

UltraNBA

Neural Bundle Adjustment for Pose Refinement in 3D Freehand Ultrasound

V. Bugra Yesilkaynak*
 Technical University of Munich
 bugrayesilkaynak@gmail.com

Magdalena Wysocki*
 Technical University of Munich
 Munich Center for Machine Learning
 magdalena.wysocki@tum.de

Nassir Navab
 Technical University of Munich
 Munich Center for Machine Learning
 nassir.navab@tum.de

Vanessa Gonzalez Duque*
 LumaVision GmbH
 vanessa.duque@lumavision.com

Mohammad Farid Azampour
 Technical University of Munich
 Munich Center for Machine Learning
 mf.azampour@tum.de

Diana Mateus
 Ecole Centrale de Nantes
 Laboratoire LS2N Nantes
 diana.mateus@ec-nantes.fr

Abstract

*Imprecise tracking presents a major challenge in the reconstruction of 3D freehand ultrasound volumes, as even small errors can lead to significant misalignment. Calibration inaccuracies and reliance on noisy sensors further exacerbate this issue. State-of-the-art approaches typically align pixel intensities across overlapping frames. However, misalignment in ultrasound sweeps, which depend on the sensor's position, often result in inconsistent intensities for the same spatial location, challenging the reliability of these methods. To address these challenges, we propose **UltraNBA**, a novel implicitly neural bundle-adjusting framework for ultrasound. By leveraging the spatial consistency of acoustic tissue properties instead of plain intensity alignment, UltraNBA corrects tracking errors while capturing stable anatomical and physical representations, yielding higher-quality reconstructions. Our method supports single and multiple sweeps, offering versatility in real-world clinical scenarios. Experimental results demonstrate a reduction in tracking errors, accompanied by enhanced image quality for rendering new frames. We make code and data open-source at <https://github.com/MrGranddy/UltraNBA>.*

1. Introduction

Skilled sonographers are trained to mentally reconstruct 3D anatomical structures from sequences of 2D ultrasound im-

ages. Through extensive clinical experience, they learn to infer the spatial relationships between image slices based on subtle cues such as probe motion, anatomical continuity, and acoustic patterns. This mental reconstruction is a cognitively demanding task that relies on their ability to integrate fragmented 2D views into a coherent 3D understanding of the scanned region. Crucially, sonographers intuitively associate spatial correspondences across successive frames. Supporting and automating this complex mental process could significantly reduce cognitive load and improve both diagnostic accuracy and workflow efficiency.

Simultaneously reconstructing the 3D anatomical volume from 2D ultrasound images and accurately localizing the corresponding probe poses presents a complex, intertwined challenge [35]. This problem parallels the classical chicken-and-egg dilemma in computer vision: recovering 3D scene structure requires known camera poses, while localizing cameras depends on reliable correspondences derived from the reconstruction. However, ultrasound imaging introduces additional complexities such as speckle noise, tissue deformation, variable acoustic properties, and probe-induced artifacts, all of which complicate establishing consistent spatial correspondences and robust tracking during reconstruction.

In computer vision, classical methods such as Structure from Motion (SfM) [13, 26] and Simultaneous Localization and Mapping (SLAM) [11, 21] address this challenge by performing local registration followed by global geometric bundle adjustment (BA) on both the scene structure

*These authors contributed equally.

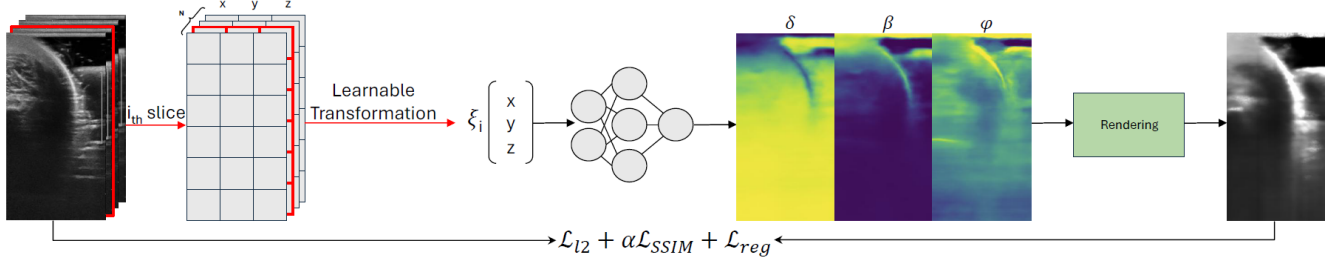


Figure 1. UltraNBA jointly optimizes pose refinement and implicit volume learning for 3D freehand ultrasound reconstruction. Here, δ is the attenuation coefficient, β is the reflectance, and ϕ is the scattering intensity, which define the ultrasound rendering properties.

and camera poses. Yet, these approaches cannot be directly transferred to ultrasound due to the domain-specific challenges mentioned above. As a result, ultrasound reconstruction requires tailored methods that can handle noisy, deformable, and acoustically complex data while jointly estimating probe poses and anatomical structure.

Three-dimensional freehand ultrasound (3D US) is a common and practical approach for reconstructing volumetric data from sequential 2D B-mode images [20, 27]. In this method, clinicians manually sweep a conventional ultrasound probe to acquire sequential 2D images, which are subsequently reconstructed into a 3D volume. This technique offers a flexible and cost-effective alternative to fixed three-dimensional matrix probes, facilitating the imaging of extensive or anatomically complex regions. It is widely employed in clinical applications including musculoskeletal assessment [22], obstetric monitoring [34], and image-guided interventions [29].

Despite its versatility, freehand ultrasound faces a critical challenge: its reliance on external tracking systems to determine the spatial position and orientation of the probe. Inaccurate tracking caused by sensor misalignment, magnetic interference, or mechanical drift can significantly degrade the quality of reconstructed volumes [1]. Another set of methods aim at sensorless reconstruction [35]. These methods avoid the use of external trackers and rely solely on the image content, typically assuming correlation between speckle patterns to find correspondences [5]. However, this assumption is often violated in ultrasound imaging due to variable acoustic properties, tissue deformation, and probe-induced artifacts that alter intensity patterns even at the same anatomical location. Moreover, since ultrasound sweeps are acquired along trajectories that do not guarantee overlap between consecutive frames, the necessary spatial correspondences for accurate registration may not exist. As a result, reconstruction approaches based primarily on image intensities can lead to suboptimal alignment, causing blurred anatomical structures, misaligned tissue layers, and targeting errors.

In this work, we propose UltraNBA, a novel implicit neural method that jointly optimizes probe pose refinement

and 3D ultrasound reconstruction by relaxing dependency on error-free tracking hardware. UltraNBA builds a continuous representation of the volume in an acoustic properties space that remains constant across acquisitions, overcoming limitations of intensity-based co-registration. By interpolating between potentially non-overlapping frames, the method ensures improved alignment and volume consistency. This is achieved through a tailored neural representation incorporating ultrasound-specific priors and frequency-based modulation, enabling simultaneous refinement of probe poses and reconstruction of high-fidelity 3D volumes.

Our main contributions are:

- We introduce UltraNBA, the first implicit neural method to jointly optimize pose refinement and 3D volume reconstruction in freehand ultrasound.
- We propose a learning strategy that integrates ultrasound-specific priors and frequency-based modulation to stabilize training and improve fidelity.
- We demonstrate consistent improvements in pose accuracy and image quality across multiple real and synthetic tracking scenarios.

2. Related Work

Neural Radiance Fields. In computer vision, the joint optimization of 3D scene structure and camera poses is a well-established problem, traditionally tackled using Structure from Motion (SfM) [2], Simultaneous Localization and Mapping (SLAM) [16], or bundle adjustment techniques [3, 9]. One notable example is Bundle Adjustment for Radiance Fields (BARF) [17], which integrates pose refinement and scene reconstruction through a coarse-to-fine photometric alignment strategy. BARF demonstrated that neural fields can be optimized even with inaccurate initial poses by leveraging photometric consistency across views.

While these advances are promising in natural image domains, ultrasound imaging introduces distinct challenges. Ultrasound signals are highly view-dependent, dominated by speckle noise, and vary with probe orientation, tissue

contact, and pressure. Methods developed for natural images, which rely on consistent appearance and rich texture, do not transfer well to this domain. Inspired by neural radiance field (NeRF) [19] models, recent methods such as UltraNeRF [32] attempt to learn implicit 3D representations from ultrasound data. UltraNeRF adapts NeRF to model acoustic responses instead of light fields, capturing view-dependent reflectivity and internal structures of tissue. However, like standard NeRFs, UltraNeRF assumes known poses during training and relies heavily on pixel-level intensity supervision. This limits its robustness in clinical freehand scenarios where tracking is noisy or unreliable.

Sensorless Freehand 3D Ultrasound. Freehand 3D ultrasound reconstruction has traditionally relied on image-based co-registration to align successive 2D frames by maximizing pixel correspondence. Early sensorless methods addressed this by estimating rigid motion through normalized cross-correlation and speckle decorrelation, enabling reconstruction without external tracking across irregularly sampled and non-parallel frames [8, 14]. Other approaches, such as kernel-regression [6] and global patch-matching [7], fuse overlapping frames into a coherent volume by matching local intensity patterns or texture features. While sensorless freehand 3D ultrasound reconstruction removes reliance on external tracking, it remains fundamentally constrained by the requirement for correlated image content across frames.

Tracked Freehand 3D Ultrasound. Tracked freehand ultrasound systems use external sensors such as optical [15] or electromagnetic trackers [10] to record the position and orientation of the probe during scanning. In contrast to sensorless methods, which rely entirely on estimating motion from image content, tracked approaches provide independent measurements of probe movement. This allows for more stable and reliable reconstructions, especially in regions with poor image correlation or when frames lack sufficient spatial overlap. Early work in this area focused on integrating tracked probe positions to reconstruct three-dimensional volumes from two-dimensional slices, often relying on calibration between the probe and the tracking system to ensure spatial consistency [18, 36].

Despite the advantages of tracked freehand ultrasound systems, pose measurements obtained from external sensors are often subject to errors arising from sensor noise, calibration inaccuracies, and environmental interference [15, 28]. These pose inaccuracies can lead to misalignment in the reconstructed volumes, resulting in blurred anatomical structures and reduced diagnostic quality. Therefore, pose refinement techniques are essential to correct these initial estimates by optimizing probe positions based on the acquired image data.

Pose refinement in Freehand 3D Ultrasound. Given that pose errors are largely unavoidable in tracked freehand

ultrasound due to sensor noise, calibration inaccuracies, and environmental interference, various strategies have been proposed to refine probe trajectories and enhance reconstruction accuracy. Hybrid tracking systems [12] that combine optical and electromagnetic sensors have been developed to improve tracking precision, while phantom-based calibration [24] and error filtering techniques [4] aim to reduce drift and correct systematic errors. However, these methods rely on expensive hardware and remain susceptible to line-of-sight occlusions or magnetic distortions, limiting their practicality in clinical environments.

Recent efforts have explored deep learning-based pose refinement [1, 23, 33], leveraging image and motion information to improve tracking accuracy. While promising, these methods operate primarily in the pixel intensity space, depending on correlation patterns between successive frames. As a result, they are sensitive to image noise, motion ambiguity, and typically require large, fixed-motion datasets that limit generalizability.

To address these limitations, we introduce UltraNBA, a patient-specific deep learning method that refines probe poses without relying on large-scale datasets. Unlike traditional approaches that operate in the pixel intensity space and rely on frame-to-frame correlations, UltraNBA learns in a latent space that reflects underlying physical properties of the anatomy, which are more consistent across frames and scanning conditions. The method simultaneously optimizes pose correction and 3D reconstruction within an implicit neural representation, enabling continuous volumetric reconstructions from freehand sweeps performed at varying speeds. While previous implicit volume methods [35] also model 3D structure, they remain limited by their dependence on image intensity alone. In contrast, UltraNBA’s representation facilitates more robust reconstruction by leveraging acoustic properties that remain more consistent across frames than raw image intensities, making it more reliable under challenging acquisition conditions.

3. Methodology

Reconstructing 3D freehand ultrasound volumes requires both an accurate scene representation and precise probe tracking. However, existing implicit models, such as UltraNeRF [32], rely on voxel coordinates derived from tracked probe positions, making them highly sensitive to calibration and motion errors. These inaccuracies lead to image misalignments, degrading the quality of reconstructed volumes.

To address these issues, we introduce UltraNBA, a novel implicit neural reconstruction method that jointly learns the ultrasound volume representation and refines poses. Our key innovation is a learnable pose refinement module, inspired by BARF [17], which refines the estimated probe

poses $\mathbf{T} \in SE(3)$ through a transformation function:

$$\mathbf{T}' = c_{\xi_i}(\mathbf{T}) = \exp(\hat{\xi}_i) \mathbf{T}, \quad (1)$$

where $\xi_i \in \mathbb{R}^6$ is a learnable refinement in $\mathfrak{se}(3)$, composed of translation $\delta \mathbf{t} \in \mathbb{R}^3$ and rotation $\delta \boldsymbol{\omega} \in \mathbb{R}^3$ for each I_i slice in the volume. The wedge operation (\wedge) maps ξ_i to its corresponding $\mathfrak{se}(3)$ Lie algebra element, forming a skew-symmetric matrix that parameterizes rigid body transformations in $SE(3)$. This adjustment refines the input camera parameters, improving spatial alignment.

UltraNBA represents the 3D ultrasound scene as a coordinate-based volumetric implicit function:

$$f_\theta : \mathbb{R}^{3+6L} \rightarrow \mathbb{R}^3 \\ (\delta, \beta, \phi) = f_\theta(\mathbf{q}, \psi(\mathbf{q})), \quad \mathbf{q} = (x, y, z) \in \mathbb{R}^3 \quad (2)$$

Here, f_θ maps spatial coordinates \mathbf{q} to the attenuation coefficient δ , reflectance β , and scattering intensity ϕ , which define the ultrasound rendering properties. The input coordinates are encoded using a positional encoding function ψ with L frequency bands:

$$\psi(\mathbf{q}) = (\sin(2^l \pi \mathbf{q}), \cos(2^l \pi \mathbf{q}))_{l=0}^{L-1}. \quad (3)$$

The final input to the network is the concatenation of the original coordinates and their encoded representations, allowing the model to capture both low- and high-frequency spatial details essential for ultrasound volume reconstruction.

By jointly optimizing both the scene representation and pose refinement, UltraNBA effectively reduces misalignment artifacts, producing high-fidelity 3D ultrasound volumes even from imperfect tracking data.

3.1. Training objective

The model is trained to minimize a combination of the following losses:

1. **L2 Loss:** Following the ultrasound volume rendering formulation introduced in UltraNeRF[32], we reconstruct intensity maps using our rendering equation and minimize the L2 loss between the generated and ground truth intensity maps:

$$\mathcal{L}_{L2} = \|I - \hat{I}\|_2^2, \quad (4)$$

where I is the rendered intensity map and \hat{I} is the ground truth.

2. **Structural Similarity Loss:** To enforce perceptual similarity, we apply the SSIM loss between the rendered and ground truth intensity maps:

$$\mathcal{L}_{SSIM} = 1 - \text{SSIM}(I, \hat{I}). \quad (5)$$

3. **Regularization Loss:** To improve stability and enforce smoothness in the estimated ultrasound parameters, we use a regularized version of Ultra-NeRF [30] with regularization term consisting of two components:

- (a) **Local Normalized Cross-Correlation (LNCC) Penalty:** To encourage consistency between the estimated scatter amplitude and attenuation coefficients, we apply a localized NCC penalty:

$$\mathcal{L}_{LNCC} = \text{LNCC}(A_{\text{scatter}}, \delta), \quad (6)$$

where A_{scatter} is the scatter amplitude, δ is the attenuation coefficient.

- (b) **Total Variation (TV) Penalty:** To reduce noise and encourage spatial smoothness, we penalize variations in the scatter amplitude, weighted by the reflection coefficient:

$$\mathcal{L}_{TV} = \sum_{i,j} (\beta_{\max} - \beta) \cdot (|A_{\text{scatter}}^{i+1,j} - A_{\text{scatter}}^{i,j}| \\ + |A_{\text{scatter}}^{i,j+1} - A_{\text{scatter}}^{i,j}|) \quad (7)$$

where β_{\max} is a scaling factor based on the reflection coefficient β .

The regularization loss is then formulated as:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{LNCC}} \mathcal{L}_{\text{LNCC}} + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}, \quad (8)$$

where λ_{LNCC} and λ_{TV} control the weighting of the regularization terms.

The total loss function, with α a weighting hyperparameter, is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{L2} + \alpha \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{reg}}. \quad (9)$$

Our intensity reconstruction follows adapted rendering equations from UltraNeRF [31, 32], which build on prior ray-based simulation models [25]. These equations take the attenuation coefficient δ , reflectance β , and scattering intensity ϕ , predicted by f_θ , and apply ultrasound-specific rendering to generate the intensity maps. The reconstructed intensity maps serve as the predicted images, to which our loss functions are applied.

3.2. Gradient of pose refinements

We show the propagation of loss gradients to the learnable pose parameters ξ_i . From the total loss $\mathcal{L}_{\text{total}}$, the gradient w.r.t. input encoding is $\partial \mathcal{L}_{\text{total}} / \partial (\mathbf{q}, \psi(\mathbf{q}))$. For simplicity, we consider only $\psi(\mathbf{q})$, giving $\partial \mathcal{L}_{\text{total}} / \partial \mathbf{q} = (\partial \mathcal{L}_{\text{total}} / \partial \psi(\mathbf{q})) \cdot (\partial \psi(\mathbf{q}) / \partial \mathbf{q})$.

Coordinates \mathbf{q} are derived via a standard transformation from world coordinates \mathbf{p} as $\mathbf{q} = g(\mathbf{T}_i, \mathbf{p})$, where $g(\cdot)$ follows the standard NeRF ray sampling procedure, transforming world coordinates into the probe's local frame before projection. Since \mathbf{T}_i is parameterized by ξ_i , gradients propagate as $\partial \mathcal{L}_{\text{total}} / \partial \xi_i = (\partial \mathcal{L}_{\text{total}} / \partial \mathbf{q}) \cdot (\partial \mathbf{q} / \partial \mathbf{T}_i) \cdot (\partial \mathbf{T}_i / \partial \xi_i)$.

With \mathbf{T}_i defined as $\exp(\hat{\xi}_i)$, the gradient $\partial\mathbf{T}_i/\partial\xi_i$ remains well-formed, enabling pose refinement through gradient descent.

3.3. Training & Implementation Details.

We employ a similar frequency band activation strategy to BARF[17], gradually enabling higher-frequency components in the positional encoding as training progresses. Early stages prioritize low-frequency information for stable optimization, while higher frequencies are introduced over time to refine spatial details and refine pose errors.

Our architecture follows the original NeRF design with modifications for ultrasound data. The model is trained with a batch size of 1024 rays per iteration, using $L = 16$ frequency bands for positional encoding. We employ the Adam optimizer, setting an initial learning rate of 10^{-4} for the network, exponentially decaying to 10^{-5} over 500K iterations. Pose refinement is optimized separately with an initial learning rate of 10^{-4} , exponentially decaying to 10^{-5} over 500K iterations. In our experiments λ_{LNCC} is 1.e-3, λ_{TV} is 1.e-5 and α is 0.75.

4. Experiments and results

Our experiments were performed on four ultrasound recordings of two participants ($Participant_1 = P_1$, $Participant_2 = P_2$) at two different acquisition frequencies ($F_1 = 10\text{MHz}$, $F_2 = 9.5\text{MHz}$). Each recording consists of 5–6 overlapping sweeps of the lower limb, captured from knee to ankle, leading to more than 1744 ± 60 B-mode images with optical tracking per recording. To ensure precise tracking, we employed the OptiTrack system with six redundant cameras. To reduce pressure-related variability, the legs were submerged in water, and all scans were performed using a 2–10 MHz VERMON probe with an Aixplorer Supersonic Imagine ultrasound machine.

4.1. Correcting noisy tracking data

Freehand ultrasound tracking is prone to noise due to sensor limitations and hand tremors, causing discontinuities in the reconstructed volume. Our method corrects this noise by refining both rotation and translation, improving pose accuracy. To validate the estimated refinements, we perform two experiments. We first correct synthetically introduced noise. Second, we correct the original real tracking noise in the data from unknown sources of errors such as camera calibration, sensor inaccuracies, mechanical drift, and temporal synchronization issues.

The synthetic perturbations were calculated as follows: Each transformation matrix $\mathbf{T} \in \text{SE}(3)$ is parameterized by an element $\xi \in \mathfrak{se}(3)$, where $\xi = (\mathbf{t}, \omega)$ consists of a translation component $\mathbf{t} \in \mathbb{R}^3$ and a rotation component $\omega \in \mathbb{R}^3$. A perturbation $\delta\xi$ is applied to the transformation

Table 1. Pose refinement comparison between UltraNBA and baseline across datasets and noise levels (A-D). Δx (mm) and $\Delta\theta$ ($^\circ$) represent changes in translation and rotation error, respectively. Positive values indicate error reduction. Average rows are computed using absolute values.

P	N	Δx (mm)	$\Delta\%$	$\Delta\theta$ ($^\circ$)	$\Delta\%$
$P_1 - F_1$	A	0.03 ± 0.15	13	1.59 ± 3.73	25
	B	0.06 ± 0.27	13	1.57 ± 3.65	25
	C	0.06 ± 0.16	25	4.86 ± 7.39	35
	D	0.02 ± 0.27	4	4.41 ± 7.40	32
$P_1 - F_2$	A	0.01 ± 0.15	4	1.63 ± 3.64	25
	B	0.06 ± 0.27	13	1.54 ± 3.57	24
	C	0.05 ± 0.16	21	4.52 ± 7.67	33
	D	0.03 ± 0.28	6	4.35 ± 7.60	32
$P_2 - F_1$	A	0.01 ± 0.15	4	1.33 ± 3.61	21
	B	0.05 ± 0.27	10	1.26 ± 3.66	20
	C	0.04 ± 0.16	17	4.10 ± 7.45	30
	D	0.03 ± 0.28	6	3.84 ± 7.39	28
$P_2 - F_2$	A	0.01 ± 0.14	4	1.79 ± 3.59	28
	B	0.06 ± 0.27	6	1.61 ± 3.54	25
	C	0.03 ± 0.15	8	4.18 ± 7.70	30
	D	0.03 ± 0.28	6	4.11 ± 7.63	29
Average A		0.015	6.25	1.585	24.7
Average B		0.058	10.5	1.495	23.5
Average C		0.045	17.7	4.415	32.0
Average D		0.028	5.5	4.178	30.2

\mathbf{T} following the expression:

$$\tilde{\mathbf{T}} = \exp(\delta\xi)\mathbf{T},$$

where $\delta\xi = \begin{bmatrix} \delta\mathbf{t} \\ \delta\omega \end{bmatrix}$, is sampled from: $\delta\mathbf{t} \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_3)$, $\delta\omega \sim \mathcal{N}(0, \sigma_r^2 \mathbf{I}_3)$ with standard deviations σ_r and σ_t defining the noise strength. The sampled perturbations are mapped to $\text{SE}(3)$ via the exponential map and applied to \mathbf{T} . We use two levels for rotation, $\sigma_r = 0.07$ and $\sigma_r = 0.15$, corresponding to approximately 6° and 13° , and two levels for translation, $\sigma_t = 0.15$ and $\sigma_t = 0.3$, corresponding to perturbations of approximately 0.24 mm and 0.48 mm. Combining the noise levels leads to four noise setups were introduced to each recording, for a total of 16 experiments.

Figure 2 compares the translation and rotation errors on the $P_2 - F_2$ data at initialization with noise and after UltraNBA training. Our results indicate a consistent reduction in rotation errors across all noise levels, while translation errors either decrease or show minor variations, highlighting the greater impact of rotational misalignment in volumetric reconstruction. Notably, when the initial translation error is significant, our method effectively corrects it, demonstrating robust pose refinement. This trend is further supported by Table 1, which shows that our approach consistently improves alignment across different data and noise setups through iterative tracking error correction.

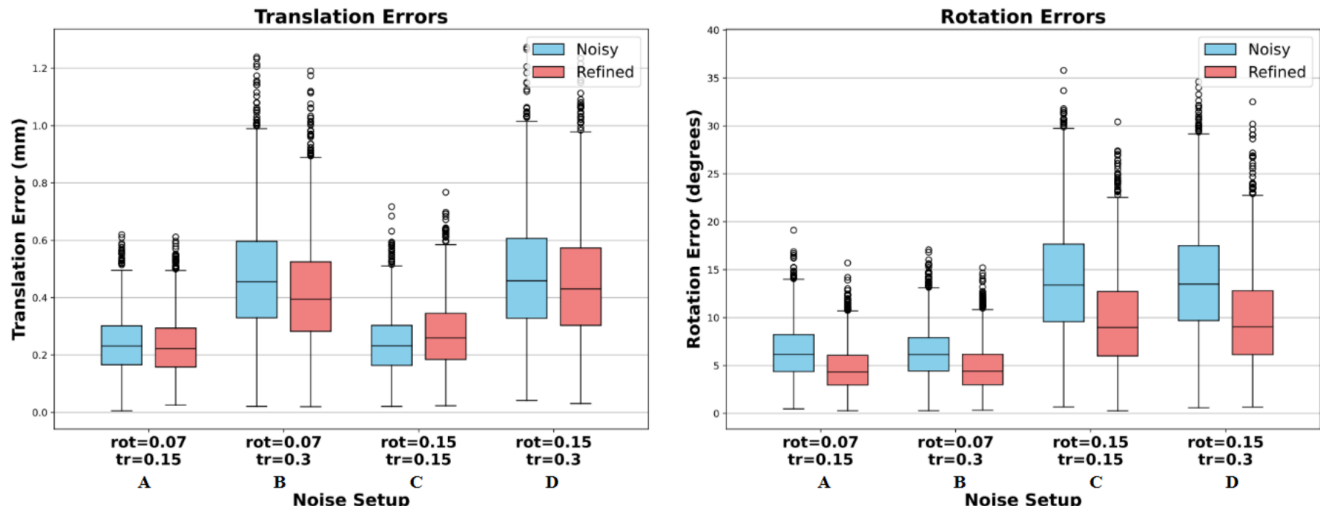


Figure 2. Comparison of translation and rotation errors on the Participant2-F2 data before and after **UltraNBA** training. Error bars show mean and standard deviation across noise conditions. The horizontal axis represents noise setups (rotation: 0.07, 0.15 \rightarrow 6°, 13°; translation: 0.15, 0.3 \rightarrow 0.24 mm, 0.48 mm). The vertical axis represents translation error as the displacement norm and rotation error as the angle derived from the relative rotation matrix trace.

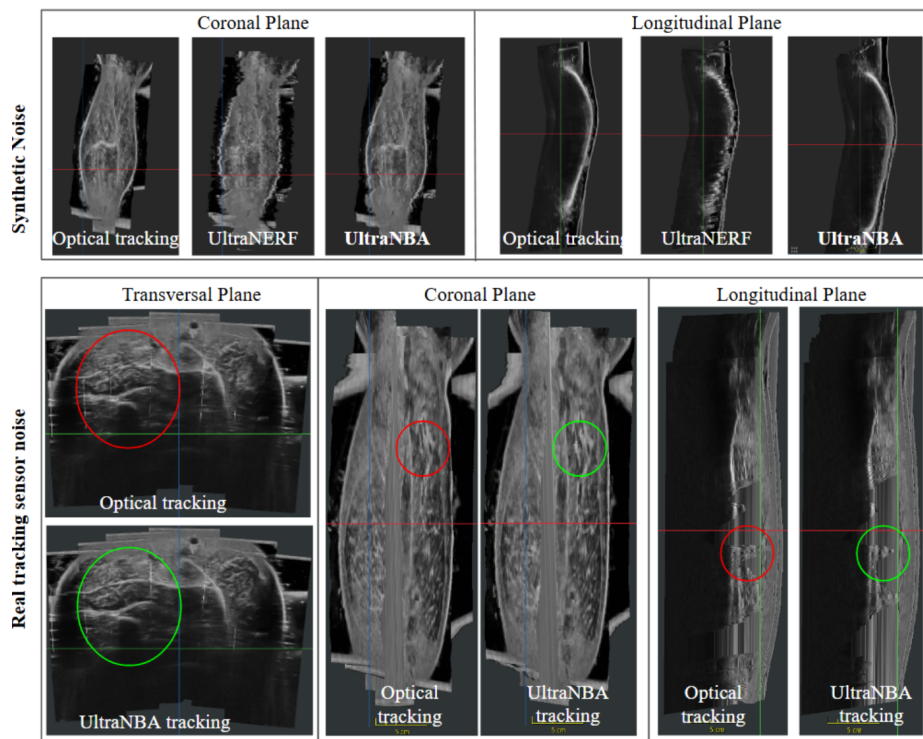
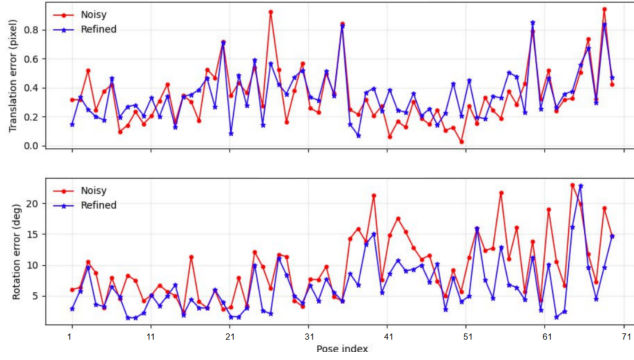


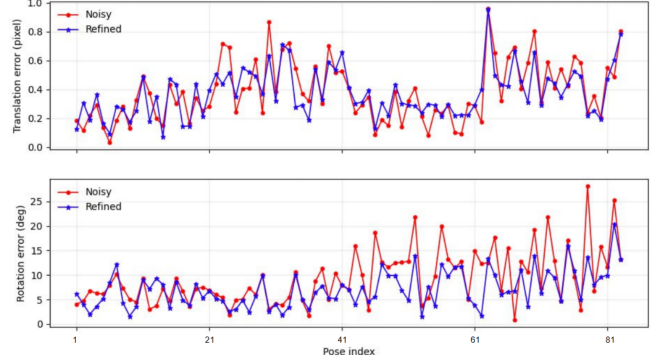
Figure 3. Comparison of pose refinement using **UltraNBA**. The first row illustrates **synthetic** noise added to the original optical tracking, while the second row shows **real** tracking noise caused by inherent tracking inaccuracies.

Figure 4 presents a robustness evaluation of the proposed **UltraNBA** method on the P1-F1 and P2-F1 datasets, across B synthetic noise level and unseen test frames. Subfigures-(a)(b) evaluate the pose errors on test images, isolated from training, across five different sweeps. Errors are presented

for translation and rotation independently, due to their difference in measured unit. Each segment corresponds to a specific sweep, revealing that initial pose errors (red) exhibit variable patterns and magnitudes across sequences. After refinement (blue), the errors are significantly reduced



(a) Tracking error correction for 71 test images for P1-F1-Noise B



(b) Tracking error correction for 83 test images for P2-F1-Noise B

Figure 4. Robustness analysis of Ultra-Nerf and the proposed **ULTRA-NBA** method for Participant1-F1, Participant2-F1 recording with 5 sweeps (sequences). Remark: one pixel length represents 0.25 mm in practice.

and stabilized across all sweeps, confirming the method’s ability to generalize to unseen frames and handle diverse motion patterns. The consistent performance across sequences of different participants and different frequencies emphasizes the adaptability of UltraNBA to different scan trajectories and anatomical variability, supporting its practical use in real-world ultrasound workflows.

4.2. Quantitative evaluation of pose refinement and rendering quality

Table 2. Image quality comparison using PSNR (dB) and SSIM between UltraNeRF and UltraNBA. Higher values indicate better image fidelity. Averages are computed per noise level; bold indicates the better value per method comparison.

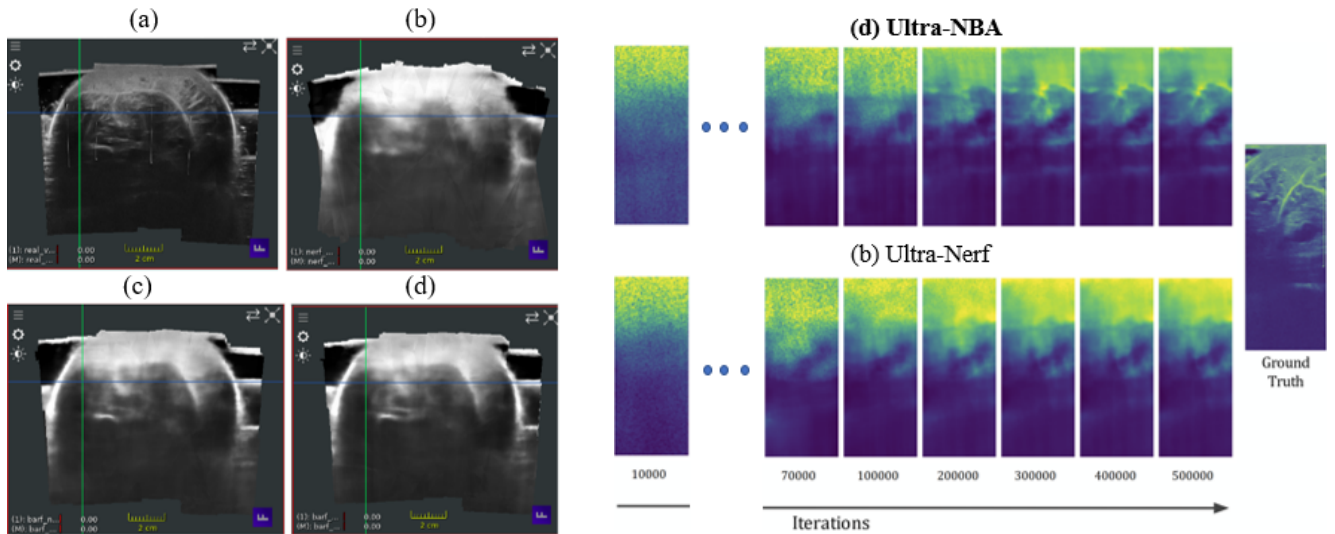
P	N	UltraNeRF		UltraNBA	
		PSNR (\uparrow)	SSIM (\uparrow)	PSNR (\uparrow)	SSIM (\uparrow)
$P_1 - F_1$	A	14.40	0.61	16.74	0.67
	B	14.08	0.59	16.41	0.66
	C	12.79	0.54	15.67	0.63
	D	13.17	0.56	15.76	0.64
$P_1 - F_2$	A	15.37	0.65	17.98	0.69
	B	15.45	0.65	17.85	0.69
	C	14.04	0.63	17.25	0.68
	D	14.19	0.63	17.40	0.68
$P_2 - F_1$	A	12.96	0.54	14.75	0.60
	B	12.85	0.53	14.62	0.59
	C	11.78	0.50	14.12	0.58
	D	11.98	0.50	14.01	0.58
$P_2 - F_2$	A	16.49	0.44	18.44	0.46
	B	16.50	0.43	18.23	0.46
	C	15.46	0.43	17.80	0.45
	D	15.52	0.43	17.90	0.45
Average	A	14.80	0.56	16.98	0.60
	B	14.72	0.55	16.78	0.60
	C	13.52	0.53	16.21	0.58
	D	13.72	0.54	16.27	0.58

We evaluated our method across diverse experimental conditions, as summarized in Table 1 and Table 2. These include variations in participants, scanning trajectories, and noise configurations to simulate both realistic and challenging tracking conditions. In the average part of the table, we can see that for the different noise setups configurations (A,B,C,D), in a patient independent evaluation, UltraNBA consistently reduced both translation and rotation errors, demonstrating robust pose refinement even under high noise levels. In addition to geometric accuracy, our method significantly improved image reconstruction quality of the implicit learned rendered image, as reflected by higher PSNR and SSIM scores compared to UltraNeRF. These results confirm that UltraNBA not only effectively compensates for synthetic tracking distortions but also generalizes well to real-world freehand ultrasound scenarios, enhancing both the reliability and fidelity of volumetric reconstructions.

4.3. Qualitative evaluation of pose refinement

Numerical improvements alone do not fully capture the impact of pose refinement on 3D volume reconstruction. We qualitatively assess UltraNBA’s ability to restore structural continuity and mitigate distortions caused by both synthetic and real-world tracking errors.

Figure 3 showcases tracking noise correction, with the first row depicting synthetic noise and the second-row addressing real sensor errors. Misalignments between adjacent slices create visible discontinuities, particularly at structural boundaries. UltraNBA effectively restores smooth transitions while preserving fine details, focusing corrections on anatomical edges while leaving homogeneous regions largely unchanged. This selective correction explains why numerical errors persist, UltraNBA prioritizes structural integrity over absolute error minimization. For real-world sensor tracking, UltraNBA mitigates subtle misalignments, reducing jitter and discontinuities while



(a) Render evaluation: (a)B-mode + Optical tracking, (b)Ultra-Nerf Rendering+ Noisy tracking, (c)Ultra-NBA rendering with not frequency, (d)Ultra-NBA rendering with frequency.

(e) Rendering improvement of one single frame at different iterations steps of the training for Ultra-Nerf and **Ultra-NBA** compared with the Ground Truth on the right side.

Figure 5. Rendering performance analysis of Ultra-Nerf and the proposed **ULTRA-NBA** method with and without coarse-to-fine frequency learning for Participant2-F2.

maintaining spatial consistency without introducing artificial blur.

Figure 5 illustrates the rendering performance of UltraNBA compared to UltraNeRF under tracking noise. Subfigure (a) in the left shows qualitative comparisons, where noisy input trajectories (b) cause visible distortions and loss of anatomical detail in UltraNeRF reconstructions. UltraNBA mitigates these effects through learned pose correction (c), with further enhancement observed when applying our frequency-based modulation strategy (d), which sharpens edges and improves structural clarity. Subfigure (b), on the right, provides a temporal rendering progression of a representative frame throughout training. While UltraNeRF remains stuck in a blurry reconstruction due to incorrect tracking, UltraNBA progressively refines both the tracking and the volume, resulting in a sharper and more accurate representation. These results demonstrate the effectiveness of both the joint pose optimization and the frequency-guided learning strategy in producing accurate and consistent reconstructions from noisy inputs.

5. Conclusion

In this work, we address two main challenges in 3D freehand ultrasound: pose refinement for improving tracking accuracy, and low-resolution rendering in learned implicit representations like UltraNeRF. We introduced UltraNBA, a novel neural bundle adjustment framework that jointly performs robust pose refinement and high-fidelity rendering for 3D freehand ultrasound imaging.

UltraNBA significantly improves both alignment accuracy and reconstruction quality. Specifically, it achieves up to a 35 % reduction in rotation error and 25 % in translation error across various noise levels, compared to baseline initializations. UltraNBA demonstrates notable improvements in image fidelity over UltraNeRF, achieving a maximum PSNR gain of 3.21dB (from 14.04 to 17.25dB in case P1-F2, noise level C) and a maximum SSIM improvement of 0.06 (from 0.61 to 0.67 in case P1-F1, noise level A), highlighting its effectiveness in producing high-quality ultrasound renderings. Its ability to handle both single and multi-sweep acquisitions further highlights its robustness and suitability for real-world clinical workflows. To the best of our knowledge, this is the first method to explicitly refine probe poses using neural fields in the context of 3D freehand ultrasound. Moreover, UltraNBA is data-efficient and requires no external tracking hardware, making it ideal for practical deployment. Future work will extend UltraNBA’s adaptability to broader clinical use cases, including integration with IMU-equipped probes and incorporation of additional modalities for multi-modal diagnostics.

6. Acknowledgments

This work was partially supported by the HINAV project funded by the Bavarian State Ministry for Economic Affairs, Regional Development and Energy.

References

- [1] Chrissy A Adriaans, Mark Wijkhuizen, Lennard M van Karnebeek, Freija Geldof, and Behdad Dashtbozorg. Trackerless 3d freehand ultrasound reconstruction: A review. *Applied Sciences*, 14(17):7991, 2024.
- [2] Andrea Alfarano, Luca Maiano, Lorenzo Papa, and Irene Amerini. Estimating optical flow: A comprehensive review of the state of the art. *Computer Vision and Image Understanding*, page 104160, 2024.
- [3] Hatem Alismail, Brett Browning, and Simon Lucey. Photometric bundle adjustment for vision-based slam. In *Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part IV 13*, pages 324–341. Springer, 2017.
- [4] Benjamin Busam, Patrick Ruhkamp, Salvatore Virga, Beatrice Lentes, Julia Rackerseder, Nassir Navab, and Christoph Hennersperger. Markerless inside-out tracking for 3d ultrasound compounding. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings*, pages 56–64. Springer, 2018.
- [5] Jian-Feng Chen, J Brian Fowlkes, Paul L Carson, and Jonathan M Rubin. Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test. *International Journal of Imaging Systems and Technology*, 8(1):38–44, 1997.
- [6] Xiankang Chen, Tiexiang Wen, Xingmin Li, Wenjian Qin, Donglai Lan, Weizhou Pan, and Jia Gu. Reconstruction of freehand 3d ultrasound based on kernel regression. *Biomedical engineering online*, 13:1–15, 2014.
- [7] Weijian Cong, Jian Yang, Danni Ai, Hong Song, Gang Chen, Xiaohui Liang, Ping Liang, and Yongtian Wang. Global patch matching (gpm) for freehand 3d ultrasound reconstruction. *Biomedical engineering online*, 16:1–26, 2017.
- [8] Juliette Conrath and Catherine Laporte. Towards improving the accuracy of sensorless freehand 3d ultrasound by learning. In *International Workshop on Machine Learning in Medical Imaging*, pages 78–85. Springer, 2012.
- [9] Amaël Delaunoy and Marc Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1493, 2014.
- [10] DEO Dewi, MHF Wilkinson, TLR Mengko, IKE Purnama, PMA Van Ooijen, AG Veldhuizen, NM Maurits, and GJ Verkerke. 3d ultrasound reconstruction of spinal images using an improved olympic hole-filling method. In *International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering 2009*, pages 1–5. IEEE, 2009.
- [11] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsdslam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014.
- [12] Hengtao Guo, Xuanang Xu, Sheng Xu, Bradford J Wood, and Pingkun Yan. End-to-end ultrasound frame to volume registration. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 56–65. Springer, 2021.
- [13] Richard Hartley. *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [14] R James Housden, Andrew H Gee, Graham M Treece, and Richard W Prager. Sensorless reconstruction of unconstrained freehand 3d ultrasound data. *Ultrasound in medicine & biology*, 33(3):408–419, 2007.
- [15] Richard James Housden, Graham M Treece, Andrew H Gee, and Richard W Prager. Calibration of an orientation sensor for freehand 3d ultrasound and its use in a hybrid acquisition system. *Biomedical engineering online*, 7:1–13, 2008.
- [16] San Jiang, Cheng Jiang, and Wanshou Jiang. Efficient structure from motion for large-scale uav images: A review and a comparison of sfm tools. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:230–251, 2020.
- [17] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021.
- [18] Laurence Mercier, Thomas Langø, Frank Lindseth, and D Louis Collins. A review of calibration techniques for freehand 3-d ultrasound systems. *Ultrasound in medicine & biology*, 31(4):449–471, 2005.
- [19] 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. *Communications of the ACM*, 65(1):99–106, 2021.
- [20] Mohammad Hamed Mozaffari and Won-Sook Lee. Freehand 3-d ultrasound imaging: a systematic review. *Ultrasound in medicine & biology*, 43(10):2099–2124, 2017.
- [21] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.
- [22] Anna Pichiecchio, Francesco Alessandrino, Chandra Borlototto, Alessandra Cerica, Cristina Rosti, Maria Vittoria Raciti, Marta Rossi, Angela Berardinelli, Giovanni Baranello, Stefano Bastianello, et al. Muscle ultrasound elastography and mri in preschool children with duchenne muscular dystrophy. *Neuromuscular Disorders*, 28(6):476–483, 2018.
- [23] Raphael Prevost, Mehrdad Salehi, Simon Jagoda, Navneet Kumar, Julian Sprung, Alexander Ladikos, Robert Bauer, Oliver Zettinig, and Wolfgang Wein. Deep learning-based 3d freehand ultrasound reconstruction with inertial measurement units. *Medical Image Analysis*, 2018.
- [24] Matteo Ronchetti, Julia Rackerseder, Maria Tirindelli, Mehrdad Salehi, Nassir Navab, Wolfgang Wein, and Oliver Zettinig. Pro-tip: phantom for robust automatic ultrasound calibration by tip detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 84–93. Springer, 2022.
- [25] 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 *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 510–518, Cham, 2015. Springer International Publishing.
- [26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [27] Ole Vegard Solberg, Frank Lindseth, Hans Torp, Richard E Blake, and Toril A Nagelhus Hernes. Freehand 3d ultrasound reconstruction algorithms—a review. *Ultrasound in medicine & biology*, 33(7):991–1009, 2007.
- [28] Matthew Toews and William M Wells. Phantomless auto-calibration and online calibration assessment for a tracked freehand 2-d ultrasound probe. *IEEE transactions on medical imaging*, 37(1):262–272, 2017.
- [29] J Nerney Welch, Jeremy A Johnson, Michael R Bax, Rana Badr, and Ramin Shahidi. A real-time freehand 3d ultrasound system for image-guided surgery. In *2000 IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No. 00CH37121)*, pages 1601–1604. IEEE, 2000.
- [30] M. Wysocki. Neural radiance fields for ultrasound imaging. Master’s thesis, Technische Universität München, 2023.
- [31] Magdalena Wysocki. Neural radiance fields for ultrasound imaging. Master’s thesis, Technische Universität München, Munich, Germany, 2023.
- [32] Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: Neural radiance fields for ultrasound imaging. In *Medical Imaging with Deep Learning*, pages 382–401. PMLR, 2024.
- [33] Zhongnuo Yan, Xin Yang, Mingyuan Luo, Jiongquan Chen, Rusi Chen, Lian Liu, and Dong Ni. Fine-grained context and multi-modal alignment for freehand 3d ultrasound reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 340–349. Springer, 2024.
- [34] Murat Yassa, Memis Ali Mutlu, Pinar Birol, Taha Yusuf Kuzan, Erkan Kalafat, Canberk Usta, Emre Yavuz, Ilkhan Keskin, and Niyazi Tug. Lung ultrasonography in pregnant women during the covid-19 pandemic: an interobserver agreement study among obstetricians. *Ultrasonography*, 39(4):340, 2020.
- [35] Pak-Hei Yeung, Linde Hesse, Moska Aliasi, Monique Haak, Weidi Xie, Ana IL Namburete, INTERGROWTH 21st Consortium, et al. Implicitvol: Sensorless 3d ultrasound reconstruction with deep implicit representation. *arXiv preprint arXiv:2109.12108*, 2021.
- [36] Honggang Yu, Marios S Pattichis, Carla Agurto, and M Beth Goens. A 3d freehand ultrasound system for multi-view reconstructions from sparse 2d scanning planes. *Biomedical engineering online*, 10:1–22, 2011.