

# NeRF-US : Removing Ultrasound Imaging Artifacts from Neural Radiance Fields in the Wild

Rishit Dagli <sup>1,2</sup>

RISHIT@CS.TORONTO.EDU

Atsuhiro Hibi <sup>2,3,4</sup>

ATSUHIRO.HIBI@MAIL.UTORONTO.CA

Rahul G. Krishnan <sup>1,5</sup>

RAHULGK@CS.TORONTO.EDU

Pascal N. Tyrrell <sup>2,4,6</sup>

PASCAL.TYRRELL@UTORONTO.CA

*Departments of <sup>1</sup> Computer Science; <sup>2</sup> Medical Imaging, University of Toronto, Canada*

*<sup>3</sup> Division of Neurosurgery, St Michael’s Hospital, Unity Health Toronto, Canada*

*<sup>4</sup> Institute of Medical Science; Departments of <sup>5</sup> Laboratory Medicine and Pathobiology; <sup>6</sup> Statistical Sciences, University of Toronto, Canada*

[rishitdagli.com/nerf-us/](http://rishitdagli.com/nerf-us/)

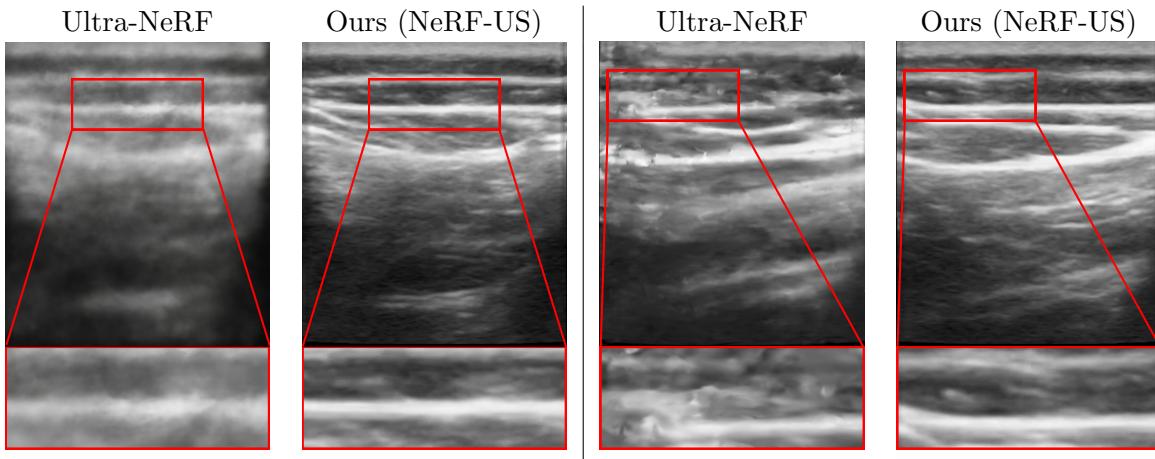


Figure 1: Comparison of novel ultrasound views rendered by Ultra-NeRF (Wysocki et al., 2024) (left) and our NeRF-US (right). These reconstructions are of the human knee featuring an anteroposterior view (left) and a lateral view (right). Our approach, NeRF-US, significantly improves the quality of the reconstructions, making these suitable for any subsequent downstream clinical tasks.

## Abstract

Current methods for performing 3D reconstruction and novel view synthesis (NVS) in ultrasound imaging data often face severe artifacts when training NeRF-based approaches. The artifacts produced by current approaches differ from NeRF floaters in general scenes because of the unique nature of ultrasound capture. Furthermore, existing models fail to produce reasonable 3D reconstructions when ultrasound data is captured or obtained casually in uncontrolled environments, which is common in clinical settings. Consequently,

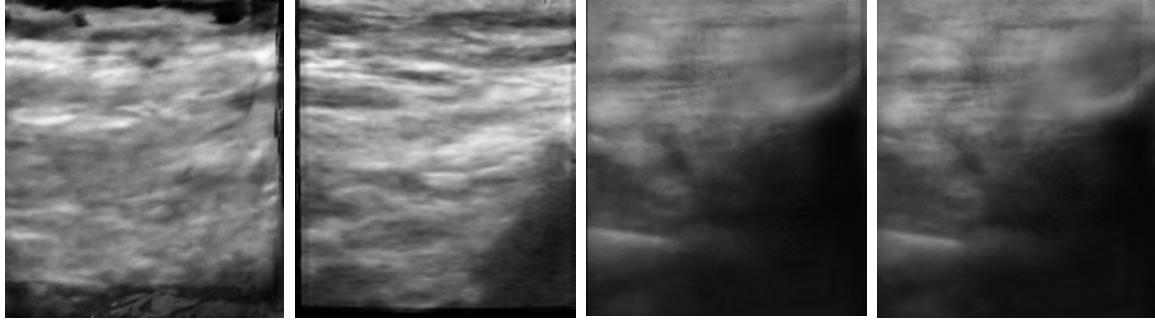


Figure 2: We observe the existence of multiple NeRF artifacts in ultrasound imaging which is a common challenge in medical NeRF-based methods.

existing reconstruction and NVS methods struggle to handle ultrasound motion, fail to capture intricate details, and cannot model transparent and reflective surfaces. In this work, we introduced NeRF-US, which incorporates 3D-geometry guidance for border probability and scattering density into NeRF training, while also utilizing ultrasound-specific rendering over traditional volume rendering. These 3D priors are learned through a diffusion model. Through experiments conducted on our new “Ultrasound in the Wild” dataset, we observed accurate, clinically plausible, artifact-free reconstructions.

## 1. Introduction

All imaging systems benefit from capturing the 3D geometry of the scenes being imaged. Capturing the 3D geometry of scenes is crucial in medical imaging, not only for accurate diagnosis but also for effective treatment planning. This is particularly vital in conditions like hemophilic arthropathy, where joint bleeds lead to synovial proliferation and effusion, and the measurement of synovial fluid volume is a common diagnostic step. Among the key indicators of disease activity in the joints is synovial recess distention (SRD), which may arise from various factors such as accumulation of blood or synovial fluid (Heilmann et al., 1996). Similarly, diagnosing conditions such as scoliosis requires understanding the volumetric properties of the spine (Kadoury et al., 2007, 2009). Producing robust and accurate 3D representations while performing view synthesis from 2D medical imaging techniques is an important problem that can be used for planning and downstream clinical tasks (Thomas W. Hash, 2013; Jackson et al., 1988; Ji et al., 2011).

We are interested in the problem of 3D reconstruction and novel view synthesis for ultrasound imaging. We focus our efforts on working with ultrasound imaging because it is one of the most cost-effective and accessible forms of medical imaging. We are interested in performing this task on casually captured or in the wild ultrasounds since all ultrasounds captured by clinicians are casually captured. Furthermore, performing this task with ultrasound imaging is also a very challenging task due to the nature of how ultrasounds are captured. This task has traditionally been performed by training clinicians to mentally assemble a 3D image. There have been some digital techniques towards the creation of such 3D ultrasound images which relied on utilizing advanced wobbler probes (Morgan et al.,

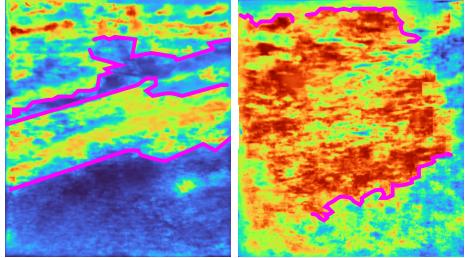


Figure 3: We show a depth map of two reconstructions produced by our approach, NeRF-US. We superimpose the depth map with tissue boundaries as shown by color-coded curves. Our approach produces accurate representations with the tissue boundaries unambiguously reconstructed.

2018), 2D transducers (Smith et al., 2002), or tracking probes (Poon and Rohling, 2005). These kinds of approaches are often based upon assembling a 3D image from multiple 2D slices and use a lot of handcrafted priors. Thus, these classes of approaches often have many limitations, primarily these approaches are often unable to produce plausible 3D reconstructions that could be used for downstream tasks (Kojcev et al., 2017). Such manual and digital approaches end up being very costly, error-prone, and irreproducible (Kojcev et al., 2017; Lyshchik et al., 2004).

Some recently learned digital approaches based on neural radiance field (Mildenhall et al., 2021) methods have also been developed that work toward the problem we highlight (Wysocki et al., 2024; Gaits et al., 2024; Li et al., 2021a). These kinds of approaches have also successfully been used in other medical contexts like reconstructing CT projections from X-ray (Corona-Figueroa et al., 2022a) and surgical scene 3D reconstruction (Wang et al., 2022; Zha et al., 2023) which have shown impressive results. However, using such methods for ultrasound imaging does not produce artifact-free 3D representations from casually captured ultrasounds.

There are a few challenges common across all these learned digital methods: the need for high-quality diverse datasets that capture intricate details like tissue interface locations in high detail, accurately modeling transparent and reflective surfaces, specular surface rendering, and delineating boundaries between different tissues. We qualitatively show some NeRF artifacts that appear when Ultra-NeRF is adopted (Wysocki et al., 2024) on ultrasound images in the wild in Figure 2. These challenges are common across all medical NeRF-based methods. In contrast, our approach, NeRF-US, produces artifact-free reconstructions with minor details that are accurately reconstructed, as shown in Figure 3.

Towards the problem of producing high-quality ultrasound reconstructions in the 'wild', we propose our approach, NeRF-US. Our goal is to produce a 3D representation given a set of ultrasound images taken in the wild and their estimated camera or ultrasound probe positions. We first train a 3D denoising diffusion model which serves as geometric priors for the reconstruction. We then train a NeRF model that takes in a 3D vector (denoting positions in 3D) and learns a 5D vector (attenuation, reflectance, border probability, scattering density, and scattering intensity) following the success of Ultra-NeRF (Wysocki et al., 2024). While training this NeRF, we incorporate the geometric priors from the diffusion model to guide the outputs for border probability and scattering density. This allows our approach, NeRF-US, to accurately generate 3D representations for ultrasound imaging in the 'wild', as shown in Figure 1. We also evaluate our approach with a new dataset for this task, which we will release publicly. To the best of our knowledge, ours is the first approach

that handles this challenging task of reconstructing noniconic, nonideal ultrasound images in the ‘wild’ and also outperforms previous methods.

Our approach is universal in ultrasound imaging and our experiments include human spine and knee joint ultrasound. We observe that current evaluation and management strategies for painful musculoskeletal episodes in patients with hemophilia often rely on subjective evaluations, leading to discrepancies between patient-reported symptoms and imaging findings. Ultrasonography (US) has emerged as a valuable tool for assessing joint health due to its accessibility, safety, and ability to detect soft tissue changes, including joint recess distension, with precision comparable to magnetic resonance imaging (MRI). However, accurately interpreting US images, especially in identifying and quantifying recess distention, remains challenging for clinicians. Providing clinicians with a 3D representation of recess distension acquired from US cine loops could greatly enhance their ability to appreciate the extent of fluid accumulation within the joint space. This information is crucial for guiding treatment decisions, especially regarding the initiation and optimization of hemostatic therapy and the utilization of adjunctive treatments such as physical therapy and anti-inflammatory agents. Thus, our new “Ultrasound in the wild” dataset is focused on ultrasounds around the knee for the recess distention problem.

### **Generalizable Insights about Machine Learning in the Context of Healthcare**

Addressing the challenge of casually captured ultrasounds, which is the common practice among clinicians today, NeRF-US integrates specific ultrasound properties within the NeRF framework. This integration not only surpasses existing methods in producing artifact-free 3D reconstructions, but also expands the potential applications beyond routine medical diagnostic tasks. For example, the ability to generate accurate and artifact-free 3D models from ultrasound data can significantly enhance surgical planning and postoperative evaluations. In addition, our approach aims to increase the use of cost-effective methods such as ultrasound, offering a viable alternative to more expensive imaging techniques for certain medical scenarios.

**Contributions.** Modern 3D reconstruction methods for ultrasound imaging are either overly reliant on hand-crafted priors, manual intervention, or are unable to work on ultrasound imaging in the ‘wild’. The key novelty of our approach stems from the modification of NeRF-based methods for this task.

- We propose a first-of-its-kind approach for training NeRFs on ultrasound imaging that incorporates the properties of ultrasound imaging and incorporates 3D priors through a diffusion model to reconstruct accurate images in uncontrolled environments.
- We introduce a new dataset, “Ultrasound in the Wild”, featuring real, non-iconic ultrasound imaging of the human knee. This dataset will serve as a resource for benchmarking NVS performance on ultrasound imaging under real-world conditions and will be made publicly available.
- Our approach demonstrates improved qualitative and quantitative performance by significantly reducing or eliminating ultrasound imaging artifacts. This leads to more accurate 3D reconstructions compared to other medical and non-medical NeRF-based

methods. Furthermore, we observe clinically plausible 3D reconstructions from ultrasound imaging in the ‘wild’.

We open-source our code and data on our project webpage.

## 2. Related Works

**Implicit Neural Representations and View Synthesis.** Historically, the field of novel view synthesis relied on traditional techniques like image interpolation (Chaurasia et al., 2013; Chen and Williams, 2023, 1993; Debevec et al., 2023, 1996; Fitzgibbon et al., 2005; Seitz and Dyer, 1996; McMillan and Bishop, 2023, 1995) and light field manipulation (Sloan et al., 1997; Levoy and Hanrahan, 1996) to generate new views without needing to understand the geometry of the scene. These methods worked best with densely sampled scenes, which limited their usability. The advent of learned techniques, popularized using image blending (Flynn et al., 2016; Hedman et al., 2018) and multiplane images (Li et al., 2020; Mildenhall et al., 2019). Further progress involved creating explicit 3D scene representations through meshes (Debevec et al., 1998; Riegler and Koltun, 2020; Shan et al., 2013), point clouds (Qi et al., 2017; Aliev et al., 2020; Bui et al., 2018; Grossman and Dally, 1998), and voxel grids (Kutulakos and Seitz, 1999; Penner and Zhang, 2017; Seitz and Dyer, 1999), offering a better understanding of scene geometry but with many challenges in model accuracy and robustness.

The introduction of neural implicit representations (INRs) marked a change in representing 3D scenes, offering a more adaptable and comprehensive method for encoding both the geometric and appearance aspects of scenes (Sitzmann et al., 2019, 2020). Utilizing neural networks as a foundation, this approach allowed for the implicit modeling of scene geometry, with the capacity to generate novel views through ray-tracing techniques. By interpreting a scene as a continuous neural-based function, INRs map 3D coordinates to color intensity. Among the diverse neural representations for 3D reconstruction and rendering, the success of deep learning has popularized methods like Level Set-based representations, which map spatial coordinates to a signed distance function (SDF) (Park et al., 2019; Jiang et al., 2020; Yariv et al., 2020) or occupancy fields (Mescheder et al., 2019).

One of the most popular INR methods is the Neural Radiance Field (NeRF) (Mildenhall et al., 2021), an alternative implicit representation that directly maps spatial coordinates and viewing angles to local point radiance. NeRFs can also perform novel view synthesis through differentiable volumetric rendering, utilizing only RGB supervision and known camera poses. Furthermore, there have also been many extensions of NeRF-based methods for dynamic scenes (Pumarola et al., 2020; Gafni et al., 2020; Park et al., 2020), compression (Takikawa et al., 2022), editing (Liu et al., 2021; Jiakai et al., 2021; Liu et al., 2022), and more.

**NeRFs in the wild.** Applying NeRFs in uncontrolled environments often results in the emergence of artifacts, often referred to as *floaters*. These are small, disconnected regions in volumetric space that inaccurately represent parts of the scene when viewed from different angles, appearing as blurry clouds or distortions in the scene which we demonstrate in Figure 4. These artifacts are primarily observed under conditions such as sub-optimal camera registration (Warburg et al., 2023), sparse input sets (Liu et al., 2023; Warburg et al.,



Figure 4: Floater artifacts often appear in NeRF-based models for casually-captured videos.

2023), strong view-dependent effects (Liu et al., 2023), and inaccuracies in scene geometry estimation. Additionally, the very nature of NeRF’s approach, which reconstructs scene geometry and texture from 2D projections of a 3D scene causes information loss.

To mitigate floater artifacts in NeRF reconstructions, several strategies focus on improving camera pose estimation and scene geometry. Mip-NeRF 360 (Barron et al., 2022) and RegNeRF (Niemeyer et al., 2022) enhance alignment and consistency, with the former also incorporating a distortion loss to prevent floaters. Techniques such as Nerfbusters (Warburg et al., 2023) and ViP-NeRF (Somraj and Soundararajan, 2023) leverage visibility information to better reconstruct scenes with sparse data, while CleanNeRF (Liu et al., 2023) offers a method to separate view-dependent effects for more accurate geometry estimation. Additionally, previous works have also explored post-processing methods to remove floaters without modifying the original NeRF framework (Jambon et al., 2023; Wirth et al., 2023; Goli et al., 2023).

**NeRFs for Medical Imaging.** The application of NeRFs in medical imaging is confronted with unique challenges that stem from the inherent properties of medical imaging data and the demand for accurate representations. These include the need for detailed inner structure representation, the handling of ambiguous object boundaries, the significance of color density variations in medical images, and the adaptation to different imaging principles compared to traditional NeRF applications (Wang et al., 2024).

Recent advances have demonstrated that NeRF-based techniques, such as MedNeRF (Corona-Figueroa et al., 2022b) and UMedNeRF (Hu et al., 2024), as effective in reconstructing CT projections from single X-ray views, showcasing their potential to minimize exposure to ionizing radiation in medical imaging. ACNeRF (Sun et al., 2024) further enhances reconstruction quality through improved alignment and pose correction.

There have also been other specialized NeRFs either for a particular body part or a kind of imaging technique. In brain imaging, advances such as 3D reconstruction from MRI scans using NeRFs (Iddrisu et al., 2023) aim to improve diagnostics and patient care by providing more precise and less invasive diagnostic methods. Similarly, in dental and maxillofacial imaging, Masked NeRF (Zhou et al., 2023) has been introduced to address challenges related to skull CBCT reconstructions, offering refined methods to mitigate artifacts and improve pose estimation accuracy. For cardiovascular imaging, there has been work on 3D reconstruction of coronary angiography images (Zha et al., 2022; Maas et al., 2023) and

improved diagnostic capabilities. Similarly, NeRF-based models have been developed for feet (Zha et al., 2022), chest CT imaging (Corona-Figueroa et al., 2022b; Hu et al., 2024; Sun et al., 2024; Maas et al., 2023), and abdominal surgical planning (Wang et al., 2022).

Furthermore, there has been similar work to ours, leveraging ultrasounds to perform 3D reconstruction. Particularly, there have been a few studies performing 3D reconstruction on ultrasound images (Li et al., 2021b; Yeung et al., 2021; Gu et al., 2022; Song et al., 2022). Ultra-NeRF (Wysocki et al., 2024) is another popular technique that encodes the physics of ultrasound into volume rendering and demonstrates promising results for liver and spine ultrasounds. However, such methods still pose multiple challenges in adopting NeRF-based methods for ultrasounds. These methods qualitatively produce multiple ultrasound imaging artifacts, are unable to capture intricate details like tissue boundaries and all ultrasound imaging performed by clinicians inherently has some body motion in it which these methods do not account for. These limit such methods from being applied in the wild. However, our method, NeRF-US, tackles these challenges common across all current studies.

### 3. Preliminary

#### 3.1. Diffusion Models

The denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) in the forward process takes an input image  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ , and progressively in the  $T$  steps add Gaussian noise to the image. This is implemented using a Markov chain of  $T$  steps.

$$\begin{aligned} q(\mathbf{x}_t | \mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \mu_t = \sqrt{1 - \beta_t} \cdot \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \\ &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \end{aligned} \quad (1)$$

At each step of the Markov chain, the forward process adds a Gaussian noise with variance  $\beta_t$  to  $\mathbf{x}_{t-1}$ ,  $\bar{\alpha}_t = \prod_{s=0}^t 1 - \beta_s$ ,  $\beta_t$  is a hyper-parameter representing a variance schedule, and produces a new latent variable  $\mathbf{x}_t$ . The reverse process of diffusion models aims to recover the data distribution from the Gaussian noises by approximating the posterior distribution  $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$  as,

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \quad (2)$$

This only requires approximating the mean  $\mu_\theta(\mathbf{x}_t, t)$  by training some neural network  $\epsilon_\theta(\mathbf{x}_t, t)$ , this network can be trained using the optimization objective,

$$\mathcal{L}_t = \mathcal{E}_{\mathbf{x}_0, t, \epsilon} \left[ \left\| \epsilon - \theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right] \quad (3)$$

#### 3.2. Neural Radiance Fields

Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021) use a single 5-D coordinate as input  $(x, y, z, \theta, \phi)$  representing the spatial location and viewing angle, and outputs  $(r, g, b, \sigma)$  representing color intensities and volume density. NeRFs usually output different representations for the same point when viewed from different camera angles which allows us to capture various lighting effects as well. To train these networks without ground truth

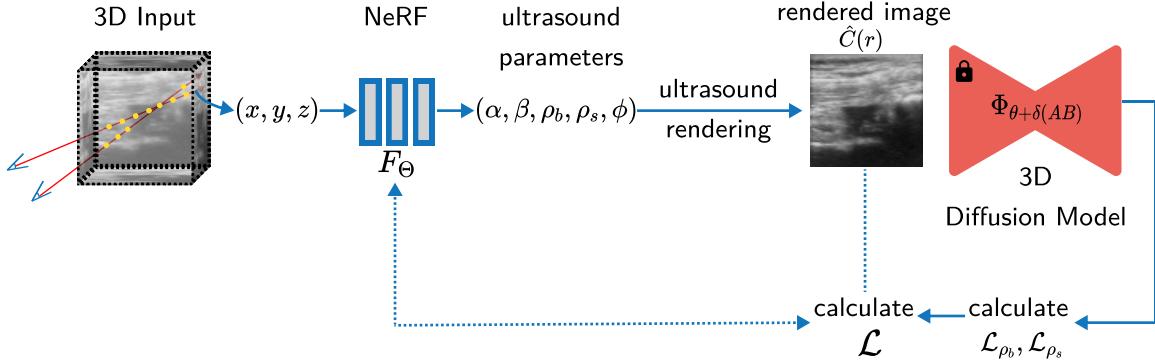


Figure 5: **An overview of how our method works.** We train a NeRF model (Section 4.2) that uses ultrasound rendering to convert the representations into a 2D image after which we infer through a 3D diffusion model (Section 4.1) which has geometry priors through which we calculate a modified loss definition to train the NeRF.

density and color, we sample pixels from the original images, using ray marching. For a given pixel we have the ray,

$$r(t) = o + td \quad (4)$$

where  $o$  represents the origin and  $d$  the direction, which we can sample at timesteps  $t$ . To map these back to an image we can integrate these rays (differentiable rendering),

$$C(r) = \int_{t_n}^{t_f} \underbrace{T(t)}_{\text{transmittance}} \cdot \underbrace{\sigma(r(t))}_{\text{density}} \cdot \underbrace{c(r(t), d)}_{\text{color}} dt \quad (5)$$

We can now train the network by simply computing  $L_2$  loss since from Equation (5) we have a way to map the neural field output back to a 2D image.

## 4. NeRF-US

Our goal is to produce a 3D representation given a set of ultrasound images taken in the ‘wild’ and their camera positions. Our approach on a high level involves modifying volumetric rendering while training NeRFs, and using a Diffusion Model to produce clean artifact-free representations as summarized in Figure 5. We show our approach for training the diffusion model (Section 4.1) and our approach on training the NeRF model (Section 4.2). Notice that, while our training approaches utilize data around the human knee, our work is not limited to this single application. It can be extended right out-of-the-box to any kind of medical ultrasound imaging. We limit our experiments to just the human knee to be able to collect sufficient data and to be able to evaluate our approach satisfactorily.

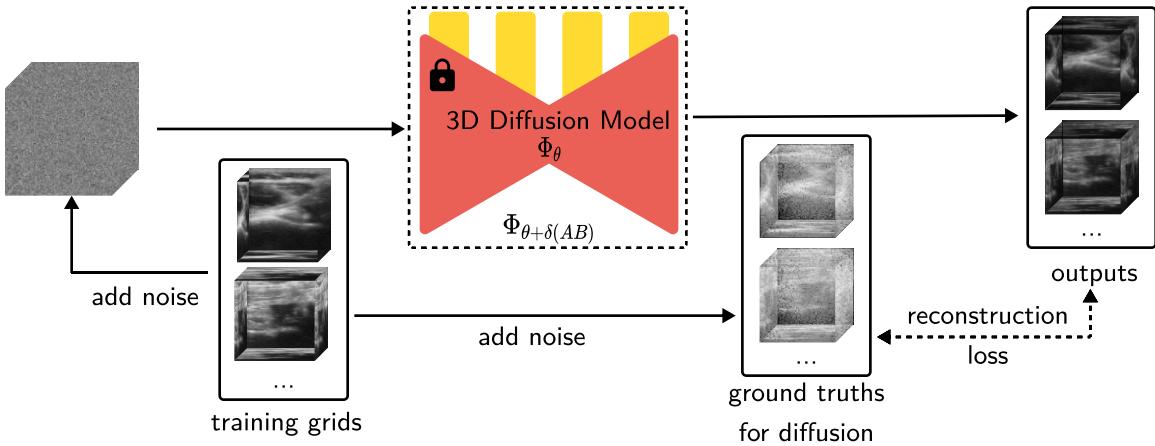


Figure 6: **Training our Diffusion Model.** An overview of how our diffusion model is finetuned, we use  $32^3$ -sized patches to LoRA-finetune a 3D diffusion model trained on ShapeNet (Chang et al., 2015).

#### 4.1. Training the Diffusion Model

The first step of our approach relies on the training of a 3D diffusion model, which can serve as geometric priors for our NeRF model (Section 4.2).

Similar to Nerfbusters (Warburg et al., 2023), this diffusion model produces an  $32 \times 32 \times 32$  occupancy grid,  $x$ . We use the 3D diffusion model  $\Phi$  with parameters  $\theta$  trained on ShapeNet (Chang et al., 2015), from Nerfbusters (Warburg et al., 2023). Based on LoRA (Hu et al., 2022), our fine-tuning process now utilizes the same loss function as the original DDPM training, with the low-rank update applied to the model’s parameters,

$$\mathcal{L}_{\text{FT}} = \|x - \Phi_{\theta+\delta(AB)}(\bar{x}_t + (1 - \bar{\beta}_t), t)\|_2^2 \quad (6)$$

where  $x$  is the target occupancy grid,  $\bar{x}_t$  is the noised version of  $x$  at timestep  $t$ ,  $\delta$  is a scaling factor determining the magnitude of adaptation and  $AB$  represents a low-rank update to the original weights, and  $\bar{\beta}_t$  controls the noise level as per the noise schedule. The model  $\Phi_{\theta+\delta(AB)}$  indicates the fine-tuned model. We make sure to only update the parameters of  $A$  and  $B$  matrices are updated during this process, leaving the original model parameters  $\theta$  fixed.

We finetune the 3D diffusion model on a small dataset of voxels around the human knee generated synthetically. From these synthetic knees, we extracted localized, cube-bounded patches that represent the area of interest for ultrasound imaging. We particularly ensure that all the voxels we sample have one of their ends on human skin mimicking ultrasound imaging conditions. These cubes are variably sized proportional to the knee model’s bounding volume and are voxelized and scaled. We finetune the model  $\Phi$  as summarized in Figure 6, on these  $32 \times 32 \times 32$ -sized voxels to produce an adapted model  $\Phi_{\theta+\delta(AB)}$  which we use as our 3D geometric priors.

## 4.2. Training the NeRF

Following previous ultrasound representations (Salehi et al., 2015; Wysocki et al., 2024), the parameter vector for some position  $q = (x, y, z)$  is  $[\alpha(q), \beta(q), \rho_b(q), \rho_s(q), \phi(q)]$  where  $\alpha$  is the attenuation,  $\beta$  is the reflectance,  $\rho_b$  is border probability,  $\rho_s$  is the scattering density, and  $\phi$  is the scattering intensity. Just like standard NeRF models, we employ an MLP to learn the mapping,

$$\mathbf{F}_\Theta : q \rightarrow [\alpha, \beta, \rho_b, \rho_s, \phi]$$

Ultra-NeRF (Wysocki et al., 2024) for the volume rendering defines the intensity as

$$I(r, t) = I_0 \cdot \prod_{n=0}^{t-1} [(1 - \beta(r, n)) \cdot G(r, n)] \cdot \exp \left( - \int_{n=0}^{t-1} (\alpha \cdot f \cdot dt) \right), \quad (7)$$

where  $\beta(r, t)$  represents the reflection coefficient,  $I_0$  represents the initial unit intensity,  $G(r, t)$  represents a boundary mask, and  $(\alpha f dt)$  represents the loss of energy due to attenuation at each step of the propagation with  $f$  as the frequency. Composing these intensities a 2D ultrasound image is generated which can be trained using the standard NeRF techniques.

We particularly modify this training process inspired by the area of work that incorporate diffusion models in the NeRF training (Wynn and Turmukhambetov, 2023; Warburg et al., 2023; Chen et al., 2023; Gu et al., 2023; Yang et al., 2023), however incorporating these with the rendering process is not straightforward. We do so leveraging the diffusion model we trained  $\Phi_{\theta+\delta(AB)}$  to refine key ultrasound parameters: border probability  $\rho_b$ , scattering density  $\rho_s$  for which we can efficiently use the 3D geometric priors. Given a set of ultrasound parameters  $\mathcal{G}_i = \{\rho_b, \rho_s\}$  for each voxel  $i$ , we process these through the diffusion model to generate predictions on the expected values of these parameters. We then denote the two new loss terms as,

$$\begin{aligned} \mathcal{L}_{\rho_b} &= \frac{1}{N} \sum_{i=1}^N |\mathcal{G}_i^{\rho_b} - m_i^{\rho_b}|^2 \\ \mathcal{L}_{\rho_s} &= \frac{1}{N} \sum_{i=1}^N |\mathcal{G}_i^{\rho_s} - m_i^{\rho_s}|^2, \end{aligned} \quad (8)$$

where  $\rho_b$  represents the border probability,  $\rho_s$  represents the scattering density,  $N$  represents the number of points being sampled,  $m_i^{\rho_b}$  represents the value of  $\rho_b$  from the output of the diffusion model  $\Phi_{\theta+\delta(AB)}$  for a voxel  $i$ , and  $m_i^{\rho_s}$  represents the value of  $\rho_s$  from the output of the diffusion model  $\Phi_{\theta+\delta(AB)}$  for a voxel  $i$ . We find that formulating the two loss components as shown in Equation (8) to perform particularly well since many of the underlying objects captured by ultrasound imaging can often contain opaque, translucent, and transparent in the same scene (note: that our voxels are rather small and often only contain a single kind of object).

We now write our final loss definition for the NeRF training as

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \underbrace{\left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2}_{\text{move output close to ground-truth image}} + \overbrace{\lambda_{\rho_b} \mathcal{L}_{\rho_b}}^{\text{move output close to correct border prob.}} + \overbrace{\lambda_{\rho_s} \mathcal{L}_{\rho_s}}^{\text{move output close to correct scattering dens.}} \quad (9)$$

where  $\mathcal{R}$  is the set of rays in a given batch,  $\lambda_{\rho_b}$ ,  $\lambda_{\rho_s}$  are the weighting factors,  $C(\mathbf{r})$  is the ground truth frame,  $\hat{C}(\mathbf{r})$  is the frame obtained after rendering which we get by Equation (7). It is also straightforward to notice that incorporating our approach does not require any overhead during inference. Furthermore, once the diffusion model is trained, there is very little overhead (caused by inferencing the diffusion model) while training the NeRF.

## 5. Experiments

### 5.1. Ultrasound in the wild Dataset

We collect the data following a standardized protocol using a hand-held ultrasound device (Butterfly iQ+ by Butterfly Network Inc., Burlington, MA, USA<sup>1</sup>). We capture video and mid-sagittal images, including the suprapatellar longitudinal view of the suprapatellar recess of the knee. Using this, our pilot dataset captures 10 unique casual sweeps at 30 FPS on a subject around the human knee with at least 85 frames in each sweep. All sweeps in this dataset have been captured by the authors on healthy knees following institutional REB (Research Ethics Board) guidelines; we provide details about REB guidelines in Appendix D. Following this, the camera registration is performed by COLMAP (Schönberger et al., 2016; Schönberger and Frahm, 2016). To create testing frames, we use every 8th frame. A popular way to evaluate such models is to use another camera across a different trajectory to create the test frames, often with two imaging equipment attached to each other. However the nature of how ultrasounds are captured in the wild make it quite difficult to build a setup that could allow us to collect this data.

### 5.2. Experimental Setup

We evaluate the performance of our approaches based on the novel view synthesis quality on the “Ultrasound in the wild” Dataset (human knee) we introduced in Section 5.1 and the “phantom dataset” (human spine) (Wysocki et al., 2024). The generated images are compared with the ground truth test view images to calculate a few commonly used quantitative metrics that have become the standard way to evaluate NeRFs: PSNR, MS-SSIM (Wang et al., 2003), and LPIPS (Zhang et al., 2018). We report the average values of these quantitative metrics across multiple frames, we follow the evaluation from most previous techniques (Mildenhall et al., 2021). We compare our model with the baseline models of ImplicitVol (Yeung et al., 2021), Gu et al. (2022), DCL-Net (Guo et al., 2020), Ultra-NeRF (Wysocki et al., 2024), as well as with standard NeRF methods: Original NeRF (Mildenhall et al., 2021), Instant-NGP (Müller et al., 2022), TensoRF (Chen et al.,

---

1. <https://www.butterflynetwork.com/iq-plus>

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
DCL-Net (Guo et al., 2020)	14.38 ( $\pm 0.89$ )	0.4418 ( $\pm 0.16$ )	0.5618 ( $\pm 0.17$ )
ImplicitVol (Yeung et al., 2021)	18.42 ( $\pm 0.83$ )	0.6182 ( $\pm 0.16$ )	0.5482 ( $\pm 0.06$ )
Original NeRF (Mildenhall et al., 2021)	14.90 ( $\pm 1.06$ )	0.5516 ( $\pm 0.03$ )	0.5371 ( $\pm 0.04$ )
Gu et al. (2022)	15.01 ( $\pm 1.13$ )	0.5724 ( $\pm 0.18$ )	0.5092 ( $\pm 0.07$ )
Instant-NGP (Müller et al., 2022)	22.73 ( $\pm 4.06$ )	0.7399 ( $\pm 0.11$ )	0.2842 ( $\pm 0.08$ )
TensoRF (Chen et al., 2022)	25.14 ( $\pm 6.16$ )	0.7548 ( $\pm 0.14$ )	0.2716 ( $\pm 0.11$ )
Nerfacto (Tancik et al., 2023)	15.04 ( $\pm 1.76$ )	0.4631 ( $\pm 0.05$ )	0.4645 ( $\pm 0.07$ )
Gaussian Splatting (Kerbl et al., 2023)	22.33 ( $\pm 4.59$ )	0.7746 ( $\pm 0.13$ )	0.2683 ( $\pm 0.10$ )
Ultra-NeRF (Wysocki et al., 2024)	24.14 ( $\pm 2.54$ )	0.8312 ( $\pm 0.15$ )	0.2573 ( $\pm 0.16$ )
Ours (NeRF-US)	27.37 ( $\pm 1.83$ )	0.8412 ( $\pm 0.13$ )	0.2325 ( $\pm 0.14$ )

Table 1: **Quantitative Results.** We show quantitative comparisons between our NeRF-US against the baselines of our model on our “Ultrasound in the wild” dataset. We report the average PSNR  $\uparrow$ , SSIM  $\uparrow$ , LPIPS  $\downarrow$  metrics across all scenes. The best, second best, and third best results for each metric are color coded.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Original NeRF (Mildenhall et al., 2021)	21.36 ( $\pm 1.27$ )	0.5100 ( $\pm 0.12$ )	0.2415 ( $\pm 0.14$ )
Instant-NGP (Müller et al., 2022)	26.13 ( $\pm 1.35$ )	0.5052 ( $\pm 0.05$ )	0.2324 ( $\pm 0.06$ )
TensoRF (Chen et al., 2022)	27.16 ( $\pm 3.17$ )	0.5178 ( $\pm 0.03$ )	0.2342 ( $\pm 0.05$ )
Nerfacto (Tancik et al., 2023)	26.28 ( $\pm 2.63$ )	0.5124 ( $\pm 0.04$ )	0.2353 ( $\pm 0.04$ )
Gaussian Splatting (Kerbl et al., 2023)	29.72 ( $\pm 4.27$ )	0.5248 ( $\pm 0.12$ )	0.2214 ( $\pm 0.11$ )
Ultra-NeRF (Wysocki et al., 2024)	28.12 ( $\pm 2.35$ )	0.5314 ( $\pm 0.09$ )	0.2182 ( $\pm 0.10$ )
Ours (NeRF-US)	29.24 ( $\pm 2.13$ )	0.5306 ( $\pm 0.05$ )	0.1963 ( $\pm 0.13$ )

Table 2: **Quantitative Results.** We show quantitative comparisons between our NeRF-US against the baselines of our model on the “phantom” dataset from Wysocki et al. (2024). However, the captures in this dataset are not in the wild. We report the average PSNR  $\uparrow$ , SSIM  $\uparrow$ , LPIPS  $\downarrow$  metrics across all scenes. The best, second best, and third best results for each metric are color coded.

2022) and Nerfacto (Tancik et al., 2023). We also compare our method with standard Gaussian Splatting (Kerbl et al., 2023). We provide more details about the implementation in Appendix A.

### 5.3. Results and Discussion

**Quantitative Results.** We demonstrate the quantitative results of our approach against other methods for novel views on our “Ultrasound in the wild” dataset. We compare our

approach quantitatively across methods that are specialized for 3D ultrasound reconstruction as well as state-of-the-art standard 3D reconstruction methods which are not specific to ultrasound or medical imaging. We particularly compare with state-of-the-art 3D reconstruction methods which are not focused on ultrasound to demonstrate our approach’s significance for reconstruction such sound-based imaging. We show these quantitative results and comparisions in Table 1 and Table 2. We also demonstrate the quantitative results on the “phantom” dataset (Wysocki et al., 2024). However, this dataset does not feature in the wild captures.

**Qualitative Results.** We present the qualitative results of novel view synthesis in Figure 7. We particularly show that other approaches tend to reconstruct ultrasound scenes with severe geometric artifacts. The rendered ultrasound scenes are blurry or torn apart along the moving trajectory. We particularly also notice that other approaches especially Ultra-NeRF (Wysocki et al., 2024) produce reconstructions where tissue borders and separation are not captured properly due to a lot of uncertainty in these areas by all previous approaches. We also demonstrate some of these issues in Figure 8. Our approach particularly alleviates these issues with the reconstruction of 3D ultrasound representations.

#### 5.4. Ablations

We provide an in-depth analysis motivating our two-fold training approach highlighted in Section 4 by ablating each component and evaluating the use of each component. We provide ablations in Figure 9 and Table 3.

**Border Probability Guidance (w/o  $\mathcal{L}_{\rho_b}$ ).** To implement this we simply set  $\lambda_{\rho_b}$  to 0 for the entire training process in Equation (9). We find that the border probability guidance is useful for accurately modeling tissue interface locations and border locations; if we disable it during training, we observe that many objects blend into each other and have discontinuities in the representation space.

**Scattering Density Guidance (w/o  $\mathcal{L}_{\rho_s}$ ).** To implement this we simply set  $\lambda_{\rho_s}$  to 0 for the entire training process in Equation (9). We find that the scattering density guidance is useful to accurately model bodily objects; if we disable it during the training, we observe that many microstructural features are missing from the reconstructions. For example, in Figure 9, many minute details are entirely missing when we do not use scattering density guidance.

**Ultrasound Rendering (w/o  $I(t)$ ).** Technically, learning a NeRF model without the ultrasound rendering (Equation (7)) is possible, to implement this we still calculate the rendered 2D image with ultrasound rendering since the results from this are used to calculate  $\mathcal{L}_{\rho_b}$  and  $\mathcal{L}_{\rho_s}$  terms, however, we parallelly also learn another model which is trained with the objective function as Equation (9) but uses standard volumetric rendering to calculate  $\hat{Cr}$  and we report the metrics for this NeRF model. We find that using ultrasound rendering is useful to model ultrasound effects on the image; if we disable it during the training, we observe that while individual frames look decently reconstructed, the 3D geometry of the underlying objects is improper due to the nature of how ultrasounds capture data.

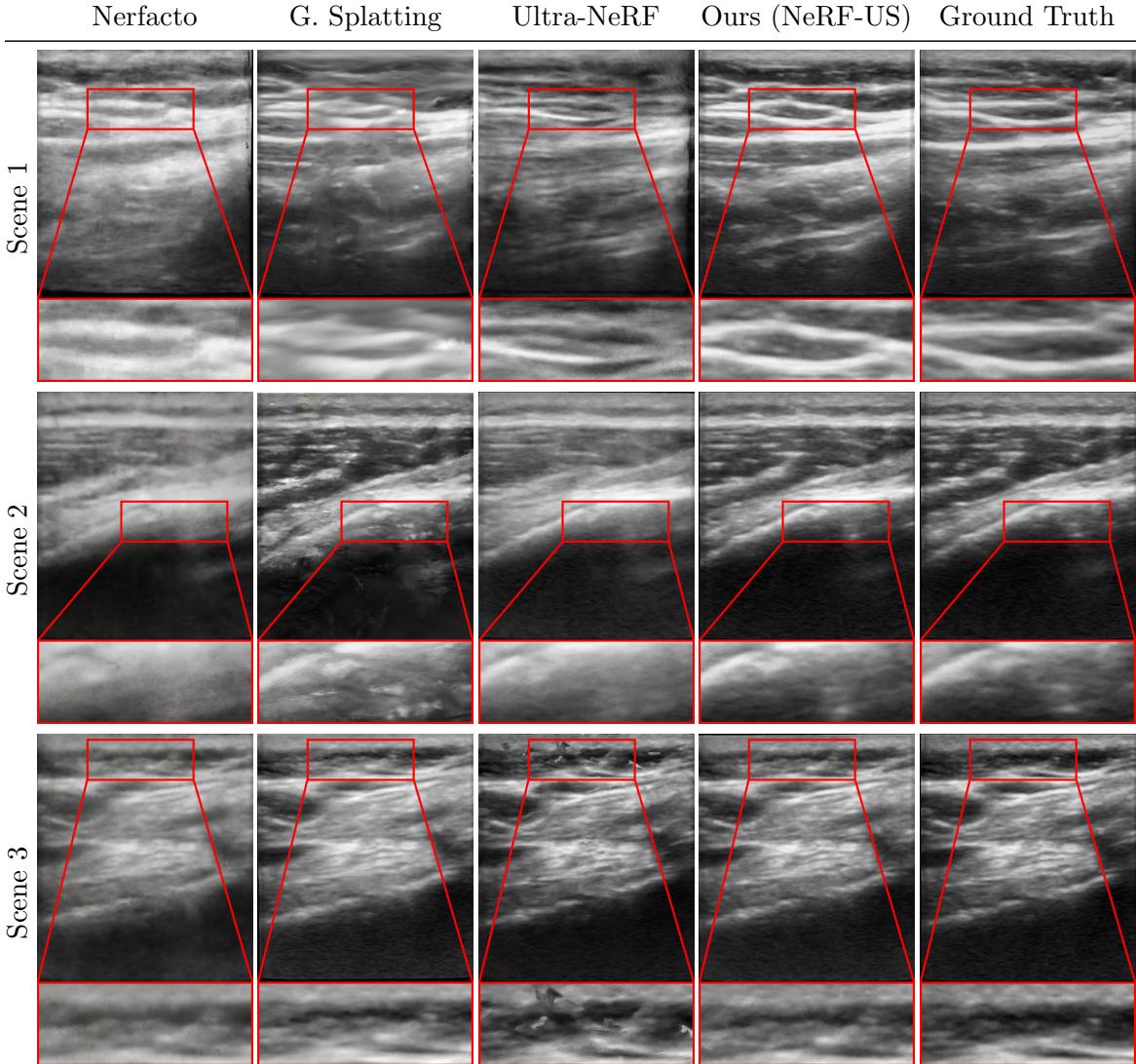


Figure 7: **Qualitative Results.** We demonstrate the results of our method and compare it qualitatively with Nerfacto (Tancik et al., 2023), Gaussian Splatting (Kerbl et al., 2023), and Ultra-NeRF (Wysocki et al., 2024). Our approach, NeRF-US, produces accurate and high-quality reconstructions as compared to the baseline models on novel views (**best viewed with zoom**).

## 6. Limitations

Although our method produces compelling reconstructions of ultrasound imaging, there are several limitations and avenues for future work. First, the complexity of motion in our scenes is limited to simple movements, while our approach is accurately able to reconstruct ultrasounds taken in the natural setting which always has some movement, our method fails

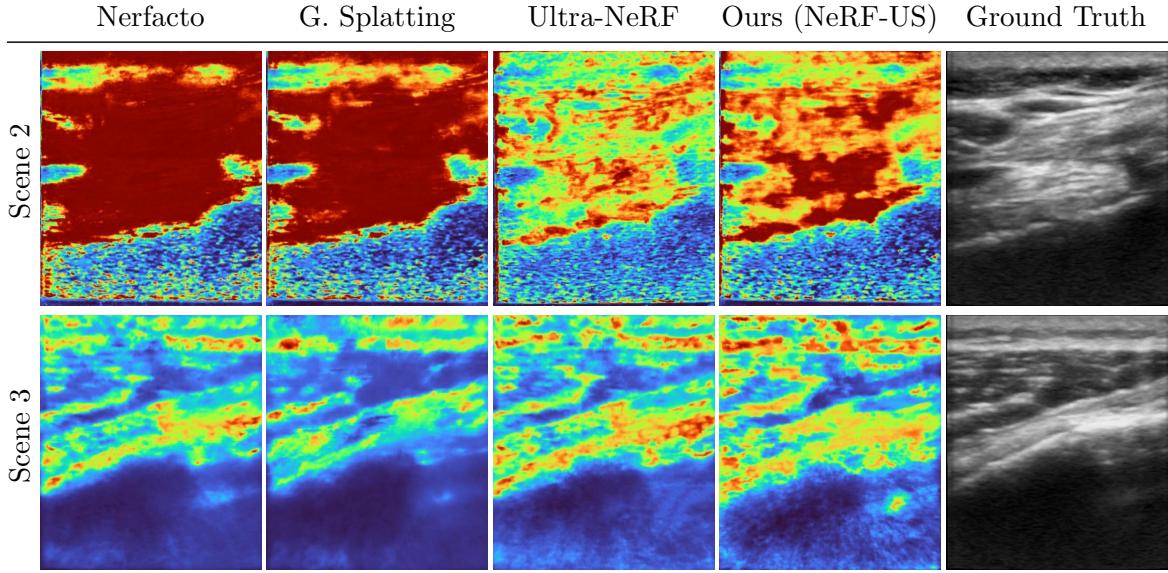


Figure 8: **Qualitative Results.** We demonstrate the results of depth maps produced from our method and compare them qualitatively with Nerfacto (Tancik et al., 2023), Gaussian Splatting (Kerbl et al., 2023), and Ultra-NeRF (Wysocki et al., 2024) (**best viewed in color and with zoom**).

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Ultra-NeRF (Wysocki et al., 2024)	24.14 ( $\pm 2.54$ )	0.8312 ( $\pm 0.15$ )	0.2573 ( $\pm 0.16$ )
Ours (NeRF-US)	27.37 ( $\pm 1.83$ )	0.8412 ( $\pm 0.13$ )	0.2325 ( $\pm 0.14$ )
<i>Ablation Study</i>			
Ours (NeRF-US)			
w/o $\mathcal{L}_{\rho_b}$	24.89 ( $\pm 1.72$ )	0.8208 ( $\pm 0.98$ )	0.2613 ( $\pm 0.13$ )
w/o $\mathcal{L}_{\rho_s}$	26.02 ( $\pm 1.67$ )	0.8197 ( $\pm 1.10$ )	0.2598 ( $\pm 0.08$ )
w/o $I(t)$	23.98 ( $\pm 1.46$ )	0.8065 ( $\pm 1.23$ )	0.2591 ( $\pm 0.11$ )

Table 3: **Ablation Study.** We show ablations of our approach: w/o  $\mathcal{L}_{\rho_b}$ , w/o  $\mathcal{L}_{\rho_s}$ , and w/o the rendering  $I(t)$ . The **best**, **second best**, and **third best** results for each metric are color coded.

to reconstruct accurately when there are immensely complex full-body motions. We believe that our method will directly benefit from the progress on dynamic reconstruction methods that use dynamic representations eg. HyperNeRF (Park et al., 2021), NeRF-DS (Yan et al., 2023), Dynamic 3D Gaussians (Luiten et al., 2023). Moreover, currently our diffusion model is fine-tuned on voxels specific to a body region, while fine-tuning the diffusion model is

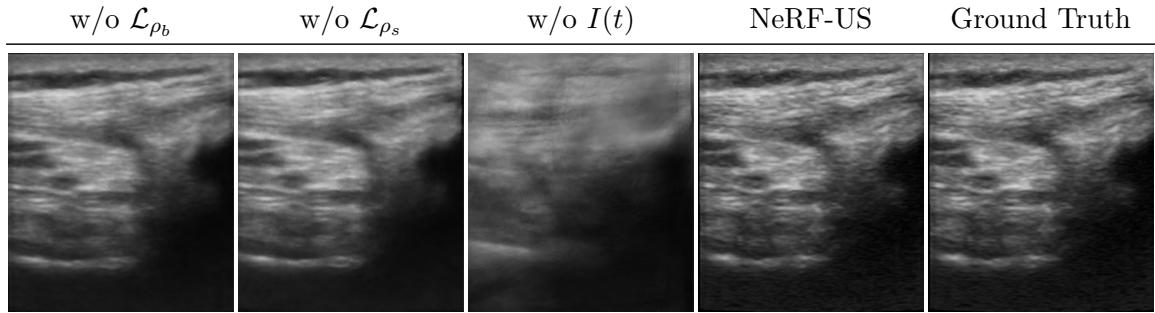


Figure 9: **Ablation Study.** We assess the qualitative impact of removing each of the components of our pipeline:  $\mathcal{L}_{\rho_b}$ ,  $\mathcal{L}_{\rho_s}$ , and  $I(t)$ . Our approach without the border probability guidance ( $\mathcal{L}_{\rho_b}$ ) fail to reconstruct minor details in the scene especially it is unable to clearly segregate multiple objects in the image. Removing the scattering density guidance ( $\mathcal{L}_{\rho_s}$ ) produces very blurred reconstructions. Our approach without the ultrasound rendering produces a lot of artifacts in the reconstruction and fails to capture most aspects of the reconstruction. (**Best viewed with zoom**).

very cheap (< 5 hours on A100-80), and we believe that an interesting avenue for future work would be to build a generalized diffusion model. Lastly, our diffusion model is fine-tuned on a dataset of synthetic voxels (notice that our NeRF model is evaluated and trained on real-life data only), and we currently do not experiment with a real-life dataset from 3D captures. Furthermore, an interesting avenue for future works is to explore the effect of training the diffusion model with other modalities, we particularly believe it might be promising to explore training the diffusion model on Magnetic Resonance Imaging (MRI) scans to reconstruct ultrasound imaging.

## 7. Conclusion

Our work builds a NeRF-based technique that can perform accurate view synthesis and 3D reconstruction on ultrasound imaging. Our work is the first to tackle the problem of performing view synthesis and 3D reconstructions on ultrasound imaging data collected in its natural in the wild form as opposed to works that tackle this problem on simulated data or very heavy ultrasound capture mechanisms. We also observe accurate artifact-free outputs for ultrasound imaging with our method. We will release the code and data to facilitate future research in this area.

## Acknowledgments

This research was enabled in part by support provided by the Digital Research Alliance of Canada <sup>2</sup>. This research was supported in part with Cloud TPUs from Google’s TPU

---

2. <https://alliancecan.ca/>

Research Cloud (TRC)<sup>3</sup>. The resources used to prepare this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute<sup>4</sup>. We thank Wysocki et al. (2024) for readily sharing the data they collected for their work. We thank anonymous reviewers of the MLHC Conference for their insightful suggestions which we incorporate in this work.

---

3. <https://sites.research.google/trc/about/>  
4. <https://vectorinstitute.ai/partnerships/current-partners/>

## References

- Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 696–712, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58542-6.
- Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, June 2022.
- Giang Bui, Truc Le, Brittany Morago, and Ye Duan. Point-based rendering enhancement via deep learning. *The Visual Computer*, 34:829–841, 2018.
- Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015.
- Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Trans. Graph.*, 32(3), jul 2013. ISSN 0730-0301. doi: 10.1145/2487228.2487238. URL <https://doi.org/10.1145/2487228.2487238>.
- Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, page 333–350, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19823-6. doi: 10.1007/978-3-031-19824-3\_20. URL [https://doi.org/10.1007/978-3-031-19824-3\\_20](https://doi.org/10.1007/978-3-031-19824-3_20).
- Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. URL <https://arxiv.org/abs/2304.06714>.
- Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH ’93, page 279–288, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897916018. doi: 10.1145/166117.166153. URL <https://doi.org/10.1145/166117.166153>.
- Shenchang Eric Chen and Lance Williams. *View Interpolation for Image Synthesis*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 978400708978. URL <https://doi.org/10.1145/3596711.3596757>.
- Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert P. H. Shum, and Chris G. Willcocks. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray, 2022a.

Abril Corona-Figueroa, Jonathan Frawley, Sam Bond Taylor, Sarath Bethapudi, Hubert P. H. Shum, and Chris G. Willcocks. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3843–3848, 2022b. doi: 10.1109/EMBC48229.2022.9871757.

Paul Debevec, Yizhou Yu, and George Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. In George Drettakis and Nelson Max, editors, *Rendering Techniques '98*, pages 105–116, Vienna, 1998. Springer Vienna. ISBN 978-3-7091-6453-2.

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96*, page 11–20, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237191. URL <https://doi.org/10.1145/237170.237191>.

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. *Modeling and Rendering Architecture from Photographs: A hybrid geometry- and image-based approach*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL <https://doi.org/10.1145/3596711.3596761>.

Andrew Fitzgibbon, Yonatan Wexler, and Andrew Zisserman. Image-based rendering using image-based priors. *International Journal of Computer Vision*, 63:141–151, 2005.

John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. <https://arxiv.org/abs/2012.03065>, 2020.

François Gaits, Nicolas Mellado, and Adrian Basarab. Ultrasound volume reconstruction from 2D Freehand acquisitions using neural implicit representations. In *21st IEEE International Symposium on Biomedical Imaging (ISBI 2024)*, page à paraître, Athènes, Greece, May 2024. IEEE Signal Processing Society and IEEE Engineering in Medicine and Biology Society. URL <https://hal.science/hal-04480668>.

Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes’ rays: Uncertainty quantification for neural radiance fields, 2023.

J. P. Grossman and William J. Dally. Point sample rendering. In George Drettakis and Nelson Max, editors, *Rendering Techniques '98*, pages 181–192, Vienna, 1998. Springer Vienna. ISBN 978-3-7091-6453-2.

Ang Nan Gu, Purang Abolmaesumi, Christina Luong, and Kwang Moo Yi. Representing 3d ultrasound with neural fields. In *Medical Imaging with Deep Learning*, 2022.

- Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, 2023. URL <https://arxiv.org/abs/2302.10109>.
- Hengtao Guo, Sheng Xu, Bradford Wood, and Pingkun Yan. Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 463–472. Springer, 2020.
- Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph.*, 37(6), dec 2018. ISSN 0730-0301. doi: 10.1145/3272127.3275084. URL <https://doi.org/10.1145/3272127.3275084>.
- HH Heilmann, K Lindenhayn, and HU Walther. Synovial volume of healthy and arthrotic human knee joints. *Zeitschrift fur Orthopadie und ihre Grenzgebiete*, 134(2):144–148, 1996.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Jing Hu, Qinrui Fan, Shu Hu, Siwei Lyu, Xi Wu, and Xin Wang. Umednerf: Uncertainty-aware single view volumetric rendering for medical neural radiance fields, 2024.
- Khadija Iddrisu, Sylwia Malec, and Alessandro Crimi. 3d reconstructions of brain from mri scans using neural radiance fields. In *International Conference on Artificial Intelligence and Soft Computing*, pages 207–218. Springer, 2023.
- DW Jackson, LD Jennings, RM Maywood, and PE Berger. Magnetic resonance imaging of the knee. *The American Journal of Sports Medicine*, 16(1):29–38, 1988.
- Clément Jambon, Bernhard Kerbl, Georgios Kopanas, Stavros Diolatzis, George Drettakis, and Thomas Leimkühler. Nerfshop: Interactive editing of neural radiance fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(1), 2023.
- Songbai Ji, David W Roberts, Alex Hartov, and Keith D Paulsen. Real-time interpolation for true 3-dimensional ultrasound image volumes. *Journal of Ultrasound in Medicine*, 30(2):243–252, 2011.
- Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021.

Chiyu "Max" Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Niessner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Samuel Kadoury, Farida Cheriet, Jean Dansereau, and Hubert Labelle. Three-dimensional reconstruction of the scoliotic spine and pelvis from uncalibrated biplanar x-ray images. *Clinical Spine Surgery*, 20(2):160–167, 2007.

Samuel Kadoury, Farida Cheriet, and Hubert Labelle. Personalized x-ray 3-d reconstruction of the scoliotic spine from hybrid statistical and image-based models. *IEEE Transactions on Medical Imaging*, 28(9):1422–1435, 2009. doi: 10.1109/TMI.2009.2016756.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Risto Kojcev, Ashkan Khakzar, Bernhard Fuerst, Oliver Zettinig, Carole Fahkry, Robert DeJong, Jeremy Richmon, Russell Taylor, Edoardo Sinibaldi, and Nassir Navab. On the reproducibility of expert-operated and robotic ultrasound acquisitions. *International journal of computer assisted radiology and surgery*, 12:1003–1011, 2017.

K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 307–314 vol.1, 1999. doi: 10.1109/ICCV.1999.791235.

Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96*, page 31–42, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237199. URL <https://doi.org/10.1145/237170.237199>.

Honggen Li, Hongbo Chen, Wenke Jing, Yuwei Li, and Rui Zheng. 3d ultrasound spine imaging with application of neural radiance field method. In *2021 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4, 2021a. doi: 10.1109/IUS52206.2021.9593917.

Honggen Li, Hongbo Chen, Wenke Jing, Yuwei Li, and Rui Zheng. 3d ultrasound spine imaging with application of neural radiance field method. In *2021 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2021b.

Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 178–196, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

Hao-Kang Liu, I-Chao Shen, and Bing-Yu Chen. Nerf-in: Free-form nerf inpainting with rgb-d priors. *arxiv preprint arXiv:2206.04901*, 2022.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rkgz2aEKDr>.

Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields, 2021.

Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Clean-nerf: Reformulating nerf to account for view-dependent observations, 2023.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023.

Andrej Lyshchik, Valentina Drozd, Susanne Schloegl, and Christoph Reiners. Three-dimensional ultrasonography for volume measurement of thyroid nodules in children. *Journal of ultrasound in medicine*, 23(2):247–254, 2004.

Kirsten WH Maas, Nicola Pezzotti, Amy JE Vermeer, Danny Ruijters, and Anna Vilanova. Nerf for 3d reconstruction from x-ray angiography: Possibilities and limitations. In *VCBM 2023: Eurographics Workshop on Visual Computing for Biology and Medicine*, pages 29–40. Eurographics Association, 2023.

Leonard McMillan and Gary Bishop. Plenoptic modeling: an image-based rendering system. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH ’95, page 39–46, New York, NY, USA, 1995. Association for Computing Machinery. ISBN 0897917014. doi: 10.1145/218380.218398. URL <https://doi.org/10.1145/218380.218398>.

Leonard McMillan and Gary Bishop. *Plenoptic Modeling: An Image-Based Rendering System*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL <https://doi.org/10.1145/3596711.3596758>.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 38(4), jul 2019. ISSN 0730-0301. doi: 10.1145/3306346.3322980. URL <https://doi.org/10.1145/3306346.3322980>.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, dec 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL <https://doi.org/10.1145/3503250>.

Matthew R Morgan, Joshua S Broder, Jeremy J Dahl, and Carl D Herickhoff. Versatile low-cost volumetric 3-d ultrasound platform for existing clinical 2-d systems. *IEEE transactions on medical imaging*, 37(10):2248–2256, 2018.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), jul 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530127. URL <https://doi.org/10.1145/3528223.3530127>.

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5480–5490, June 2022.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan Goldman, Steven Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. <https://arxiv.org/abs/2011.12948>, 2020.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Trans. Graph.*, 36(6), nov 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130855. URL <https://doi.org/10.1145/3130800.3130855>.

Tony C. Poon and Robert N. Rohling. Comparison of calibration methods for spatial tracking of a 3-d ultrasound probe. *Ultrasound in Medicine & Biology*, 31(8):1095–1108, 2005. ISSN 0301-5629. doi: <https://doi.org/10.1016/j.ultrasmedbio.2005.04.003>. URL <https://www.sciencedirect.com/science/article/pii/S0301562905001754>.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. <https://arxiv.org/abs/2011.13961>, 2020.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017.

Gernot Riegler and Vladlen Koltun. Free view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 623–640, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58529-7.

Mehrdad Salehi, Seyed-Ahmad Ahmadi, Raphael Prevost, Nassir Navab, and Wolfgang Wein. Patient-specific 3d ultrasound simulation based on convolutional ray-tracing and

appearance optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. URL <https://api.semanticscholar.org/CorpusID:9711821>.

Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 501–518, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9.

Steven M Seitz and Charles R Dyer. View morphing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 21–30, 1996.

Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision*, 35:151–173, 1999.

Qi Shan, Riley Adams, Brian Curless, Yasutaka Furukawa, and Steven M. Seitz. The visual turing test for scene reconstruction. In *2013 International Conference on 3D Vision - 3DV 2013*, pages 25–32, 2013. doi: 10.1109/3DV.2013.12.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Peter-Pike Sloan, Michael F Cohen, and Steven J Gortler. Time critical lumigraph rendering. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 17–ff, 1997.

Stephen W Smith, Warren Lee, Edward D Light, Jesse T Yen, Patrick Wolf, and Salim Idriss. Two dimensional arrays for 3-d ultrasound imaging. In *2002 IEEE Ultrasonics Symposium, 2002. Proceedings.*, volume 2, pages 1545–1553. IEEE, 2002.

Nagabhushan Somraj and Rajiv Soundararajan. Vip-nerf: Visibility prior for sparse input neural radiance fields. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH ’23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701597. doi: 10.1145/3588432.3591539. URL <https://doi.org/10.1145/3588432.3591539>.

Sheng Song, Yunqian Huang, Jiawen Li, Man Chen, and Rui Zheng. Development of implicit representation method for freehand 3d ultrasound image reconstruction of carotid vessel. In *2022 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2022.

Mengcheng Sun, Yu Zhu, Hangyu Li, Jiongyao Ye, and Nan Li. Acnerf: enhancement of neural radiance field by alignment and correction of pose to reconstruct new views from a single x-ray. *Physics in Medicine & Biology*, 69(4):045016, 2024.

Towaki Takikawa, Alex Evans, Jonathan Tremblay, Thomas Müller, Morgan McGuire, Alec Jacobson, and Sanja Fidler. Variable bitrate neural fields. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David Mcallister, Justin Kerr, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701597. doi: 10.1145/3588432.3591516. URL <https://doi.org/10.1145/3588432.3591516>.

II Thomas W. Hash. Magnetic resonance imaging of the knee. *Sports Health*, 5(1): 78–107, 2013. doi: 10.1177/1941738112468416. URL <https://doi.org/10.1177/1941738112468416>. PMID: 24381701.

Xin Wang, Shu Hu, Heng Fan, Hongtu Zhu, and Xin Li. Neural radiance fields in medical imaging: Challenges and next steps, 2024.

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery, 2022.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirly-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Hołyński, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *International Conference on Computer Vision (ICCV)*, 2023.

Tristan Wirth, Arne Rak, Volker Knauthe, and Dieter W Fellner. A post processing technique to automatically remove floater artifacts in neural radiance fields. In *Computer Graphics Forum*, volume 42, page e14977. Wiley Online Library, 2023.

Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *CVPR*, 2023.

Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: Neural radiance fields for ultrasound imaging. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and

Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 382–401. PMLR, 10–12 Jul 2024. URL <https://proceedings.mlr.press/v227/wysocki24a.html>.

Z. Yan, C. Li, and G. Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8285–8295, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00801. URL <https://doi.ieee.org/10.1109/CVPR52729.2023.00801>.

Guandao Yang, Abhijit Kundu, Leonidas J Guibas, Jonathan T Barron, and Ben Poole. Learning a diffusion prior for nerfs. *arXiv preprint arXiv:2304.14473*, 2023.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020.

Pak-Hei Yeung, Linde Hesse, Moska Aliasi, Monique Haak, Weidi Xie, Ana IL Namburete, et al. Implicitvol: Sensorless 3d ultrasound reconstruction with deep implicit representation. *arXiv preprint arXiv:2109.12108*, 2021.

Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: Neural attenuation fields for sparse-view cbct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 442–452. Springer, 2022.

Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 13–23, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43996-4.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Chaochao Zhou, Syed Hasib Akhter Faruqui, Abhinav Patel, Ramez N Abdalla, Michael C Hurley, Ali Shaibani, Matthew B Potts, Babak S Jahromi, Leon Cho, Sameer A Ansari, et al. Robust single-view cone-beam x-ray pose estimation with neural tuned tomography (nett) and masked neural radiance fields (mnerf). *arXiv preprint arXiv:2308.00214*, 2023.

## Appendix A. Implementation Details

### A.1. Training the Diffusion Model

The pre-trained diffusion model is a UNet-based diffusion model designed for  $32 \times 32 \times 32$  data cubes. The model utilizes 3D convolutions with a base channel count of 32, employing channel multipliers of (1, 2, 4, 8) across resolution levels. We use two residual blocks per resolution level, with self-attention mechanisms strategically applied at  $16\times$  and  $8\times$  downsampling. The model also incorporates time embeddings. The model features a down-sampling path that progressively reduces spatial dimensions while increasing channel count, a bottleneck with residual connections and attention mechanisms, and an upsampling path with skip connections. The pre-trained model also uses scale conditioning through a learned embedding.

We provide more details on how we perform LoRA (Hu et al., 2022) adaption to the denoising diffusion model. Our implementation closely follows the original LoRA, which demonstrated the adaption approach for language models, and draws inspiration from the HuggingFace Diffusers implementation, which demonstrates adaption for text-to-image (T2I) models.

We use the 3D diffusion model  $\Phi$  with parameters  $\theta$  trained on ShapeNet (Chang et al., 2015) which works on  $32 \times 32 \times 32$  occupancy grid,  $x$ , from Nerfbusters (Warburg et al., 2023). Based on LoRA (Hu et al., 2022), for a given layer in  $\Phi$ , we introduce trainable low-rank matrices. For the weight matrix  $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$  in the model, LoRA decomposes the adaptation into two low-rank matrices,  $A \in \mathbb{R}^{d_{\text{out}} \times r}$  and  $B \in \mathbb{R}^{r \times d_{\text{in}}}$ , where  $r \ll \min(d_{\text{out}}, d_{\text{in}})$  representing the rank controlling adaption. The updated weights  $W'$  can be computed as,

$$W' = W + \delta(AB) \quad (10)$$

where  $\delta$  is a scaling factor that determines the magnitude of the adaptation and  $AB$  represents a low-rank update to the original weights.

Our fine-tuning process now utilizes the same loss function as the original DDPM training, with the low-rank update applied to the model parameters, giving us Equation (6).

### A.2. NeRF Optimization Objective

Our optimization objective for the NeRF was defined in Equation (9), we can write this optimization objective as,

$$\begin{aligned} \mathcal{L} &= \sum_{\mathbf{r} \in \mathcal{R}} \overbrace{\left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2}^{\text{photometric loss}} + \lambda_{\rho_b} \mathcal{L}_{\rho_b} + \lambda_{\rho_s} \mathcal{L}_{\rho_s} \\ &= \sum_{\mathbf{r} \in \mathcal{R}} \left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \lambda_{\rho_b} \left( \frac{1}{N} \sum_{i=1}^N |C(i)^{\rho_b} - \Phi_{\theta+\delta(AB)}(C(i))^{\rho_b}|^2 \right) \\ &\quad + \lambda_{\rho_s} \left( \frac{1}{N} \sum_{i=1}^N |C(i)^{\rho_s} - \Phi_{\theta+\delta(AB)}(C(i))^{\rho_s}|^2 \right) \end{aligned} \quad (11)$$

We can notice that while our loss formulation is different than the Density Score Distillation Sampling (Warburg et al., 2023), however it can be still looked upon as adding a regularizer to the original NeRF photometric loss that penalizes an incorrect border probability and scattering density. Unlike Ultra-NeRF (Wysocki et al., 2024), we do not use a SSIM loss, we observe that with the new definition of  $\mathcal{L}$ , using the SSIM loss does not demonstrate any benefits over using the photometric loss term. Furthermore, we also observe that setting  $\lambda_{\rho_b} > \lambda_{\rho_s}$  usually works well towards reconstructing high-quality scenes.

### A.3. NeRF MLP

Following previous work on NeRFs like mip-NeRF (Barron et al., 2022), we use an MLP ( $F_\Theta$ ) with 8 layers and 256 hidden layer units with ReLU and a skip connection between the input and the fifth layer. To ensure numerical stability and physiologically plausible results, we enforce the attenuation parameter to assume continuous positive values, while reflectance, border probability, scattering density, and scattering intensity are confined within the range of  $[0, 1]$ . We do not employ additional MLP directional components for camera viewing directions  $(\theta, \phi)$ . This decision stems from the recognition that our MLP, does not directly assimilate information from the camera viewing angles. Rather, this aspect is effectively addressed by the ultrasound rendering of the reconstruction pipeline. Following NeRF (Mildenhall et al., 2021), we also use positional encodings to train the MLP.

### A.4. Training Hyperparameters

All of our final code was optimized for a 1 x A100-80 GB GPU. For our Diffusion model, we trained our model for  $30K$  steps with a batch size of 32 on the  $32 \times 32 \times 32$  resolution. We use the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$  with an initial learning rate of  $10^{-4}$ . We use cosine decay for the learning rate. We use a rank of 4 for the LoRA update weights. Our NeRF implementation is based upon Nerfstudio (Tancik et al., 2023). We trained our model for  $300K$  iterations with a batch size of  $2^{12}$ . We use the RAdam optimizer (Liu et al., 2020) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use a learning rate that starts at  $5 \times 10^{-4}$  and decays exponentially to  $5 \times 10^{-5}$ , and  $\epsilon = 10^{-8}$ . We however, notice that the results become pretty accurate after the first  $100K$  iterations.

### A.5. Code and Data

To foster future work in this direction we open-source our code and Ultrasound in the wild dataset which can be found on our project page at: [rishitdagli.com/nerf-us/](http://rishitdagli.com/nerf-us/).

### A.6. Baseline Models

Here we share the details for the baseline models we compare our approach with in Table 1. Our baseline models for Original NeRF (Mildenhall et al., 2021), Instant-NGP (Müller et al., 2022), TensoRF (Chen et al., 2022), Nerfacto (Tancik et al., 2023), and Gaussian Splatting (Kerbl et al., 2023) are trained with Nerfstudio (Tancik et al., 2023) implementations using the details we list below. Just like rest of our code all of our baseline models were trained on 1 x A100-80 GB GPU.

**DCL-Net.** We trained this baseline model for 300 epochs with a batch size of  $2^5$  using the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-14}$ . We use cosine decay starting with the learning rate of  $5 \times 10^{-5}$  and a warmup of 5 epochs. Furthermore, all the underlying images are resized to  $224 \times 224$ -sized image.

**ImplicitVol.** We construct this baseline model with an MLP of 5 layers with 128 neurons. We trained this baseline model for  $10K$  epochs and also use NeRF-style positional encodings. Since using this technique involves first estimating 2D plane locations, as suggested by ImplicitVol in their work we first estimate these plane positions using PlaneInVol (Wang et al., 2021). We use the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$  with an initial learning rate of  $10^{-3}$ . We use the standard multi-step schedule with each of its stages at every 10 epochs with  $\gamma = 0.9954$ . We incorporate a window size of  $k = 5$  for calculating the SSIM loss.

**Original NeRF.** We trained this baseline model for  $300K$  iterations with a batch size of  $2^{12}$ . We use the RAdam optimizer (Liu et al., 2020) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use a learning rate that starts at  $5 \times 10^{-4}$  and decays exponentially to  $5 \times 10^{-5}$ , and  $\epsilon = 10^{-8}$ . We also use gradient scaling to ensure that gradients near the camera are scaled down. We also use hierarchical sampling with 64 coarse samples and 128 importance samples for fine field evaluation. We use the NeRF defaults for all other hyperparamters.

**Gu et al. (2022).** We trained this baseline model for  $20K$  iterations with a batch size of  $2^2$ . We use the Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of  $10^{-4}$ ,  $\epsilon = 10^{-8}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and a meta learning rate of  $10^{-5}$ . We also use LookAhead with 10 steps. We use an  $\omega = 240$  for the SIREN activation. We use the defaults from Gu et al. (2022) for all other hyperparamters.

**Instant-NGP.** We trained this baseline model for  $30K$  iterations with a batch size of  $2^{12}$ . We use the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use an initial learning rate of  $10^{-2}$  and  $\epsilon = 10^{-15}$ . We exponentially decay the learning rate from  $10^{-2}$  to  $10^{-4}$  and use no warmup. We use a grid resolution of 128 with 4 grid levels for the multiresolution hash encoding. We also use a resolution of  $2^{11}$  for the hashmap used by the MLP. We perform the sampling from the rays in between  $[5 \times 10^{-2}, 10^3]$ . We use the Instant-NGP defaults for all other hyperparamters.

**TensoRF.** We trained this baseline model for  $30K$  iterations with a batch size of  $2^{12}$ . We use the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use an initial learning rate of  $10^{-2}$  and  $\epsilon = 10^{-8}$ . We exponentially decay the learning rate from  $10^{-3}$  to  $10^{-4}$  and use no warmup. We also use TV regularization while training the field. We use the TensoRF defaults for all other hyperparamters.

**Nerfacto.** We trained this baseline model for  $30K$  iterations with a batch size of  $2^{12}$ . We use the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use an initial learning rate of  $10^{-2}$  and  $\epsilon = 10^{-15}$ . We exponentially decay the learning rate from  $10^{-2}$  to  $10^{-4}$  and use no warmup. We use a vector with 6 parameters representing  $SO(3) \times R(3)$  map for the camera optimizer. We use the Nerfacto defaults for all other hyperparamters.

**Gaussian Splatting.** We trained this baseline model for  $30K$  iterations with a batch size of  $2^{12}$ . We use the Adam optimizer (Kingma and Ba, 2017) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$

for the Gaussian means network. For the Gaussian means network exponentially decay the learning rate from  $1.6 \times 10^{-4}$  to  $1.6 \times 10^{-6}$ , and  $\epsilon = 10^{-15}$ . We set the threshold for frustum culling Gaussians to  $5 \times 10^{-3}$ . For rendering we use the EWA volume splatting with a  $[0.3, 0.3]$  screen space blurring kernel. Following recent popular work with Gaussian Splatting, we also apply a regularization loss when the ratio of a Gaussian’s maximum to minimum scale exceeds a threshold. We use the Gaussian Splatting defaults for all other hyperparamters.

**Ultra-NeRF.** We trained this baseline model for  $100K$  iterations with a batch size of  $2^{12}$ . We use the Adam optimizer (Liu et al., 2020) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use a learning rate that starts at  $10^{-4}$  and decays exponentially in  $250K$  steps with a decay of 0.1, and  $\epsilon = 10^{-8}$ . For the loss calculation we use a window size of  $k = 7$  and  $\lambda = 0.7$ . We use the NeRF defaults for all other hyperparamters.

## Appendix B. Evaluation Metrics

To evaluate the quality of our rendered images, we employ three standard metrics commonly used in the assessment of such models.

**PSNR** quantifies the ratio between the maximum possible signal power and the power of distorting noise, with higher values indicating better reconstruction quality.

**SSIM** evaluates the perceived quality of images by considering structural information changes, with values closer to 1 indicating higher similarity.

**LPIPS** is a perceptual metric leveraging deep neural networks trained on human judgments, aims to capture perceptual similarities between image patches, with lower scores indicating greater perceptual similarity.

## Appendix C. Additional Qualitative Results

We present additional qualitative results reconstructed using our method on the Ultrasound in the wild dataset we introduced and show novel views in Figure 10. Furthermore, we encourage the reader to visit our project website for rendered videos with the camera moving around the scene in azimuth with a fixed elevation angle.

## Appendix D. Ethics Statement

Our approach produces very compelling qualitative and quantitative results for reconstructing ultrasound imaging in the wild. However, in its current form, the authors do not give any theoretical guarantees while using this method and we strongly suggest readers looking to adopt this method to go through the limitations as well. Our data was collected with appropriate institutional guidelines on using medical data from humans. We obtained a waiver from our institution’s Research Ethics Board, as all data collected was part of a self study (PNT) for this proof of concept.

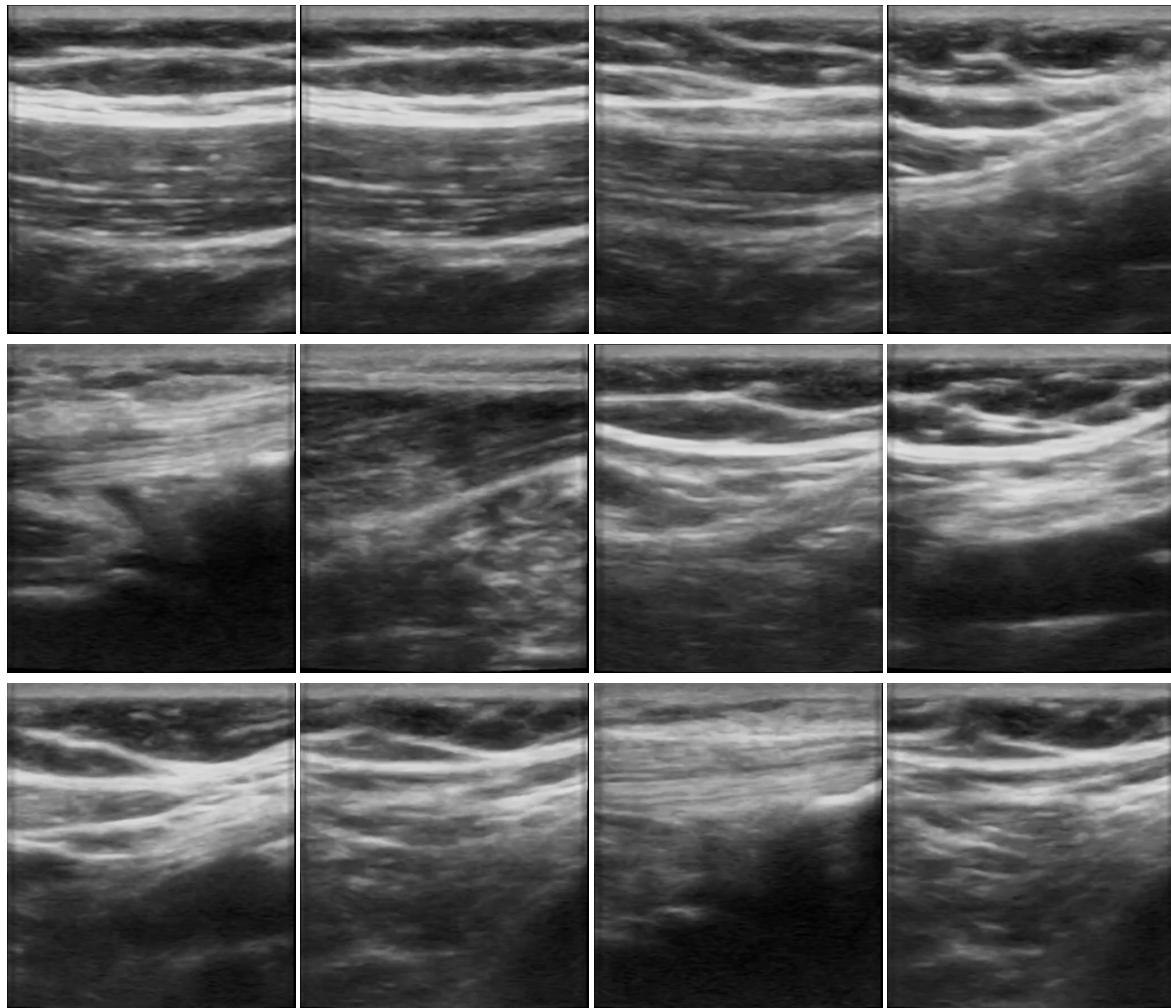


Figure 10: We present additional results using NeRF-US. In general, we observe high-quality artifact free reconstructions with our approach.