Efficient neural reconstruction for freehand 3-D ultrasound imaging and visualization in Augmented reality

François Gaits, Fabien Vidal, Adrian Basarab, Nicolas Mellado

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier XX.XXXX/ACCESS.202X.XXXXXXXX

Efficient neural reconstruction for freehand 3-D ultrasound imaging and visualization in Augmented reality

FRANÇOIS GAITS[1], and FABIEN VIDAL[1,2], and ADRIAN BASARAB[3], and NICOLAS MELLADO[1],
[1]Institut de Recherche en Informatique de Toulouse, UMR CNRS 5505, Université de Toulouse, France
[2]Department of gynecologic surgery, Clinique Croix du Sud Ramsey Santé, Quint Fonsegrives, France
[3]Université de Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, Villeurbanne, France

Corresponding author: François Gaits (e-mail: francois.gaits@ irit.fr).

ABSTRACT 3-D ultrasound reconstruction associated with augmented reality allows physicians to explore a region of interest in 3-D in an intuitive and user-friendly way, while leveraging the advantages of 2D ultrasound imaging: simple, low cost and non-ionizing. It may assist many clinical tasks, such as practician training, procedure assistance or visualization of tissues difficult to interpret through 2D visualization. Recently, new unsupervised deep learning techniques based on a continuous description of the 3-D field, showed promising results in terms of 3-D model estimation, robustness to noise and uncertainty, and efficiency. Inspired by these approaches, the objective of this work is to propose a 3-D ultrasound reconstruction method based on neural implicit representations, adapted to the challenges of an augmented reality pipeline. Results on simulated and experimental data show the precision and efficiency of the reconstruction compared to state-of-the-art neural and traditional reconstructions.

INDEX TERMS Ultrasound reconstruction, Neural implicit representation, 3-D ultrasound imaging, Unsupervised deep learning, Augmented reality

I. INTRODUCTION

Ultrasound imaging is one of the most used medical imaging modalities, due to its relatively low-cost, ease-of-use and non-ionizing nature, the strong appeal of this modality has lead to many works aiming to improve the accuracy and interpretability of ultrasound images [1], [2]. The standard procedure in most clinical applications is to acquire temporal series of 2-D images, i.e., slices, of the examined tissues. This considerably limits the ability of ultrasound imaging to represent the 3-D geometry of organs, and motivates an intensive literature on 3-D ultrasound imaging. 3-D ultrasound volumes could represent a powerful tool to assist physicians in number of applications such as, for example, organ volume measurement [3], medical procedure assistance [4] or fetal examination [5]. 3-D ultrasound imaging can be obtained through several techniques. The most usual are based on mechanically-driven moving 1-D or 2-D arrays [6], [7]. However, the first option results into low-rate volumes and low-functionality, while the latter suffers from technological limitations related to the high channel count and limited field-of-views.

An alternative to the above solutions is to reconstruct 3-D ultrasound volumes from a collection of 2-D slices acquired by manually moving a 1D ultrasound probe. To perform such a reconstruction, the position of each 2-D image in space needs to be known. Three main approaches are usually employed to obtain these positions: mechanical guiding of the probe [8], [9], computation of the relative position of the slices [10] or tracking the probe using six degrees of freedom (6DOF) position sensors, e.g., optical or magnetic trackers [11]. The particular interest of 3-D freehand ultrasound has been shown in a number of applications, e.g., [12]–[14].

Building on these approaches, the integration of 3-D freehand ultrasound with augmented reality (AR) shows promising perspectives in medical imaging and surgical applications [15]–[17]. This combination enables real-time visualization of internal structures in a 3-D spatial context, improving the accuracy and efficiency of diagnosis [18] and interventional procedures [16], [19], [20] in per-operative scenarios. In the latter scenarios, the tracking of the ultrasound probe and the resulting images enable precise repositioning within a global scene, so that the tissue reconstruction can be displayed to the physician in real-time. However, this aim of uninterrupted, real-time combination of 3-D freehand ultrasound and AR presents several challenges, among which a key issue is the need to minimize reconstruction time. This constraint is necessary to ensure seamless integration and interactive feedback during surgical procedures [15], [20]. In addition, the readability of the reconstructed volumes is critical [17], requiring adapted algorithms to enhance volume

clarity and reduce noise in the low resolution and high-noise ultrasound modality.

In this article, we consider a scenario in which an ultrasound probe is tracked in three dimensions with a magnetic sensor within an AR setup. The latter already provides an interesting tool for visualization of the 2-D slices acquired. The main objective here is to provide an even better visualization of the medium under examination by reconstructing a volume from the 2-D slices acquired and their 3-D positions, and to display it in AR in the same time frame. This scenario poses two main challenges: to be able to obtain a faithful reconstruction in interactive time, and for this reconstruction to be visually convincing and noise-free. To this effect, this paper presents a novel approach that employs an implicit representation of the underlying volume. This approach leverages recent advances in Neural Implicit Representations (NIR) to create a volume from acquired and localized ultrasound images. Inspired by other NIR applications, our approach uses a randomly initialized network that is trained in an unsupervised fashion on the ultrasound images of the scene only.

More precisely, this article presents an end-to-end pipeline that employs NIR to achieve three main contributions: (i) a rapid, efficient and accurate volume reconstruction, (ii) a lightweight and continuous representation, and (iii) a significant reduction of the speckle noise that naturally contaminates ultrasound images.

II. BACKGROUND AND RELATED WORKS

A. 3-D ULTRASOUND IMAGE RECONSTRUCTION

Since the seminal works on 3-D ultrasound reconstruction, algorithms based on voxel grids have been adopted as the basis of most algorithms [11], [21]. They can be classified in three categories: pixel-based (PBM), voxel-based (VBM) and function-based methods (FBM).

PBM usually consist in two steps: bin-filling and hole-filling [22]. In the bin-filling step, the pixel values of each input 2-D image are assigned to one or multiple voxels of the 3-D volume, based on their position. In the hole-filling step, the voxels with no value assigned are filled using various methods such as fast marching [23] or olympic filling [22].

In contrast to PBM, VBM directly assign a value to each voxel of the volume of interest (VOI) based on the neighboring input images. This approach has the advantage of creating a fully filled 3-D grid in one step and is the most common method in 3-D ultrasound reconstruction [11]. However, VBM are usually sensitive to noise and spatial inconsistency, i.e., widely separated inputs or region imbalance. The most standard VBM algorithm is the voxel-nearest neighbor (VNN) [24], which assigns to each voxel the value of the closest pixel in the dataset. Complementarily, the distance-weighted (DW) [24], [25] method aggregates the values of multiple near pixels and weight them using their distance, allowing a smoother and more robust reconstruction.

Finally, FBM use the input dataset to estimate functions that describe the values of each voxel in the volume of interest (VOI). Commonly used functions range from polynomial to radial basis functions and Bézier spline [26]. Less commonly used in ultrasound imaging, FBM suffer from large computational time [21], but benefit from strong properties in regard to robustness to noise or missing data.

The main challenges of these methods are related to their real-time capabilities [21], the difficulty to fill large gaps in data, or to overcome ultrasound imaging limitations such as low spatial resolution and signal-to-noise ratio, or variable depth penetration of the ultrasound waves [12], [27].

Another important challenge of these approaches is related to the resulting large voxel grid, which prevents the reconstructed volumes from being transmitted effectively to remote devices such as AR headsets. Another difficulty is the compromise between computational efficiency and quality of the final volumes, in particular in ultrasound images highly degraded by speckle noise.

B. NEURAL IMPLICIT REPRESENTATION

One of the most prominent NIR applications is novel view synthesis. It consists in generating an image representing of an unseen point-of-view from a collection of photographs. This field has recently observed important advances thanks to the use of NIR to learn intrinsic parameters of the scene [28], [29] and inputting them in a standard rendering pipeline. Following this seminal work, NIR has been shown to be effective in a number of applications such as surface reconstruction [30], semantic decomposition [31] or medical imaging [32], [33], detailed in the next section.

NIR is a family of approaches that use unsupervised deep learning to learn a volumetric function from 3-D samples [34], [35], in NIRs, a new network is initialized for each represented scene. Figure 1 illustrates the general functioning of implicit representations. In most cases, NIR uses Multilayer perceptron (MLP) [28], [30], [32] to quantify properties at any point in the VOI. The network in those applications represent a function of $\mathbb{R}^3 \rightarrow \mathbb{R}^n$ with n the number of parameters quantified by the network.

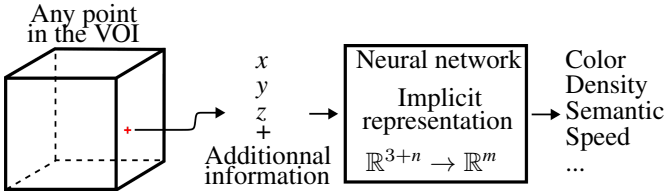

FIGURE 1. Neural implicit representations tie any set of properties in the VOI to their coordinates and additional information, forming an application of $\mathbb{R}^{3+n} \rightarrow \mathbb{R}^m$, with n the size of additional information added to the position vector, and m the size of the property vector.

The position vector inputted in the network can be augmented by additional information, or embedded to enhance expressivity. This embedding is one of the key aspects introduced by Mildenhall *et al.* [28] in their seminal work, and its impact on the ability of the network to represent details is explored and quantified in [36]. Additionally [29], [37] continued to expand this notion by proposing other types of encoding and providing a reflection on the order of the elements in the input vector. This embedding phase is the first, non-learnable, layer of the network. Its role is to increase the dimension of the position vector, allowing the network to represent more finely spatial variations.

The most widely used embedding function in NeRF-inspired networks [28], [30], and in particular in medical applications [32], [33], is the frequency embedding [28], [36]. Each element of the input vector $\mathbf{x}$, denoted by x_i in (1), is mapped from $\mathbb{R}$ to $\mathbb{R}^{2L}$ (with L being a meta-parameter) using the following γ function:

$$\gamma(\mathbf{x}) = [sin(2^0\pi x_i), cos(2^0\pi x_i), \cdots, sin(2^{L-1}\pi x_i), cos(2^{L-1}\pi x_i)], \quad (1)$$

where $i \in \{0, 1, 2\}$ for 3-D volume reconstruction applications.

Many works aiming to improve the accuracy and computational efficiency of this type of network take a particular interest to this crucial embedding step [29], [37]. Its role in ultrasound volume reconstruction is studied in-depth in this work, and represents one of the major contributions of this paper.

C. NEURAL-BASED 3-D RECONSTRUCTIONS IN MEDICAL IMAGING

NIRs have a number of noteworthy applications in medical imaging, particularly related to their close relation to the 3-D tissue they represent. Early work associated the measured data in the medium with their coordinates directly [32], such that the network represented a continuous function over the VOI. Such continuous representations further allow slicing the VOI at any position and under any angle. An example of this approach on ultrasound imaging can be found in [33], [38]. Li *et al.* compared the results of a direct application of NeRF to the standard VNN algorithm for the creation of a spine volume to evaluate spine curvature [38]. Wysocki *et al.* used the output of a NeRF network as multiple parameters to render an ultrasound slice using a simulation process, and used those learned parameters to re-slice the volume [33].

In addition, numerous works span from this line of thinking, and try to leverage the ability of deep learning to achieve more complex or diverse tasks. In [39], the authors aim at learning the relative pose of each slice of MR imaging in addition to reconstructing the volume in the form of a NIR. The method proposed in [40] associates a NIR with a segmentation network to map not only the volume but a precise semantic property it contains. An interesting work in this line can be found in [32], which first proved the feasibility of the approach by using the NeRF methodology, giving rise to works in multiple medical fields such as dentistry [41], otorhinolaryngology [42] and motricity [43].

Despite these appealing applications, and while NIR is remarkably lightweight for a deep learning application, it still suffers from too long training and inference (ranging from days [28] to minutes [32]) for our application. Moreover, their application in ultrasound imaging is still a big challenge, mainly because of the intrinsic nature of ultrasound images, contaminated by a high amount of speckle noise. In this work, we introduce a 3-D ultrasound reconstruction algorithm that takes advantage from NIR designing a specific learning scheme and network architecture, to create a lightweight, denoised (despeckled) volumetric representation in a significantly reduced amount of time.

Algorithm 1: Training process of the network

Data:
coordinates, *values* :=Ultrasound acquisitions
totalEpochs := 5,000
batchSize := 50,000

see section III-A
network := randomly initialized MLP

Result:
network := Trained network representing the volume

see section III-B2;
$S_1 \leftarrow subsampleData(1, coordinates)$; // 1%
$S_2 \leftarrow subsampleData(10, coordinates)$; // 10%
$S_3 \leftarrow subsampleData(50, coordinates)$; // 50%

see section III-B1;
Function `batchSelection(epoch)`**:**
  if $epoch \leq 750$ **then**
    $batch \leftarrow selectRandom(S_1, batchSize)$
  else if $epoch \leq 1500$ **then**
    $batch \leftarrow selectRandom(S_2, batchSize)$
  else if $epoch \leq 2500$ **then**
    $batch \leftarrow selectRandom(S_3, batchSize)$
  else
    $batch \leftarrow selectRandom(coordinates, batchSize)$
  return *batch*;

for $epoch \leftarrow 1$ **to** *totalEpochs* **do**
  $batch \leftarrow batchSelection(epoch)$
  $expected \leftarrow getValues(batch, values)$
  $estimated \leftarrow evaluateNetwork(network, batch)$
  see section III-A1;
  $loss \leftarrow MSE(expected, estimated)$
  $network \leftarrow backpropagation(network, loss)$

III. PROPOSED METHOD

In the context of our AR application, the ultrasound probe is 3-D tracked. This serves a double objective. First, it provides the relative position between the 2-D slices, and consequently the 3-D positions of each individual pixel. Second, the absolute positions of the slices in the 3-D world are available, thus allowing their placement in the original positions through the AR headset. From this collection of 2-D ultrasound images, the proposed method reconstructs a 3-D volume, which is then overlaid with the real-world location of the imaged tissue.

In contrast to standard use of NIR, we observed that standard design improvements (e.g., embedding function, network architecture and loss function) lead to a reduction of the performances and quality of the reconstruction from ultrasound images. This section presents the steps taken to adjust those aspects and use NIR efficiently with ultrasound imaging.

More precisely, we find that the embedding used, the network architecture and the order and diversity of the sample provided to the network during its training process is of great importance in order to achieve fast and qualitative reconstructions. This training process design from an initialized network to the scene representation constitutes the core of our proposed method, which is outlined in Algorithm 1 and

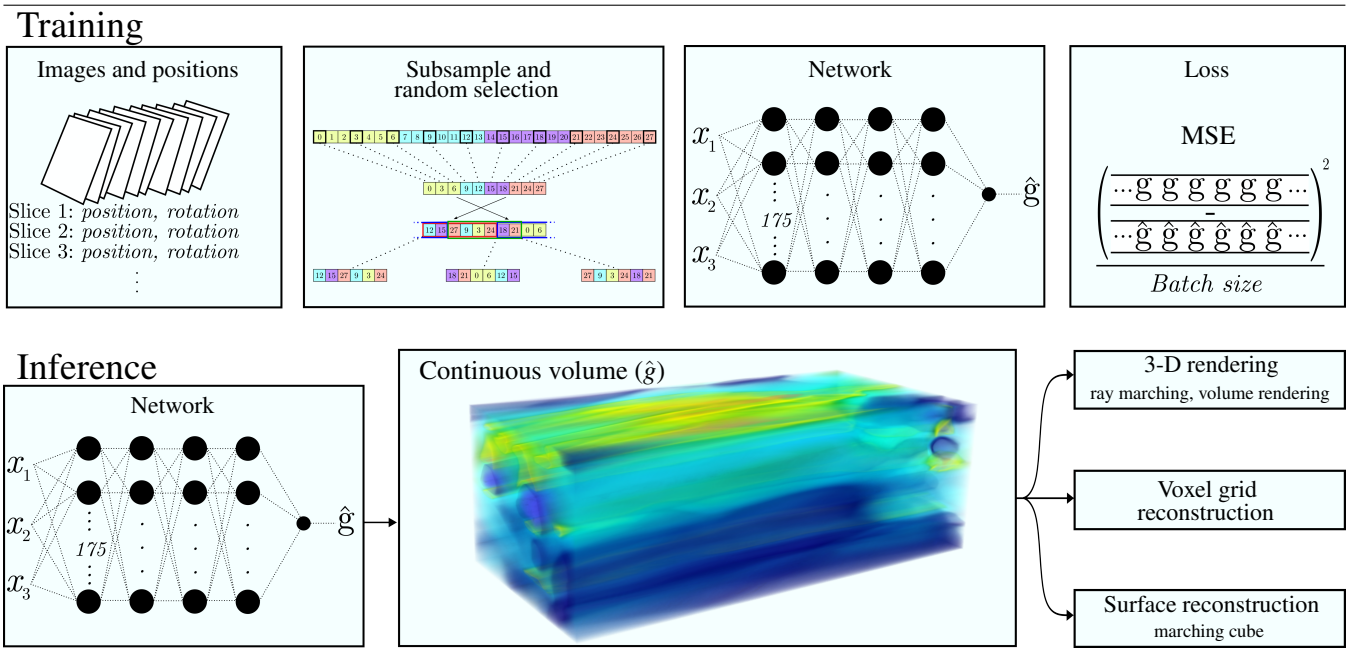

FIGURE 2. Main steps of the proposed 3-D ultrasound reconstruction method. Given a collection of ultrasound slices and their position in the 3-D world, a dedicated network is trained to represent the underlying imaged medium. From a collection of images and positions, each pixel represents a point (x_1, x_2, x_3) associated with a gray value g. A batch is then formed thanks to subsampling and random selection and the coordinate sent to the network to obtain the cooresponding inferences $\hat{g}$. The inferences and the real values for all the batches are then compared with the MSE function to establish the loss and train the network.

Figure 2. Each key point is subsequently discussed in the following subsections.

A. NETWORK

The objective is to develop a lightweight and fast framework. To achieve this goal, the network is analyzed hereafter under three aspects: its usage, its input and its architecture.

1) Network usage

Once trained, the network represents a volumetric function Φ_θ such that

$$\Phi_\theta(\mathbf{x}) = \hat{g}_\mathbf{x}, \tag{2}$$

where $\mathbf{x}$ is a 3-D vector containing the 3-D coordinates in the VOI and $\hat{g}_\mathbf{x} \in \mathbb{R}$ is the estimated value of an ultrasound image at position $\mathbf{x}$.

This function can be seen as a volume of infinite resolution, as any 3-D point $\mathbf{x}$ is associated with a value (within the spatial bounds of the dataset). However, given that it is a linear compounding of its weights and biases, the frequency of the variations of the values associated to $\mathbf{x}$ are limited [36]. This is a key aspect of our method, as constraining the network's capacity to represent high frequencies encourages it to circumvent the undesirable noise present in the original data, and especially the speckle.

From Φ_θ, there are three possible approaches to obtain an exploitable ultrasound volume: (i) use it directly through a volume rendering process such as ray-marching [44], (ii) discretize a given isovalue using marching cube [45] or a similar algorithm, (iii) discretize the function in an arbitrarily sized voxel grid. Given the flexibility of the network, each of those solutions can be employed in different use cases to represent the same volume. In this work, we primarily utilize voxel grids and isosurface discretization to compute metrics and provide a clear visualization of the reconstructed volumes. Our final AR pipeline is illustrated in Figure 12.

2) Inputs and embeddings

The embedding of the input coordinates into a larger vector representation allows the NIR-based networks to represent finer details. However, in the case of ultrasound images, high frequencies are mostly dominated by the speckle noise. Therefore, such fine details are not suitable in 3-D ultrasound reconstructions, and consequently as network outputs. Conversely, we propose to feed the coordinates directly in the network, after only a normalization of their respective elements between -1 and 1.

Figure 3 shows a comparison of the impact of different embeddings in ultrasound imaging. One acquired image not used in the training process is shown in comparison to the ones obtained by re-slicing the 3-D reconstructed volumes. In contrast to standard novel view synthesis applications that accumulate multiple network results in a single pixel, our method displays the inferences by visualizing each of them, yielding unwanted artifacts when applying a frequency embedding. Other embeddings, such as hash-grid embedding [29] allow the network to recreate the speckle noise originally present in the given examples and maintain the high-noise nature of the acquisition. On the other hand, passing directly the coordinates to the network makes it more complex to represent noise, and the effect of generalizing the shape of the objects present and to get rid of the speckle noise.

3) Architecture

Having a small-scale architecture speeds up the learning process by reducing the time required for each training epoch

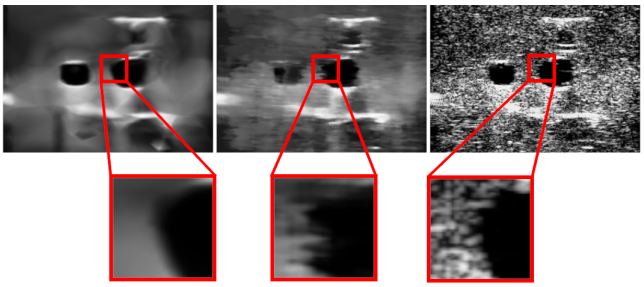

FIGURE 3. Comparison of volume reslicing at a known position using different embeddings. From left to right: direct input, frequency embedding [28] and base ultrasound image.

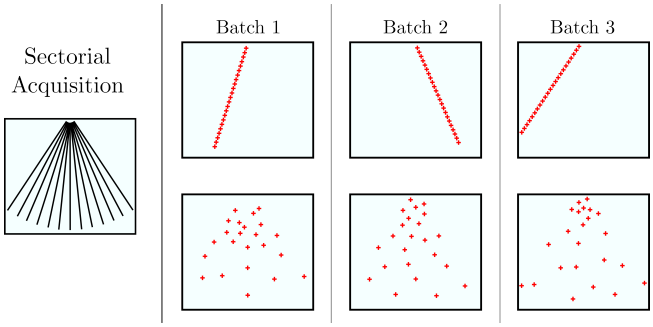

FIGURE 4. 2-D illustration of the selection of a learning batch. From the original acquisition, 3 learning batches of 23 points are formed. Top: An image is randomly selected and its 23 pixels form the learning batch, Bottom: 23 points are randomly selected from all pixels of the dataset.

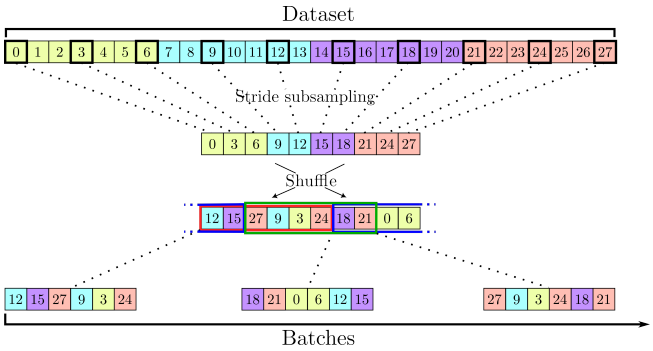

FIGURE 5. Illustration of the selection process of batches of size 6 from a dataset composed of 28 pixels from 4 images of 7 pixels. First, the whole dataset is subsampled by a factor of 3 and shuffled. The batches are then created using a sliding circular window to ensure each selected sample is used equally.

and its inference time. Furthermore, as presented in Section III-A1 and explored in Section IV-C1, the architecture selected should be minimal while allowing representation of any dataset.

For this reason, we use an architecture consisting of four layers of 175 fully connected neurons with ReLU activation. The optimizer used is Adam, with a learning rate of 5×10^{-3}. All experiments conducted in this work use this architecture. For all tested cases, it offers a good balance between processing speed and denoising, compared to alternative architectures investigated in Sections IV-C1 and IV-D.

B. TRAINING PROCESS

The network's parameters are randomly initialized for each dataset, and trained in an unsupervised manner. Based on the given 2-D ultrasound images collection, it to represent the function shown in (2) at each point in the VOI. To achieve this, the network is given examples in the form of pixels from ultrasound images, $(x_1, x_2, x_3, g_{\mathbf{x}})$, where $\mathbf{x} = (x_1, x_2, x_3)$ are the coordinates of the pixel in the VOI and $g_{\mathbf{x}}$ is its acquired gray value.

At each learning step, the network is given a set of coordinates $\mathbf{x} = (x, y, z)$ to produce an estimate of the gray level $\hat{g_{\mathbf{x}}}$ at that point in space. The estimates are then compared to the true value using the Mean Squared Error (MSE), yielding the following optimization problem:

$$\arg\min_{\theta} ||\Phi_\theta(\mathbf{x}) - g_{\mathbf{x}}||_2^2. \tag{3}$$

MSE is the standard loss function used in NIR applications. While many works on medical NIR (see Table 1) are using the Structural Similarity Index Measure (SSIM) or a weighted sum of both, MSE has significantly less computational overhead and is more compatible with the changes to batch formation made in this article.

Indeed, the proper selection of the coordinates that make up the batches has a major impact on the properties of the results and the efficiency of the network. We design a batch selection process, to enable the network to learn the desired properties as rapidly as possible.

1) Random selection of the samples

Given the constraints of our application, the proposed network needs to be able to generalize and capture the general shapes present in the VOI as early as possible during the training process.

Most of the existing approaches inherit directly from the novel view synthesis approach and use a single ultrasound image as a batch for each learning step [32], [33], [40]. However, in the case of ultrasound images, this reduces the representativeness of a batch to a single plane (i.e. the ultrasound image) through the volume. This process makes the network oscillate at each step between the information present in each image, at the expense of global information, as illustrated by Figure 4. Furthermore, it provides too many fine details in the form of closely related pixels, which the network is not able to learn.

To address this challenge, we ignore the structure of the ultrasound images, and rather select random pixels among all the available points, forming a batch. However, pure random selection has two main drawbacks: it can be very computationally expensive (randomly generating a large number samples at each iteration) and does not guarantee a balanced representation of the samples at any point during the learning phase.

For these reasons, the array of points considered for a given learning step is shuffled once, at the time of its creation. The selections are then performed by slicing the shuffled array in a section of the size of a batch. The slicing is performed by progressing in the array without overlap, circularly. The *i*-th

TABLE 1. Summary of the key components of some of the most prominent NeRF-inspired medical reconstruction.

Source	Modality	Input	Embedding	Architecture (Total number of neurons)	Loss	Time	Output
Implic-itVol[32]	Sensorless Ultrasound	Slices	Frequency	MLP 5 Layers of 128 neurons (640)	SSIM	10 000 epochs	Density and positions
Ultra-NeRF[33]	Ultrasound	Slices	Frequency	MLP 8 Layers 256 neurons (2048)	SSIM + MSE	Unspecified	5 render parameters
Carotid Imaging[40]	Ultrasound and segmentation	Slices	Frequency	MLP 9 Layers of 256 neurons (2304)	MSE + Semantic loss	300 000 epochs	Density and classification
IREM[46]	MRI	Voxels	Random frequency	MLP 13 Layers of 256 neurons and 2 Layers of 512 (4352)	MSE	Unspecified	Density
Ours	Ultrasound	Subsampling and Random selection	Direct	MLP 4 Layers of 175 neurons (700)	MSE	5 000 epochs, 30 seconds	Density

batch is then composed of the elements between the indices $i \times k + 1 \mod n$ and $(i+1) \times k \mod n$ in the shuffled array, with k the size of a batch and n the total number of samples. This ensures efficiency and representativeness while being decoupled from the original structure of the data set. This process is illustrated in Figure 5.

2) Data subsampling

When the random selection of points is applied to the entire set of available points to constitute the batches, it may be necessary to perform numerous iterations before the network has been trained on the entirety of the VOI.

To address this issue, we propose to form the batches based on a subsample of all the available points (i.e. not all samples are considered for random selection but only a smaller subset), which size increases as learning progresses.

This results in the network having to generalize strongly in order to fit the few samples, and then to refine its representation as the proportion of the size of the subset increases. Furthermore, a restricted, well distributed set ensures that each batch will represent the whole VOI.

The creation of the subsets is performed by selecting every n^{th} pixels in each images with $\lfloor n = \frac{1}{proportion} \rfloor$, ensuring that the whole VOI is represented by each subset. The selection process of a batch from the initial dataset is represented in Figure 5.

In a typical application, the batches are drawn from 4 subsets: S_1, S_2, S_3, three subsamples of the full set (representing 1%, 10%, and 50% of the data, respectively), and the full data itself. Given a fixed number of epochs, the network will first learn from S_1 for the first 15% of training, followed by S_2 for another 15%, and then S_3 for 20%. The network then finishes its training on the entire dataset to capture structures that may have been missed during this generalization step.

IV. RESULTS

The results of our method are presented under three different aspects that are key to our application: accuracy through the comparison to ground truth available in simulated scenes, perception, by studying Contrast-to-Noise Ration (CNR) and Signal-to-Noise Ratio (SNR) in the reconstructed volumes, and training time.

Additionally, we conduct a qualitative analysis of various properties of the volumes reconstructed by our method, such as resolution in relation to network size and shape profile. This enables a more comprehensive understanding of the behavior of our method.

Comparisons with two existing methods are also conducted. The first comparative method is the standard voxel-based reconstruction Distance Weighted (DW) [24], [25].

The second, based on the methods [32], [33], [40] (see Table 1), is a reconstruction network with an MLP of size 8 by 256, using single ultrasound images as batches and using a frequency encoding, referred to as Network Baseline (NB). Despite its significantly longer train time, due to its heavy architecture, NB is trained until convergence in the following results.

Finally, we present an ablation study in which we desactivate each step of the proposed method to study its impact on the results. Some of the ablation cases correspond to existing methods, providing a larger comparison base for our method.

The results are generated from three different scenes presented in Figure 6:

- Shapes: A collection of two 15-mm sided and two 10-mm sided cubes and a sphere of 15-mm diameter.
- Pelvis: A dataset simulated from a patient pelvic MRI constituted of two orthogonal sweeps.
- Nerve: A real linear acquisition performed on a CAE blue phantom [47], a phantom used for vessel and nerve puncture training.

Additionally, we provide the sizes of the datasets:

- Shapes: 210 images of 192×256 pixels, giving 10,321,920 samples, covering a VOI of 7×7×7 cm^3, i.e. 686,000 voxels at 0.5 mm spatial sampling rate in each dimension.
- Pelvis: 240 images of 500×600 pixels, giving 72,000,000 samples, covering a VOI of 6×6×5 cm^3, i.e. 360,000 voxels at 0.5 mm spatial sampling rate in each dimension.
- Nerve: 243 images of 192×512 pixels, giving 23,887,872 samples, covering a VOI of 9×6×4 cm^3, i.e.

432,000 voxels at 0.5 mm spatial sampling rate in each dimension.

Furthermore, all simulated images include a random positional error and angular error at each slice, with a range of $\pm$ 0.1 mm and $\pm$ 0.03 rad respectively. The simulations were obtained by drawing scatterers in three dimensions over the ground truth volumes and extracting individual slices of coherent scatterers for each image, in accordance with the methodology outlined in [48]. We use this approach because it provides spatially coherent speckle, a typical property observed in real datasets and represents a significant challenge for reconstruction techniques.

A. RECONSTRUCTED VOLUMES

Figure 6 shows the reconstruction results obtained using the proposed method, NB and DW reconstruction on the 4 main datasets.

The DW reconstruction is faithful to the original data and reproduces the speckle noise introduced by the ultrasound imaging process. Similar to the real acquisition, the simulation process described in [48] produces spatially coherent slices, resulting in bumps and wrinkles on the surfaces and noise in the low signal areas.

The NB performs well at reproducing the details present in the shapes and is very close to the original data, as expected, thanks to the contribution of the embedding. However, many problems are raised by this method. First, the importance given to high frequencies contribute in favor of speckle noise and produce results noisier than DW in certain instances. Furthermore, in the areas with fewer samples, the network tends to reproduce the central structures due to the periodic nature of sine and cosine function.

On the other hand, the proposed method tends to summarize the shapes and smooth the surfaces, producing more visually convincing results and significantly reducing the impact of noise.

B. QUANTITATIVE RESULTS

This section quantifies the errors and properties of the reconstruction method.s This is key to measuring the efficiency of the approach and its acceptability for further use. In this section, all methods are given the necessary time to produce optimal results, as detailed in Section IV-B3.

The following sections consider two main aspects to quantify the reconstruction results: accuracy with respect to the ground truth, if available, and perception through the use of Signal-to-Noise Ratio (SNR) and Contrast-to-Noise Ratio (CNR).

All results presented in these sections and throughout the article for all methods were computed on a normalized voxel grid of 0.5 mm size, which represents the density of the reconstruction.

1) Accuracy

The first step is to quantify the errors introduced by the reconstructions in order to assess their acceptability in real applications. To do this, we compare the obtained reconstruction with the ground truth of the imaged medium, which is only possible on simulated datasets.

TABLE 2. Accuracy metrics for volumes reconstruction on datasets with known ground truth

		NB	DW	Ours
Shapes	MSE ↓	513.59	414.23	**260.66**
	MAE ↓	14.799	13.82	**8.02**
	NCC ↑	0.90	0.93	**0.94**
	SSIM ↑	0.53	0.81	**0.88**
Pelvis	MSE ↓	1105.06	1040.98	**976.16**
	MAE ↓	23.74	**23.40**	23.53
	NCC ↑	0.874	0.881	**0.914**
	SSIM ↑	0.621	0.623	**0.642**

Table 2 presents the results for four metrics: Mean Square Error (MSE), Mean Absolute Error (MAE), Normalized Cross-Correlation (NCC) and Structure Similarity Index Measure (SSIM). These results highlight the competitiveness of our method, which consistently produces comparable or better values across all metrics. On the Shapes dataset, the network takes full advantage of its shape generalization and noise removal properties, showing significant improvement on all metrics.

More importantly, on the more complex and large scene Pelvis, our method still shows satisfactory results, but is closely followed by the other methods, or even outperformed by the DW for the MAE. Indeed, as it is geared towards fast and generalized datasets, our method loses its edge on a dataset with a very large number of samples and details to be represented faithfully, but is still able to perform its reconstruction correctly. Furthermore, the visual aspect which is not captured by those metrics is more in line with our application.

2) Perception

To represent the perceptual properties of the reconstructions, we use two well-known metrics, SNR and CNR. Both measures have been realized on blocks of voxels extracted from inside and outside a region of interest and computed using the following formulas: $SNR = 20 \times \log_{10}(\frac{\mu_{in}}{\sigma_{in}})$ and, $CNR = 20 \times \log_{10}(\frac{|\mu_{in}-\mu_{out}|}{\sqrt{\sigma_{in}^2+\sigma_{out}^2}})$ with μ and σ being respectively the mean and the standard deviation of the regions.

TABLE 3. Perception metrics for volumes reconstruction on datasets with known ground truth

		NB	DW	Ours
Shapes	SNR ↑	48.80	31.82	**56.51**
	CNR ↑	23.86	22.93	**32.82**
Pelvis	SNR ↑	21.54	23.37	**30.63**
	CNR ↑	17.84	18.47	**27.32**
Nerve	SNR ↑	22.72	23.55	**25.64**
	CNR ↑	11.34	12.64	**14.57**

Table 3 shows these metrics for the four datasets. Similar to the accuracy metrics, the proposed method is able to match or outperform the baselines, with a marked convergence in

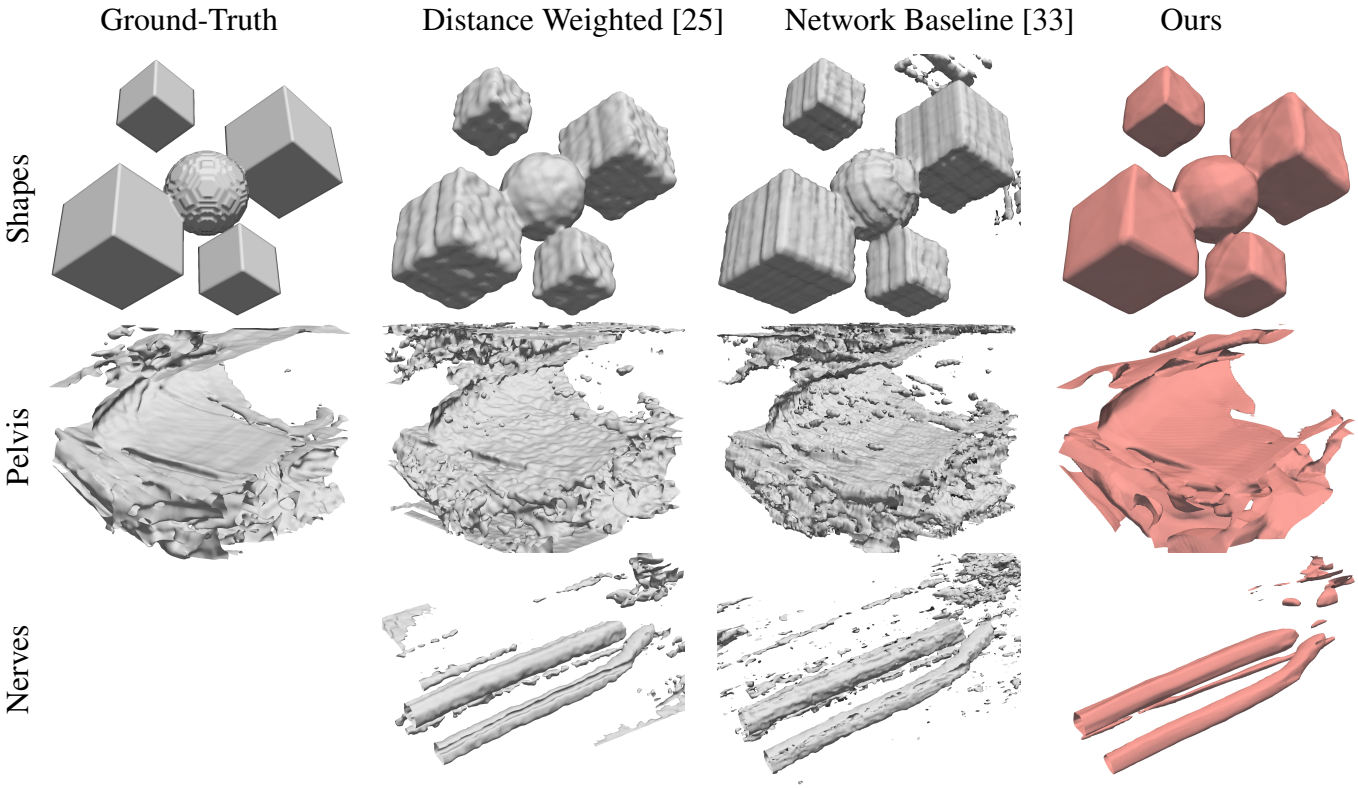

FIGURE 6. Reconstructed volume examples. Volumes are represented as a voxel grid with a resolution of 0.5mm, for visualization purposes an appropriate isosurface has been hand-picked. The ground-truth is provided by the explicit representations or MRI used to simulate the input images and is not available for an experimental dataset such as Nerves

values for more complex datasets. This confirms the bias of the network towards simple shapes.

3) Computational time

For all the results, both the proposed and NB networks were run until convergence was reached. In the case of NB, this duration depends on the dataset, since the size of the images, their spatial distribution, and their number have a direct impact on the size and efficiency of a batch. For standard datasets, i.e. of reasonable size and typical ultrasound image resolution, the learning time for NB in our implementation is about 5 minutes, as illustrated by Figure 7 on the Pelvic dataset, our most complex one.

DW reconstruction is also susceptible to large variations in reconstruction time depending on the dataset, and depends on the number of voxels used to represent the volume, as each voxel involves a specific computation, resulting in reconstruction times ranging from 30 seconds to several minutes on our datasets.

One of the advantages of our subsampling and random selection for batch construction is that learning times are consistent across all datasets, at 30 seconds to convergence.

C. RECONSTRUCTION PROPERTIES

1) Ability to represent details

To study the ability of the network to represent fine details in the scene, we consider a test scene constituted of cubes of various sizes. We then simulate ultrasound images

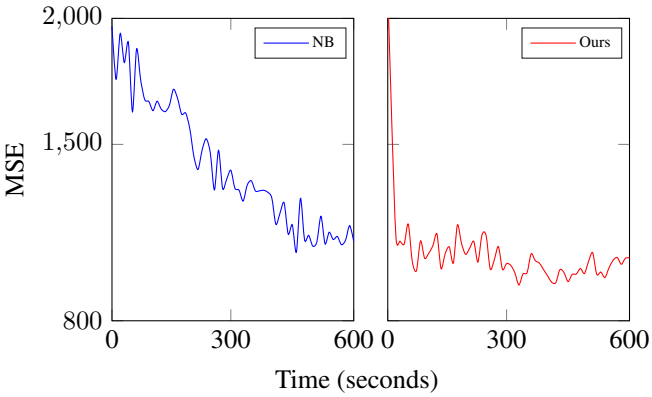

FIGURE 7. Average (10 runs) MSE over time with respect to ground-truth for the scene Pelvis

densely packed to create a dataset and use multiple network configurations to learn this scene. The networks represent all combinations of 2, 4 and 8 layers of 64, 128 and 256 perceptron. Figure 8 shows the resulting volumes sliced at a known position, on the smallest cubes in the scene.

The first observation is the degradation of details in the input compared to the ground truth, the smallest cube of 1.25 mm being barely noticeable in the input images.

We also notice that networks with only two layers fail to represent even large shapes. However, the same networks and other small networks seem to be better at identifying the

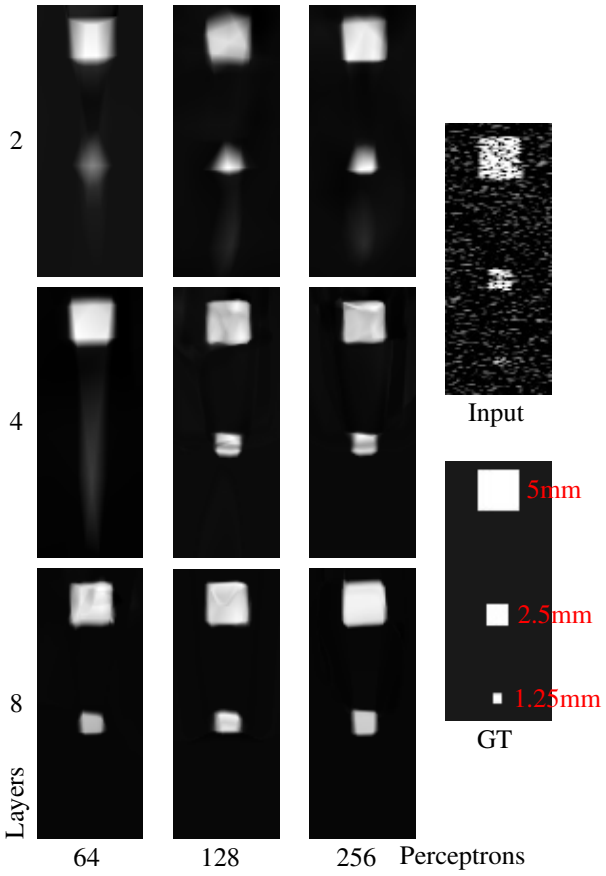

FIGURE 8. Comparison of the ability to represent between different sizes of network at convergence (20 000 iterations). Reproduction of a selected subset of the whole VOI with three cubes of sides 5, 2.5 and 1.25 mm.

largest cube. The most likely explanation for this is that the networks compensate for their inability to faithfully represent the shapes by loosely pointing out the areas of higher density.

Conversely, the larger networks are able to capture the two cubes almost perfectly and treat the 1,25mm cube as background noise. Once the network reaches a certain threshold (i.e., 4 layers and 128 perceptrons, or 8 layers and 64 perceptrons, for a total of 512 perceptrons) and avoids abnormal values such as only 2 layers, its reconstruction varies only slightly and is subject more to the random component of the learning phase than to the inability to represent the medium.

In our experiments, we observed that 4 layers of 128 perceptrons occasionally encounter difficulties in representing the most complex datasets, exhibiting some artifacts. Based on this observation, we empirically identified 175 perceptrons per layer as an optimal value for a network of minimal size that is suitable for use in any situation. Our results are based on this architecture.

2) Clear edge profile

Next, we highlight the continuous aspect of the network by examining the boundaries of a well-defined shape at a small discretization step. In this experiment, the network is queried every 0.1 mm when approaching a cube of 10 mm side in the Shapes scene. For the DW comparison, each value is the one contained in the voxel encompassing the query point.

This experiment is performed at 100 locations and by querying 100 networks trained from different random initialization. Results are reported in Figure 9.

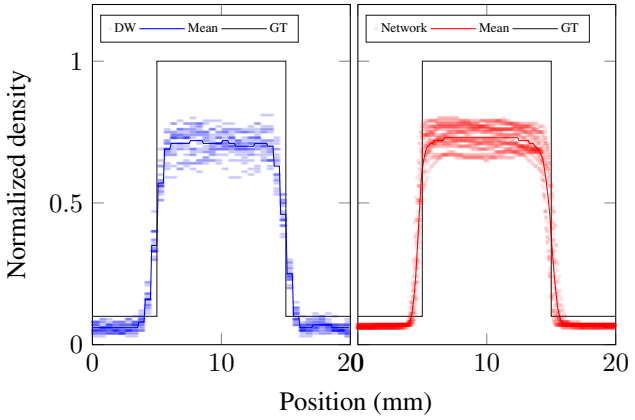

FIGURE 9. Evolution of the values when traversing the edge of a cube. Values and average over 100 runs.

This experiment shows the consistency of the network across all runs and the high variance in the DW reconstruction. Furthermore, the adequacy of the reconstruction is suggested by the similarity of the mean curves of the two methods and their intensity peaks, the averaging effect of the DW reconstruction being reproduced by the network trying to accommodate the peaks and valleys of the signal. One can also observe the blurring of the boundary in all methods, mostly introduced by the speckle noise.

D. ABLATION STUDY

We study hereafter the impact of each key feature of the proposed method on the metrics presented in the previous sections, and their impact on the learning time. We note that our method with disabled or modified steps becomes similar to other methods, as per Table 1.

As our method is geared toward fast reconstruction, the metrics are computed at an equivalent time of 20 seconds of training on the same machine using the same implementation.

1) Input slices

We remove the random selection mechanism and form each batch with a randomly selected image among the dataset. One point of interest is the fact that a typical ultrasound image may be constituted of much more pixels than our fixed batch size, making the epoch time and training dependent on the properties of the dataset. The result of this study is reported in Table 4 under the column "image input".

This ablation clearly shows the contribution of the random selection proposed in this work, especially in a short time scenario, where the network may not even have the time to be exposed to all samples when using a random image selection. This puts random selection as one of the main contribution of this article as an enhancement of efficient learning. Figure 10 illustrates how the image input impairs the network's ability to learn. As demonstrated in Figure 4, at each step the network is encouraged to learn only one slice of the space, which

TABLE 4. Quantitative results (mean and standard deviation over 10 runs with 20s learning time) for volume reconstructions with various ablation scenarios .

		Image input	Frequency embedding	Large network	SSIM + MSE	UltraNeRF [33]	Ours
Shapes	MSE ↓	697.6 (± 94.8)	626.9 (± 32.2)	**353.0** (± 67.6)	500.4 (± 114.7)	522.3 (± 102.2)	493.5 (± 133.5)
	MAE ↓	16.2 (± 4.16)	20.5 (± 0.98)	**11.8** (± 2.72)	16.5 (± 4.19)	13.5 (± 3.23)	17.0 (± 4.35)
	NCC ↑	0.868 (± 0.019)	0.927 (± 0.008)	0.936 (± 0.012)	0.932 (± 0.007)	0.905 (± 0.016)	**0.941** (± 0.005)
	SSIM ↑	0.950 (± 0.008)	0.969 (± 0.003)	0.972 (± 0.006)	0.968 (± 0.004)	0.958 (± 0.008)	**0.973** (± 0.003)
	SNR ↑	43.40 (± 3.07)	47.73 (± 1.57)	**51.07** (± 3.38)	46.80 (± 4.63)	49.79 (± 1.85)	47.39 (± 3.40)
	CNR ↑	23.36 (± 2.98)	26.16 (± 1.21)	34.85 (± 4.10)	30.65 (± 2.54)	20.71 (± 3.90)	**36.05** (± 2.58)
Pelvis	MSE ↓	1852.2 (± 376.2)	1215.4 (± 149.3)	1412.6 (± 554.4)	**992.8** (± 126.8)	1603.3 (± 232.00)	1188.1 (± 360.5)
	MAE ↓	32.1 (± 3.75)	25.3 (± 1.98)	27.4 (± 6.59)	23.2 (± 1.85)	30.3 (± 2.69)	**22.5** (± 4.09)
	NCC ↑	0.753 (± 0.023)	**0.898** (± 0.004)	0.894 (± 0.01)	0.890 (± 0.009)	0.752 (± 0.017)	0.896 (± 0.009)
	SSIM ↑	0.794 (± 0.027)	0.896 (± 0.006)	0.888 (± 0.011)	0.886 (± 0.011)	0.773 (± 0.035)	**0.898** (± 0.012)
	SNR ↑	26.24 (± 2.35)	28.78 (± 1.66)	33.41 (± 3.25)	**37.11** (± 3.23)	25.18 (± 1.44)	35.38 (± 2.26)
	CNR ↑	19.39 (± 1.81)	23.10 (± 0.86)	26.19 (± 2.01)	24.82 (± 2.86)	16.81 (± 1.76)	**26.24** (± 2.07)
Nerve	SNR ↑	32.25 (± 2.77)	29.26 (± 0.84)	32.39 (± 2.26)	31.03 (± 31.03)	30.65 (± 2.24)	**32.91** (± 2.12)
	CNR ↑	18.55 (± 3.42)	17.61 (± 0.97)	**19.75** (± 1.46)	17.03 (± 2.00)	19.15 (± 1.84)	19.12 (± 1.98)

results in significant errors for the remainder of the volume. This leads to high variance and challenging convergence.

2) Embedding of the input

Most of the NeRF-inspired methods use frequency embedding, as reported in Table 1. Here we proceed with our random input selection but encode the 3-dimensional inputs in a 63-dimensional frequency embedding as used in [32], [33]. The result of this study is reported in Table 4 under the column "Frequency embedding".

Frequency embedding has the property of enhancing the ability of the network to represent fine details and variations. As such, for many scenarios, this feature contributes greatly to produce a stable result and be highly evaluated by many metrics but comes with the cost of a larger input layer, implying a slightly slowed down learning.

More importantly for our application and as demonstrated in Section III-A2, frequency embedding introduces artifacts in the volumes and pushes the network to learn the noise present in the dataset, a behavior that most metrics struggle to represent.

The two spikes in the loss curve depicted in Figure 10 correspond to the distinct phases of learning with different levels of resolution, as outlined in the methodology section. As expected, frequency embedding accelerates convergence but raise two issues, as illustrated in Figure 10: the introduction of artifacts in the reconstruction and a tendency towards overfitting to the noise. Note that other learning schemes such as the one presented in Section IV-D5 still present artifacts and noise.

3) Network size

The network architecture used in our methods is minimal, as opposed to typical NeRF implementations, with the notable exception of ImplicitVol [32]. Here, we study the implications of having a larger network on the quality of the results at a fixed time. The architecture used for comparison contains 8 layers of 256 perceptrons. The result of this study is reported in Table 4 under the column "Large network".

In all experiments, a large network produces comparable or slightly better results than the proposed method. These results are correlated with the content of Figure 8, where after a certain threshold value, the network results are comparable.

However, a larger network in our application has several drawbacks. First, more than doubling the number of neurons means a significant increase in both memory size and iteration time (iterations are about 52% longer in our implementation). The former lessen the interest of the method for transmission to the AR device, while the latter may lead to less efficient convergence on a strict time budget, but is largely offset by the additional degrees of freedoms offered. Second, when the network has access to more degrees of freedom and the ability to represent fine details, its convergence may try to accommodate noise and imprecision, further away from our goal as illustrated in Figure 11 where a large network is trained for 30 minutes and distorts to fit the noise. Third, as illustrated in Figure 10, a larger network has a less predictable convergence and lead to higher variance.

4) Loss function

Among the available choices for the loss function, the literature is dominated by MSE and SSIM. This study compares our method with the weighted MSE and SSIM sum proposed in [33], as pure SSIM loss tends to push the network to converge toward an empty result due to the relatively restricted size of the objects in the VOI. The result of this study is reported in Table 4 under the column "SSIM + MSE".

As with the preceding sections, the results presented here are closely related to those of the proposed method. The principal disadvantage of this approach is the additional computational burden of calculating two losses, particularly the complex SSIM. In our experiments, the iterations using the SSIM + MSE loss were about 23% longer. This loss in number of iteration seem to offset the benefits of this more tailored

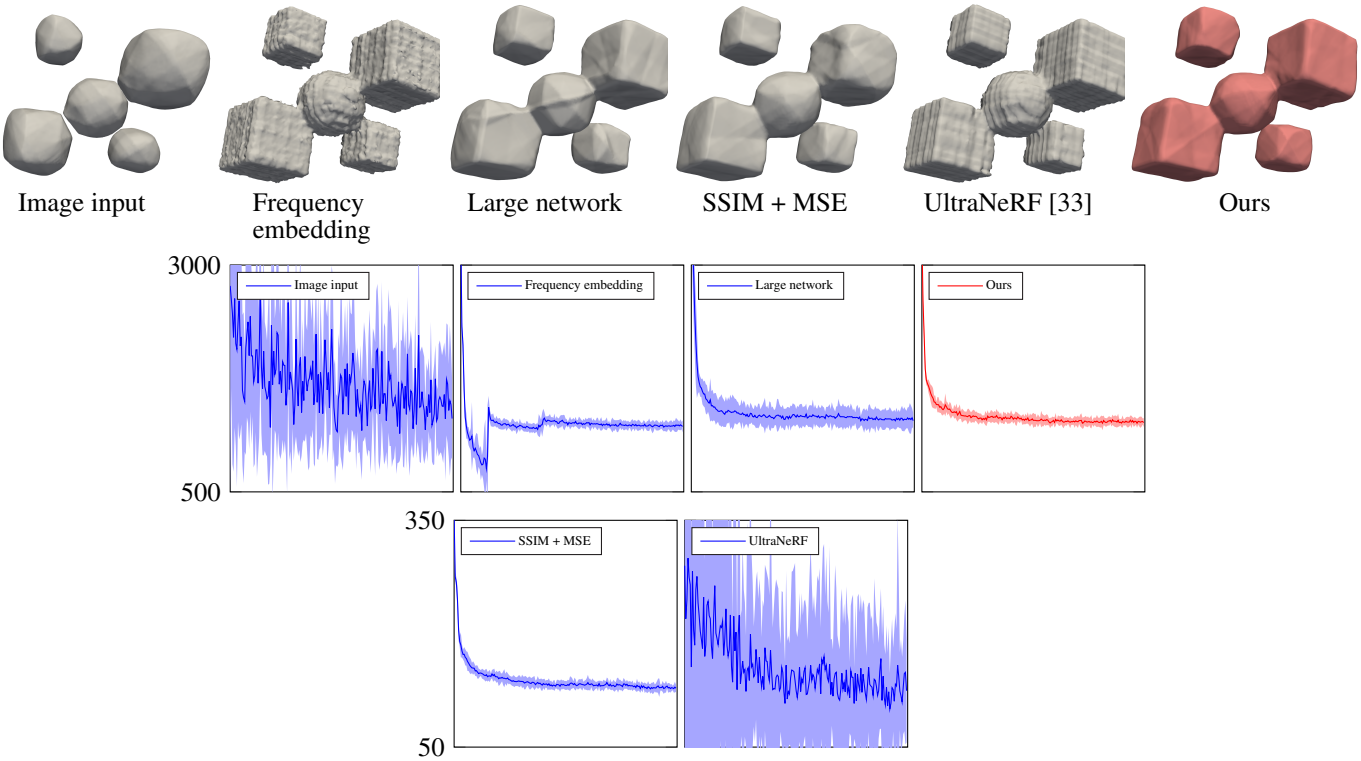

FIGURE 10. **Top: Example volumes after 20 seconds of learning. Middle: Mean loss evolution (10 runs) over the first 20 seconds for the Shapes for the MSE-based loss. Bottom: Mean loss evolution (10 runs) for SSIM-based loss.**

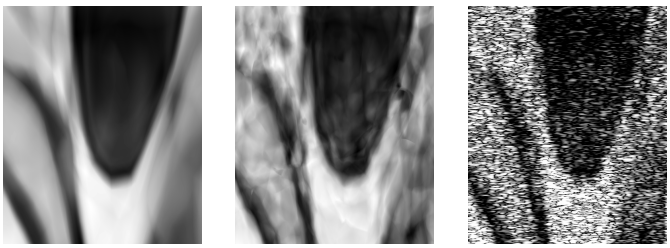

FIGURE 11. **Illustration of the overfit of a large network over a long time period. Left: Typical result of our method (20 seconds), middle: result after 30 minutes of learning with a large network, right: input image**

loss.

5) Combination of ablations

The modifications made to our base method are largely inspired by other methods described in the literature. Consequently, multiple ablations correspond to a state-of-the-art network and methodology, namely an 8 by 256 layers network, with image input, frequency embedding and the SSIM and MSE mixed loss. This is equivalent to the method employed in UltraNeRF [33] without their rendering process, which is beyond the scope of this study. The results in Table 4 for this study are reported under "UltraNeRF [33]".

This version combines a number of promising features, but it is hindered by the accumulated bloat of its components in terms of speed. In particular, as demonstrated in Section IV-D1, it lacks efficiency due to the image input format. The most notable outcome of this experiment is that it performs better than pure image input, suggesting that its other features are able to improve the results.

6) Conclusion on ablation study

This ablation study primarily demonstrates the efficacy of our batch construction method. The image input is not optimized for a reduced time scenario and is highly contingent upon the quality of the acquisition. The other features yield comparable results in terms of metrics, and the selection among them is influenced by the temporal and visual necessities of our application: a more seamless reconstruction without frequency embedding, less noise at convergence, and a more lightweight representation with a smaller network and accelerated learning through the use of the simple MSE loss.

E. APPLICATION TO AUGMENTED REALITY

This section presents the application of our methods in an AR scenario. Figure 12 illustrates the display of a reconstructed 3-D volume to assist practitioners, enabling easier and more precise operations.

Our custom software employs a Hololens 2 [49] to automate the entire process, enabling the visualization of the procedure in AR. The initial step involves acquiring spatially localized ultrasound images, which are then displayed in real-time at their respective locations. This allows the practitioner to observe the entire image collection as the reconstruction process unfolds. Subsequently, the volume is presented after the short ($\approx$ 30sec) reconstruction delay, to facilitate the guidance during any procedure on this medium.

This scenario can be applied to many procedures [16], [19], [20] such as tumor localization, cardiac interventions

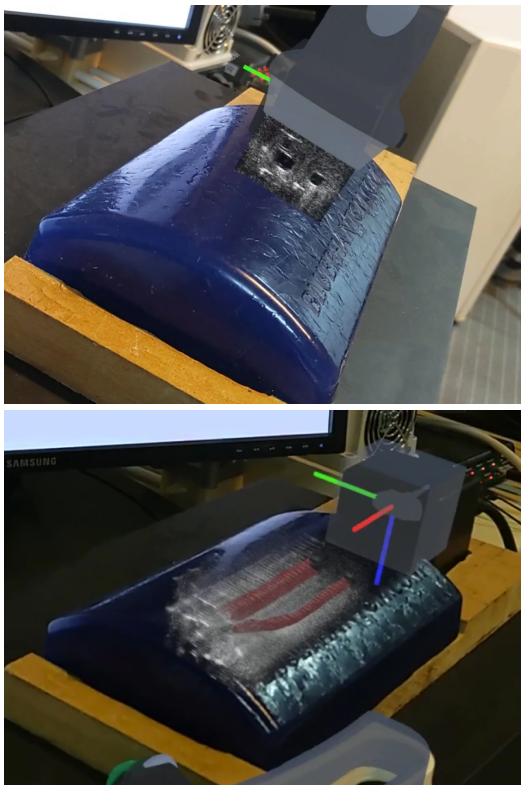

FIGURE 12. Display of the reconstructed volume in an AR framework. Top: Acquisition of the freehand ultrasound images, Bottom: reconstruction overlapped with acquisitions.

to visualize and navigate catheter placement, or obstetrics to monitor fetal development and assist in amniocentesis.

F. CODE AND DATASETS

The code for the volume generation and the public datasets are available at github.com/STORM-IRIT/Neural-Ultrasound-Field

V. CONCLUSION AND PERSPECTIVES

This work showed that the 3-D freehand ultrasound reconstruction under the constraint of an augmented reality pipeline problem can be efficiently addressed using NIR and unsupervised deep learning. The method presented in this paper produces smooth volumes in tens of seconds with accuracy comparable or better than state-of-the-art method.

This article puts forward the unique nature of the 3-D ultrasound volume reconstruction, associated with its associated set of difficulties and specificities that must be accounted for. Taking these challenges into account, this work proposes encouraging results on the use of NIR for efficient volume reconstruction and enhancement.

Promising perspectives remain to be explored, as they have been shown to be fruitful by this article:

- Design a dedicated embedding function in order to enhance shape generalization and accelerate learning, as the NIR community already demonstrated feasible on other applications.
- Automatically determine the dimension of the network for each scene, based on the specific properties of the dataset (e.g. total number of images, resolution, volume covered...).
- Formally quantify the representation capacity of given network dimensions and correlate it to the ultrasound systems resolutions, noise level and imprecision.
- Explore more tailored loss function, as the one used in the denoising literature, and their impact on both the results and the length of the reconstruction process.
- Study more in depth the impact and implication of the many network meta parameters, such as subset and batch sizes, iterations per subset, learning rate or number of epoch.

ACKNOWLEDGEMENTS

This project partially was funded by the SSLAM project (JCJC ANR-22-CE23-0004) from French National Research Agency. The authors thank Gauthier Bouyjou for his help on the AR platform.

REFERENCES

[1] Mostafa Amin Naji, Iman Taghavi, Erik Vilain Thomsen, Niels Bent Larsen, and Jørgen Arendt Jensen. Underestimation of flow velocity in 2-d super-resolution ultrasound imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, pages 1–1, 2024.

[2] Zhenxuan Zhang, Chengjin Yu, Heye Zhang, and Zhifan Gao. Embedding tasks into the latent space: Cross-space consistency for multi-dimensional analysis in echocardiography. *IEEE Transactions on Medical Imaging*, 43(6):2215–2228, 2024.

[3] Aaron Fenster and Dónal B Downey. Three-dimensional ultrasound imaging and its use in quantifying organ and pathology volumes. *Analytical and bioanalytical chemistry*, 377:982–989, 2003.

[4] Shyam Natarajan, Leonard S. Marks, Daniel J.A. Margolis, Jiaoti Huang, Maria Luz Macairan, Patricia Lieu, and Aaron Fenster. Clinical application of a 3d ultrasound-guided prostate biopsy system. *Urologic Oncology: Seminars and Original Investigations*, 29(3):334–342, 2011. New Technology in Urology: Balancing Risk and Reward.

[5] RL Dyson, DH Pretorius, NE Budorick, DD Johnson, MS Sklansky, CJ Cantrell, S Lai, and TR Nelson. Three-dimensional ultrasound in the evaluation of fetal anomalies. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 16(4):321–328, 2000.

[6] Alessandro Ramalli, Enrico Boni, Emmanuel Roux, Hervé Liebgott, and Piero Tortoli. Design, implementation, and medical applications of 2-d ultrasound sparse arrays. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(10):2739–2755, 2022.

[7] Ji Cao, Kerem Karadayi, Ravi Managuli, and Yongmin Kim. Reconstruction error in 3d ultrasound imaging with mechanical probes. In *Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy*, volume 7629, pages 21–31. SPIE, 2010.

[8] Jeffrey Bax, Derek Cool, Lori Gardi, Kerry Knight, David Smith, Jacques Montreuil, Shi Sherebrin, Cesare Romagnoli, and Aaron Fenster. Mechanically assisted 3d ultrasound guided prostate biopsy system. *Medical physics*, 35(12):5397–5410, 2008.

[9] Bastian S. Generowicz, Stephanie Dijkhuizen, Laurens W. J. Bosman, Chris I. De Zeeuw, Sebastiaan K. E. Koekkoek, and Pieter Kruizinga. Swept-3-d ultrasound imaging of the mouse brain using a continuously moving 1-d-array—part ii: Functional imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 70(12):1726–1738, 2023.

[10] Richard Housden, Graham Treece, Andrew Gee, and Richard Prager. Calibration of an orientation sensor for freehand 3d ultrasound and its use in a hybrid acquisition system. *Biomedical engineering online*, 7:5, 02 2008.

[11] Richard W. Prager, Andrew Gee, and Laurence Berman. Stradx: real-time acquisition and visualization of freehand three-dimensional ultrasound. *Medical Image Analysis*, 3(2):129–140, June 1999.

[12] Pierrick Coupé, Pierre Hellier, Noura Azzabou, and Christian Barillot. 3d freehand ultrasound reconstruction based on probe trajectory. In James S. Duncan and Guido Gerig, editors, *Medical Image Computing and*

Computer-Assisted Intervention – MICCAI 2005, pages 597–604, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[13] Qinghua Huang, Zhaozheng Zeng, et al. A review on real-time 3d ultrasound imaging technology. *BioMed research international*, 2017, 2017.

[14] Mohammad I. Daoud, Abdel-Latif Alshalalfah, Falah Awwad, and Mahasen Al-Najar. Freehand 3d ultrasound imaging system using electromagnetic tracking. In *2015 International Conference on Open Source Software Computing (OSSCOM)*, pages 1–5, 2015.

[15] Andy Wai Kan Yeung, Anela Tosevska, Elisabeth Klager, Fabian Eibensteiner, Daniel Laxar, Jivko Stoyanov, Marija Glisic, Sebastian Zeiner, Stefan Tino Kulnik, Rik Crutzen, Oliver Kimberger, Maria Kletecka-Pulker, Atanas G Atanasov, and Harald Willschke. Virtual and augmented reality applications in medicine: Analysis of the scientific literature. *Journal of Medical Internet Research*, 23(2):e25499, February 2021.

[16] E. Z. Barsom, M. Graafland, and M. P. Schijven. Systematic review on the effectiveness of augmented reality applications in medical training. *Surgical Endoscopy*, 30(10):4174–4183, February 2016.

[17] Eigil Samset, Dieter Schmalstieg, Jos Vander Sloten, Adinda Freudenthal, Jérôme Declerck, Sergio Casciaro, Øyvind Rideng, and Borut Gersak. Augmented reality in surgical procedures. In *Human Vision and Electronic Imaging XIII*, volume 6806, pages 194–205. SPIE, 2008.

[18] Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Shiro Kadowaki, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S Corrado, Jason D Hipp, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature medicine*, 25(9):1453–1457, 2019.

[19] Christoph Rüger, Markus A. Feufel, Simon Moosburner, Christopher Özbek, Johann Pratschke, and Igor M. Sauer. Ultrasound in augmented reality: a mixed-methods evaluation of head-mounted displays in image-guided interventions. *International Journal of Computer Assisted Radiology and Surgery*, 15(11):1895–1905, July 2020.

[20] Yoshinobu Sato, Masahiko Nakamoto, Yasuhiro Tamaki, Toshihiko Sasama, Isao Sakita, Yoshikazu Nakajima, Morito Monden, and Shinichi Tamura. Image guidance of breast cancer surgery using 3-d ultrasound images and augmented reality visualization. *IEEE Transactions on Medical Imaging*, 17(5):681–693, 1998.

[21] Farhan Mohamed and Chan Vei Siang. A survey on 3d ultrasound reconstruction techniques. In Marco Antonio Aceves-Fernandez, editor, *Artificial Intelligence*, chapter 4. IntechOpen, Rijeka, 2019.

[22] 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.

[23] Tiexiang Wen, Qingsong Zhu, Wenjian Qin, Ling Li, Fan Yang, Yaoqin Xie, and Jia Gu. An accurate and effective fmm-based approach for freehand 3d ultrasound reconstruction. *Biomedical Signal Processing and Control*, 8(6):645–656, 2013.

[24] Robert Rohling, Andrew Gee, and Laurence Berman. A comparison of freehand three-dimensional ultrasound reconstruction techniques. *Medical Image Analysis*, 3(4):339–359, 1999.

[25] C.D. Barry, C.P. Allott, N.W. John, P.M. Mellor, P.A. Arundel, D.S. Thomson, and J.C. Waterton. Three-dimensional freehand ultrasound: Image reconstruction and volume analysis. *Ultrasound in Medicine & Biology*, 23(8):1209–1224, 1997.

[26] Qinghua Huang, Yanping Huang, Wei Hu, and Xuelong Li. Bezier interpolation for 3-d freehand ultrasound. *IEEE Transactions on Human-Machine Systems*, 45:385–392, 06 2015.

[27] Hyungil Moon, Geonhwan Ju, Seyoun Park, and Hayong Shin. 3d freehand ultrasound reconstruction using a piecewise smooth markov random field. *Computer Vision and Image Understanding*, 151:101–113, 2016. Probabilistic Models for Biomedical Image Analysis.

[28] 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. *CoRR*, abs/2003.08934, 2020.

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022.

[30] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction, 2023.

[31] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023.

[32] 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.

[33] Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: Neural radiance fields for ultrasound imaging, 2023.

[34] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2018.

[35] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020.

[36] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2019.

[37] Shaowen Xie, Hao Zhu, Zhen Liu, Qi Zhang, You Zhou, Xun Cao, and Zhan Ma. Diner: Disorder-invariant implicit neural representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6143–6152, 2023.

[38] 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, 2021.

[39] Junshen Xu, Daniel Moyer, Borjan Gagoski, Juan Eugenio Iglesias, P. Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Nesvor: Implicit neural representation for slice-to-volume reconstruction in mri. *IEEE Transactions on Medical Imaging*, 42(6):1707–1719, 2023.

[40] 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.

[41] Hai Li, Hongjia Zhai, Xingrui Yang, Zhirong Wu, Yihao Zheng, Haofan Wang, Jianchao Wu, Hujun Bao, and Guofeng Zhang. Imtooth: Neural implicit tooth for dental augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2837–2846, 2023.

[42] You Zhang, Yuxiang Wang, and Zhiyao Duan. Hrtf field: Unifying measured hrtf magnitude representation with neural fields, 2023.

[43] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *CoRR*, abs/2008.04451, 2020.

[44] K. Perlin and E. M. Hoffert. Hypertexture. *SIGGRAPH Comput. Graph.*, 23(3):253–262, jul 1989.

[45] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, August 1987.

[46] Qing Wu, Yuwei Li, Lan Xu, Ruiming Feng, Hongjiang Wei, Qing Yang, Boliang Yu, Xiaozhao Liu, Jingyi Yu, and Yuyao Zhang. Irem: high-resolution magnetic resonance image reconstruction via implicit neural representation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 65–74. Springer, 2021.

[47] Elevate Healthcare regional anesthesia ultrasound training block. https://elevatehealth.net/shop/product/regional-anesthesia-ultrasound-training-block/. Accessed: 2024-07-15.

[48] François Gaits, Nicolas Mellado, Gauthier Bouyjou, Damien Garcia, and Adrian Basarab. Efficient stratified 3d scatterer sampling for freehand ultrasound simulation. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, page 1–1, 2023.

[49] Hololens 2 augmented reality headset hololens 2. https://www.microsoft.com/en-us/hololens. Accessed: 2024-07-17.

FRANÇOIS GAITS received the bachelor's degree in computer engineering and the Master of Science (M.Sc.) degree in computer graphics and image analysis from the from Paul Sabatier University, Toulouse, France. He is currently pursuing a PhD degree in IRIT laboratory (UMR CNRS 5505, Paul Sabatier University Toulouse 3) on 3-D ultrasound reconstruction and ultrasound simulation. His research interests include medical imaging and artificial intelligence, with a focus on efficient computing and real-time applications.

FABIEN VIDAL is a medical doctor specialized in the field of gynecologic surgery. He also received a M.S. in radiophysics and Medical imaging from Paul Sabatier University, Toulouse, France, in 2011. He was assistant professor at Paul Sabatier University until November 2020. Since then he has worked as a gynecologic at the tertiary health facility Croix du Sud Ramsey Santé, Quint Fonsegrives, France. He is also a PhD student in IRIT laboratory (UMR CNRS 5505, Paul Sabatier University Toulouse 3). His research interests comprise endometriosis diagnosis and management, image fusion and augmented reality.

ADRIAN BASARAB received the M.S. and PhD degrees in signal and image processing from the National Institute for Applied Sciences of Lyon, France, in 2005 and 2008. Since 2009 (respectively 2016) he was assistant (respectively associate) professor at the University Paul Sabatier Toulouse 3 and a member of IRIT laboratory (UMR CNRS 5505). Since 2021, he is full professor at the University Claude Bernard Lyon 1 and member of CREATIS lab. His research interests include medical imaging and more particularly inverse problems (deconvolution, super-resolution, compressive sampling, beamforming, image registration, reconstruction and fusion) applied to ultrasound image formation, quantitative acoustic microscopy, computed tomography and magnetic resonance imaging. Adrian Basarab has been elevated to the grade of IEEE Senior Member in 2019. He was an associate editor for Digital Signal Processing from 2015 to 2020, and a member of the French National Council of Universities Section 61 - Computer sciences, Automatic Control and Signal Processing from 2010 to 2015. In 2017, he was guest co-editor for the IEEE TUFFC special issue on "Sparsity driven methods in medical ultrasound". Adrian Basarab was the head of "Computational Imaging and Vision" group of IRIT laboratory from 2018 to 2020. Since 2019, he has been a member of the EURASIP Technical Area Committee Biomedical Image & Signal Analytics. He is member of the IEEE Ultrasonics Symposium TPC since 2020, and subject editor of the journal Signal Processing (Elsevier) from January 2024.

NICOLAS MELLADO is a Full-time researcher at CNRS, IRIT, University of Toulouse since 2016. His research interests include point cloud processing, multiscale analysis and registration. Dr. Mellado received a PhD degree in Computer Science from the University of Bordeaux, France, December 2012.