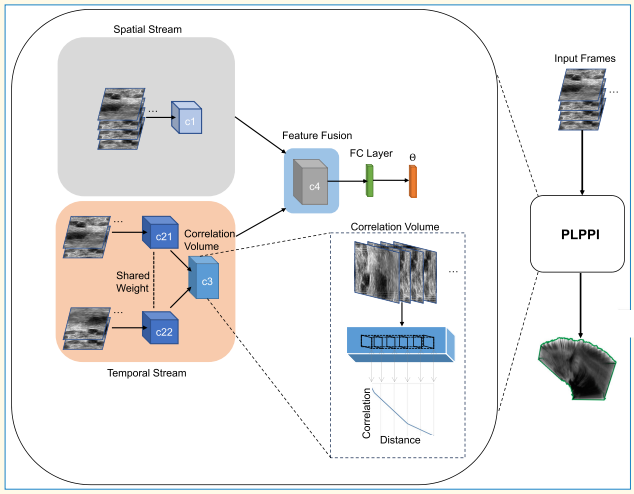
Sensorless End-to-End Freehand 3-D Ultrasound Reconstruction With Physics-Guided Deep Learning

Yimeng Dou, *Graduate Student Member, IEEE*, Fangzhou Mu, *Student Member, IEEE*, Yin Li, *Member, IEEE*, and Tomy Varghese, *Senior Member, IEEE*

***Abstract*—Three-dimensional ultrasound (3-D US) imaging with freehand scanning is utilized in cardiac, obstetric, abdominal, and vascular examinations. While 3-D US using either a "wobbler" or "matrix" transducer suffers from a small field of view and low acquisition rates, freehand scanning offers significant advantages due to its ease of use. However, current 3-D US volumetric reconstruction methods with freehand sweeps are limited by imaging plane shifts along the scanning path, i.e., out-of-plane (OOP) motion. Prior studies have incorporated motion sensors attached to the transducer, which is cumbersome and inconvenient in a clinical setting. Recent work has introduced deep neural networks (DNNs) with 3-D convolutions to estimate the position of imaging planes from a series of input frames. These approaches, however, fall short for estimating OOP motion. The goal of this article is to bridge the gap by designing a novel, physics-inspired DNN for freehand 3-D US reconstruction without motion sensors, aiming to improve the reconstruction quality and, at the same time, to reduce computational resources needed for training and inference. To this end, we present our physics-guided learning-based prediction of pose information (PLPPI) model for 3-D freehand US reconstruction without 3-D convolution. PLPPI yields significantly more accurate reconstructions and offers a major reduction in computation time. It attains a performance increase in the double digits in terms of mean percentage error, with up to 106% speedup and 131% reduction in graphic processing unit (GPU) memory usage, when compared to the latest deep learning methods.**



***Index Terms*—Deep learning, medical image processing, ultrasonic imaging.**

Received 16 July 2024; accepted 14 September 2024. Date of publication 20 September 2024; date of current version 27 November 2024. This work was supported by the National Institutes of Health under Grant 1R01HL147866. *(Corresponding author: Yimeng Dou.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Yimeng Dou and Tomy Varghese are with the Department of Medical Physics, University of Wisconsin (UW) School of Medicine and Public Health, Madison, WI 53705 USA, and also with the Department of Electrical and Computer Engineering, UW–Madison, Madison, WI 53706 USA (e-mail: ydou8@wisc.edu; tvarghese@wisc.edu).

Fangzhou Mu was with the Department of Computer Science, UW–Madison, Madison, WI 53706 USA. He is now with Nvidia Corporation, Santa Clara, CA 95051 USA (e-mail: fmu2@wisc.edu).

Yin Li is with the Department of Biostatistics and Medical Informatics, UW School of Medicine and Public Health, Madison, WI 53726 USA, and also with the Department of Computer Sciences, UW–Madison, Madison, WI 53706 USA (e-mail: yin.li@wisc.edu).

Digital Object Identifier 10.1109/TUFFC.2024.3465214

I. Introduction

ULTRASOUND (US) imaging is a principal modality in medical imaging, frequently employed as a diagnostic tool within clinical settings for the examination of soft tissue. US has several unique benefits, such as safety (nonionizing), portability, low cost, and real-time capability, making it suitable for numerous clinical imaging applications. The 3-D US can provide inexpensive and safe clinical information for improved local guidance, in comparison to other imaging modalities, such as computed tomography (CT) or magnetic resonance imaging (MRI) [1], [2], [3]. US has been used in many clinical settings, such as cardiovascular [2], obstetrics [4], surgical guidance [3], as well as research studies, including biomechanics [5], cancer [6], and brain imaging [7].

Highlights

- **We present a novel deep learning method that integrates the concept of speckle decorrelation as part of the model for freehand 3D ultrasound reconstruction.**
- **Our two-stream model separately extracts spatial and temporal data from a single freehand ultrasound sweep with 2D convolutions, followed by a correlation layer to synergize spatiotemporal cues.**
- **Compared to Chen et al. [44] and Prevost et al. [9], our model explicitly utilizes speckle decorrelation as inductive bias, leading to improved performance.**
- **Our model also contains fewer parameters, which requires less GPU memory to train.**

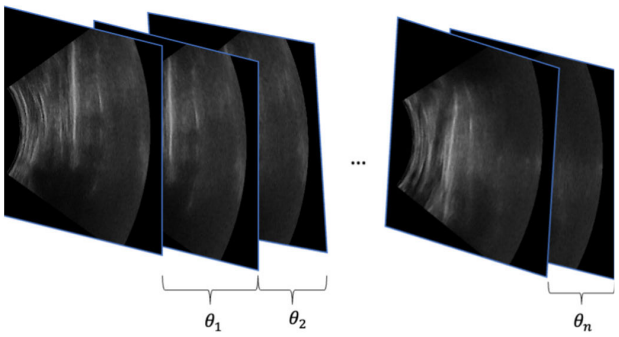


Fig. 1. Goal of our PLPPI model is to predict the relative transformation matrix or relative pose θ between two US frames.

Several methods have been developed for 3-D US reconstruction, including freehand imaging and the utilization of "wobbler" or "matrix" array transducers [8], [9], [10], [11], [12]. Unlike traditional 2-D US, 3-D volumes provide enhanced detail and accurately capture shape information from the underlying 3-D structure [13]. "Wobbler" transducers mechanically steer the US beam across various angles to gather 3-D data [13]. In contrast, "matrix" transducers steer the beam electronically [14]. However, both "wobbler" and "matrix" transducers suffer from a small field of view and low acquisition volume frame rates [14]. Freehand scanning, therefore, stands out for its simplicity and practicality in obtaining data for 3-D imaging and reconstructions [2], [9]. Motion sensors have been employed in freehand 3-D US to track transducer movement, generating a sequence of positions for US sweeps. As illustrated in Fig. 1(a), a tracking device tracks and produces the relative pose (position) transformation matrix θ between consecutive frames, thereby capturing long US sweeps. Nevertheless, external tracking devices can introduce artifacts due to optical blockages in tracking systems or magnetic disruptions in electromagnetic (EM) tracking systems, while other tracking devices, such as stepper motors, have a limited range [9]. Artifacts from external tracking devices can negatively impact the overall reconstruction quality [9]. Therefore, scanning without an external sensor has several advantages, offering benefits such as not needing calibration for each use, compatibility with existing equipment, and eliminating the need for additional hardware.

Early work in sensorless 3-D US reconstruction focused on tracking adjacent US B-mode frames by matching features extracted from local regions between frames. While these methods, including optical flow [15], [16], [17] and speckle decorrelation [18], [19], [20], were successful for in-plane motion, they struggled to estimate out-of-plane (OOP) motion. Optical flow methods rely on changes between consecutive frames to track apparent motion in the 2-D space and thus cannot handle OOP [9]. Furthermore, they often neglect longer range relationships beyond neighboring frames. On the other hand, speckle decorrelation methods employ ensemble correlation among multiple frames to determine the temporal (elevational) displacement for OOP motion. However, their reconstruction quality for OOP motion is often unsatisfactory [21]. Moreover, these methods require constructing correlation curves as data priors [22]. These curves, which define the relationship between image correlation and elevational distance, are obtained through scans of a speckle phantom, rendering them less practical for clinical settings. Toews and Wells [23] proposed to extract features with a chunk of overlapping frames, but their method still falls short in estimating OOP motion.

Machine learning-based methods, particularly deep neural networks (DNNs), have been employed to reconstruct 3-D volumes from freehand US B-mode images. Researchers have explored models involving both 2-D and 3-D convolutional neural networks (CNNs) as well as recurrent neural networks (RNNs) to estimate the relative motion—both translational and rotational—between successive US frames or to acquire the relative positioning within a sequence of US frames [9], [12], [24]. Conceptually, 3-D CNNs and RNNs can operate on the temporal axis, making them well suited for capturing temporal dynamics in a series of US frames [24]. However, these methods are inadequate to address the complexity of US imaging. Freehand scans exhibit significant variability due to physiological motion over time, which is further complicated by different scanning paths. Even within the same scanning sequence, as shown in Fig. 2 using deep contextual learning network (DCL-Net) [24], small changes in OOP motion could lead to significant changes in spatial appearance and contribute to the variability in temporal information. Prior deep learning methods often overlooked the underlying physical processes, placing an additional burden on the network. While models, such as DCL-Net, could learn the physical properties [24], they tend to overly smooth the scanning path, as shown in Fig. 3.

To address this challenge, we introduce a lightweight, physics-influenced deep learning model for reconstructing

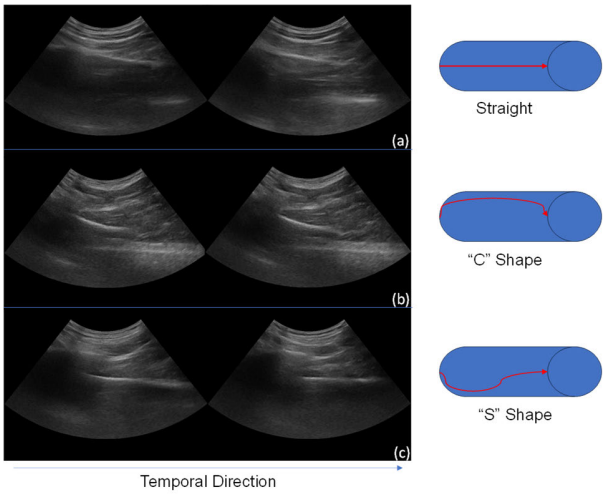


Fig. 2. B-mode images depict variations in appearance of the same anatomy when acquired through different freehand scanning paths. The same anatomy is imaged under (a) scanning path along a straight line (linear), (b) scanning path following a "C" shape, and (c) scanning path with an "S" shape. For each path, we show an initial image and an image acquired 0.5 s after the initial image. The corresponding trajectory of the transducer is also shown. Note that the B-mode images for these three scanning paths are visually distinct. Our model is designed to leverage these variations for estimating the transducer pose.

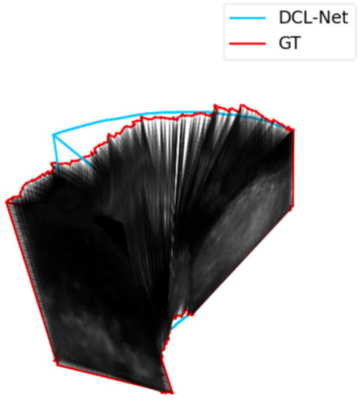


Fig. 3. Reconstruction trajectory in blue with DCL-Net is overly smoothed with 3-D convolution when compared to GT.

3-D US images in the freehand mode without the need for motion tracking devices. Our model enables the use of multiple frames as input with limited impact on graphic processing unit (GPU) memory usage (MGU) and speed. This multiframe implementation can reduce the accumulated prediction errors and is effective in mitigating prediction errors for longer sequences. Inspired by both the motion-based and deep learning methods, our model separates the learning for spatial and temporal information through a two-stream network. It employs learnable operators to capture data priors for depicting temporal connections and further integrates a physical model to ease the complexity of learning tasks. The result is a highly flexible model that is adaptive to different scanning paths.

A key feature of our model is the correlation operator, defined between pairs of feature maps to model temporal information, akin to the concept of speckle decorrelation. Instead of using CNN or RNN architectures, our method employs 2-D convolutions to extract frame-level features and leverages the correlation operator for temporal modeling. This approach results in a lightweight model with fewer parameters and lower memory usage. To capture longer range temporal information, we also propose sampling multiple frames within the input sequence during training.

We have conducted extensive experiments to evaluate our approach using experimentally acquired datasets on tissue-mimicking phantom and human subjects rather than simulated datasets. We compared our proposed model to several baseline networks that utilize 3-D CNNs and RNNs, including DCL-Net [24], convolutional long short-term memory (ConvLSTM) [12], and deep contextual-contrastive network (DC^2)-Net [11]. Using only 2-D convolutions, our physics-guided learning-based prediction of pose information (PLPPI) model achieves robust empirical results in reconstruction quality, outperforms all baseline methods, and excels at estimating OOP. Furthermore, PLPPI demonstrates up to 106% speedup and 131% reduction in MGU compared to the baseline methods.

A preliminary version of this work was presented at the 2023 IEEE International Ultrasonics Symposium (IUS) [25]. This expanded journal version enhances our previous work fourfold. First, it incorporates multiple frames for training, in contrast to [25], which used two adjacent frames. Second, this version analyzes two real-world datasets, whereas work [25] relied on a synthetic 3-D transrectal US dataset [26]. Third, it includes a new comparison with ConvLSTM [12] and DCL-Net [24]. Finally, it features ablation studies to evaluate the effectiveness of the newly added components.

II. Related Work

A. Freehand US Reconstruction With Physical Properties

Speckle, also called "incoherent interference artifact," arises from the interference phenomenon due to incoherent scattering from the propagation of the US pulse, i.e., when the average distance between microscopic scatterers is smaller than the resolution cell, coupled with a limited field of view at focus [18], [27]. In US imaging, speckles from inhomogeneous volumes of tissues remain correlated as long as they are within the US beam [27], [28]. This indicates that correlation values between patches from different frames are influenced by their elevational separation. Laporte and Arbel [22] suggested the use of speckle decorrelation to determine OOP motion. This is based on the principle that speckle correlation diminishes with distance, providing insights into the relative positioning. Tuthill et al. [27] demonstrated the feasibility of estimating the separation between consecutive US frames with notable precision, both in phantom and in vivo breast imaging. A theoretical decorrelation curve with frame spacing corresponding to the correlation coefficient is generated using a speckle phantom. Chang et al. [18] proposed a method for detecting speckle regions first by block matching since intensity autocorrelation is accurate only in regions with fully developed speckles (FDSs). However, in practice, Gee et al. [28] observed in real tissue that the correlation value can vary. They further proposed an adaptive speckle decorrelation scheme that can be utilized even in regions without FDS. However, the direct relationship between intensity autocorrelation and displacement is complex and tissue-dependent [22]. Laporte and Arbel [22] then proposed a statistical learning-based method that learns

the correlation using synthetic datasets before applying these to real images. Still, relying only speckle is challenging and may not provide accurate freehand US reconstruction results [21]. In contrast, our proposed method utilizes speckle patterns, enabling the learning model to infer the temporal relationship more effectively.

B. Video Recognition via Two-Stream Neural Network

A 3-D US scanning sequence is conceptually like a video loop, with each slice a frame in the video. This data structure can be inherently decomposed into spatial and temporal components. To aggregate spatial and temporal information, two-stream models are often considered for video recognition [29]. These models, such as those in [30] and [31], include one spatial and one temporal branch, with both branches fused before the fully connected layer. These methods focus on feature extraction for temporal aggregation or the combination of nonlocal neural networks [32]. Other methods rely on 3-D CNN and its variants, such as convolutional 3-D (C3D) [33], or inflating 2-D convolution as 3-D [34]. SlowFast [35] utilizes two pathways to obtain both local and long-term dependencies.

To reduce the overhead associated with 3-D convolution, methods, such as temporal shift module (TSM) [36], (2 + 1)D [37], expand 3-D (X3D) [38], channel-separated convolutional networks (CSNs) [39], and hybrid 2-D and 3-D CNN HybridSN [40], have been proposed. Inspired by those models, we use 2-D convolutions as a core component of our model to enhance efficiency while maintaining performance comparable to 3-D counterparts. Our model also adopts a two-stream approach, utilizing speckle decorrelation for temporal feature extraction.

C. Machine Learning for Freehand US Reconstruction

Several studies have utilized neural networks for 3-D freehand US reconstruction [9], [10], [11], [12], [24], [41], [42]. Prevost et al. [9] leveraged image-based algorithms to estimate relative interplane frame positions directly using CNNs [9]. Guo et al. [24] proposed a deep contextual learning network for reconstruction with 3-D CNNs. However, sensorless methods often show limited reconstruction accuracy due to accumulative drift and estimation errors, especially for nonmonotonic OOP motion [10]. Luo et al. [10] proposed deep motion network (MoNet) to improve image-based reconstruction performance by using an additional IMU sensor to incorporate velocity between frames. Luo et al. [41] proposed online learning framework (RecON), a self-learning method, using LSTM to predict the US image location. This method integrates a differentiable volumetric reconstruction technique to improve reconstruction results, and a discriminator to differentiate the reconstructed volume from randomly selected real volumes [12], [41]. Chen et al. [43] used a 3-D CNN to predict the initial position and then LSTM for refinement to estimate the pose location.

However, all prior methods rely on the assumption of temporal consistency and experience degraded performance with irregular scanning paths between frames. In contrast, our method explicitly learns physical patterns from observed speckle between nearby frames to aid in motion tracking.

III. Methodology

We propose the PLPPI model, a physics-inspired DNN to address the challenging task of freehand 3-D US reconstruction. Given n consecutive frames of a freehand 3-D US scan, PLPPI takes every frame of a US sweep as input and predicts their relative poses $\Theta = [\theta_1, \ldots, \theta_n]$, where θ_i comprises both rotation and translation.

A key innovation of PLPPI is the integration of a correlation operator within the deep model to compute a correlation activation volume in the learned feature space, representing interframe patch similarity. Here, a patch represents a subset of the feature space extracted by 2-D convolutions. This design is inspired by: 1) the physical process of speckle formation, where imaging planes closer in elevational distance are more correlated, and 2) the classic reconstruction method with speckle decorrelation, which seeks to uncover relative position changes in the imaging plane. The correlation of features within the patches depends on the percentage of overlap (i.e., the similarity of features), revealing elevational changes, as shown in Fig. 4(b). Importantly, establishing patch correspondence in the rich feature space allows our model to overcome measurement noise, generalize well on challenging real-world scans, and rectify drift and inconsistency within sequential estimates.

A. Overview of Network Design Model

Fig. 4 provides an overview of our model, comprising a spatial stream and a temporal stream, followed by a fusion module and prediction head. Given a sequence of input frames, the spatial stream aggregates intraframe spatial context using learned parameters from 2-D convolutions, whereas the temporal stream extracts interframe motion cues by constructing the correlation volume. Outputs from both streams are subsequently fused to provide a holistic view of the spatiotemporal scan volume, enabling accurate prediction of the relative transducer poses between adjacent frames.

The two-stream architecture of our model separates out the spatial and temporal reasoning in 3-D US reconstruction, thereby simplifying the learning task compared to learning the combined spatiotemporal features. In addition, this modular design allows us to incorporate speckle decorrelation as a physics prior into the network, unifying the strengths of physics and learning-based approaches. By decomposing into spatial and temporal branches, the network introduces more nonlinearities due to the additional rectified linear units (ReLUs) in the temporal branch. The additional nonlinearities also increase the functional complexity [37]. We will now provide a detailed explanation of each component of our model next.

B. Temporal Stream

The 3-D convolutions only extract features present at the same spatial location in successive input frames, which is limited when substantial changes in motion and appearance between frames exist, such as variations in scanning paths. To address this, we propose using only 2-D convolutions and building a correlation volume on top of 2-D outputs. This

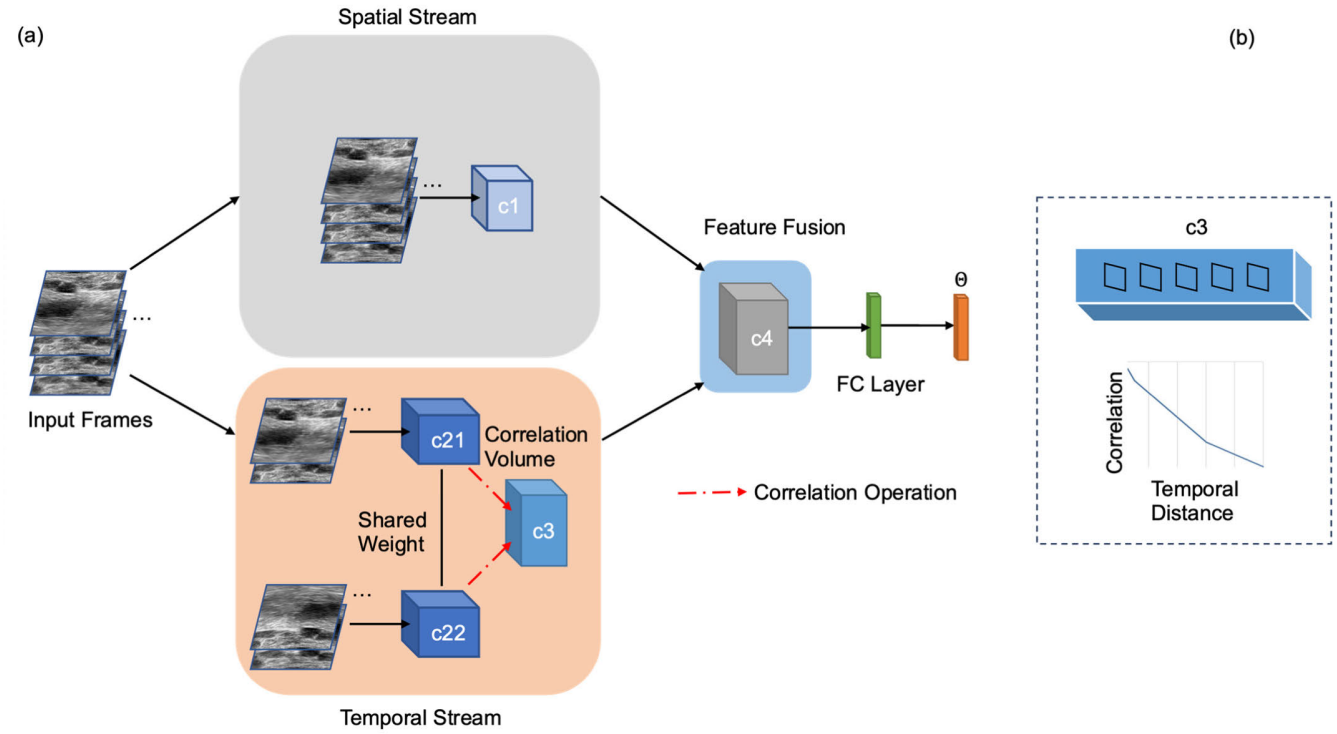


Fig. 4. (a) Schematic of the spatial and temporal stream in the correlation operation of our proposed network. (b) Importantly, the correlation operation in the temporal stream (in orange) extracts temporal information by performing speckle decorrelation. Consider the corresponding patches labeled as the black bounding boxes in the feature space extracted from a series of US frames. For contents in these patches, the correlation is related to the percentage of overlap, assuming FDS. Large elevational motion leads to a small degree of overlap, resulting in a decrease in correlation corresponding to the temporal travel distance, as indicated in (b). The blocks (labeled c_1, c_{21}, c_{22}, c_3, and c_4) are feature maps extracted using 2-D convolutions. In the temporal stream, the input image stack is divided into two subvolumes, which then pass through a 2-D convolution layer with shared weight to obtain c_{21} and c_{22}.

design is more flexible for extracting temporal features during rapid motion.

The correlation operation models temporal information by computing patch similarity between two dense feature maps. The correlation coefficient from the correlation operation, i.e., the correlation activation volume, is calculated by taking the dot product between all pairs of patches from each feature map [28], [44]. For two feature maps c_{21} and c_{22}, the correlation volume cv is represented as

$$\mathrm{cv}(x_1, x_2) = \sum_{s \in [-\mathrm{pp}] \times [-\mathrm{pp}]} c_{21}(x_1 + s)^T \cdot c_{22}(x_2 + s) \quad (1)$$

where x_1 and x_2 are the patch locations centered at c_{21} and c_{22}, respectively. The parameter s is the index of summation, where p is the maximum displacement between x_1 and x_2. The size of each feature map $c \in R^{H \times W \times D}$ and the squared patch size is $K = 2p + 1$. Features within the same receptive field become less correlated over distance if an elevational change is presented. Since speckle is correlated locally [28], we limit the computation of cv locally within the range of K, which also reduces computational complexity. Parameters p and K are hyperparameters, and the impact of p and K is studied in the ablation study section. A detailed analysis can be found in [44]. The size of cv would be $H/2 \times W/2 \times K^2$ for our model.

The correlation volume is then passed through two 2-D convolution operations to generate c_3 in Fig. 4. Our model can handle more than two images by evenly splitting the input image stacks and concatenating them into separate subvolumes. After computing their correlation, the two subvolumes would pass through a 2-D convolution layer to generate two dense feature maps c_{21} and c_{22}.

C. Spatial Stream

The spatial branch, on the other hand, extracts spatial features from the frames stacked together using a 2-D convolution layer. The spatial component, represented by the appearance of individual frames, conveys information about the scenes and objects depicted in the scans [29]. Our model captures spatial appearance at the early stage through 2-D convolution and later combines output features with the correlation volume to fuse spatial and temporal information. The resulting spatiotemporal feature volume is then further processed. For the spatial stream, the input frames are concatenated into a single volume and passed through a 2-D convolution layer to obtain spatial features c_1 in Fig. 4.

D. Feature Fusion and Prediction

The temporal information from correlation activation is fused with local spatial feature activation volume from the spatial branch to generate spatiotemporal volume. To combine temporal and spatial features, the upsampling operation for temporal features is needed to match the shape of spatial

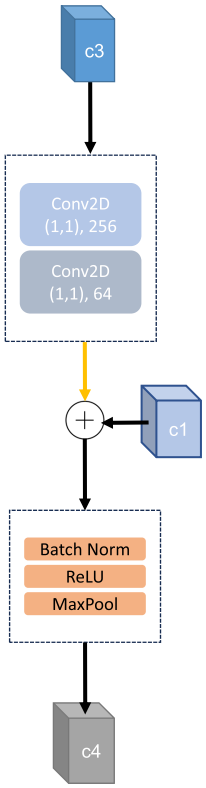


Fig. 5. Schematic illustrating the generation of spatial–temporal features c_4 by adding correlation volume c_3 to spatial feature maps c_1. Yellow arrow indicates the bilinear upsampling operation.

features. Instead of using deconvolution, our model uses bilinear upsampling to avoid checkerboard artifacts [45]. After summing the correlation volume to the spatial feature maps, the result is then passed to a residual network next (ResNeXt) model [46]. After that, the weights are passed through a fully connected layer to predict the relative translation and rotation between each input frame parametrized as a quaternion. This is illustrated in Fig. 5.

E. Training

The loss function for our PLPPI model consists of three components; the first two terms follow the same loss setup from DCL-Net and ConvLSTM [12], while the third term incorporates triplet ranking loss (TRL) [11]. The first loss component is the mean squared error (MSE) between the estimated $\hat{\theta}$ and the ground truth θ transformation parameters. However, using MSE alone would lead to oversmoothing [11], [12]. The Pearson correlation loss [11] is then implemented as the second component as

$$\left|\left|\hat{\theta}-\theta\right|\right|^2+\left(1-\frac{\mathrm{Cov}\left(\hat{\theta},\theta\right)}{\sigma\left(\hat{\theta}\right)\sigma\left(\theta\right)}\right) \tag{2}$$

where Cov represents covariance operation and σ represents the standard deviation.

Following [11], we also implemented TRL, allowing the model to learn discriminative feature embedding by penalizing estimated samples $\hat{\theta}$ in terms of distance in feature space. For a feature map v_m, we enforce a constraint that the embedding of v_m should be closer to a positive embedding v_n than a negative embedding v_q as

$$\max\left(0, d(v_m-v_n)+M-d\left(v_m-v_q\right)\right) \tag{3}$$

where M is the margin and d is the function for calculating distance. The positive embedding would be closer to the sample than the negative embedding. This loss is given by

$$L_{\max}=\max\left(0,-y\cdot\left(\left||v_m-v_n\right||^2-\left||v_m-v_q\right||^2\right)+M\right). \tag{4}$$

For TRL, y is the adaptive binary label as

$$y=\mathrm{sign}\left(\left||\theta_m-\theta_n\right||^2-\left||\theta_m-\theta_q\right||^2\right) \tag{5}$$

where θ represents the corresponding ground truth transformation parameters to the feature vectors. The final loss function becomes

$$L=\lambda_1\left|\left|\hat{\theta}-\theta\right|\right|^2+\left(1-\frac{\mathrm{Cov}\left(\hat{\theta},\theta\right)}{\sigma\left(\hat{\theta}\right)\sigma\left(\theta\right)}\right)+\max\left(0,-y\cdot\left(\left||v_m-v_n\right||^2-\left||v_m-v_q\right||^2\right)+M\right). \tag{6}$$

The hyperparameter λ_1 is set to 5 following [11]. Similarly, M is the margin parameter and is set to 0.25 as in [11].

IV. Experiments

A. Phantom Data

We collected datasets from a tissue-mimicking abdominal phantom (Model 057 A, CIRS Inc., Norfolk, VA, USA) using a linear array transducer (10L4) on a Sequoia system (Siemens, Mountain View, CA, USA). Transducer movement is tracked with a 6-D EM tracking system (NDI 3D Guidance, Northern Digital Inc., Waterloo, ON, Canada), with the tracking sensor attached to the US transducer using a tracking bracelet (CIVCO Medical Solutions, Kalona, IA, USA) to ensure that the separation between transducer and image plane remains constant. The EM tracking system is calibrated by the manufacturer. During acquisition, the transducer is perpendicular to the scanning plane, and the scanning path is mostly linear along the phantom contour, from distal to proximal and vice versa. The dataset contains a total of 75 scans. The final image is cropped to 224 × 224 pixels. The distance range of the dataset ranges from 30 to 94 mm.

B. Human Subject Data

In addition, we utilize an open-source dataset from [47] to validate our model. This dataset comprises freehand US data on volunteer forearms, containing 228 3-D US volumes acquired from 19 volunteers using a curvilinear transducer (4DC7-3/40) and an Ultrasonix system (BK, Herlev, Denmark) at 20 frames/s. Tracking position data on the transducer were collected using an NDI Polaris Vicra (Northern Digital Inc., Canada). The sequence lengths of the dataset range from 100 to 200 mm, for training, validation, and testing, and the image size is reshaped from 480 × 640 pixels to 224 × 224 pixels. The path of the freehand US scanning sequences includes a straight line, "C" and "S" shape along a distal-to-proximal direction. A detailed description of this dataset is presented in [48].

Stage	Kernel Dimension	Output Size
Raw Clip	-	4*224*224 (spatial) 2*224*224 (temporal)
c1	Stride 2, Padding 3, 7x7, 64	64*112*112
c21, c22	Stride 2, Padding 3, 7x7, 32	32*112*112
c3	[Stride 1, 1x1, 221] [Stride 1, 1x1, 64] Bilinear Up sampling	64 *112 *112
c4	ResNext50	2048*7*7
FC	-	3*6

Fig. 6. Network architecture used for the experiments.

C. PLPPI Implementation

For spatiotemporal feature fusion, we use summation rather than concatenation, as it provides better performance. We use ResNeXt [49] and explore cardinality for feature extraction. Data are separated into training/test/validation using a 60%/20%/20% split for both phantom and human forearm data. Adam [50] optimizer is used to optimize our PLPPI model with a weight decay of 0.01 and 0.00026 as the learning rate. During training, we set the number of epochs to 300 and the batch size to 36. Each input frame is set to be four and is selected randomly within the range of ± 10 frames. The network is implemented using the PyTorch library [51]. All networks, including the baseline networks, were trained on a NVIDIA A40 GPU (Nvidia, Santa Clara, CA, USA). The detailed architecture of our model is shown in Fig. 6.

D. Evaluation Metrics

For both experiments, we evaluate and compare all methods using metrics reported in other freehand US reconstruction publications, including distance error (DE), final drift (FD) [9], frame error (FE) [11], and MGU. These metrics are defined as follows.

1) *DE:* It measures the average reconstructed DE between predicted and ground truth frames for each scanning sequence by using the corresponding corner points of each image frame.
2) *FE:* It provides the mean individual FE using the relative pose information between the prediction and ground truth [11].
3) *FD:* It measures the distance of the final frame from the prediction and ground truth in terms of Euclidean distance, and is also used to measure the drift for larger motion errors [9].
4) *MGU:* It assess GPU memory efficiency for each model.
5) *Average Inference Time (AIT):* It assess the inference time of each model.

E. Baseline Methods

Our baseline networks include methods using the ResNeXt backbone with 2-D convolution (2-D ResNeXt) [49], DCL-Net and DCL-Net with two frames as input ($N = 2$) [24], ConvLSTM [12], and DC^2-Net [11]. We use ResNeXt-50 [49] for the residual network blocks. Our model is denoted as PLPPI, while PLPPI* denotes our model that uses consecutive frames. PLPPI* and PLPPI utilize the same method of using two subvolumes for the temporal branch as described in Section III-B. The difference is that PLPPI* uses consecutive frames, while PLPPI uses random frame selection within a range, so a given frame selected could either be past or future frames in the temporal dimension. We train 2-D ResNeXt with the same random frame selection as our model with four input frames, where 2-D ResNeXt is the backbone of our network. We also train standalone 2-D ResNeXt with two frames as input for comparison to 3-D ResNeXt and our model.

V. Results

A. Comparison to Baseline Methods

Table I demonstrates that PLPPI outperforms baseline models for both datasets, including the state-of-the-art model DC^2-Net. The reconstruction quality with our model PLPPI, assessed by quantitative measurements DE, FD, and FE surpasses that of most baseline models for both datasets, except FE with the human forearm dataset. In addition, our model uses less GPU memory than baseline methods. Notably, decoupling spatial and temporal information at an earlier stage helps with propagating information with a deeper model. PLPPI also improves training and inference performance in terms of memory and time usage since the 2-D operation is less time-consuming than 3-D and the 2-D convolution layer has fewer learnable parameters [36].

We further visualized the reconstruction results of the baseline networks compared to PLPPI. Fig. 7 displays the best and worst cases, filtered using FD values, and reconstructed and visualized from the human forearm dataset. These scans present challenges for all methods, particularly in the worst case, where the trajectory reverses direction. Our model follows the scan trajectory more accurately than other methods, even though ConvLSTM and DC^2-Net were developed for learning complicated sequences.

Utilizing multiple frames could also help mitigate reconstruction errors that are accumulated early in the prediction. As shown in Fig. 8, with $N = 4$, PLPPI shows a reduction in reconstruction error.

B. Capturing OOP Motions

Table II presents the average error between the ground truth and model prediction for each degree of freedom for OOP motion. PLPPI achieves a lower average error in most categories when compared to other baseline methods. Furthermore,

TABLE I

QUANTITATIVE RESULTS COMPARING MODEL PERFORMANCE USING TESTING SETS INCLUDE (a) 15 PHANTOM AND (b) 46 HUMAN FOREARM SCANS. THE RESULTS ARE REPORTED WITH THE MEAN ± STD. **BOLD** REPRESENTS THE BEST RESULTS, AND UNDERLINE DENOTES THE SECOND BEST. DE: DISTANCE ERROR, FD: FINAL DRIFT, FE: FRAME ERROR, MGU: GPU MEMORY USAGE, AIT: AVERAGE INFERENCE TIME IN GPU TIME AFTER TEN ITERATIONS USING A SWEEP WITH 104 IMAGE FRAMES FOR THE FOREARM DATASET AND FRAMES FOR THE PHANTOM DATASET, AND APD: AVERAGE PERCENTAGE DIFFERENCE COMPARED TO PLPPI

(a)

Model	DE (mm)	APD (%)	FD (mm)	APD (%)	FE (mm)	APD (%)	MGU (MiB)	APD (%)	AIT (s)	APD (%)
ConvLSTM	4.27±2.17	32.07	6.07±3.78	28.84	0.15± 0.05	0	29206	131.36	3.19	106.00
DCL-Net (N=2)	4.56±2.31	38.43	6.60±4.05	36.98	0.17±0.05	12.50	10330	52.26	1.51	42.57
DCL-Net	4.41±2.24	35.20	6.47±3.93	35.06	0.16± 0.05	6.45	10330	52.26	1.31	28.82
DC^2-Net	4.23±2.13	31.11	6.10±3.75	29.32	0.16± 0.05	6.45	10330	52.26	1.35	31.76
2D ResNeXt	4.30±2.21	32.75	6.15±3.86	30.28	0.16± 0.05	6.45	**5894**	-2.61	**0.78**	-22.72
PLPPI*	3.13±1.36	1.29	**4.35±2.15**	-4.27	**0.13±0.03**	-6.90	6050	0	0.98	0
PLPPI	**3.09±1.40**	0	4.65±2.25	0	0.15±0.04	0	6050	0	0.98	0

(b)

Model	DE (mm)	APD (%)	FD (mm)	APD (%)	FE (mm)	APD	MGU (MiB)	APD (%)
ConvLSTM	15.35±7.37	18.31	22.78±11.03	23.42	**0.46±0.23**	-4.26	30861	126.63
DCL-Net (N=2)	20.86±7.56	48.11	23.30±11.18	25.67	**0.46±0.23**	-4.26	10681	45.40
DCL-Net	15.54±7.31	19.57	22.86±10.62	23.74	0.47±0.24	-4.26	10681	45.40
DC^2-Net	15.39±7.28	18.61	22.55±7.27	22.39	0.51± 0.24	6.06	10681	45.40
2D ResNeXt	14.71±6.10	14.18	21.15±10.96	16.03	0.50±0.25	4.08	**6729**	-2.97
PLPPI*	13.44±7.24	5.93	19.86±11.66	9.78	0.49±0.26	0	6932	0
PLPPI	**12.77±6.84**	**0**	**18.01±10.44**	0	0.48±0.26	0	6932	0

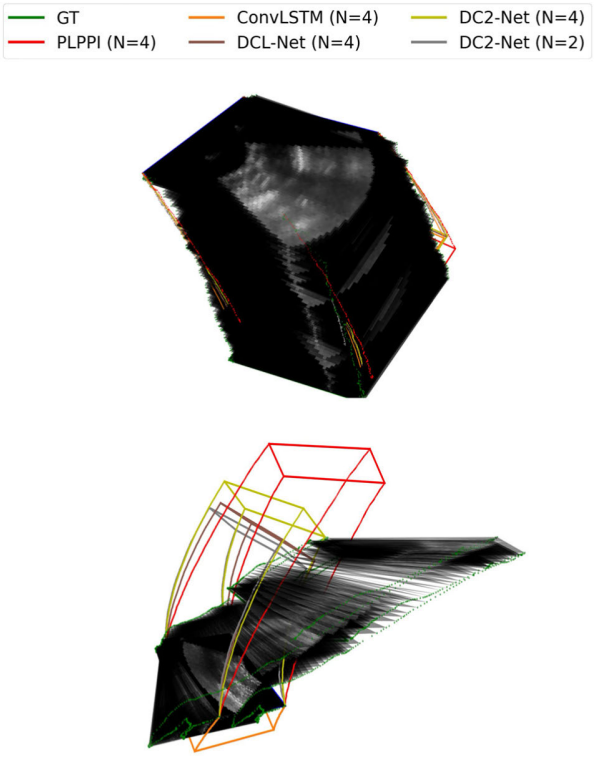


Fig. 7. Visualize the best and worst human forearm cases from top to bottom based on the FD. Green lines represent the ground truth reconstruction given the US images.

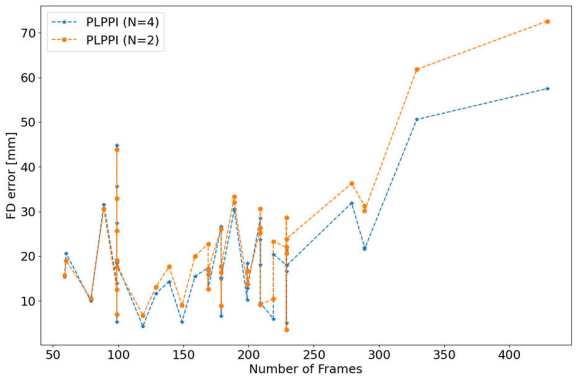


Fig. 8. FD error (mm) plotted as a function of the number of frames, either using two frames (in orange) or four frames (in blue). PLPPI using multiple frames can effectively reduce the reconstruction error, especially for longer sequences.

PLPPI yields a more accurate estimation of rotations in OOP motion. To demonstrate PLPPI's capability in capturing large OOP motion, we plot the motion changes along the transducer movement as translation Z in Fig. 9, using the case with the worst FD score from the forearm dataset. We compare the motion changes Z per frame for PLPPI and baseline methods. Our model captures individual frame motion better compared to baseline methods, confirming the earlier reconstruction results. Despite using four frames as input, PLPPI is less prone to oversmoothing and can better handle motion with varying velocity compared to baseline methods.

Fig. 10 illustrates OOP motion utilizing speckle decorrelation, where areas that appear brighter denote a larger elevational distance to the subsequent frame. Such patterns, characterized by increased brightness in the bottom right,

TABLE II

ROOT-MEAN-SQUARED ERROR (RMSE) AND ROOT-MEAN-SQUARED DEVIATION (RMSD) FOR TRANSLATION Z (ALONG THE MOVEMENT DIRECTION OF THE TRANSDUCER) AND ROTATION (PITCH, YAW, AND ROLL) ARE CALCULATED BETWEEN THE GROUND TRUTH AND THE MODEL'S PREDICTION ON THE TEST SETS. (a) FOREARM. (b) PHANTOM DATA. THESE METRICS ARE COMPUTED BY COMPARING THE PREDICTION AND GROUND TRUTH FOR EACH INDIVIDUAL FRAMES WITHIN THE TEST SETS

(a)

Model	Z (mm)	Pitch (degree)	Yaw (degree)	Roll (degree)
PLPPI	**0.62±0.40**	**5.95±3.58**	**6.22±3.34**	**6.99±3.86**
DC^2-Net	0.66±0.46	6.29±4.69	6.57±3.33	7.10±3.84
DCL-Net	0.63±0.45	6.35±3.72	6.29±3.34	7.08±3.85
ConvLSTM	0.62±0.44	6.18±3.68	6.25±3.44	7.02±3.85

(b)

Model	Z (mm)	Pitch (degree)	Yaw (degree)	Roll (degree)
PLPPI	**0.23±0.07**	**3.48±1.58**	**3.74±2.01**	**2.91±1.11**
DC^2-Net	0.29±0.10	3.77±1.48	3.93±1.92	3.06±1.09
DCL-Net	0.31±0.10	3.58±1.64	3.78±2.20	2.99±1.15
ConvLSTM	0.28±0.10	3.60±1.54	3.77±2.18	2.98±1.12

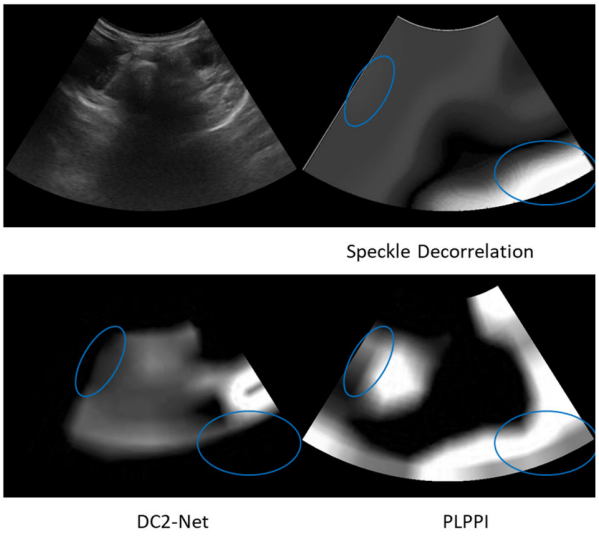


Fig. 9. Motion of the transducer per frame for PLPPI and baseline methods is labeled as Z. The x-axis represents the frame number in the scanning sequence with the worst FD prediction from the forearm dataset, while the y-axis indicates translation Z in millimeters. The red line shows the parameter prediction from PLPPI, and the blue line represents the ground truth.

align with the C-shaped data scanning path, where one side exhibits less motion compared to the other. The bottom two images show the attention map for the OOP Roll motion along the six degrees of freedom. Compared to DC^2-Net, the heatmap generated by our method closely matches the speckle decorrelation pattern, suggesting that our introduced temporal branch enhances the model's ability to learn OOP motion.

VI. DISCUSSION

Now, we analyze the impact of hyperparameters and additional components using the forearm dataset.

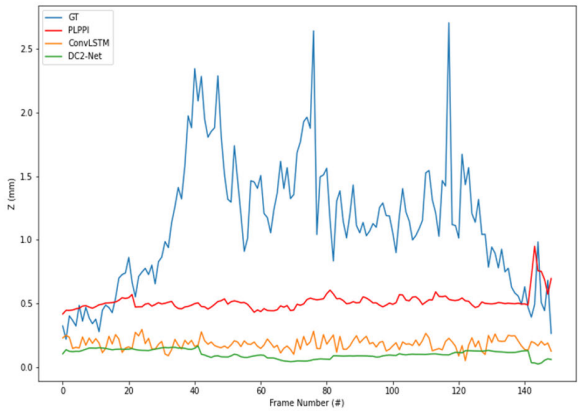


Fig. 10. Representation of heat maps related to OOP roll motion. The top left shows the image in question, and the top right shows the speckle decorrelation map. This map is computed between the US image on the left and its successive frame using the method described in [16].

A. Ablation Study

The performance impact of each component of the model is shown in Table III. First, we observe that adding the temporal stream helps the model's ability to learn temporal information, indicating that our correlation operation effectively extracts frame order and temporal information from US speckle. Next, the random selection of frames during training improves generalization at test time. Using the temporal stream alone has advantages over using only the spatial stream. Furthermore, incorporating TRL results in reductions in DE and FD, demonstrating the benefit of this loss function. This part of the ablation study justifies the key components of our model.

We now analyze the effects of K, the maximum distance to compute correlation on the feature map. Since correlated speckle only exists locally, with an increase in K, the result worsens, as shown in Table IV. The results also suggest that $K = 21$ provides both the lowest average DE and the FD error for the forearm dataset. However, for the phantom dataset $K = 21$, it provides the lowest average DE, but not the lowest FD error. This suggests that the optimal value of K can vary for different datasets. However, using $K = 21$ rather than a higher number reduces the computational load for the correlation operation.

Next, we examine the impact of the number of input frames using results in Table V. For the forearm dataset, four input frames provide the best results in terms of DE, FD, and FE. Table V also concurs with results reported in [48], indicating that additional frames may not significantly improve performance as sufficient long-term dependency information may not be present.

To study the effects of hyperparameters on our model, we vary M and λ_1. Tables VI and VII show how these parameters impact the model performance. First, we perform the ablation study with $\lambda_1 \in [1, 5]$ and M equal to 0.2. The results in Table VI show that setting λ_1 equal to 2 gives a better result as TRL could help with overcoming the oversmoothing effect while reducing the overall DE. Next, we experiment further for M with $\lambda_1 = 2$, as shown in Table VII. The performance increases from $M = 0.1$ to $M = 0.2$ and then

TABLE III
QUANTITATIVE RESULTS INDICATING INFLUENCE OF EACH NETWORK COMPONENT TO THE OVERALL PERFORMANCE

Spatial Stream	Random frame selection	Temporal Stream	Triplet Ranking Loss	DE (mm)	FD (mm)	FE (mm)
✓				15.67±7.39	21.46±12.87	0.50±0.26
		✓		13.59±7.67	20.64±12.44	0.50±0.26
✓	✓			14.71±6.10	21.15±10.96	0.50±0.25
✓		✓		13.50±7.34	20.75±11.92	0.49±0.26
✓	✓	✓		13.23±7.20	20.00±11.71	0.51± 0.26
✓		✓	✓	13.44±7.24	19.86±11.66	0.49±0.26
✓	✓	✓	✓	12.77±6.84	18.01±10.44	0.48±0.26

TABLE IV
QUANTITATIVE RESULTS FOR DIFFERENT VALUES OF THE PARAMETER *K*, THE LOCAL MAXIMUM RANGE USED FOR COMPUTING FEATURE CORRELATION FROM BOTH DATASETS

K	Forearm dataset		Phantom dataset	
	DE (mm)	FD (mm)	DE (mm)	FD (mm)
21	12.77±6.84	18.01±10.44	3.09±1.40	4.65±2.25
25	13.59±7.60	20.89±12.15	3.13±1.36	4.15±2.15
29	13.50±7.40	20.79±11.71	3.19±1.63	4.20±2.60

TABLE V
QUANTITATIVE ANALYSIS FOR THE IMPACT OF THE INPUT NUMBER OF FRAMES

Frame	DE (mm)	FD (mm)	FE (mm)
2	14.13±7.84	22.32±13.14	0.50±0.25
4	12.77±6.84	18.01±10.44	0.48±0.26
6	13.66±7.38	20.87±12.42	0.48±0.25
8	13.35±7.17	20.18±11.40	0.48±0.26

TABLE VI
QUANTITATIVE RESULTS SHOWING THE IMPACT OF HYPERPARAMETERS *M*

M	DE (mm)	FD (mm)	FE (mm)
0.1	13.10±7.17	19.49±11.38	0.49±0.26
0.2	12.77±6.84	18.01±10.44	0.48±0.26
0.3	13.16±7.18	19.47±11.44	0.48±0.26
0.4	13.25±7.33	19.95±11.22	0.49±0.26

TABLE VII
QUANTITATIVE RESULTS SHOWING THE IMPACT OF HYPERPARAMETERS λ_1

λ	DE (mm)	FD (mm)	FE (mm)
1	13.16±7.12	19.40±11.33	0.49±0.26
2	12.77±6.84	18.01±10.44	0.48±0.26
3	13.13±7.36	19.54±11.22	0.50±0.25
4	13.43±7.49	20.15±10.96	0.50±0.25
5	13.07±7.12	19.65±11.56	0.50±0.25

starts to decrease afterward, showing that $M = 0.2$ is a good hyperparameter value.

VII. CONCLUSION

In this article, we presented a novel deep learning approach using constraints derived from US speckle decorrelation to develop a physics-guided learning-based model for 3-D reconstructions. PLPPI leverages a 2-D correlation operation to capture correspondence between frames and decouples spatial and temporal information for learning in a two-stream model.

Extensive experiments were conducted to compare PLPPI to several baseline deep learning methods. The results show that PLPPI significantly outperforms other baseline approaches, particularly for different scanning protocols and sweep sequences with large OOP motion. Further model ablation demonstrates the importance of the correlation operation and effectiveness of our method for complicated sweep sequences. Our experiments also reveal performance degradation when large OOP motion is presented due to sudden movement either by the patient or physician.

One of the limiting factors of our approach is the size of the dataset since only limited freehand US datasets are publicly available. Almost all of them are relatively small, including the forearm dataset and dataset collected on an abdominal phantom. Developing a resource-efficient training scheme that requires less data could also be beneficial for future studies.

Using sensorless 3-D US volume reconstruction from conventional 2-D freehand scans could substantially enhance diagnostic processes, treatment planning, and the continuous visual monitoring of anatomical structures. Moreover, it could improve the precision of image-guided interventional procedures, including biopsies and thermal ablation [52]. There is ample room for improvement, especially for OOP motion. One potential solution would be to refine the transducer trajectory, such as using pose graph optimization—a graph-based optimization previously designed for estimating the unknown trajectory of robots, to optimize the path. Neural radiance field (NeRF) has been utilized for US reconstruction. However, NeRF requires scanning the same region multiple times, which can introduce artifacts that have to be addressed [53]. Dagli et al. [54] also presented issues associated with NeRF under real-world conditions with complex motion. A complete discussion of this aspect is beyond the scope of this article.

We are confident that our method will offer new insights into freehand 3-D US reconstruction and its many clinical applications.

ACKNOWLEDGMENT

The 3-D US dataset used in this article was obtained from Zenodo titled “Freehand ultrasound without external trackers,” Created by Qi Li, last modified on November 9, 2022. The authors include attributions to the following citations: data citation [47] and publication citation [48].

REFERENCES

[1] W. Wein, A. Ladikos, B. Fuerst, A. Shah, K. Sharma, and N. Navab, "Global registration of ultrasound to MRI using the LC_2 metric for enabling neurosurgical guidance," in *Advanced Information Systems Engineering*, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds., Berlin, Germany: Springer, 2013, pp. 34–41, doi: 10.1007/978-3-642-40811-3_5.

[2] A. Gopal, M. Schnellbaecher, Z. Shen, L. Boxt, J. Katz, and D. King, "Freehand three-dimensional echocardiography for determination of left ventricular volume and mass in patients with abnormal ventricles: Comparison with magnetic resonance imaging," *J. Amer. Soc. Echocardiography*, vol. 10, no. 8, pp. 853–861, Oct. 1997, doi: 10.1016/s0894-7317(97)70045-2.

[3] Y. Xiao et al., "Evaluation of MRI to ultrasound registration methods for brain shift correction: The CuRIOUS2018 challenge," *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 777–786, Mar. 2020, doi: 10.1109/TMI.2019.2935060.

[4] L. F. Gonçalves, W. Lee, J. Espinoza, and R. Romero, "Three- and 4-dimensional ultrasound in obstetric practice: Does it help?" *J. Ultrasound Med.*, vol. 24, no. 12, pp. 1599–1624, Dec. 2005, doi: 10.7863/jum.2005.24.12.1599.

[5] S. Engell, J. J. Triano, J. R. Fox, H. M. Langevin, and E. E. Konofagou, "Differential displacement of soft tissue layers from manual therapy loading," *Clin. Biomechanics*, vol. 33, pp. 66–72, Mar. 2016, doi: 10.1016/j.clinbiomech.2016.02.011.

[6] E. Saegusa-Beecroft et al., "Three-dimensional quantitative ultrasound for detecting lymph node metastases," *J. Surgical Res.*, vol. 183, no. 1, pp. 258–269, Jul. 2013, doi: 10.1016/j.jss.2012.12.017.

[7] W. Hu, S. Zhu, F. Briggs, and M. M. Doyley, "Functional ultrasound imaging reveals 3D structure of orientation domains in ferret primary visual cortex," *NeuroImage*, vol. 268, Mar. 2023, Art. no. 119889, doi: 10.1016/j.neuroimage.2023.119889.

[8] E. Roux, F. Varray, L. Petrusca, C. Cachard, P. Tortoli, and H. Liebgott, "Experimental 3-D ultrasound imaging with 2-D sparse arrays using focused and diverging waves," *Sci. Rep.*, vol. 8, no. 1, Jun. 2018, Art. no. 1, doi: 10.1038/s41598-018-27490-2.

[9] R. Prevost et al., "3D freehand ultrasound without external tracking using deep learning," *Med. Image Anal.*, vol. 48, pp. 187–202, Aug. 2018, doi: 10.1016/j.media.2018.06.003.

[10] M. Luo, X. Yang, H. Wang, L. Du, and D. Ni, "Deep motion network for freehand 3D ultrasound reconstruction," 2022, *arXiv:2207.00177*.

[11] H. Guo, H. Chao, S. Xu, B. J. Wood, J. Wang, and P. Yan, "Ultrasound volume reconstruction from freehand scans without tracking," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 3, pp. 970–979, Mar. 2023, doi: 10.1109/TBME.2022.3206596.

[12] M. Luo et al., "Self context and shape prior for sensorless freehand 3D ultrasound reconstruction," 2021, *arXiv:2108.00274*.

[13] W. Yang, A. Ingle, and T. Varghese, "Comparison of three dimensional strain volume reconstructions using SOUPR and wobbler based acquisitions: A phantom study," *Med. Phys.*, vol. 43, no. 4, pp. 1615–1626, Apr. 2016, doi: 10.1118/1.4942814.

[14] Q. Wang, Q.-H. Huang, J. T. W. Yeow, M. R. Pickering, and S. Saarakkala, "Quantitative analysis of musculoskeletal ultrasound: Techniques and clinical applications," *BioMed Res. Int.*, vol. 2017, no. 1, pp. 1–2, 2017, doi: 10.1155/2017/9694316.

[15] Y. Xie, H. Liao, D. Zhang, L. Zhou, and F. Chen, "Image-based 3D ultrasound reconstruction with optical flow via pyramid warping network," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3539–3542, doi: 10.1109/EMBC46164.2021.9630853.

[16] J. Porée, M. Baudet, F. Tournoux, G. Cloutier, and D. Garcia, "A dual tissue-Doppler optical-flow method for speckle tracking echocardiography at high frame rate," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2022–2032, Sep. 2018, doi: 10.1109/TMI.2018.2811483.

[17] J. Luo and E. E. Konofagou, "A fast normalized cross-correlation calculation method for motion estimation," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 57, no. 6, pp. 1347–1357, Jun. 2010, doi: 10.1109/TUFFC.2010.1554.

[18] R.-F. Chang et al., "3-D U.S. frame positioning using speckle decorrelation and image registration," *Ultrasound Med. Biol.*, vol. 29, no. 6, pp. 801–812, Jun. 2003, doi: 10.1016/s0301-5629(03)00036-x.

[19] A. A. Azar, H. Rivaz, and E. Boctor, "Speckle detection in ultrasonic images using unsupervised clustering techniques," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 8098–8101, doi: 10.1109/IEMBS.2011.6091997.

[20] L. H. Frich, K. L. Lambertsen, J. Hjarbaek, J. S. Dahl, and A. Holsgaard-Larsen, "Musculoskeletal application and validation of speckle-tracking ultrasonography," *BMC Musculoskeletal Disorders*, vol. 20, no. 1, p. 192, May 2019, doi: 10.1186/s12891-019-2562-8.

[21] A. Lang, P. Mousavi, G. Fichtinger, and P. Abolmaesumi, "Fusion of electromagnetic tracking with speckle-tracked 3D freehand ultrasound using an unscented Kalman filter," *Proc. SPIE*, vol. 7265, pp. 399–410, Feb. 2009, doi: 10.1117/12.813879.

[22] C. Laporte and T. Arbel, "Learning to estimate out-of-plane motion in ultrasound imagery of real tissue," *Med. Image Anal.*, vol. 15, no. 2, pp. 202–213, Apr. 2011, doi: 10.1016/j.media.2010.08.006.

[23] M. Toews and W. M. Wells, "Phantomless auto-calibration and online calibration assessment for a tracked freehand 2-D ultrasound probe," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 262–272, Jan. 2018, doi: 10.1109/TMI.2017.2750978.

[24] H. Guo, S. Xu, B. Wood, and P. Yan, "Sensorless freehand 3D ultrasound reconstruction via deep contextual learning," 2020, *arXiv:2006.07694*.

[25] Y. Dou, F. Mu, Y. Li, and T. Varghese, "Sensorless end-to-end freehand ultrasound with physics inspired network," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, vol. 53, Sep. 2023, pp. 1–4, doi: 10.1109/IUS51837.2023.10307112.

[26] G. A. Sonn et al., "Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device," *J. Urology*, vol. 189, no. 1, pp. 86–92, Jan. 2013, doi: 10.1016/j.juro.2012.08.095.

[27] T. A. Tuthill, J. F. Krücker, J. B. Fowlkes, and P. L. Carson, "Automated three-dimensional U.S. frame positioning computed from elevational speckle decorrelation," *Radiology*, vol. 209, no. 2, pp. 575–582, Nov. 1998, doi: 10.1148/radiology.209.2.9807593.

[28] A. H. Gee, R. J. Housden, P. Hassenpflug, G. M. Treece, and R. W. Prager, "Sensorless freehand 3D ultrasound in real tissue: Speckle decorrelation without fully developed speckle," *Med. Image Anal.*, vol. 10, no. 2, pp. 137–149, 2006, doi: 10.1016/j.media.2005.08.001.

[29] C. R. Chen et al., "Deep analysis of CNN-based spatio-temporal representations for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6161–6171, doi: 10.1109/CVPR46437.2021.00610.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2014, pp. 1–4. Accessed: Mar. 16, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/00ec53c4682d36f5c4359f4ae7bd7ba1-Abstract.html

[31] L. Wang et al., "Temporal segment networks for action recognition in videos," 2017, *arXiv:1705.02953*.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.

[34] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," 2017, *arXiv:1705.07750*.

[35] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," 2019, *arXiv:1812.03982*.

[36] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092, doi: 10.1109/ICCV.2019.00718.

[37] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2018, *arXiv:1711.11248*.

[38] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," 2020, *arXiv:2004.04730*.

[39] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," 2019, *arXiv:1904.02811*.

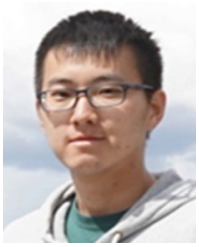
[40] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3D-2D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.

[41] M. Luo et al., "RecON: Online learning for sensorless freehand 3D ultrasound reconstruction," *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102810, doi: 10.1016/j.media.2023.102810.

[42] K. Miura, K. Ito, T. Aoki, J. Ohmiya, and S. Kondo, "Pose estimation of 2D ultrasound probe from ultrasound image sequences using CNN and RNN," in *Proc. Int. Workshop Adv. Simplifying Med. Ultrasound*, Strasbourg, France. Berlin, Germany: Springer, Sep. 2021, pp. 96–105, doi: 10.1007/978-3-030-87583-1_10.
[43] X. Chen, H. Chen, Y. Peng, L. Liu, and C. Huang, "A freehand 3D ultrasound reconstruction method based on deep learning," *Electronics*, vol. 12, no. 7, p. 1527, Mar. 2023, doi: 10.3390/electronics12071527.
[44] P. Fischer et al., "FlowNet: Learning optical flow with convolutional networks," 2015, *arXiv:1504.06852*.
[45] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016, doi: 10.23915/distill.00003.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
[47] Q. Li et al., *Freehand Ultrasound Without External Trackers*. Genève, Switzerland: Zenodo, Nov. 2022, doi: 10.5281/zenodo.7740734.
[48] Q. Li et al., "Long-term dependency for 3D reconstruction of freehand ultrasound without external tracker," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 3, pp. 1033–1042, Mar. 2024, doi: 10.1109/TBME.2023.3325551.
[49] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2017, *arXiv:1611.05431*.
[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, *arXiv:1711.05101*.
[51] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.
[52] R. M. Pohlman et al., "Differential imaging of liver tumors before and after microwave ablation with electrode displacement elastography," *Ultrasound Med. Biol.*, vol. 47, no. 8, pp. 2138–2156, Aug. 2021, doi: 10.1016/j.ultrasmedbio.2021.03.027.
[53] Y. Dou and T. Varghese, "Pitfalls with neural radiance fields for 3D freehand ultrasound reconstruction," in *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jun. 2024, pp. 1–4.
[54] R. Dagli, A. Hibi, R. G. Krishnan, and P. N. Tyrrell, "NeRF-US: Removing ultrasound imaging artifacts from neural radiance fields in the wild," 2024, *arXiv:2408.10258*.

Yimeng Dou (Graduate Student Member, IEEE) received the B.S. degree in biomedical engineering from the University of California, Daivs, CA, USA, in 2017, and the M.S. degree in biomedical data science from the University of Wisconsin–Madison, Madison, WI, USA, in 2024, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

From 2018 to 2020, he was a Junior Research Specialist with the University of California. His research mainly focuses on 3-D and volume representation, dynamic 3-D novel view synthesis, and its application to healthcare and novel imaging systems.



Fangzhou Mu (Student Member, IEEE) received the Ph.D. degree in computer sciences from the University of Wisconsin–Madison, Madison, WI, USA, in 2023.

He is currently a Senior Deep Learning Algorithm Engineer at NVIDIA, Santa Clara, CA, USA. His research interests include computer vision, focusing on multimodal learning, and computational photography.

Yin Li (Member, IEEE) received the Ph.D. degree in computer science from Georgia Tech, Atlanta, GA, USA, in 2017.

He is an Assistant Professor with the Department of Biostatistics and Medical Informatics and an Affiliate Faculty Member with the Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA. He was a Postdoctoral Fellow with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His work was covered by MIT Tech Review, WIRED U.K., New Scientist, BBC, and Forbes. His primary research focus is computer vision. He is also interested in the applications of vision and learning in healthcare.

Dr. Li was a co-recipient of the Best Student Paper Awards at MobiHealth in 2014 and IEEE Face and Gesture 2015, and the Best Demo Nominee at ECCV 2020. He has been serving as the Area Chair for the top vision and AI conferences, including CVPR, ICCV, ECCV, IJCAI, and NeurIPS.

Tomy Varghese (Senior Member, IEEE) received the B.E. degree in instrumentation technology from the University of Mysore, Mysore, India, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the University of Kentucky, Lexington, KY, USA, in 1992 and 1995, respectively.

From 1988 to 1990, he was employed as an Engineer at Wipro Information Technology Ltd., Mumbai, India. From 1995 to 2000, he was a Postdoctoral Research Associate at the Ultrasonics Laboratory, Department of Radiology, University of Texas Medical School, Houston, TX, USA. He is currently a Professor with the Department of Medical Physics, University of Wisconsin–Madison, Madison, WI, USA. His current research interests include elastography, ultrasound imaging, ultrasonic tissue characterization, detection and estimation theory, statistical pattern recognition, signal and image processing, and deep learning applications in medical imaging.

Dr. Varghese is a fellow of the American Institute of Ultrasound in Medicine (AIUM) and a member of the American Association of Physicists in Medicine (AAPM) and Eta Kappa Nu.