framework is designed to work with a set of unstructured images capturing a scene centered around opaque objects. In contrast, medical ultrasound imaging involves capturing images that pierce through objects, making the definition of surfaces more challenging compared to natural images. Therefore, the capturing process and ray-tracing methodology differ significantly between traditional and medical imaging techniques. Moreover, the observation angles in medical ultrasound imaging are limited due to the physical constraints of the scanning device. At the same time, an object captured by a traditional camera can be observed from less restricted viewing angles. Additionally, in medical ultrasound imaging, the acquired images tend to be noisy, and the noise exhibits a statistical distribution. As a result,

maintaining precise correspondence and consistency in terms of pixel intensities and appearances becomes more complex compared to traditional imaging modalities. These limitations impose additional complexities when applying NeRF to medical ultrasound imaging and necessitate careful consideration and adaptation of the framework to account for the unique characteristics and constraints of the medical imaging domain. In the next chapter, we will explore the application of NeRF in the context of medical imaging.

4. Related Work

While the previous chapter presents the literature related to NeRF in computer vision, this chapter specifically focuses on recent developments concerning NeRF in medical imaging. It begins by introducing recent research in implicit neural representation in the context of medical imaging, followed by a selection of relevant papers that explore the application of NeRF in this domain.

4.1. Implicit Neural Representation in Medical Imaging

Implicit neural representations have emerged as a promising approach to representing 3D medical imaging data as continuous functions encoded within the weights of neural networks. Several important papers have contributed to advancing this representation in medical imaging, leading to a wide range of applications. These methods can be divided into two main categories. For different medical imaging tasks, the first category involves using implicit neural representations, such as SIREN [45]. This approach focuses on leveraging the expressive power of implicit neural representations for the following tasks:

- **3D Reconstruction:** ImplicitVol [55] presents a sensorless 3D reconstruction method. An essential part of its framework is a novel approach for representing 3D reconstruction from ultrasound images as a continuous implicit neural function. In this approach, a neural network maps 3D Cartesian coordinates to intensities in B-mode images. Hence, a neural network represents the 3D volume implicitly as a continuous volumetric function parametrized by the weight of a neural network.

- **Segmentation:** Khan et al. [24] propose use implicit neural representation to represent segmentation as a continuous function instead of discrete segmentation maps. In this approach, a neural network parameterizes segmentation map as a continuous function that maps a coordinate in a volume to occupancy value to indicate whether it belongs to an organ.

- **Registration:** Wolterink et al. [51] utilize differentiability of implicit neural representation to represent a transformation function in deformable medical image registration. This work defines each transformation between an image pair as a

parametrized neural network. Since finding unique parameters to represent deformation is an ill-posed task, the authors propose directly imposing regularization on the weights of neural networks representing the deformation function.

- **High-Resolution Reconstruction:** IREM [53] leverages the continuity of implicitly learned neural representations to achieve super-resolution reconstruction of brain MRI images. In this work, a fully-connected neural network is trained to represent a high-resolution MRI of an entire anatomy from which one can sample a 2D slice by mapping 3D coordinates in the volume to image intensities.

The second category combines simple implicit neural representations with complex medical imaging acquisition models and aims to improve representation accuracy by combining the flexibility of implicit neural representations. These methods consider unique features of medical imaging. For example, Reed et al. [37] explore the physical principles of the acquisition process behind CT to use implicit neural representation in their 4D-CT reconstruction pipeline. In this work, the implicit neural representation serves as a fixed prior for the 3D scene and is combined with a parametric motion field that accurately estimates the temporal evolution of the scene in question. Xu et al. [54] proposes using multiple fully-connected neural network to represent implicit neural representation of bias field, noise variance and volume intensities. These representations are then fused to enhance 3D magnetic resonance (MR) reconstruction from motion corrupted 2D slices.

4.2. NeRF in Medical Imaging

While there is a significant body of work on implicit neural representation in medical imaging, the application of NeRF in the medical domain is still an emerging research direction. However, several papers have addressed some of the domain differences between medical and conventional images to allow the application of NeRF for medical imaging. The majority of available literature, however, focuses on reconstruction from CT or MRI. For instance, MedNeRF [7] presents a NeRF framework designed explicitly for reconstructing CT projections from X-ray data. The authors harness the power of NeRF to generate high-quality volumetric representations of CT scans using a collection of 2D X-ray projections. MedNeRF showcases promising results in accurately reconstructing the volumetric structure of organs and tissues from X-ray images. Idrissu et al. [18] extend the idea presented in MedNeRF to enable the reconstruction from brain MRI scans. To this end, they propose incorporating a radiation attenuation response from MRI in the generative radiance field. However, more than direct application of these methods to medical ultrasound is needed due to the distinct characteristics of

CT, MRI, and ultrasound imaging, specifically in their respective image formation models. Despite the potential advantages of employing neural radiance fields (NeRF) in medical ultrasound, to the best of my knowledge, the application of NeRF to ultrasound imaging has been addressed in only one study by Li et al. [27]. The authors employ the NeRF algorithm to reconstruct 3D ultrasound spine data and evaluate spinal curvature measurements based on the reconstructed results. However, it is essential to note that the authors do not address ultrasound imaging's physical properties. Specifically, a volumetric rendering method that respects the physics of ultrasound has yet to be addressed.

5. Methodology

This chapter starts by explaining the problem setting and the underlying assumptions. Then offers a detailed description of the proposed solution for novel view synthesis in ultrasound imaging using the NeRF framework. In particular, it presents critical aspects of novel view rendering for ultrasound. Firstly, a general two-step framework and a network architecture are introduced. Secondly, a differential volumetric rendering process tailored explicitly for ultrasound is defined. Lastly, the focus shifts to the regularization of the parameters space.

5.1. Problem Setting

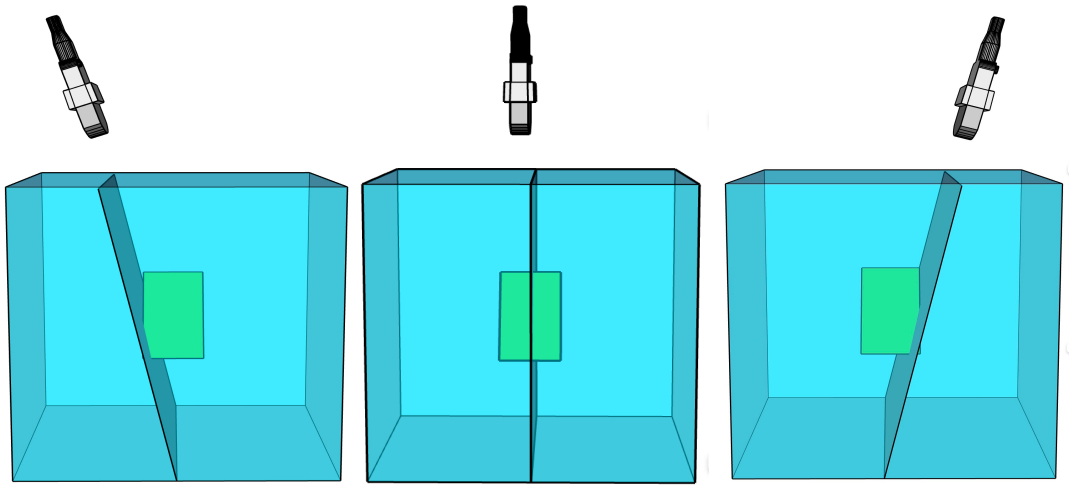

Figure 5.1.: Visualization of probe orientation and its impact on the scanned ROI. Tilting the probe not only alters the observation angle but also modifies the path that an acoustic wave must travel to reach the ROI.

To address the novel view rendering for ultrasound, the focus of this study is to propose an implicit neural representation of the 3D ROI, allowing for view-dependent rendering. The challenge involves limited information about the imaged ROI, primarily

derived from B-mode sweeps. Notably, the observation of the ROI occurs from multiple viewing points, depending on the probe orientation, as depicted in Figure 5.1. The assumptions relevant to this study are presented in the following section.

5.2. Assumptions

The work presented in this thesis assumes:

- the scene observed is static, and there are no deformations present;
- the tracking of a probe is accurate and that the tracing error is negligible;
- any point rendered from a novel observation point has been observed in the training data from at least one, different observation point;
- tissue characteristics are unknown a priori, which presents a challenge for this task;
- it is possible to estimate the number of tissue types by categorizing them based on similar tissue characteristics.

5.3. Neural Radiance Fields for Ultrasound Imaging

Similar to the original NeRF framework, the method comprises two modules: a neural network that maps coordinates to a multidimensional parameter vector representing a 3D scene and a volumetric rendering function that takes the parameters for each position along a ray and produces the corresponding intensity for the B-mode image. Figure 5.2 provides an overview of the complete pipeline. The method builds upon ray-based ultrasound models commonly employed in ultrasound simulation. These models typically assume that prior knowledge of physical parameters and semantic tissue segmentation is available for rendering.

However, in general, the need for per-pixel tissue labels and information about tissue characteristics poses a difficulty. Tissue characteristics are approximated from other modalities or general parameter estimations for a specific tissue type, and detailed tissue labeling requires additional modalities. This thesis addresses these challenges by training a neural network to restore an intermediate physical parameter space for ray-based ultrasound modeling. Regressing continuous space of physical tissue parameters allows for the prediction of parameters for infinitesimal probe movements during rendering.

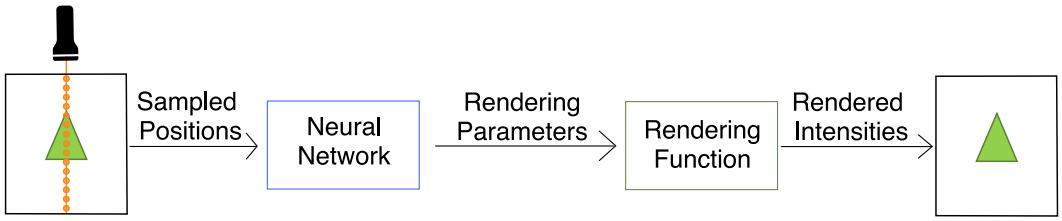

Figure 5.2.: Overview of the complete method. The method consists of two modules: a neural network and a continuous volumetric rendering function. At each sampled point, the neural network generates a vector of rendering parameters, which are subsequently mapped by the rendering function to pixel intensities.

Furthermore, the goal is to obtain a coherent physical parameter space by assuming spatial continuity of these parameters, ensuring a smooth transition between neighboring regions. The following sections provide a detailed discussion of the method and implementation.

Network Architecture

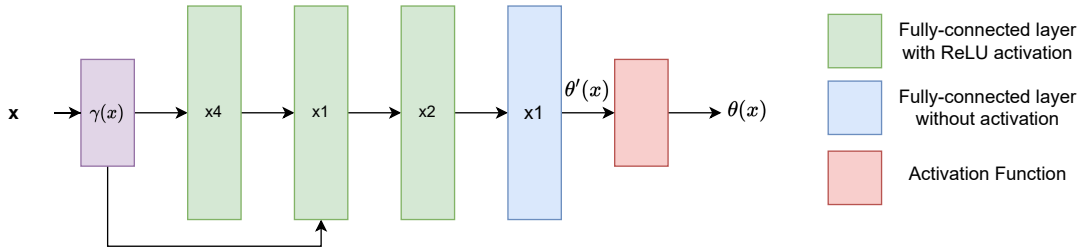

Figure 5.3.: Schematic visualization of the neural network module.

Figure 5.3 depicts the network architecture utilized to represent a 3D scene. The MLP network approximates the complex and nonlinear function that maps 3D positions to parameter values, enabling the rendering of realistic B-mode images. By stacking multiple layers of nonlinear functions, utilizing Rectified Linear Unit (ReLU) activations, the MLP can capture an increasingly intricate and complex mapping from Cartesian coordinates to the parameter space. Increasing the number of layers in a neural network can enhance its ability to approximate increasingly complex functions. However, deeper models may suffer from computational inefficiencies and overfitting. Conversely, shallower models may be insufficiently expressive to capture the

complexity of the radiance field. To strike a balance between model capacity and computational efficiency, a recommended approach, as proposed in DeepSDF [33], is to utilize an eight-layer architecture. Such an architecture provides a judicious trade-off between model capacity and computational resources. Following NeRF, the network incorporates a skip connection that concatenates the input with the activation of the fifth layer. However, unlike the original NeRF architecture, no additional layers are included to account for viewing direction. Instead, the rendering module handles the viewing direction dependency of the final B-modes. The sigmoid activation function is employed to ensure that the final output parameters fall within the range of 0 and 1, except for the attenuation coefficient, which may take any positive value. The absolute value activation function is applied to the attenuation coefficient to accommodate this requirement.

5.3.1. Ultrasound Volumetric Rendering

The ultrasound volume rendering model presented in this thesis expands upon a ray-tracing-based formulation that has been previously discussed in the relevant literature, such as the study conducted by Salehi et al. [40] that employ this formulation for simulating patient-specific B-mode images for known anatomy and pre-defined tissue characteristics. Additionally, this model considers the ultrasound ray physics outlined in Section 2.2.1. A key strength of this model is its inherent flexibility, which enables it to accurately depict B-mode intensities arising from both reflection and scattering phenomena. It differentiates between view-dependent ultrasonic effects such as occlusions, large-scale reflections, and attenuation combined with speckle patterns. This flexibility makes it particularly advantageous for neural rendering for ultrasound. Following, we provide a formal definition for a ray-based rendering used in this thesis.

For each scan-line r, Equation 5.1 defines a recorded ultrasound echo $E(r,t)$, measured at distance t from the transducer, as a sum of reflected $R(r,t)$ and backscattered $B(r,t)$ echo intensity:

$$E(r,t) = R(r,t) + B(r,t) \tag{5.1}$$

Additionally, log compression is employed to enhance the reflected part, effectively representing the dynamic range of the original image:

$$R(r,t) = \ln(1 + \gamma \cdot R'(r,t)) \cdot \ln(1+\gamma) \tag{5.2}$$

The reflected energy before the log compression is defined by:

$$R(r,t) = |I(r,t) \cdot \beta(r,t)| \cdot PSF(r) \otimes G(r',t') \tag{5.3}$$

Where $I(r,t)$ is the remaining intensity at the distance t, $\beta(r,t)$ denotes the reflection coefficient, and $PSF(r)$ signifies a predefined 2D point-spread function. G(r, t) admits the value 1 for points at the boundary and 0 otherwise. In the model employed in this thesis, a constant value of 1 is assumed, simplifying the equation to solely involve the intensity reflection coefficient for computing the reflection:

$$R(r,t) = |I(r,t) \cdot \beta(r,t)| \tag{5.4}$$

The propagation of intensity is tracked along each scan-line, and the residual energy $I(r,t)$ is modeled by considering the intensity loss attributed to reflection β at the boundaries, as well as the attenuation compensated through the application of an undisclosed time-gain compensation (TGC) function. The final formulation for $I(r,t)$ assumes an initial unit intensity $I_0(r,0)$ and accounts for the intensity loss occurring at each incremental step dt.

The resulting equation can be further simplified by representing the compensated attenuation α with a single parameter, considering that the TCG acts as a scaling factor

$$I(r,t) = I_0 \cdot \prod_{n=0}^{t-1} [(1 - r_{coeff}(r,n))] \cdot \exp^{(-\int_{n=0}^{t-1}(\alpha \cdot f \cdot dt))} \tag{5.5}$$

Consequently, the values of α correspond to the physical attenuation up to an unknown scaling factor.

The backscattered energy $B(r,t)$ originating from the scattering medium is influenced by the remaining energy $I(r,t)$ and a2D map of scattering points $T(r,t)$:

$$B(r,t) = I(r,t) \cdot PSF(r) \otimes T(r',t') \tag{5.6}$$

The map $T(r,t)$ is learnt using a generative model inspired by [56]:

$$T(r,t) = H(r,t) \cdot \phi(r,t) \tag{5.7}$$

Within this model, the function $H(r,t)$ assumes a value of 1 when the query point corresponds to a scattering point, and 0 otherwise. This function is sampled from the Bernoulli distribution, with the scattering density ρ_s as its parameter. It encapsulates the inherent uncertainty regarding the observance of the scattering effect of each scattering point. The intensity associated with a scattering point is determined by its amplitude ϕ, which is sampled from a normal distribution with a mean of ϕ and a unit variance, effectively capturing the variability in intensity levels.

This model is not fully differentiable due to the sampling process from a Bernoulli distribution. However, despite this limitation, the loss can still be back-propagated, even though the sampled variables do not directly contribute to updating the network

weights. In order to achieve full differentiability across the entire model, one approach is to approximate the Bernoulli distribution with smooth relaxation. One such technique is Relaxed Bernoulli [28], where a low temperature value can substitute the original discrete distribution, ensuring differentiability throughout the rendering process.

5.3.2. Ray Sampling

Without loss of generality, the presented ray sampling here primarily focuses on a linear probe; however, the sampling method can also be extended to a curvilinear probe. T he first step involves defining a ray to sample a position within the ultrasound image. The ray, denoted as **r**, is characterized by its origin **o** and direction vector **d**. Each ray corresponds to a single scan line in the ultrasound image. The ray origin is positioned at the contact point of the probe with the surface for the respective scan line. The direction vector aligns with the elevation and coincides with the scan line. With the origin and direction of each ray determined, it becomes possible to sample points along the ray using a step **t**. Consequently, a ray can be defined as $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$. This thesis assumes equidistant sampling, as depicted in Figure 5.4.

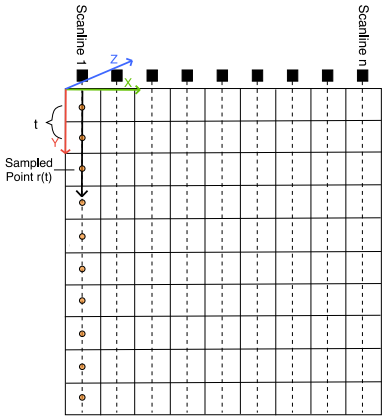

Figure 5.4.: Visualization of the rays definition and ray sampling. For a ray defined as $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ points are sampled equidistantly and coincide with pixel positions in the 3D space.

5.3.3. Parameters Space Regularization

The problem of finding a set of parameters for the rendering function that best fits the observed data is a classical inverse problem. However, the formulation presented

in the previous section allows for degenerate solutions, primarily stemming from the inherent ambiguity in representing the final pixel intensity within the parameter space. In machine learning, regularization is a popular method used to constrain the problem and improve the search space of possible solutions. Regularization helps to avoid overfitting and encourages models to find simpler solutions that generalize well. This thesis proposes a method of regularizing the parameters space based on the physical characteristics of the rendering parameters. The proposed method aims to improve the accuracy of the rendering function considering the representation of underlying topology by incorporating physical constraints on the estimated parameters.

Correlation between attenuation and scattering

Based on the underlying physics, a correlation between the physical properties of the tissue exists. For instance, the Pearson correlation coefficient indicates that attenuation and scattering are highly correlated. To exploit this relationship, we propose using a correlation measure on the maps of the regressed parameters to constrain the network to maximize the correlation between the attenuation and scattering amplitude. Specifically, using Local Normalized Cross Correlation (LNCC) to measure the correlation between the attenuation and scattering amplitude maps. The formula for this measurement is defined as follows:

$$\text{LNCC}(x,y) = \frac{\sum_{i,j} w(i,j)[I(x+i,y+j) - \mu(x,y)][J(x+i,y+j) - \nu(x,y)]}{\sqrt{\sum_{i,j} w(i,j)[I(x+i,y+j) - \mu(x,y)]^2 \sum_{i,j} w(i,j)[J(x+i,y+j) - \nu(x,y)]^2}}$$

In this formula, $I(x,y)$ and $J(x,y)$ are the intensity values of the two images being compared at pixel location (x,y), and $w(i,j)$ is the weighting function applied to each pixel location within a local window centered at (x,y). The mean values of I and J within the local window are denoted by $\mu(x,y)$ and $\nu(x,y)$, respectively. -

Local smoothness of attenuation and scattering

In ultrasound imaging, it is common to assume that tissues are locally homogeneous. This assumption suggests that one can expect a low local variation in tissue characteristics and, therefore, in the rendering parameters. Specifically, it is reasonable to anticipate a minimal gradient in both the attenuation and scattering maps, given that the physical properties of the tissue remain relatively constant in a particular area. To exploit this attribute and further constrain the solution space, we propose using Total Variation (TV) regularization on the scattering map. However, traditional TV regularization does not account for abrupt changes in the underlying tissue properties

that arise from imaging multiple tissue types. To address this limitation, we introduce a weighted version of TV regularization, which presents as follows:

$$TV(\phi_s) = \sum_{i,j} \gamma_{i,j}(|\phi_{s_{i+1,j}} - \phi_{s_{i,j}}| + |\phi_{s_{i,j+1}} - \phi_{s_{i,j}}|) \tag{5.8}$$

$$\text{where } \gamma_{i,j} = r_{max} - r_{i,j}$$

In this equation, the penalty function incorporates a weight factor, denoted as $\gamma(i,j)$, contributing to the optimization process two-fold. On the one hand, it relaxes the constraint imposed by the regressed reflection coefficient, allowing for some degree of flexibility in the reconstruction. On the other hand, it imposes a constraint on the value of the reflection coefficient based on the variability observed in the scattering attenuation map. The correlation between attenuation and scattering indicates that enforcing a low total variation on the scattering map results in a similar outcome on the attenuation map due to the bidirectional correlation between the physical properties that underlie these maps.

Tissue clustering

The assumption underlying the estimation of rendering parameters is that they reflect the physical characteristics of the tissue. It is reasonable to expect that it is possible to cluster tissue types in the parameter space. This clustering is further used to globally constrain the variation in the parameter space, specifically in the attenuation map. To achieve this, we propose a two-step approach. This approach is presented in Algorithm 1.

Algorithm 1 Attenuation Clustering Penalty

1: Step 1: Cluster Attenuation
2: Choose the number of clusters, k
3: Cluster attenuation
4: Step 2: Penalize Distance to the centroids
5: **for** each pixel (i,j) in attenuation map **do**
6: Calculate the penalty factor, $\gamma_{i,j}$ as a distance between the attenuation value and the centroid
7: **end for**

In the first step, attenuation is clustered using a method such as k-means. This thesis uses k-means clustering to group similar attenuation values. In this step, we must decide on the number of expected clusters, denoted as k, which is a hyper-parameter.

In the second step, the obtained centroids are used to calculate a penalty factor that penalizes distance from the centroid.

One limitation of the clustering-based penalty is that it may result in a trivial solution where attenuation values are equal to zero. Furthermore, allowing for sufficient variation in cluster centroids can facilitate the network's ability to learn distinct tissue characteristics. We introduce an additional penalty factor that penalizes the distance between cluster centroids to address these issues. However, directly computing the distances between centroids would prevent gradient propagation, so we compute new centroid values based on the labeled attenuation map. Then, we calculate the pairwise distances between the centroids obtained from clustering and the re-computed centroids and use them in the final penalty. The tissue clustering penalty function is as follows:

$$L_c = \frac{1}{N}\sum_{i=1}^{N}(\alpha_i - centroids)^2 \tag{5.9}$$

5.3.4. Training Details

Loss Function

A combination of a primary loss function and supporting regularization terms are utilized for training. The primary loss function employed is the Structural Similarity Index Metric (SSIM), a perceptual loss that considers differences in luminance, contrast, and structure between two images. However, during the initial stages of training, it is observed that using the pure SSIM loss was challenging since the training commenced with no prior knowledge of the final intensity values. A warm-up stage is introduced to address this, during which a simple photogrammetric loss, specifically the L2 loss measuring pixel-wise intensity differences, is utilized. Following the warm-up phase, the L2 loss acts as an auxiliary loss that encourages rendering images with intensities closer to the ground truth B-mode images. After the warm-up stage, regularization terms are incorporated. During the early stages of training, clustering regularization is not employed as it requires considerable time for the clustering method to establish clusters. The clustering penalty is first introduced during the final stages of training to refine the results and maintain global consistency among the regressed parameters for the same tissue types. The final loss function is defined by a combination of the SSIM loss, the L2 auxiliary loss, and additional regularization terms as follows:

$$\mathcal{L} = w_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + w_{\text{L2}}\mathcal{L}_{\text{L2}} + w_{\text{LCC}}\mathcal{L}_{\text{LCC}} + w_{\text{TV}}\mathcal{L}_{\text{TV}} + w_{\text{C}}\mathcal{L}_{\text{C}} \tag{5.10}$$

Where,

$$\mathcal{L}_{\text{SSIM}} := \text{Structural Similarity Index Metric}$$
$$\mathcal{L}_{\text{L2}} := \text{L2 Loss}$$
$$\mathcal{L}_{\text{LCC}} := \text{LCC Penalty Factor}$$
$$\mathcal{L}_{\text{TV}} := \text{Total Variation Penalty Factor}$$
$$\mathcal{L}_{\text{C}} := \text{Clustering Penalty Factor}$$

$$w_{\text{SSIM}} := \text{Weight for SSIM component}$$
$$w_{\text{L2}} := \text{Weight for L2 component}$$
$$w_{\text{LCC}} := \text{Weight for LCC component}$$
$$w_{\text{TV}} := \text{Weight for TV component}$$
$$w_{\text{C}} := \text{Weight for Clustering component}$$

Table 5.1 presents weights used for the training.

Table 5.1.: Weights per loss component

w_{SSIM}	w_{L2}	w_{LCC}	w_{TV}	w_{C}
1.0	1e-1	1e-2	1e-6	1e-2

Positional Encoding

The positional encoding function denoted as γ in Figure 5.3 maps from a lower dimensional space to a higher dimensional space. A recent study by Rahaman et al. [36] reveals that deep neural networks are often biased towards learning lower frequency functions. However, the researchers also demonstrate that the input data can be preprocessed by mapping it to a higher-dimensional space using high-frequency functions before feeding it into the network. This preprocessing technique improves the model's ability to fit high-frequency variations data. It suggests that careful selection of preprocessing techniques can mitigate the limitations of deep networks and enhance their performance in dealing with high-frequency variations. In NeRF, it has been shown that using positional encoding can improve the quality of the rendered images. Following NeRF, the positional encoding function is defined as:

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)) \tag{5.11}$$

In this thesis, an embedding size of L=10 is used.

6. Experiments and Results

This chapter presents the results of experiments validating the method proposed in this thesis. It begins with a description of the datasets utilized in these experiments, emphasizing their relevance and distinctive features. Subsequently, it explicates the quantitative metrics and qualitative validation methods employed to evaluate the effectiveness of the proposed approach. Afterward, it presents the experimental setup and procedure, providing a detailed account of the steps taken to evaluate the methods. Finally, it presents the outcomes of these experiments and an analysis of the results. The chapter concludes with a discussion of the limitations of the method.

6.1. Dataset

In this thesis, experiments were conducted using datasets collected explicitly for the purpose of novel view synthesis of B-mode images that consist of B-mode images with respective pose annotations. These datasets were designed to support the assumption of multiple observations per ROI from various viewing perspectives. Each ROI was observed from a set of positions that varied in both position and orientation of the probe, as illustrated in Figure 6.1. The collected datasets consist of multiple sweeps and Table 6.1 provides information regarding the number of sweeps and the test-train split. Each sweep differed in the angle between the acquisition direction and the probe orientation, with the probe being tilted from left to right during the acquisitions. Each sweep was tracked using robotic tracking to ensure accurate pose estimations. Further details regarding each dataset are explained in their respective sections.

Table 6.1.: Test-train split per dataset type

Dataset Type	**Train**		**Test**		**Total**	
	Sweeps	B-mode	Sweeps	B-mode	Sweeps	B-mode
Synthetic Liver	4	800	3	600	7	1200
Spine Phantom	8	1200	4	600	12	1800
Exvivo Spine	4	3400	1	800	5	4200

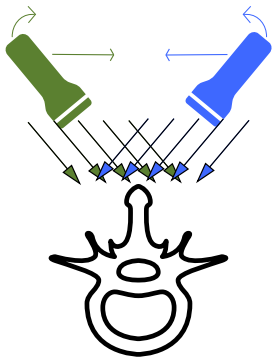

Figure 6.1.: Illustration of various viewing points with respect to a ROI. In this illustration, two probes are moved along a trajectory parallel to the surface, scanning the vertebra from different sides. During the acquisitions, the probes was tilted tilted relative to the trajectory, allowing for variations in viewing angles. In the experiments, the viewing angles have been fixed within a sweep and adjusted for different sweeps.

6.1.1. Synthetic Dataset

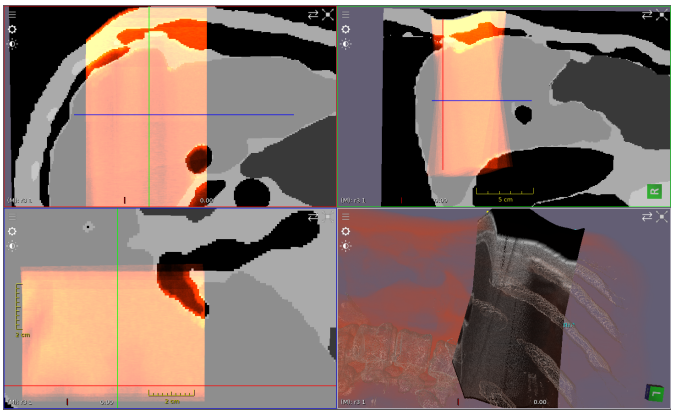

Figure 6.2.: Illustration of the labels utilized for simulating B-mode images and simulated B-mode sweeps. The highlighted region represents the area captured during the simulation, with stronger intensities indicating greater overlap between sweeps. In the bottom right corner, all simulated sweeps are displayed, incorporating ribs that contribute to the presence of acoustic shadows.

The synthetic dataset utilized in this study is a simulated collection of B-mode images. These images were generated using specialized software available within the ImFusion Suite [3]. The software employed labels derived from CT scans to simulate the B-mode images. The method used for simulation requires known properties of simulated tissue. The respective values for acoustic tissue parameters are shown in Table 6.2. The virtual scanning process covered the liver region, as depicted in Figure 6.2. Notably, the simulation also accounted for the presence of ribs, resulting in bone reflections and acoustic shadows. Each simulated sweep varies in terms of the acquisition orientation. The simulated images have a depth of 140 mm, and the virtual probe has a width of 80 mm. With a resolution of 256x512 pixels, the B-mode images offer an axial resolution of 0.31 mm, an elevational resolution of 0.27 mm, and a lateral resolution of 0.5 mm. The dataset encompasses a specific number of B-mode images, sweeps, and a test-train split, detailed in Table 6.1.

Table 6.2.: Tissue properties used in simulations

Tissue Type	**Attenuation Coefficient (dB/cm/MHz)**	**Acoustic Impedance (MRayls)**	**Scattering Density (cm^{-1})**	**Scattering Amplitude (unitless)**
Lung	1.64	0.2	0.5	0.5
Fat	0.48	1.38	0.8	0.5
Water	0.18	1.61	0.001	0.0
Kidney	0.2	1.62	0.4	0.6
Muscle	0.49	1.62	0.53	0.51
Background	0.54	0.3	0.3	0.2
Liver	0.4	1.65	0.2	0.4
Soft Tissue	0.54	1.63	0.64	0.64
Bone	2.0	7.8	1.0	0.8

6.1.2. Phantom Dataset

The lumbar spine phantom dataset was collected on a gelatine-based phantom. This phantom is a synthetic model of a bone connected by a soft material simulating discs. In our dataset, we defined trajectories that support capturing the region encompassing two vertebrae. Like the synthetic dataset, the probe's orientation varies between sweeps

[3] ImFusion GmbH, Munich, Germany, software version 2.42

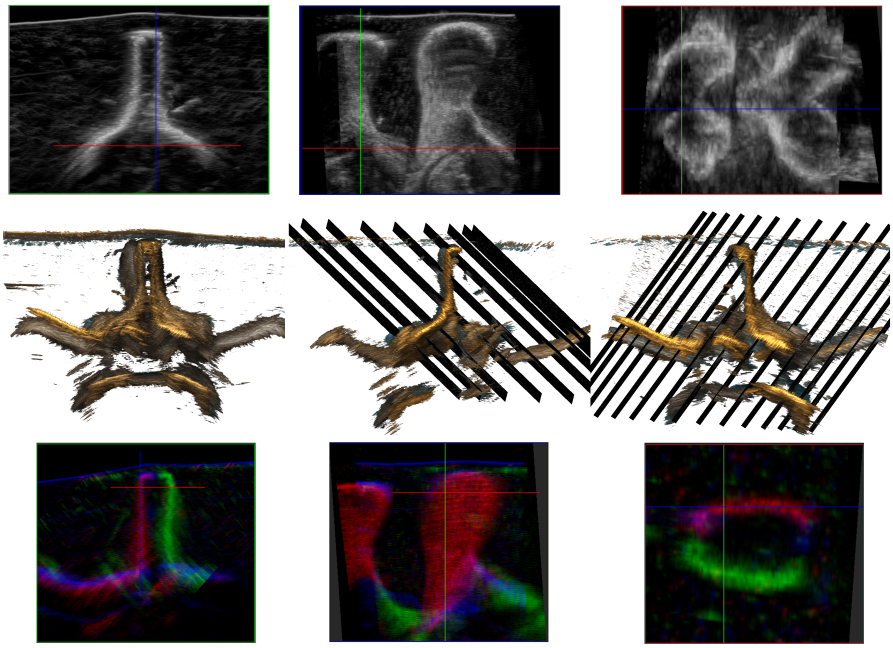

Figure 6.3.: Visualization of the phantom dataset. In the top row, white intensities are displayed, representing the max-value compounding of all-angle images. The middle row provides a visual representation of the observation directions with respect to the anatomy of the lumbar spine phantom. In the bottom row, the red, blue, and green intensities compose the white intensities based on the view angle.

to capture the ROI from different viewing positions. The sweeps were acquired using a linear probe (Cephasonics Cicada LX ultrasound machine and Piezo Composite Linear probe) connected to a robotic manipulator (KUKA LBR iiwa 7 R800). Real-time tracking information was captured using ImFusion Suite [3].

The acquired phantom data consists of B-mode images of the lumbar spine phantom. The B-mode images were captured in a paramedian sagittal orientation, resulting in occlusions from the spinous process on both sides of the vertebrae, depending on the acquisition orientation, similar to the illustration in Figure 6.1. The B-mode images have a depth of 100 mm and a width of 38 mm. With a resolution of 164x1300 pixels, the B-mode images offer an axial resolution of 0.22 mm, an elevational resolution of 0.07 mm, and a lateral resolution of 0.5 mm. The dataset include a specific number of B-mode images, sweeps, and a test-train split, further detailed in Table 6.1. Figure 6.3

shows an example of the acquired sweeps of the lumbar spine phantom. The figure provides a visual representation of the captured ultrasound sweeps, showcasing the details and structures of the lumbar spine phantom.

6.1.3. Ex-vivo Dataset

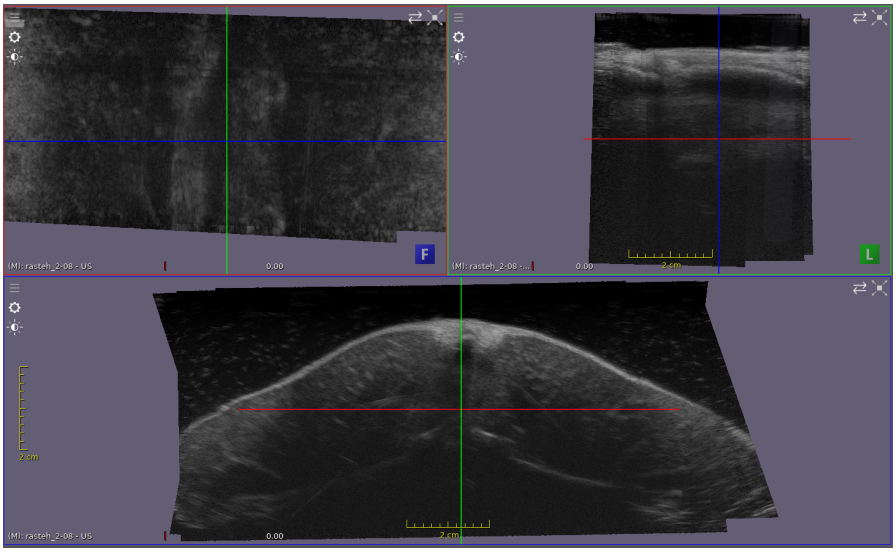

Figure 6.4.: Exvivo data: The figure displays overlaid ultrasound sweeps acquired from the ex-vivo experiment. The phantom used in this study consists of multiple tissue types, such as bone, fat, and muscle, each possessing distinct acoustic properties. The sweeps were obtained using a tilted probe, similarly to the spine phantom, and the resulting effect is visualized in the bottom image.

An animal meat phantom was utilized for the experiments involving ex-vivo B-mode images. The phantom was a lamb spine surrounded by tissues. Lamb phantoms have been widely acknowledged in the literature as effective models for training in needle insertion due to their similarity to human tissues in echogenicity [41]. The spine was placed in a container filled with water to ensure proper probe connection to the surface while tilting the probe. This arrangement avoided any issues arising during imaging due to the lack of contact between the probe and the phantom surface. The ultrasound sweeps were acquired using the same hardware and software setup described in Section 6.1.2. Similar to the spine phantom, the B-mode images were captured in a paramedian sagittal orientation to capture the effect of occlusions caused by vertebrae. The B-mode images had a depth of 60 mm, while the probe itself had

a width of 38 mm. With a resolution of 164x780 pixels, these B-mode images offered an axial resolution of 0.22 mm, an elevational resolution of 0.07 mm, and a lateral resolution of 0.5 mm. Details about the number of sweeps and train-test split can be found in Table 6.1. Figure 6.4 shows example of images acquired on the exvivo phantom.

6.2. Evaluation Methods

6.2.1. Baseline Setup

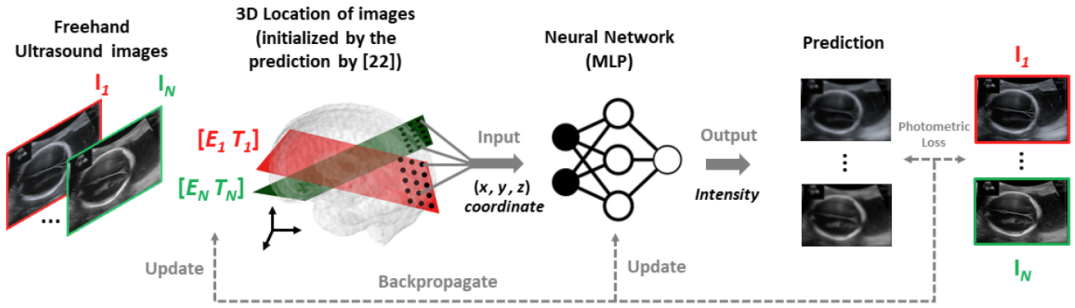

Figure 6.5.: ImplicitVol pipeline. The method uses a direct mapping from 3D coordinates to intensity values for 3D ultrasound reconstruction.

A novel view synthesis is an emerging field in medical ultrasound imaging. To the best of our knowledge, there is currently no benchmark method available for evaluating the performance of such methods. Therefore, in this thesis, we propose evaluating the method's performance by comparing it to two existing methods that can generate new B-mode images of a ROI based on multiple ultrasound sweeps of the same ROI. The first method, in the thesis referred to as the "method without rendering," is adapted from the ImplicitVol framework proposed by Yeung et al. in their work on volume reconstruction [55]. This method employs the MLP network for the task of coordinate-to-intensity mapping and has been used by the authors for volume reconstruction as presented in Figure 6.5. To obtain the baseline, we trained the MLP network introduced in Section 5.3 on coordinate-to-intensity mapping for the same number of epochs as the complete model with the rendering method presented in this thesis. The second baseline method adopted in this thesis involves B-mode generation through the compounding of ultrasound sweeps. Firstly, a volume is compounded using all training sweeps, resulting in a compounded volume (Figure 6.6). Subsequently, the compounded volume is resliced to simulate ultrasound sweeps using the tracking

information from testing sweeps. The ImFusion Suite [3], facilitates the compounding and reslicing steps.

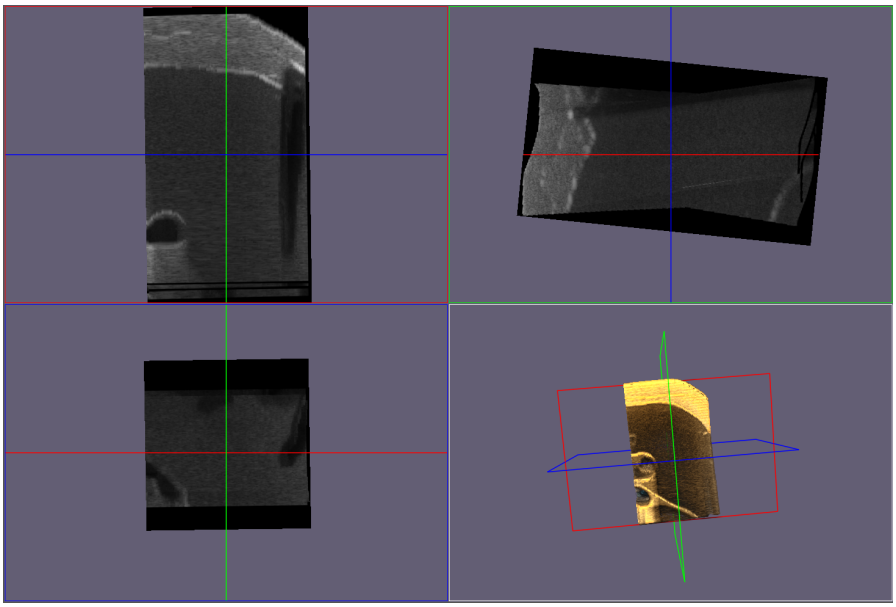

Figure 6.6.: Volume compounding visualization: The figure illustrates the compounded volume, where multiple ultrasound sweeps are combined to create a volume. This compounded volume serves as a basis for reslicing and generating new sweeps using tracking information. The bottom right image displays the compounded volume itself, while the remaining images depict axial, sagittal, and coronal views of the volume, showcasing an effect of multiple sweeps on the final volume.

6.2.2. Evaluation on Downstream Tasks

Volume Compounding

As discussed in Section 2.4, volume compounding is a significant method for reconstructing and visualizing anatomical 3D structures. An essential aspect of the novel views generated using the method proposed in this thesis is the ability to create a compounded volume from the rendered B-mode images of the same quality as real B-mode frames. Therefore, evaluating the rendered sweeps on this task becomes crucial to understand whether rendered B-mode images represent the anatomy in the same way as a real ultrasound system observes it. We compare compounding from real B-mode

images and rendered B-mode images to analyze rendering quality for representing the real anatomy. First, we render sweeps from the novel views using tracking information derived from the training sweeps. Subsequently, we create per-sweep compounded volumes from the rendered B-mode images and the corresponding sweeps for real B-mode images. The evaluation methods employed for assessing the performance of the generated compounded volumes will be further discussed in subsequent sections of this thesis.

Confidence Maps

Confidence maps discussed in Section 2.5 are significant measures in quantifying the uncertainty associated with the information in B-mode images. This uncertainty is often linked to acoustical phenomena, such as acoustic shadows, which the method presented in this thesis aims to preserve. Preserving the uncertainty becomes crucial in rendered images as we expect the rendered images to represent the same information as images captured with an ultrasound device. Therefore, it is essential for the rendered views to accurately capture and retain the inherent uncertainty of the underlying acoustical phenomena. To evaluate the performance of the rendered views in terms of confidence map calculation, we compute confidence maps for each rendered testing sweep. These computed confidence maps are then compared to the ground truth confidence maps, enabling a thorough assessment of the accuracy of the rendered images in preserving the uncertainty associated with the acoustical observations. Detailed discussions regarding the evaluation methodology and results will be presented in subsequent sections of this thesis.

6.2.3. Quantitative Metrics

Jaccarrd Index

The Jaccard Index (J) is a widely used metric for evaluating segmentation tasks in medical imaging [3]. Formally it is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.1}$$

where A and B are two sets, and |A| and |B| represent the cardinality of the sets A and B, respectively. In this thesis, we apply this metric to evaluate the quality of confidence maps computed for rendered images. To compute the J, each confidence map is first transformed into a binary segmentation map, where 0 indicates low certainty and 1 represents high certainty. The binary confidence map is then computed at multiple thresholds $j = 0.1, 0.2, ..., 0.9$. For the evaluation, we report both median and mean

Jaccard similarity scores over all possible threshold values. Formally the evaluation metrics are defined as follows:

Let S be a set of n Jaccard Indices, i.e., $S = J(Y_1, Y_1'), J(Y_2, Y_2'), ..., J(Y_n, Y_n')$. Then, the mean J over S is:

$$\bar{J}(S) = \frac{1}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} J(Y_{ij}, Y_{ij}') \tag{6.2}$$

where $j(Y_i, Y_i')$ is the J between a binary confidence map of a target B-mode (Y_i) and a confidence map of a rendered B-mode (Y_i') a a threshold j .

Let S′ be a set of nm Jaccard Indices, i.e., $S' = J(Y_{11}, Y_{11}'), J(Y_{12}, Y_{12}'), ..., J(Y_{mn}, Y_{mn}')$. Then, the median J over S is:

$$\text{med}(J(S')) = \begin{cases} \frac{1}{2}\left(S'_{\left(\frac{nm}{2}\right)} + S'_{\left(\frac{nm}{2}+1\right)}\right) & \text{if } nm \text{ is even} \\ S'_{\left(\frac{nm+1}{2}\right)} & \text{if } nm \text{ is odd} \end{cases} \tag{6.3}$$

where S′ is sorted in non-decreasing order.

Mutual Information

Mutual Information (MI) is an important measure of independence between two random variables [8]. In the field of medical image registration, it is a popular metric used to assess the similarity or correspondence between images. One of the key assumptions of MI is that similar tissues in one image will correspond to tissues in another image, even if the pixel intensities differ in absolute values [34]. Therefore, MI quantifies the amount of information shared between two images, reflecting their similarity or dissimilarity. The formula for calculating MI between two discrete images, or random variables, is as follows:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{6.4}$$

where $I(X;Y)$ represents the mutual information, $H(X)$ is the entropy of variable X, $H(Y)$ is the entropy of variable Y, and $H(X,Y)$ is the joint entropy of variables X and Y. Alternatively, the mutual information can be defined using the probability mass function (PMF) for the discrete random variable:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{6.5}$$

To calculate MI between two images or random variables as defined by the above euqations, several steps are involved. First, the joint probability distribution needs to be

estimated by creating a two-dimensional histogram that represents the joint distribution. Next, the marginal probability distribution of each image is computed by summing over the corresponding dimensions of the joint distribution. The entropies of each image are then computed by summing the probabilities and multiplying them by their logarithms. Finally, the joint entropy is calculated by summing the joint distribution multiplied by the logarithm of the joint distribution.

Mean Square Error

Mean Square Error (MSE) is a commonly used metric to evaluate the performance of an estimator or model by measuring the average squared difference between the predicted values and the true values. It provides a quantitative measure of the overall deviation between the predicted and actual values.Mathematically, MSE is computed as the average of the squared differences between each predicted value ($\hat{y}$) and its corresponding true value (y):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{6.6}$$

The MSE metric is particularly useful in regression tasks, where the goal is to predict continuous values.

6.2.4. Qualitative Methods

Qualitative methods in medical imaging involve the subjective assessment of the performance of a given method through the evaluation of its output. In this thesis, a qualitative assessment was deployed for both the novel view rendering and the output of the volume compounding, as described in Section 6.2.2. The primary qualitative method employed in this thesis is visual inspection.

For the task of novel view rendering, the images were evaluated based on the domain gap between the rendered images and the ground truth data. Key aspects examined in the rendered images include image quality, similarity to B-mode images, and accuracy in representing anatomical structures. The comparative evaluation approach is adopted during the visual inspection, wherein the rendered images and ground truth data are displayed side by side. This process allows for the identification of any deviations between the renderings and the B-mode images, enabling the interpretation of the utility of the rendered images.

In the case of volume compounding, the compounded volumes are visualized side by side to evaluate whether the method preserves the topology of anatomical structures depending on the viewing direction. This qualitative assessment aims to determine the

accuracy and reliability of the rendering method in preserving the overall anatomical structure across different viewing angles.

6.3. Results and Discussion

6.3.1. Novel View Rendering

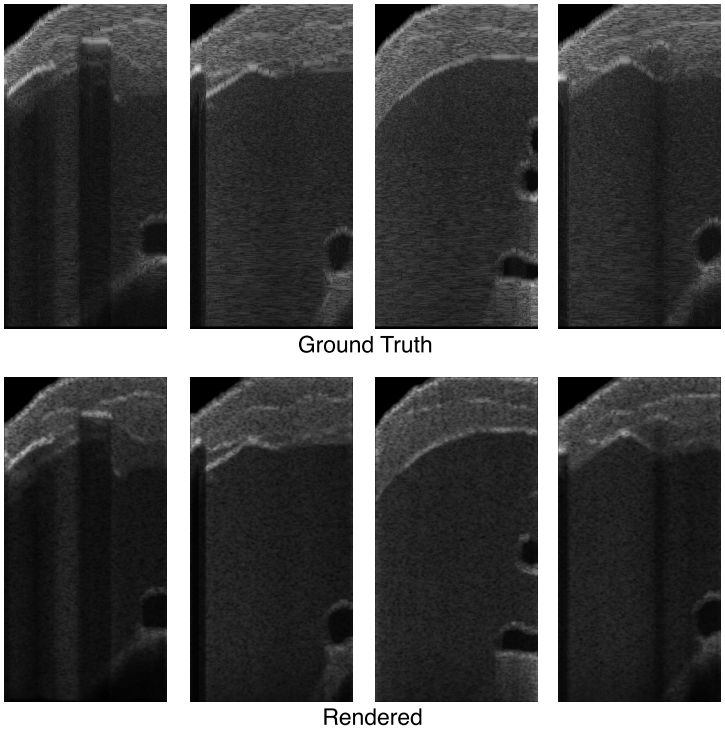

Figure 6.7.: Examples of rendered B-mode images and ground truth B-mode images from the synthetic dataset.

The primary goal of the method presented in this thesis is to achieve novel view rendering that preserves view-dependent acoustical phenomena. As described in Section 6.1, each dataset comprises sweeps that vary in terms of probe orientation. This variation in probe orientation provides a diverse set of viewing directions for each respective ROI. The method was tested on the complete sweeps. The method takes poses from the testing sweeps as input and renders corresponding B-mode images for each input pose. Figure 6.7 displays a selection of randomly sampled rendered B-mode images and corresponding ground truth B-mode images from the simulated dataset.

The visual comparison between real and rendered B-mode images demonstrates that the renderings accurately preserve anatomical structures, effectively capturing the boundaries between different tissue types. Furthermore, the rendering process enables the correct generation of view-dependent acoustic shadows, accurately representing their presence in the rendered images. However, it is necessary to note that the rendered images do not perfectly match the ground truth data regarding B-mode appearance. One notable difference lies in the noise pattern resulting from scattering, which varies between the rendered images and the ground truth data. Figure 6.8 displays a set of

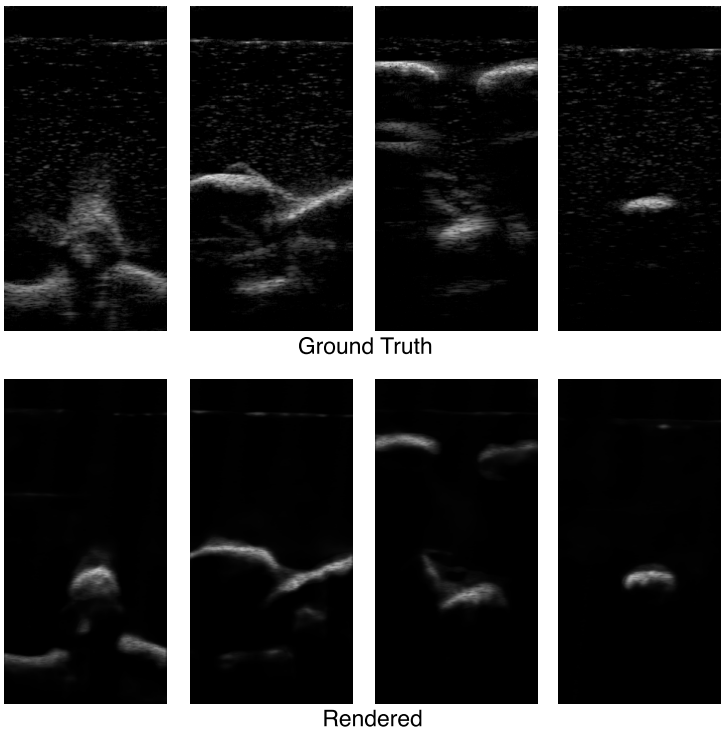

Figure 6.8.: Examples of rendered B-mode images and ground truth B-mode images from the phantom dataset

randomly selected rendered B-mode images and their corresponding ground truth B-mode images obtained from the phantom dataset. Consistent with the findings on the synthetic dataset, the rendered images accurately preserve the anatomical structure, maintaining the boundaries between different tissue types; however, the rendered images do not replicate the appearance observed in real B-mode images. Furthermore, the rendered images exhibit a tendency towards oversmoothing, resulting in a loss of fine details and subtle features present in the ground truth images. Additionally, the

rendered images do not capture complex acoustic phenomena, such as reverberations, commonly observed in real B-mode images. These acoustic phenomena, which arise from multiple reflections of sound waves within tissues, contribute to B-mode images' overall appearance and texture. Therefore in their presence, the domain gap between real and rendered images increases. The results obtained from the ex-vivo dataset,

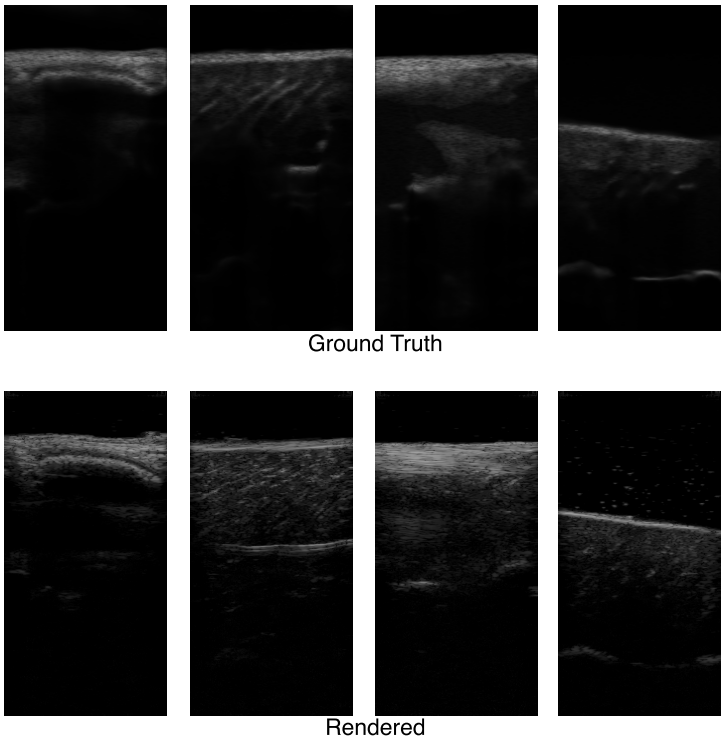

Figure 6.9.: Examples of rendered B-mode images and ground truth B-mode images from the exvivo dataset.

as depicted in Figure 6.9, align with the observations made for both the synthetic and phantom datasets. Specifically, the method successfully identifies the anatomical structure within the rendered images but fails to accurately replicate the appearance observed in real-world ultrasound imaging. This discrepancy is particularly noticeable in tissues with elongated structures, where the rendered images fail to capture the fine details and nuances of their appearance. The method's limitations become more evident when attempting to represent muscle fibers, as the rendered images lack the texture associated with fibers.

Comparison to Baseline Methods

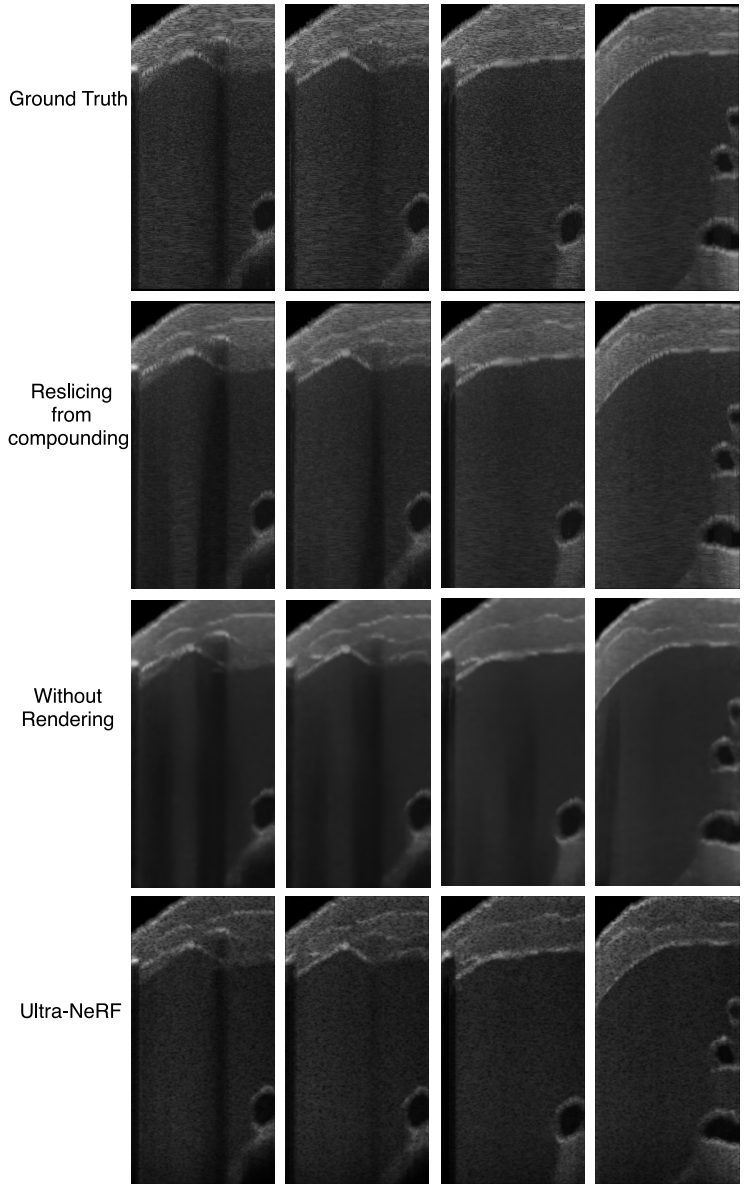

Figure 6.10.: Comparison of B-mode images rendered with physics-inspired rendering and baseline methods, highlighting the advantage of accurate acoustic shadow reconstruction.

Figure 6.10 shows examples of visual comparison of the rendered B-mode images using the proposed physics-inspired rendering method and those rendered using baseline methods. This figure demonstrates the advantages of utilizing physics-inspired rendering for novel view generation. Notably, the baseline methods fail to accurately reconstruct acoustic shadows due to their approach of averaging pixel intensities from multiple images of the same ROI. In contrast, the method presented in this thesis learns to identify the locations of strong reflections and utilizes this knowledge and the direction of observation to generate accurate shadows. This fundamental distinction allows the proposed method to outperform the baseline methods in accurately rendering view-dependent anatomical structures in the B-mode images.

6.3.2. Rendering Parameters Space

The rendering function introduced in Section 5.3.1 relies on input parameters associated with the physical characteristics of the imaged tissue. Therefore, each B-mode image can be decomposed into intermediate maps representing this image in the parameter space. These intermediate parameter maps represent values of respective tissue properties. However, in real-world scenarios, the exact values of these parameters are typically unknown, and approximations are commonly used in simulations and rendering algorithms. As a result, it becomes infeasible to analyze the regressed parameters in absolute terms when working with acquired B-mode images. The lack of precise knowledge about the true parameter values limits the ability to accurately interpret and quantify the regressed parameters. Nevertheless, in the case of the synthetic dataset, the values of these rendering input parameters are known. This is because they have been utilized in the simulation process to generate the synthetic dataset. Knowing their parameters, it becomes possible to evaluate and analyze the performance of the rendering input parameters on the synthetic dataset. An example of the decomposition of a rendered B-mode image in parameter space with respective ground truth values is shown in Figure 6.11. The following sections present an analysis of the regressed rendering parameters and their correspondence to the actual tissue characteristics in the example of synthetic B-mode images.

Attenuation Evaluation

The primary objective of the evaluation is to quantitatively compare the regressed rendering parameters with the corresponding ground truth data. However, it is essential to note that the scattering density model employed in the simulations differs from the model proposed in this thesis. As a result, the evaluation primarily emphasizes the assessment of the attenuation coefficient. For the analysis, four types of models

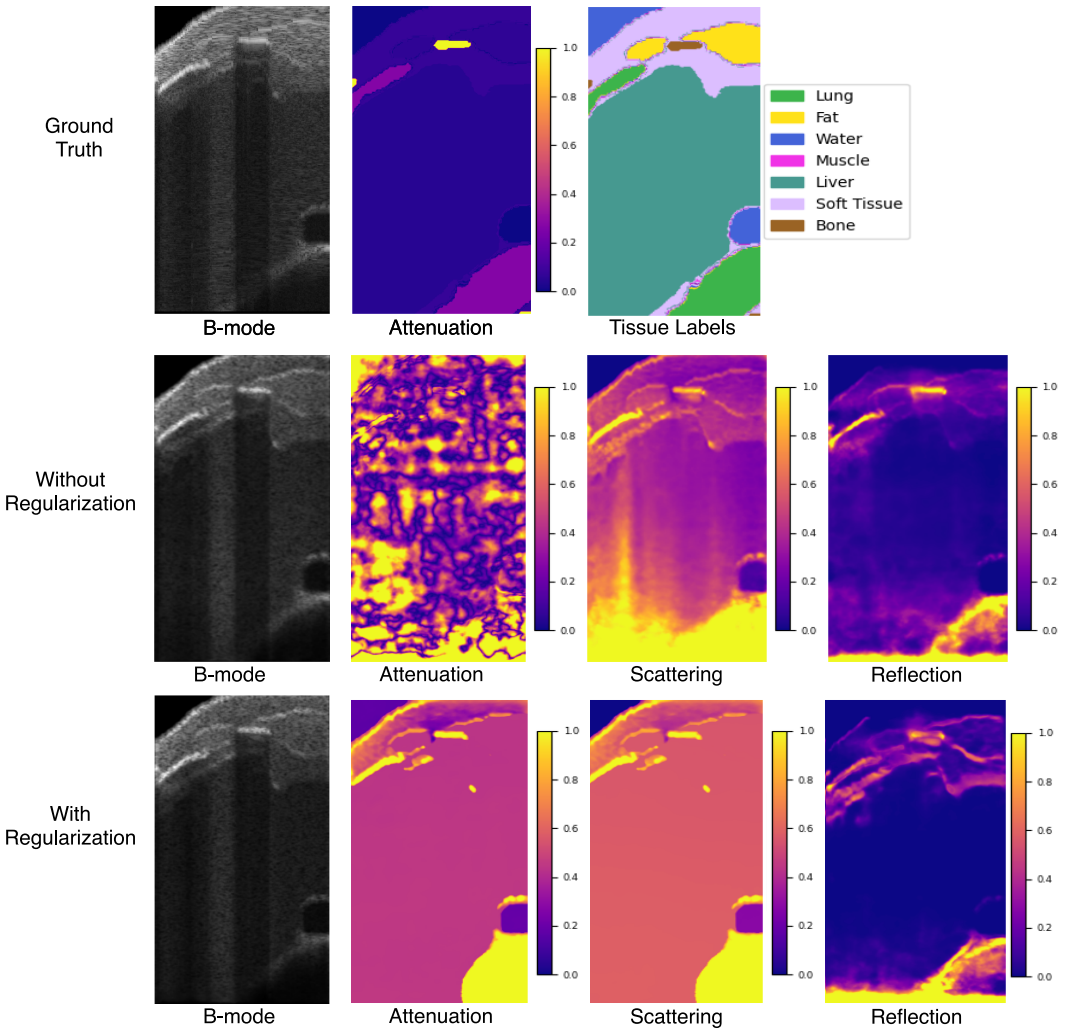

Figure 6.11.: Visualization of rendering parameters for rendering after training with and without regularization.

are considered: A model without any regularization, a model with LNCC as the regularization term, a model with TV regularization, and a model incorporating all the regularization terms discussed in the thesis. Section 5.3.1 explains that exact regression of the attenuation coefficient is infeasible due to the post-processing of B-mode images, which intensifies the image in strongly attenuated regions. the regressed and ground truth attenuation coefficients are normalized for the analysis for each testing sweep. As a result, a method of evaluating the normalized attenuation coefficient is employed. The evaluation uses two metrics: MI and MSE. Although it would be possible to

	vs. GT			
regularization type	$med(MI(S')) \uparrow$	$\bar{MI}(S) \uparrow$	$med(MSE(S')) \downarrow$	$\bar{MSE}(S) \downarrow$
w/o regularization	0.05	0.06	0.03	0.04
TV	0.26	0.21	0.03	0.04
LNCC	0.60	0.61	0.02	0.02
all penalties	0.63	0.63	0.02	0.02

Table 6.3.: MI and MSE between ground truth attenuation coefficient and learned attenuation coefficient for different regularization types

treat the estimation of the rendering coefficient as a typical regression problem, the normalization does not reverse the effects of post-processing since it might involve non-linear changes to the acquired acoustic signal. Consequently, evaluating the regressed coefficients in absolute terms remains infeasible. Therefore, the focus is on comparative evaluation, enabling a comparison between different regularization techniques. The results of this evaluation are compiled in Table 6.3. Based on the obtained results, one can observe that the regressed parameter for attenuation does not align with the ground truth values without regularization. This lack of correspondence is reflected in the MI values, which are close to zero, indicating a lack of statistical dependence between the regressed parameter and its ground truth value. This independence is visually represented in the attenuation map shown in Figure 6.11. However, when incorporating simple regularization techniques such as TV to preserve tissue continuity, there is an increase in MI values. Although the statistical dependence remains relatively weak, this regularization step contributes to a better alignment between the regressed parameter and its ground truth value. The most significant improvement in achieving statistical dependence is observed when employing LNCC as the regularization technique. This regularization method enhances the relationship between the regressed parameter and the ground truth value, resulting in a more substantial increase in the MI. The best overall results are obtained when incorporating all the regularization penalties discussed in the thesis. Each penalty focuses on different aspects and physical characteristics in the parameter space. By combining them, the resulting regressed parameter demonstrates the highest level of statistical dependence with respect to the ground truth values, indicating a stronger correlation with the ground truth values. A similar trend is observed in the MSE, where the parameter estimation error decreases with the addition of regularization. Additional decomposition of the rendered B-modes for both a method with and without regularization and all datasets is included in Appendix A.

Reflection Evaluation

This section concentrates on the evaluation of the regressed reflection coefficient. The reflection coefficient map, shown in Figure 6.11, represents higher coefficient values corresponding to tissue boundaries and lower values corresponding to the locally homogeneous tissue. This is because reflection occurs at the interface of two tissue types with distinct acoustic properties, encoding the scanned tissue's geometry. By sampling the reflection coefficient above a certain threshold, it should be possible to reconstruct the underlying anatomical structure in 3D. To validate that the reflection coefficient encodes tissue geometry, we compared the 3D reconstruction derived from the reflection coefficient with the reconstruction obtained from real B-mode images. As the spine phantom possesses a simple structure, it was chosen for visualization and is presented in Figure 6.12. Only a single view is shown to simplify the visualization of the real B-mode reconstruction. Upon examining the reconstruction generated from the reflection coefficient, one can observe that the coefficient effectively encodes the geometry of the scanned tissue.

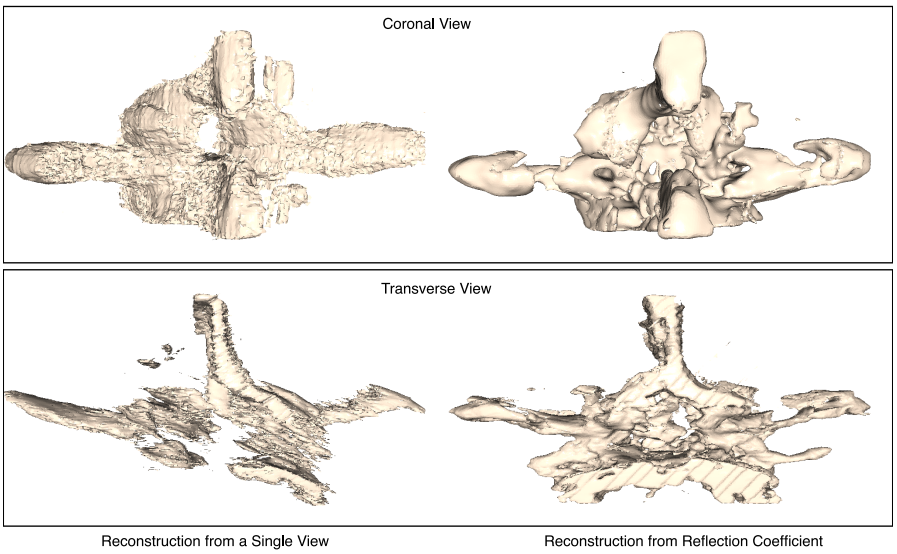

Figure 6.12.: Reconstruction from the reflection coefficient: The geometry of the tissue is encoded in the reflection coefficient.

6.3.3. Compounding Evaluation

An essential aspect of the quality of rendered B-mode images is their ability to accurately depict anatomical structures as if they were imaged with a real ultrasound probe. Therefore a sequence of rendered B-mode images that accurately represent scanned anatomy must enable the reconstruction of a 3D volume that aligns with the reconstruction obtained from real B-mode images captured using the same probe positions. To test this assumption and to assess and emphasize the significance of physics-based rendering in accurately representing view-dependent changes in imaging the anatomical structure, we compare reconstructions from different rendering methods and the ground truth data. To this end, we compounded volumes from testing sweeps from rendered and ground truth data. These compounded volumes were then exported as meshes for enhanced visualization. Figure 6.13 showcases examples of compounded volumes derived from the phantom dataset using the methods mentioned earlier. In both coronal and transverse views, it is noticeable that the absence of physics-informed rendering results in the 3D reconstruction that includes areas not observed in the ground truth reconstruction. This suggests that B-mode images generated without physic-informed rendering contain intensities not present in real B-mode images. In the reconstruction from images rendered with Ultra-NeRF one can observe how physics-informed rendering that leverages information about occlusion from the reflection coefficient. This information aids in generating B-mode images with intensities corresponding exclusively to visible anatomical structures.

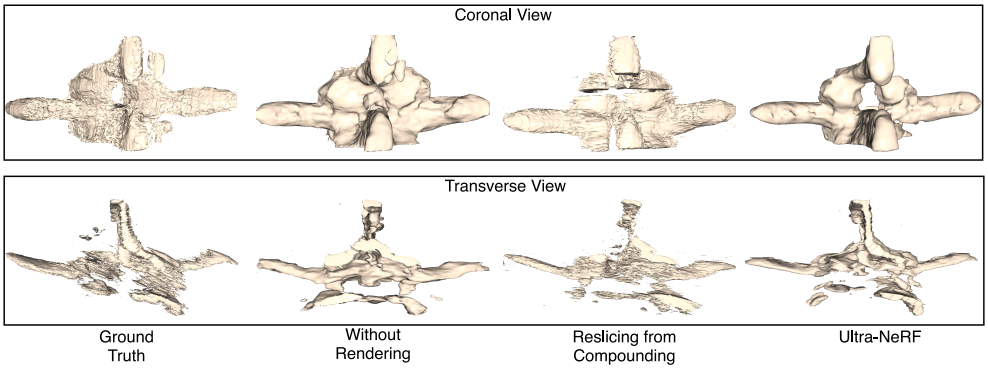

Figure 6.13.: Comparison of compounded volumes from B-mode images rendered with Ultra-NeRF and baseline methods. Because of the observation angle spinous process occludes lamina. This occlusion is reconstructed using physics-informed rendering in Ultra-NeRF, without it the occluded part of lamina is visible in a compounding.

6.3.4. Confidence Maps Evaluation

An essential objective of the rendering method presented in this thesis is that rendered B-modes accurately depict the artifacts that would be present if they were recorded with an actual ultrasound probe. These artifacts play a crucial role in altering the confidence of observations made based on the images. Therefore, assuming an accurate representation of physical phenomena, the rendered images must exhibit uncertainty equivalent to real B-mode images captured from a specific probe pose. To this assumption, we evaluate confidence maps calculated for rendered images by comparing them to respective confidence maps calculated for ground truth B-modes from the testing sets. In the experiments, an implementation of confidence maps available in ImFusion software is utilized. This implementation relies on confidence maps generated using the random walk algorithm [23]. Table 6.4 showcases an evaluation of confidence maps calculated for rendered B-mode images compared to confidence maps derived from the ground truth data of the synthetic dataset. The evaluation was also conducted for baseline

	vs. GT	
method	$med(J(S')) \uparrow$	$\bar{J}(S) \uparrow$
reslicing from compounding	0.61	0.60
without rendering	0.90	0.95
Ultra-NeRF	0.92	0.96
Ultra-NeRF with regularization	0.94	0.97

Table 6.4.: Evaluation of confidence maps and comparison with a method without rendering.

methods to provide a benchmark for comparison. The accuracy of the confidence maps was quantified using the J metric, and the results are reported. In order to facilitate this evaluation, the confidence maps are transformed into confidence masks using a range of thresholds as defined in Section 6.2.3. Between the reslicing from compounded volumes and the renderings with regularized Ultra-NeRF, we observe a 54% increase in the median J and between the simple implicit neural representation 4% increase. This evaluation highlights the benefits of employing a physics-informed rendering technique since the confidence maps calculated for B-mode images rendered with Ultra-NeRF represent more similar uncertainty as measured by J metric. The evaluation reveals that the reslicing method fails to generate B-mode images that closely resemble those captured with a real ultrasound probe. This discrepancy arises because the resliced images are a simplistic composition of all the images capturing a specific ROI. Although a simple implicit neural representation without rendering shows some improvement, it still generates artifacts inconsistent with natural physical phenomena, as depicted in

Figure 6.14. The introduction of physics-informed rendering significantly enhances the consistency of view-dependent artifacts, as exemplified by the improved depiction of shadows in Figure 6.14.

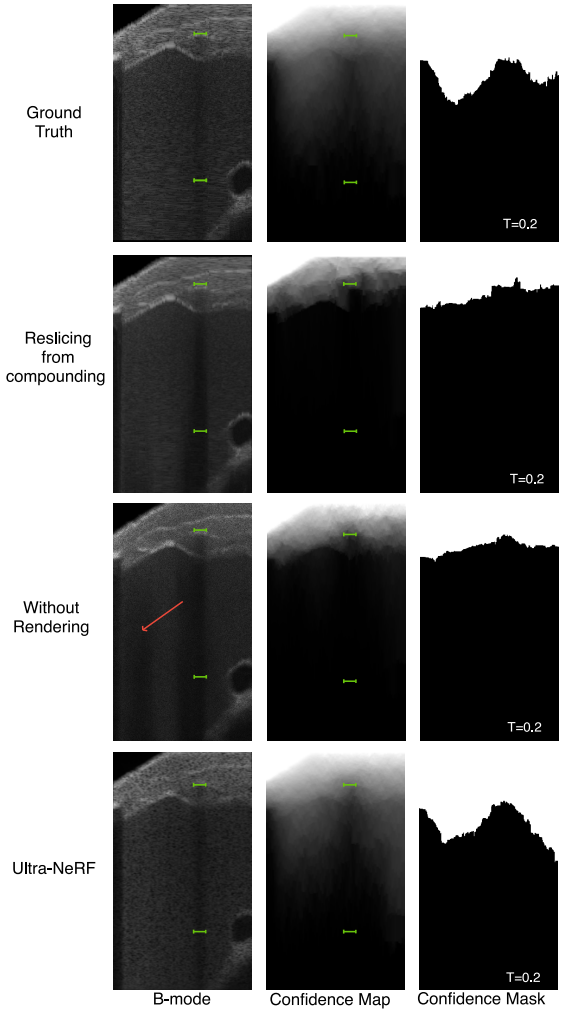

Figure 6.14.: Confidence maps evaluation: The green bar highlights the width of an acoustic shadow caused by a rib. It is evident that without the use of rendering techniques, the width of the shadow is inconsistent with the ground truth B-mode. Furthermore, the absence of rendering introduces additional artifacts, as indicated by the red arrow. These artifacts directly influence the accuracy of uncertainty estimation, as visualized by the difference in estimated confidence maps.

6.4. Limitations

The method presented in this thesis serves as a proof-of-concept for utilizing neural radiance fields in ultrasound. However, it is important to acknowledge the intrinsic limitations of this method, which partially result from assumptions detailed in Section 5.2. Specifically:

- Since the method assumes the static scene and does not model effects resulting from deformations, which are unavoidable in clinical practice due to forces from ultrasound probe pressure or other sources, such as breathing.
- The method assumes perfect tracking, which is difficult to achieve in a clinical setting. It does not compensate for the error of pose estimation, and therefore, errors in pose estimation will negatively impact the accuracy of regressed rendering parameters.
- For the clustering method, we need an estimation of the number of scanned tissue types, which may be unknown and not constant throughout the entire probe trajectory.

Additional limitations resulting from its formulation apply, particularly:

- The rendered B-mode images do not have the same texture as the real ones. This discrepancy in noise pattern arises from inherent differences in the rendering process and the underlying physical characteristics of the imaging system. In particular, the engineered PSF does not allow for recreating real noise patterns, which may vary for different tissue types and imaging systems.
- The regularization introduced in this thesis helps couple rendering parameters and reduces the number of Degrees of Freedom (DoF). However, the rendered parameters cannot be treated as real physical values due to post-processing. They can only be interpreted in a relative manner.
- The model does not consider the impact of the incidence angle on the reflection and assumes the reflection coefficient to be independent of the observation angle.
- Because the reflection coefficient has to be calculated at the tissue boundary and it is difficult to define a constant size of the tissue boundary in the image space, the method does not directly model impedance.
- The method assumes integration over a single ray in 2D. However, when working with volumes, the 2D dimensional dependencies are crucial, for example, to preserve tissue continuity and consistency of rendering parameter estimation between neighboring B-modes.

Presented limitations are essential to consider when to the application of the method in a real-world setting.

7. Outlook

The method presented in this thesis is a valuable proof-of-concept for utilizing NeRF in the ultrasound field, which opens up new avenues for future research and exploration. Based on the limitations outlined in Section 5.2, several potential directions for future investigations can be proposed:

- **Modeling Deformations**: Given that deformations are unavoidable in clinical practice, future research could focus on incorporating information about deformations into the rendering method. This would involve developing techniques to account for tissue deformations caused by factors such as ultrasound probe pressure or patient movement during imaging.
- **Pose Estimation Error Compensation**: Addressing the challenges associated with pose estimation in a clinical setting would significantly enhance the accuracy and robustness of the method. Future studies could explore novel approaches to compensate for errors in pose estimation and reduce the impact of tracking inaccuracies.
- **Enhancing Realism in Rendered Images**: Future work could explore methods to improve scattering pattern similarity to bridge the gap between rendered and real B-mode images. This could involve incorporating more realistic noise patterns and accounting for variations in noise characteristics among different tissue types and imaging systems, for example, by incorporating appearance embedding.
- **Accounting for 2D Dimensional Dependencies and Incidence Angle**: Considering the importance of 3D dependencies when working with volumetric data, future studies could focus on extending the method to explicitly integrate dependencies between consecutive frames. This would facilitate the preservation of volumetric tissue continuity. Moreover, the impact of the incidence angle on the reflection coefficient could be incorporated into the rendering process, allowing for more comprehensive modeling of ultrasound interactions with tissues.

Exploring these research directions would significantly advance the method, improving its applicability, accuracy, and clinical relevance.

8. Conclusion

The primary objective of this thesis was to propose a method that enables novel view synthesis for a ROI from the views not included in acquired ultrasound B-mode images. In particular, this study investigated NeRF in the context of ultrasound imaging and focuses on a rendering method that considers ultrasound physics. The key contributions of this thesis are as follows:

- It proposes a differentiable rendering method specifically designed to match the ultrasound image formation model and is grounded in ultrasound physics principles. This advanced method accurately models the observed echo by independently considering the contributions of backscattering and reflection phenomena. Additionally, it incorporates crucial information regarding reflections and attenuation into the ray transmission process, effectively capturing the loss of intensity based on the tissue's echogenicity. By integrating these elements, this method facilitates the end-to-end training of a deep learning approach, enabling the synthesis of B-mode images from novel perspectives that accurately reconstruct view-dependent changes in the representation of the ROI in real B-mode images.
- The proposed rendering method is incorporated into the NeRF framework, leveraging its capabilities for capturing complex scene geometry and appearance. By combining NeRF with the specific requirements of ultrasound imaging, the study demonstrates its potential for enhancing ultrasound implicit volumetric representation.
- Recognizing that the proposed rendering method involves a highly unconstrained regression problem, this study extends the rendering formulation by introducing regularization techniques in the rendering parameters space. This regularization aims to preserve tissue continuity and establish the interdependency between rendering parameters.

The proposed method has undergone testing on three datasets: simulated liver B-mode images, real B-mode images of a synthetic lumbar spine phantom, and real B-mode images of an ex-vivo spine phantom. The study evaluates B-mode images rendered from probe positions not included in the training dataset to assess the accuracy of

the rendered B-mode images in representing the anatomy from unobserved viewing points. The novel views generated by the rendering method demonstrate its capability to produce view-dependent B-mode images that accurately depict the ROI as if a real ultrasound probe had captured it. However, it is important to note that a domain gap still exists between the rendered B-mode images and real B-mode images, particularly regarding tissue texture. The method was tested on two downstream tasks: confidence map computation and volume compounding to analyze the utility of the rendered B-mode images. The experiments revealed that the method generates B-mode images that convey anatomical information comparable to real ultrasound images, particularly in accurately depicting acoustical shadows and preserving the geometric details of anatomical structures. Furthermore, to highlight the advantages of using a physics-based rendering approach for views that were not captured in the original B-modes, the study conducted a comparison against two baseline methods: a traditional technique involving reslicing planes from a compounded volume and a recently proposed method known as implicit neural representation, which encodes intensities observed in B-mode images for an entire volume using neural networks. This comparison demonstrated that, in contrast to the baseline methods, the method proposed in this thesis renderers view-dependent B-mode images with acoustic shadows that closely resemble those observed in real B-mode images. These findings underscore the efficacy of the proposed physics-based rendering method in generating synthetic B-mode images that accurately represent the underlying anatomy from novel viewing points.

By exploring the domain of physics-based rendering for ultrasound, this study introduces a novel approach that paves the way for rendering ultrasound views. This development holds significant potential for enhancing applications that depend on accurately rendering B-mode images. Specifically, synthesizing novel views offers the opportunity to render cross-sectional planes that may not have been captured during the initial ultrasonographic scanning. Consequently, these additional planes can be thoroughly analyzed even after the scanning procedure, providing healthcare professionals with invaluable insights and potentially revealing crucial information that might have been overlooked during the initial examination.

A. Additional Rendering Parameters

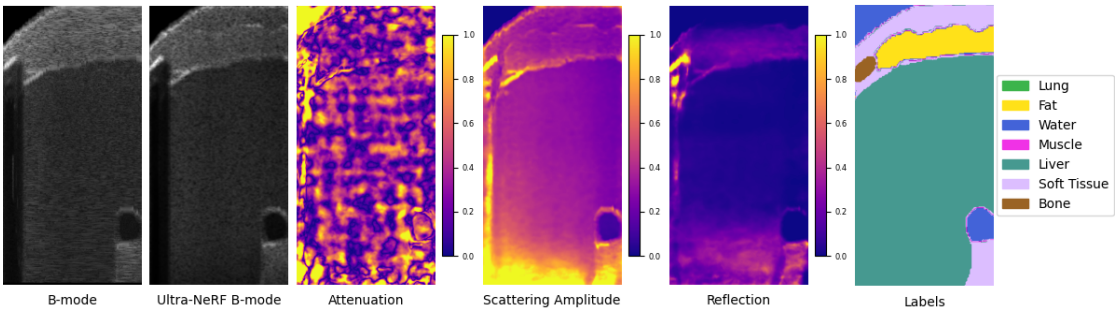

Figure A.1.: Decomposition into rendering parameters for synthetic data and B-mode rendered using Ultra-NeRF without regularization. Even though the rendered B-mode image accurately represents the anatomy we observe inconsistency in the parameter space.

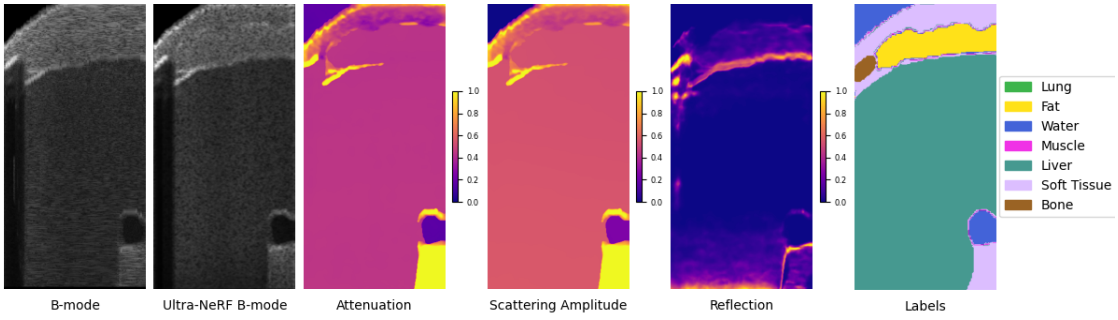

Figure A.2.: Decomposition into rendering parameters for synthetic data and B-mode rendered using Ultra-NeRF with regularization. We observe that regularization enforces consistency between the parameters space and anatomical structure such that the different tissues are visible in the intermediate parameter maps

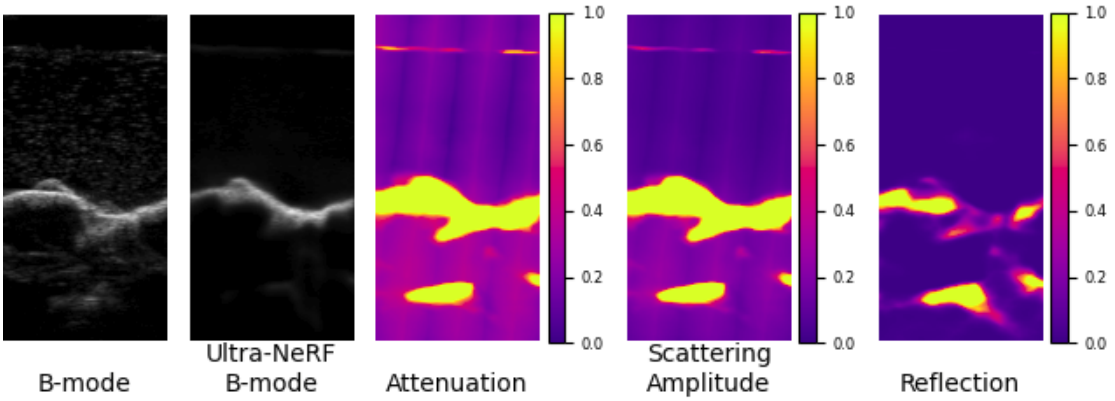

Figure A.3.: Decomposition into rendering parameters for phantom data and Ultra-NeRF with regularization. The spine structure is visible in the parameter maps.

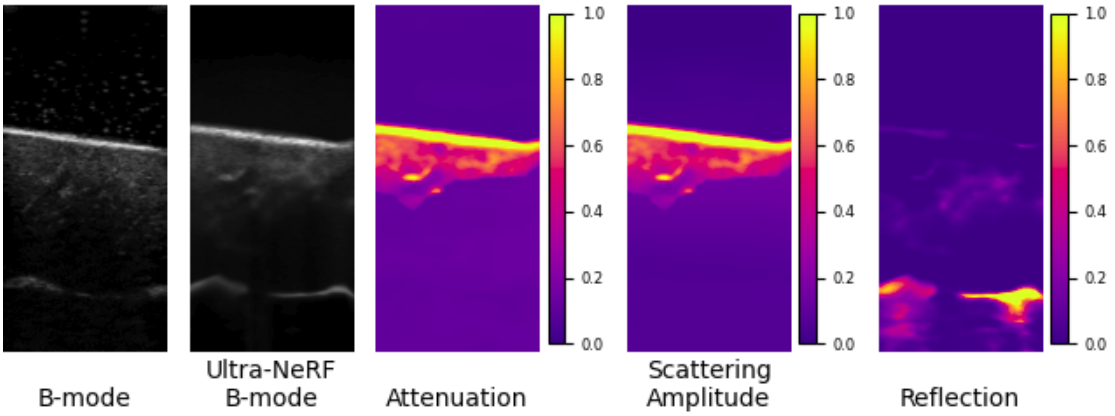

Figure A.4.: Decomposition into rendering parameters for exvivo data and Ultra-NeRF with regularization. The bone is visible in the reflection map, however the scattering and attenuation maps do not accurately represent underlying structure.

Abbreviations

MRI magnetic resonance imaging

MR magnetic resonance

CT computed tomography

NeRF neural radiance fields

Ultra-NeRF neural radiance fields for ultrasound imaging

3D three-dimensional

2D two-dimensional

ROI region of interest

MLP Multilayer Perceptron

ReLU Rectified Linear Unit

J Jaccard Index

PSF point spread function

MI Mutual Information

MSE Mean Square Error

TV Total Variation

LNCC Local Normalized Cross Correlation

SSIM Structural Similarity Index Metric

DoF Degrees of Freedom

List of Figures

List of Tables

Bibliography

[1] Kara-Ali Aliev et al. "Neural point-based graphics." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer. 2020, pp. 696–712.

[2] Penelope J Allisy-Roberts and Jerry Williams. *Farr's physics for medical imaging*. Elsevier Health Sciences, 2007.

[3] Jeroen Bertels et al. "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 92–100.

[4] Benny Burger et al. "Real-Time GPU-Based Ultrasound Simulation Using Deformable Mesh Models." In: *IEEE Transactions on Medical Imaging* 32.3 (2013), pp. 609–618. DOI: 10.1109/TMI.2012.2234474.

[5] Benjamin Busam et al. "Markerless inside-out tracking for 3d ultrasound compounding." In: *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*. Springer, 2018, pp. 56–64.

[6] Pierre Chatelain, Alexandre Krupa, and Nassir Navab. "Confidence-Driven Control of an Ultrasound Probe." In: *IEEE Transactions on Robotics* 33.6 (2017), pp. 1410–1424. DOI: 10.1109/TRO.2017.2723618.

[7] Abril Corona-Figueroa et al. "MedNeRF: Medical Neural Radiance Fields for Reconstructing 3D-aware CT-Projections from a Single X-ray." In: *arXiv preprint arXiv:2202.01020* (2022).

[8] Thomas M Cover and Joy A Thomas. "Information theory and statistics." In: *Elements of information theory* 1.1 (1991), pp. 279–335.

[9] Barbrina Dunmire et al. "Use of the acoustic shadow width to determine kidney stone size with ultrasound." In: *The Journal of urology* 195.1 (2016), pp. 171–177.

[10] Kyle Gao et al. "Nerf: Neural radiance field in 3d vision, a comprehensive review." In: *arXiv preprint arXiv:2210.00379* (2022).

[11] Rüdiger Göbl et al. "Redefining ultrasound compounding: Computational sonography." In: *arXiv preprint arXiv:1811.01534* (2018).

[12] Pierre Hellier et al. "Acoustic shadows detection, application to accurate reconstruction of 3D intraoperative ultrasound." In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2008, pp. 1569–1572.

[13] Christoph Hennersperger et al. "Computational sonography." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part II 18*. Springer. 2015, pp. 459–466.

[14] Christoph Hennersperger et al. "Towards MRI-based autonomous robotic US acquisitions: a first feasibility study." In: *IEEE transactions on medical imaging* 36.2 (2016), pp. 538–548.

[15] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[16] Peter R Hoskins, Kevin Martin, and Abigail Thrush. *Diagnostic ultrasound: physics and equipment*. CRC Press, 2019.

[17] Qinghua Huang and Zhaozheng Zeng. "A review on real-time 3D ultrasound imaging technology." In: *BioMed research international* 2017 (2017).

[18] Khadija Idrissu, Sylwia Malec, and Alessandro Crimi. "3D reconstructions of brain from MRI scans using neural radiance fields." In: *bioRxiv* (2023), pp. 2023–04.

[19] Jo/rgen Arendt Jensen. "A model for the propagation and scattering of ultrasound in tissue." In: *The Journal of the Acoustical Society of America* 89.1 (1991), pp. 182–190.

[20] Zhongliang Jiang et al. "Automatic normal positioning of robotic ultrasound probe based only on confidence map optimization and force measurement." In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 1342–1349.

[21] Zhongliang Jiang et al. "Deformation-aware robotic 3D ultrasound." In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 7675–7682.

[22] Zhongliang Jiang et al. "Towards autonomous atlas-based ultrasound acquisitions in presence of articulated motion." In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 7423–7430.

[23] Athanasios Karamalis et al. "Ultrasound confidence maps using random walks." In: *Medical image analysis* 16.6 (2012), pp. 1101–1112.

[24] Muhammad Osama Khan and Yi Fang. "Implicit Neural Representations for Medical Imaging Segmentation." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*. Springer. 2022, pp. 433–443.

[25] Antoine Leroy et al. "Rigid registration of freehand 3D ultrasound and CT-scan kidney images." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004: 7th International Conference, Saint-Malo, France, September 26-29, 2004. Proceedings, Part I 7*. Springer. 2004, pp. 837–844.

[26] M. Levoy. "Display of surfaces from volume data." In: *IEEE Computer Graphics and Applications* 8.3 (1988), pp. 29–37. DOI: 10.1109/38.511.

[27] Honggen Li et al. "3D Ultrasound Spine Imaging with Application of Neural Radiance Field Method." In: *2021 IEEE International Ultrasonics Symposium (IUS)*. IEEE. 2021, pp. 1–4.

[28] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. "The concrete distribution: A continuous relaxation of discrete random variables." In: *arXiv preprint arXiv:1611.00712* (2016).

[29] J. Meunier and M. Bertrand. "Ultrasonic texture motion analysis: theory and simulation." In: *IEEE Transactions on Medical Imaging* 14.2 (1995), pp. 293–300. DOI: 10.1109/42.387711.

[30] Mateusz Michalkiewicz et al. "Implicit Surface Representations As Layers in Neural Networks." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[31] Ben Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020. URL: http://arxiv.org/abs/2003.08934v2.

[32] Richard A Newcombe et al. "Kinectfusion: Real-time dense surface mapping and tracking." In: *2011 10th IEEE international symposium on mixed and augmented reality* (2011), pp. 127–136.

[33] Jeong Joon Park et al. "Deepsdf: Learning continuous signed distance functions for shape representation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 165–174.

[34] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. "Mutual-information-based registration of medical images: a survey." In: *IEEE transactions on medical imaging* 22.8 (2003), pp. 986–1004.

[35] Raphael Prevost et al. "Deep learning for sensorless 3D freehand ultrasound imaging." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 628–636.

[36] Nasim Rahaman et al. "On the spectral bias of neural networks." In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5301–5310.

[37] Albert W Reed et al. "Dynamic ct reconstruction from limited views with implicit neural representations and parametric motion fields." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2258–2268.

[38] Gernot Riegler and Vladlen Koltun. "Free view synthesis." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer. 2020, pp. 623–640.

[39] Robert Rohling, Andrew Gee, and Laurence Berman. "Three-dimensional spatial compounding of ultrasound images." In: *Medical Image Analysis* 1.3 (1997), pp. 177–193.

[40] Mehrdad Salehi et al. "Patient-specific 3D ultrasound simulation based on convolutional ray-tracing and appearance optimization." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 510–518.

[41] Jasmine Samuel et al. "The use of joints of meat as phantoms for ultrasound-guided needling skills: a prospective blinded study." In: *The Ultrasound Journal* 14.1 (2022), p. 14.

[42] Walter Simson et al. "Deep Learning Beamforming for Sub-Sampled Ultrasound Data." In: *2018 IEEE International Ultrasonics Symposium (IUS)*. 2018, pp. 1–4. DOI: 10.1109/ULTSYM.2018.8579818.

[43] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. "Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/b5dc4e5d9b495d0196f61d45b26ef33e-Paper.pdf.

[44] Vincent Sitzmann et al. "DeepVoxels: Learning Persistent 3D Feature Embeddings." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[45] Vincent Sitzmann et al. "Implicit neural representations with periodic activation functions." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7462–7473.

[46] Thomas L Szabo. *Diagnostic ultrasound imaging: inside out*. Academic press, 2004.

[47] Ayush Tewari et al. "Advances in neural rendering." In: *Computer Graphics Forum*. Vol. 41. 2. Wiley Online Library. 2022, pp. 703–735.

[48] Ayush Tewari et al. "State of the art on neural rendering." In: *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, pp. 701–727.

[49] Mateo Villa et al. "FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images." In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 1707–1716.

[50] Wolfgang Wein et al. "Simulation and fully automatic multimodal registration of medical ultrasound." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007: 10th International Conference, Brisbane, Australia, October 29-November 2, 2007, Proceedings, Part I 10*. Springer. 2007, pp. 136–143.

[51] Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. "Implicit neural representations for deformable image registration." In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2022, pp. 1349–1359.

[52] Robert Wright et al. "Complete fetal head compounding from multi-view 3D ultrasound." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer. 2019, pp. 384–392.

[53] Qing Wu et al. "Irem: High-resolution magnetic resonance image reconstruction via implicit neural representation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 65–74.

[54] Junshen Xu et al. "NeSVoR: Implicit Neural Representation for Slice-to-Volume Reconstruction in MRI." In: *IEEE Transactions on Medical Imaging* (2023).

[55] Pak-Hei Yeung et al. "ImplicitVol: Sensorless 3D Ultrasound Reconstruction with Deep Implicit Representation." In: *arXiv preprint arXiv:2109.12108* (2021).

[56] Lin Zhang, Valery Vishnevskiy, and Orcun Goksel. "Deep network for scatterer distribution estimation for ultrasound image simulation." In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 67.12 (2020), pp. 2553–2564.