Enhancing Free-Hand 3-D Photoacoustic and Ultrasound Reconstruction Using Deep Learning

SiYeoul Lee, Seonho Kim, MinKyung Seo, SeongKyu Park, Salehin Imrus, Kambaluru Ashok, DongEon Lee, Chunsu Park, SeonYeong Lee, Jiye Kim, Jae-Heung Yoo, and MinWoo Kim

***Abstract*— This study introduces a motion-based learning network with a global-local self-attention module (MoGLo-Net) to enhance 3D reconstruction in handheld photoacoustic and ultrasound (PAUS) imaging. Standard PAUS imaging is often limited by a narrow field of view (FoV) and the inability to effectively visualize complex 3D structures. The 3D freehand technique, which aligns sequential 2D images for 3D reconstruction, faces significant challenges in accurate motion estimation without relying on external positional sensors. MoGLo-Net addresses these limitations through an innovative adaptation of the self-attention mechanism, which effectively exploits the critical regions, such as fully-developed speckle areas or high-echogenic tissue regions within successive ultrasound images to accurately estimate the motion parameters. This facilitates the extraction of intricate features from individual frames. Additionally, we employ a patch-wise correlation operation to generate a correlation volume that is highly correlated with the scanning motion. A custom loss function was also developed to ensure robust learning with minimized bias, leveraging the characteristics of the motion parameters. Experimental evaluations demonstrated that MoGLo-Net surpasses current state-of-the-art methods in both quantitative and qualitative performance metrics. Furthermore, we expanded the application of 3D reconstruction technology beyond simple B-mode ultrasound volumes to incorporate Doppler ultrasound and photoacoustic imaging, enabling 3D visualization of vasculature. The source code for this study is publicly available at: https://github.com/pnu-amilab/US3D**

***Index Terms*— Deep learning, free-hand 3-D reconstruction, ultrasound imaging, photoacoustic imaging.**

I. Introduction

DIAGNOSTIC ultrasound (US) is a widely employed modality for examining various organs and tissues, owing to its real-time imaging capabilities. Particularly, it serves as a preferred tool for guiding medical procedures such as biopsies and injections [1]. In addition, US plays a pivotal role in dynamically monitoring the vascular system through the Doppler effect [2]. Meanwhile, the integration of photoacoustic (PA) imaging with conventional US imaging (PAUS imaging) has been extensively studied because of its promising clinical potential. This integration aims to enhance the existing benefits of ultrasound while introducing novel capabilities for both interventional and functional imaging [3].

In the PAUS system, a handheld transducer is responsible for emitting US and/or laser pulses and receiving the resulting acoustic wave signals. While this configuration allows users the flexibility to manually scan a region of interest (RoI), it is accompanied by certain limitations. Notably, the field of view (FoV) is restricted, providing only a narrow 2D cross-section of the image area, which hampers the understanding of the topological 3D structures of the target. Although some transducers are specifically designed for 3D imaging, their physical dimensions often render them impractical for handheld use [4].

An alternative approach is the 3D freehand method, which involves the sequential alignment of 2D image sections acquired from freehand sweeps with a standard transducer. Users can playback a series of 2D US image frames as a video through a CINE loop, but the visualization of the 3D rendering

Received 25 April 2025; revised 5 June 2025; accepted 10 June 2025. Date of publication 13 June 2025; date of current version 30 October 2025. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant RS-2021-NR059679, in part by the Institute of Information and communications Technology Planning and Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by the Korea government (MSIT) under Grant IITP-2025-RS-2023-00254177, and in part by the Institute of Information and communications Technology Planning and Evaluation (IITP) under the Leading Generative AI Human Resources Development grant funded by the Korea government (MSIT) under Grant IITP-2025-RS-2024-00360227. Recommended by Associate Editor R. van Sloun. *(Corresponding author: MinWoo Kim.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the local Institutional Review Board, Pusan National University IRB, under Application No. 2023_74_HR.

SiYeoul Lee, Seonho Kim, MinKyung Seo, SeongKyu Park, Salehin Imrus, Kambaluru Ashok, DongEon Lee, SeonYeong Lee, and Jiye Kim are with the Department of Information Convergence Engineering, Pusan National University, Yangsan 50612, South Korea (e-mail: gu_hong3648@pusan.ac.kr; shkim7103@pusan.ac.kr; mkseo99@pusan.ac.kr; parksk6737@pusan.ac.kr; salehin@pusan.ac.kr; ashok@pusan.ac.kr; dddlee.dev@gmail.com; jasleen37@pusan.ac.kr; jiji5000@pusan.ac.kr).

Chunsu Park is with the School of Biomedical Convergence Engineering, Pusan National University, Yangsan 50612, South Korea (e-mail: cspark5484@pusan.ac.kr).

Jae-Heung Yoo is with the Department of Orthopedic Surgery, Busan Medical Center, Busan 47527, South Korea (e-mail: jaeheung85@gmail.com).

MinWoo Kim is with the School of Biomedical Convergence Engineering, Pusan National University, Yangsan 50162, South Korea, and also with the Center for Artificial Intelligence Research, Pusan National University, Busan 46241, South Korea (e-mail: mkim180@pusan.ac.kr).

Digital Object Identifier 10.1109/TMI.2025.3579454

volume or structure of the target through accumulating frames is more straightforward. This method proves ideal in clinical practice when tissue motion is static. However, a primary technical challenge of this approach lies in estimating sweeping motion for correct 3D reconstruction. Although an external positional device can be attached to the transducer, it makes the transducer bulkier and often provides inaccurate measurements in a clinical environment due to various optical or electrical disturbances [5].

Several studies have explored the estimation of scan trajectories directly from 2D US images, eliminating the need for external sensors [6], [7], [8]. The key idea is to exploit and track tissue speckle patterns in US B-mode images. Recent approaches have employed deep learning (DL) frameworks to better leverage speckle features for predicting scan motion, typically defined by six parameters. Since the pioneering work of Prevost et al. [9], which introduced the first DL model for this task, subsequent studies have proposed advanced models [10], [11], [12], [13], [14] to improve stability and accuracy. However, their overall estimation accuracies were still insufficient for clinical practice, and they were rarely validated for long elevational scan trajectories, which are common in clinical practice.

The principal contribution of this study lies in proposing a novel DL method to elevate visualization capabilities of 3D US and PAUS system. The structure is composed of a ResNet-driven encoder [15] and an estimator with a Long Short-Term Memory (LSTM) block to extract features from given B-mode frames and estimate the motion vectors between consecutive frames, leveraging long-term memory. One part of the encoder consists of special blocks that can directly access the correlation between encoded feature maps from adjacent frames. Additionally, we designed task-specific novel global-local attention module, which effectively highlights the critical local regions for motion estimation, such as fully-developed speckle areas or high-echogenic tissue areas. Moreover, our customized loss function ensures that the network robustly learns motion without a significant distortion in specific parameter estimation.

The model was rigorously evaluated using both our proprietary dataset and a publicly available dataset. Utilizing a programmable ultrasound system enabled us to assess the model's performance across a variety of conditions, such as reductions in B-mode image speckle or the inclusion of less processed data as input. Comparing deep learning methods is inherently difficult due to the fact that each model is typically designed and optimized for a specific acquisition dataset. To mitigate this issue, we also employed an publicly open dataset [15], obtained under entirely different settings from our own. By training and testing our model on open dataset, we aimed to demonstrate that our framework is generalizable and not overly dependent on the dataset we specifically acquired.

Another significant contribution involves highlighting the versatility of the panorama technique in clinical applications, extending beyond the mere compilation of B-mode images. Through the combination of US and laser transmission sequencing, a hand-held free-scan PAUS system can obtain PA data as well as US data at once. Using positioning contents, we demonstrate that 3D reconstruction of vessels is available from either US power Doppler mode (PD-mode) data or PA data. To the best of our knowledge, this is the first reported application of this approach. These advancements have the potential to aid specialists in diagnosing diseases or precisely locating targets during the intervention process.

The main contributions of this study are summarized as follows:

- We propose MoGLo-Net, a novel network architecture combining correlation features, global-local attention, and a motion-aware supervision strategy to enhance motion estimation in free-hand 3D ultrasound and photoacoustic imaging.
- Our method is rigorously validated on both in-house and public datasets, demonstrating consistent improvements in trajectory estimation and volume reconstruction across various acquisition conditions.
- By leveraging a programmable ultrasound system, we examine the performance under diverse scenarios, including speckle suppression and raw data processing, highlighting the robustness of the proposed framework.
- We present the first demonstration of 3D vascular reconstruction from both Power Doppler and photoacoustic acquisitions using a free-hand protocol and the same underlying framework, indicating broad applicability of the method.

II. Related Works

Traditional approaches exploited tissue speckle patterns in US B-mode images, as speckle content between successive frames tends to be preserved, even in out-of-plane motions. Since the pioneering work by Trahey et al. [16], several studies have focused on estimating scan motion by analyzing the correlation between adjacent frames [6], [7]. Chang et al. [17] proposed a frame positioning method for 3D US by combining speckle decorrelation and image registration via inter-frame correlation analysis. Since the degree of speckle development varies across different regions of tissue, Gee et al. [8] proposed an adaptive speckle decorrelation method, enabling more accurate 3D ultrasound reconstruction. Laporte and Arbel [18] further advanced these approaches by introducing a learning-based decorrelation model that adapts to local tissue properties, enabling more robust motion estimation without relying on idealized speckle assumptions. These studies have demonstrated the feasibility of sensorless free-hand US through gradual improvements in various factors such as estimation accuracy, generalizability, and less-constrained scan protocols.

With the rapid advancements in DL, Prevost et al. [9] were the first to attempt 3D US volume reconstruction using a convolutional neural network (CNN). Guo et al. [10] introduced a deep contextual learning network (DCL-Net), utilizing 3D convolution to exploit the sequential context information in US scan frames, along with an innovative loss function based on correlation values. Building on this, Guo et al. [19] expanded their previous work by developing a

deep contextual-contrastive network (DC^2-Net), which applied a margin triplet loss for contrastive learning in a regression task.

Around the same time, Luo et al. [11] made key contributions to free-hand US imaging by proposing an RNN-based model with a novel self-supervised and adversarial learning strategy. This approach enabled plausible visual reconstructions, even in more challenging scanning scenarios. Luo et al. continued their active contributions with the development of MoNet, a motion network incorporating an inertial measurement unit (IMU) sensor, a lightweight sensor for capturing acceleration data [20]. Their multi-branch DL architecture leveraged both US images and IMU sensor data for improved performance. In later work, Luo et al. utilized multiple IMU sensors to further enhance reconstruction accuracy [21].

Recently, Luo et al. [14] introduced an online learning reconstruction framework (RecON), imposing constraints such as motion-weighted training loss, frame-level contextual consistency, and path-label similarity, which significantly improved the accuracy of motion estimation in complex scan motions. Other recent DL approaches have adapted popular models for US imaging. Miura et al. [13] proposed a sequential CNN-RNN structure for both relative and absolute pose estimation, demonstrating the efficacy of relative pose integration via RNNs. Inspired by pyramid warping techniques [22], Xie et al. [23] developed a network that extracts multi-scale features from US frames to better capture low-frequency B-mode information. With the increasing popularity of transformers, Ning et al. [12] applied a transformer architecture to combine local and long-range information from a CNN-based backbone encoder and IMU sensors. Li et al. [15] further identified long-term dependencies between sequential US frames and the influence of anatomical or scan protocol factors. More recently, Yan et al. [24] proposed FiMA (Fine-grained Context and Multi-modal Alignment), which leverages multi-directional state space models inspired by Mamba [25] to capture fine-grained spatio-temporal dependencies for freehand 3D ultrasound reconstruction. In parallel, Dou et al. [26] utilized a correlation operation, a key component in the field of optical flow estimation, resulting in a physics-inspired learning protocol with improved performance.

Despite these advancements, achieving sufficient accuracy for clinical applications remains a challenge. Many existing methods have not been validated on extensive, real-world datasets, nor have they adequately addressed the effects of image processing techniques, such as speckle reduction, which can distort critical information in B-mode images. Furthermore, there has been limited exploration of adapting these DL-based methods to other imaging modes, such as PD and PA imaging.

The potential of DL-based scan motion tracking systems to enable vascular visualization over large regions is significant. By integrating optimized ultrasound and laser sequencing, raw data reconstruction, and post-processing techniques, these systems can extend their utility to PA imaging and PD-mode US imaging. Such advancements promise to open new avenues for clinical applications, including more accurate visualization of vascular structures and enhanced interventional guidance.

III. Methods

The primary aim of this study is to reconstruct a 3D US volume from sequential 2D B-mode frames captured by sweeping a standard linear transducer over time. Accurate assembly of the volume hinges on the precise estimation of each frame's position, that is, the accumulation of the transducer's motion. In hand-held PAUS imaging, the real-time B-mode images serve as the foundation. Due to their distinct anatomical layers and tissue speckle patterns, they can aid in tracking motion.

A. Problem Definition

Assuming a scan sequence comprises a total of N B-mode frames, the absolute position vector $\boldsymbol{\theta}_i$ of the ith B-mode frame $\mathbf{B}_i \in \mathbb{R}^{H\times L}$ (or transducer at the time) is represented by 6-degree-of-freedom (DoF) vector, defining the 2D Euclidean plane for each image frame within the 3D space:

$$\boldsymbol{\theta}_i = \left[\theta_i^1, \theta_i^2, \theta_i^3, \theta_i^4, \theta_i^5, \theta_i^6\right]^\top \in \mathbb{R}^6, \tag{1}$$

where $\left[\theta_i^1, \theta_i^2, \theta_i^3\right]^T$ and $\left[\theta_i^4, \theta_i^5, \theta_i^6\right]^T$ denote the 3D Euclidean coordinates and the Euler angles. Here, $\theta^1, \theta^2, \theta^3$ correspond to the positions along the axial, lateral, and elevational axes, while $\theta^4, \theta^5, \theta^6$ represent the rotations around the pitch, yaw, and roll axes, respectively. For $i = 0$, $\boldsymbol{\theta}_0$ is initialized as $\mathbf{0}$.

The vector θ_i can be converted into the homogeneous transformation matrix $\mathbf{T}_i = [\mathbf{R}_i, \mathbf{t}_i; \mathbf{0}^\top, 1]$ where $\mathbf{R}_i \in \mathbb{R}^{3\times3}$ and $\mathbf{t_i} \in \mathbb{R}^3$ indicate the rotation matrix and translation vector, respectively. Then, the relative transformation $\Delta\mathbf{T}_i$ between adjacent frames $(\mathbf{B}_i, \mathbf{B}_{i+1})$ can be obtained from the transformation matrices $(\mathbf{T}_i, \mathbf{T}_{i+1})$ as $\Delta\mathbf{T}_i = \mathbf{T}_{i+1}\mathbf{T}_i^{-1}$. The cumulative product of the relative transformation matrices results in the absolute transformation matrix as

$$\mathbf{T}_{n+1} = \prod_{i=0}^{n} \Delta\mathbf{T}_i \mathbf{T}_0, \tag{2}$$

where $\mathbf{T}_0 = \mathbf{I}$. Finally, the relative position vector $\Delta\boldsymbol{\theta}_i$ is extracted from $\Delta\mathbf{T}_i$ for use in supervised learning.

Our goal is to estimate the relative motion $\Delta\boldsymbol{\theta}_i$ between two consecutive frames $(\mathbf{B}_i, \mathbf{B}_{i+1})$ using information from the two frames as well as the preceding frames. This can be formally expressed as

$$f(\{B_i\}_{i=n}^{n+s}) = \{\Delta\hat{\theta}_i\}_{i=n}^{n+s-1}, \tag{3}$$

where f is the DL model, and $\Delta\hat{\boldsymbol{\theta}}_i$ is the estimated relative motion. The estimate $\hat{\boldsymbol{\theta}}_i$ is then used to derive the absolute transformation matrix $\hat{\mathbf{T}}_i$, which is employed to map voxel positions and reconstruct the 3D volume. In the training phase, all estimates $\{\Delta\hat{\theta}_i\}_{i=n}^{n+s-1}$ are used to compute the loss and update model parameters. In contrast, during inference, only the final prediction $\Delta\hat{\theta}_{n+s-1}$ is used to update the global scan trajectory and reconstruct the 3D volume.

B. Deep Learning Framework

As shown in Fig. 1, we developed a motion-based learning network, called **MoGLo-Net** (motion-based learning network with a global-local self-attention module), to estimate the relative scan motion $\Delta\hat{\boldsymbol{\theta}}_i$ from B-mode sequences.

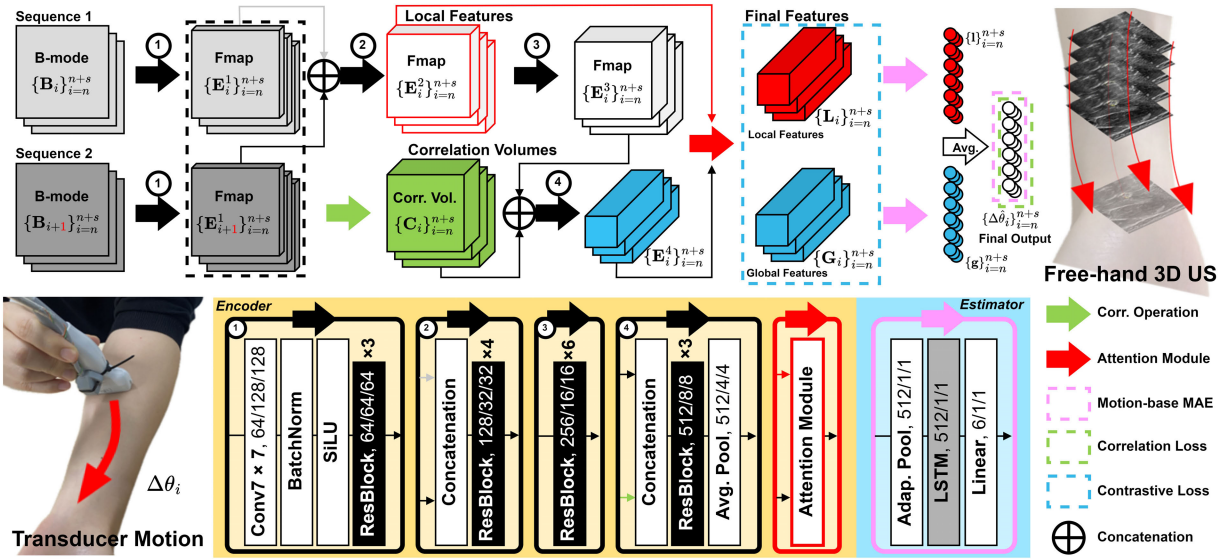

Fig. 1. Overview of our motion-based learning network with a global-local self-attention module (MoGLo-Net) structure. Trainable components are denoted by color-filled arrows, signifying neural network modules. Rectangular or cubic shapes represent 2D images or 3D tensors, respectively. The model processes two B-mode sequences and outputs the estimates of relative motion vectors $\Delta\hat{\boldsymbol{\theta}}_i$. Vectors or feature maps within dotted boxes contribute to the loss function, while the final estimates facilitate the assembly of 2D images into a 3D volume.

1) Data Preparation: Prior to feeding data into the model, two B-mode sequences were generated. The first sequence, denoted as $\{\mathbf{B}_i\}_{i=n}^{n+s}$, consists of frames from $\mathbf{B}_n$ to $\mathbf{B}_{n+s}$. The second sequence, $\{\mathbf{B}_{i+1}\}_{i=n}^{n+s}$, consists of frames from $\mathbf{B}_{n+1}$ to $\mathbf{B}_{n+s+1}$. These sequences are processed in parallel using ResNet-based Encoder Block 1 [27], ensuring consistent refinement of features. The resulting feature maps are represented as $\{(\mathbf{E}_i^1, \mathbf{E}_{i+1}^1)\}_{i=n}^{n+s}$.

2) Feature Extraction: From these features, correlation volumes $\{\mathbf{C}_i\}_{i=n}^{n+s}$ are extracted to capture relationships between successive frames. The concatenated features from the two sequences are then fed into Encoder Block 2, producing refined feature maps, denoted as $\{\mathbf{E}_i^2\}_{i=n}^{n+s}$. These features are further processed by Encoder Block 3, which outputs $\{\mathbf{E}_i^3\}_{i=n}^{n+s}$. Next, the concatenation of the correlation volumes $\{\mathbf{C}_i\}_{i=n}^{n+s}$ and the features from Encoder Block 3, $\{\mathbf{E}_i^3\}_{i=n}^{n+s}$, is passed through Encoder Block 4, yielding $\{\mathbf{E}_i^4\}_{i=n}^{n+s}$.

3) Global-Local Attention: To refine both global and local contextual information, we employed not only conventional attention but also a self-attention mechanism. These combined attention mechanisms recalibrate both global and local features while effectively highlighting critical regions for motion estimation within the local features. The resulting final features, denoted as $\{\mathbf{G}_i\}_{i=n}^{n+s}$ and $\{\mathbf{L}_i\}_{i=n}^{n+s}$, are derived from the global and local features, respectively. The final features are then contrasted using a triplet loss that leverages motion vectors to improve the model's ability to differentiate subtle variation in motion.

4) Motion Estimation: For motion estimation, the model employs two RNN-based motion estimators to predict six-dimensional relative motion vectors. The motion, denoted as $\{\Delta\hat{\mathbf{g}}_i\}$, is derived from the global features $\{\mathbf{G}_i\}$, while the motion, denoted as $\{\Delta\hat{\mathbf{l}}_i\}$, is derived from the local features $\{\mathbf{L}_i\}$. The final motion predictions, $\{\Delta\hat{\boldsymbol{\theta}}_i\}_{i=1}^{N}$, are obtained by averaging these motion vectors. The predictions are supervised using a motion-based mean absolute error (MMAE) and a correlation loss to ensure accurate motion estimation.

5) Testing Phase: The predicted relative motions $\{\Delta\hat{\boldsymbol{\theta}}_i\}_{i=1}^{N}$ are transformed into relative transformations, denoted as $\{\Delta\hat{\mathbf{T}}_i\}_{i=1}^{N}$. Using the cumulative product of these transformations, as defined in Equation (2), the absolute transformations $\{\hat{\mathbf{T}}_i\}_{i=1}^{N}$ are obtained. These absolute transformations are then used to reconstruct the 3D US volumes.

C. Correlation Operation

We utilized a patch-wise correlation operation to extract the relationship between adjacent frames, inspired by speckle decorrelation techniques [6], [7], [8] and correlation operations [26], [28]. This operation is performed on the feature maps $(\mathbf{E}_i^1, \mathbf{E}_{i+1}^1)$, which are initially refined from the B-mode sequences using Encoder Block 1 (Fig. 1).

As shown in Fig. 2, the first step in this operation is to define two RoIs, represented by dotted squares, at the same locations in both feature maps $(\mathbf{E}_i^1, \mathbf{E}_{i+1}^1)$. Within these RoIs, pairs of patches are extracted. While the patch in $\mathbf{E}_i^1$ (red filled square) remains stationary at the center of the RoI, the patch in $\mathbf{E}_{i+1}^1$ (black filled square) moves across the RoI. A correlation array is computed from all possible pairs of patches within the RoIs. By stacking these correlation arrays from multiple RoIs, the correlation volume $\mathbf{C}_i$ is generated.

This correlation volume provides valuable motion-related information, enabling the model to estimate motion more precisely. For instance, if there is no motion (i.e., a stationary frame), the correlation volume ideally contains consistent values. In the case of pure elevational motion, the correlation values generally decrease overall, with the largest value remaining at the center of the correlation arrays when the positions of the two patches coincide. Conversely, during

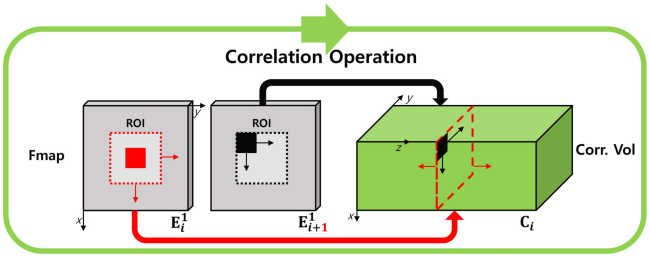

Fig. 2. Correlation operation. It generates the correlation volume $\mathbf{C}_i$ from two feature maps ($\mathbf{E}_i^1$, $\mathbf{E}_{i+1}^1$). The dotted box in each map represents the spatial RoI, and the filled box represents a 3D patch spanning all channels but covering only a portion of the spatial RoI. The red patch remains fixed at the center of the RoI, while the black patch moves across the RoI. All possible correlations between the two patches are stored in a 2D array (red dotted box) within the volume. By moving both RoIs across the feature maps, these arrays are stacked to generate the full 3D correlation volume.

lateral motion, the locations of the largest correlation values shift depending on the motion direction.

D. Global-Local Attention

In recent advancements within the fields of natural image and language processing, numerous attention mechanisms have been developed to capture dependencies between large-scale (global) and small-scale (local) contexts [29], [30], [31]. Drawing inspiration from these mechanisms, we designed a global-local attention module as a self-attention mechanism. This module highlights local features (semantics in specific regions) based on global features that summarize semantic information across the entire image, allowing the motion estimator to leverage this information efficiently.

As illustrated in Fig. 3, our attention module is formalized as $\mathbf{G}, \mathbf{L} = \eta(\mathbf{E}^2, \mathbf{E}^4)$, where $\mathbf{E}^2$ and $\mathbf{E}^4$ are the input local and global features, respectively, and $\mathbf{G}$ and $\mathbf{L}$ are the corresponding recalibrated global and local outputs. For brevity, the temporal index i is omitted in the following derivations.

First, both global and local features are recalibrated using a standard attention mechanism [32]. The local feature $\mathbf{E}_i^2 \in \mathbb{R}^{128\times H/4\times W/4}$, extracted by passing B-mode frames through several encoding layers, reduces the spatial domain from $H \times W$ to $H/4 \times W/4$ while preserving local features of the B-mode domain. For the local feature blocks, the feature map $\mathbf{E}_i^2$ is divided into 256 local feature blocks, $\{\mathbf{E}_k^2\}_{k=1}^{256}$, through a patchnization, where $\mathbf{E}_k^2 \in \mathbb{R}^{128\times H/64\times W/64}$, corresponding to a receptive field of $H/16 \times W/16$ in the original B-mode image. Then, the pooled feature map is calculated as $\mathbf{P}_\mathbf{c}^\mathbf{L} = \Gamma(\mathbf{E}^2) \in \mathbb{R}^{128}$, where $\Gamma(\cdot)$ denotes an adaptive pooling operation. The channel attention score is then obtained as:

$$\mathbf{S}_\mathbf{c}^\mathbf{L} = \sigma(\mathbf{W}_{\mathbf{c}2}^\mathbf{L}\mathbf{W}_{\mathbf{c}1}^\mathbf{L}\mathbf{P}_\mathbf{c}^\mathbf{L}) \in \mathbb{R}^{128}, \tag{4}$$

where $\mathbf{W}_{\mathbf{c}2}^\mathbf{L} \in \mathbb{R}^{128\times 8}$ and $\mathbf{W}_{\mathbf{c}1}^\mathbf{L} \in \mathbb{R}^{8\times 128}$ are the weights of a multi-layer perceptron (MLP), and $\sigma(\cdot)$ is the sigmoid function. The recalibration of each local patch block $\mathbf{E}_k^2$ is then conducted using the channel attention score:

$$\mathbf{R}_k = \mathbf{S}_\mathbf{c}^\mathbf{L} \otimes \mathbf{E}_k^2 \in \mathbb{R}^{128\times H/64\times W/64}, \tag{5}$$

where $\otimes$ denotes the Hadamard product with broadcasting. The resulting recalibrated feature blocks $\mathbf{R}_k$ capture the refined local information.

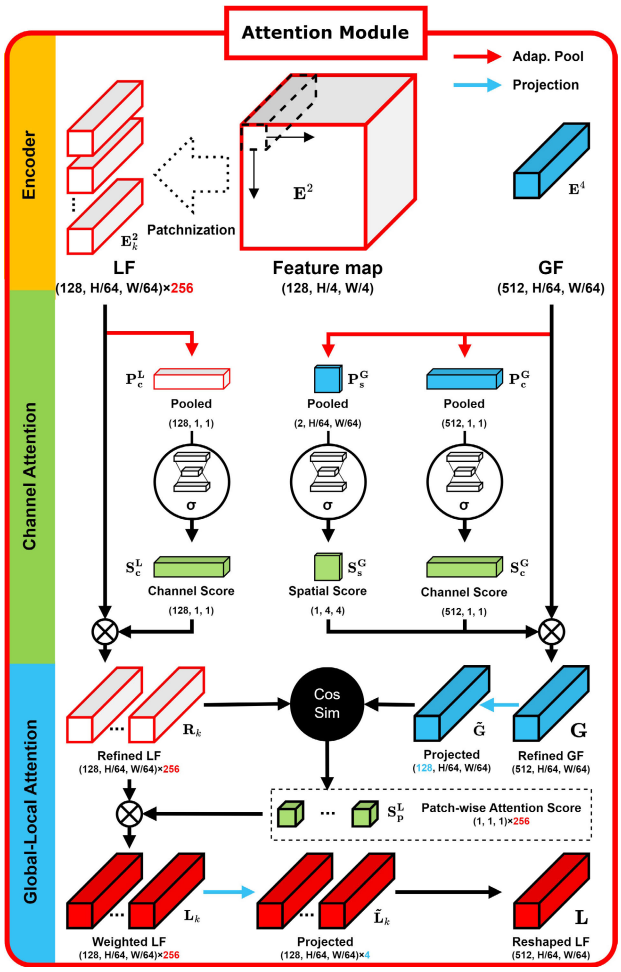

Fig. 3. Global-Local Attention Module. The module recalibrates local features $\mathbf{E}^2$ and global features $\mathbf{E}^4$ to enhance motion estimation. Local feature blocks are extracted from $\mathbf{E}^2$ and recalibrated using channel attention, resulting in refined local feature blocks $\mathbf{R}_k$. Global features are derived from $\mathbf{E}^4$ through spatial and channel attention, yielding $\mathbf{G}$, which captures semantic information across the entire image. Each local feature block is weighted based on its similarity to the recalibrated global feature $\mathbf{G}$, and the weighted local feature blocks $\mathbf{L_k}$ are projected to aggregate local information. The reshaped local feature $\mathbf{L}$ and global feature $\mathbf{G}$ serve as the final representations, which are fed into the motion estimator.

Meanwhile, the global feature $\mathbf{E}^4 \in \mathbb{R}^{512\times H/64\times W/64}$, generated by passing B-mode frames through deeper encoding layers, spatially condenses information to store comprehensive global semantic features. Thus, it effectively captures the global context across the full image. The channel attention scores for the global feature are obtained as:

$$\mathbf{S}_\mathbf{c}^\mathbf{G} = \sigma(\mathbf{W}_{\mathbf{c}2}^\mathbf{G}\mathbf{W}_{\mathbf{c}1}^\mathbf{G}\mathbf{P}_\mathbf{c}^\mathbf{G}) \in \mathbb{R}^{512}, \tag{6}$$

where $\mathbf{W}_{\mathbf{c}2}^\mathbf{G} \in \mathbb{R}^{512\times 32}$ and $\mathbf{W}_{\mathbf{c}1}^\mathbf{G} \in \mathbb{R}^{32\times 512}$ are the weights of an MLP, and $\mathbf{P}_\mathbf{c}^\mathbf{G} = \Gamma(\mathbf{E}^4) \in \mathbb{R}^{512}$ is a pooled feature derived from $\mathbf{E}^4$. Next, the spatial attention scores for the global feature are obtained as:

$$\mathbf{S}_\mathbf{s}^\mathbf{G} = \sigma(\mathbf{W}_\mathbf{s}^\mathbf{G} * \mathbf{P}_\mathbf{s}^\mathbf{G}) \in \mathbb{R}^{H/64\times W/64}, \tag{7}$$

where $\mathbf{W}_\mathbf{s}^\mathbf{G} \in \mathbb{R}^{2\times 1\times 1}$ is the kernel of a 1×1 convolutional layer, and $\mathbf{P}_\mathbf{s}^\mathbf{G} \in \mathbb{R}^{2\times H/64\times W/64}$ is the concatenation of max-pooled and average-pooled features. Here, $*$ denotes the

convolution operation. Both spatial and channel attention are then applied parallelly as:

$$\mathbf{G} = \mathbf{S}_{\mathbf{s}}^{\mathbf{G}} \otimes \mathbf{S}_{\mathbf{c}}^{\mathbf{G}} \otimes \mathbf{E}^4 \in \mathbb{R}^{512\times \mathrm{H}/64\times \mathrm{W}/64}, \tag{8}$$

where $\mathbf{G}$ represents the recalibrated global feature.

To prepare for the global-local attention operation, the global feature $\mathbf{G}$ is projected into $\tilde{\mathbf{G}} \in \mathbb{R}^{128\times \mathrm{H}/64\times \mathrm{W}/64}$ using a 1×1 convolution to match the channel dimensions. Each recalibrated local feature block $\mathbf{R}_k$ is then weighted based on its similarity to the global feature using cosine similarity:

$$\mathbf{L}_k = \Phi(\mathbf{R}_k, \tilde{\mathbf{G}}) \otimes \mathbf{R}_k, \tag{9}$$

where $\Phi(\cdot,\cdot) \in \mathbb{R}$ is the cosine similarity function, emphasizing regions critical for motion estimation.

The weighted local feature blocks $\{\mathbf{L}_k\}_{k=1}^{256}$ are projected to $\tilde{\mathbf{L}} \in \mathbb{R}^{128\times \mathrm{H}/64\times \mathrm{W}/64\times 4}$ to aggregate the local information. Finally, the reshaped local feature $\mathbf{L} \in \mathbb{R}^{512\times \mathrm{H}/64\times \mathrm{W}/64}$ is obtained from $\tilde{\mathbf{L}}$. The outputs of the global-local attention ($\mathbf{L}$ and $\mathbf{G}$) are then fed into the motion estimators.

E. Loss Functions

1) Motion-Based Mean Absolute Error (MMAE): There is a class imbalance among the elements of motion, as cases with high amplitudes (fast motion) are significantly less frequent. Moreover, estimating fast motion poses a greater challenge than estimating slow motion, primarily because the correlation between adjacent frames diminishes. To address this issue, we propose the motion-based mean absolute error (MMAE), a type of weighted metric:

$$L_{\mathrm{MMAE}} = \frac{1}{6(s+1)} \sum_{i=n}^{n+s} \sum_{k=1}^{6} \mathbf{w}_i \left| \Delta\boldsymbol{\theta}_i^k - \Delta\hat{\boldsymbol{\theta}}_i^k \right|, \tag{10}$$

where $s+1$ is the length of the B-mode sequence, and $\mathbf{w}_i = |\Delta\boldsymbol{\theta}_i| + \varepsilon$ is a weighting vector defined by the motion vector, with $|\Delta\boldsymbol{\theta}_i|$ normalized to the range [0, 1]. The error for fast motions is weighted more heavily than for slow motions, as $\mathbf{w}_i$ increases for fast motions. Here, ε is a smoothing factor used to reduce the effect of the weighting. We empirically set $\varepsilon = 2$, so the fastest motion receives approximately 1.5 times more weight than the slowest motion. This setting was found to enhance accuracy in high-motion regions while preserving overall stability.

2) Correlation Loss: To assist in model supervision, we adopted correlation loss [10], which is defined as:

$$L_{\mathrm{Corr}} = \frac{1}{6} \sum_{k=1}^{6} \left(1 - \Phi\left(\left\{ \Delta\boldsymbol{\theta}_i^k \right\}_{i=n}^{n+s}, \left\{ \Delta\hat{\boldsymbol{\theta}}_i^k \right\}_{i=n}^{n+s} \right) \right), \tag{11}$$

where $\Phi(\cdot,\cdot)$ denotes the cosine similarity function.

The correlation loss can measure the error without being affected by scale. Furthermore, it imposes a stronger penalty for incorrectly estimated directions, thereby enhancing the model's stability in capturing accurate motion.

3) Margin Triplet Loss: Although each scan can have different appearances in B-mode images due to varying scan protocols or anatomical variations, similar motions can still be found among scan frames. To further assist the model convergence, we employed contrastive learning with a margin triplet loss [33]:

$$L_{\mathrm{Triplet}} = \max(0, \mathrm{dist}(\mathbf{F}_a, \mathbf{F}_p) - \mathrm{dist}(\mathbf{F}_a, \mathbf{F}_n) + M), \tag{12}$$

where $\mathbf{F}_a$ is an anchor feature map serving as the center criterion. To determine the positive and negative feature maps, $\mathbf{F}_p$ and $\mathbf{F}_n$, which correspond to the near and far samples, the cosine similarity between labels was used.

The triplet loss L_{Triplet} encourages the model to contrast the feature maps in latent space, facilitating their convergence and improving robustness. The margin value was empirically set to $M = 0.1$ based on validation performance, as it effectively balanced the separation between similar and dissimilar motions without hindering overall convergence.

4) Final Loss Function: The final loss function is a linear combination of the individual loss components:

$$L_{\mathrm{Final}} = \alpha_1 L_{\mathrm{MMAE}} + \alpha_2 L_{\mathrm{Corr}} + \alpha_3 L_{\mathrm{Triplet}}, \tag{13}$$

where α_1, α_2, and α_3 are positive real values used to balance the magnitude of each loss function. The weighting coefficients were empirically determined as $\alpha_1 = 1$, $\alpha_2 = 1$, and $\alpha_3 = 0.005$. We chose α_2 as the reference value since L_{Corr} is inherently bounded within [0, 1]. Given that L_{MMAE} shares a comparable scale after label normalization, α_1 was set near 1 to maintain parity. The triplet loss weight α_3 was carefully tuned to a small value to prevent the contrastive term from dominating the total loss. Its final value was selected based on empirical performance, yielding optimal results during validation.

IV. Experimental Results

A. Materials

1) Dataset 1 (In-House): We used a programmable US machine (Verasonics, Vantage System 64LE) equipped with a 1D array transducer (L11-5v) to acquire US data, and the scan motion was tracked using a mounted electromagnetic sensor (Polhemus) at an acquisition rate of 60 Hz (Fig. 4). The B-mode images were reconstructed using 31 different angle plane waves to enhance their quality and had dimensions of 256×256 pixels, with a pixel resolution of 0.1484 × 0.1484 mm. The image frame rate was set at 20 Hz. We collected 54 scans, primarily from the forearms of 9 individuals, and randomly divided them into 30, 6, and 18 scans for the training, validation, and testing sets, respectively, ensuring that the subjects were distinct across sets. Each scan consisted of 900-1000 frames, capturing the entire length of the forearm through an arbitrary S-shaped trajectory. This scanning procedure was conducted three times for each forearm. The local Institutional Review Board (Pusan National Univ. IRB, 2023_74_HR) granted approval for the acquisition and use of the data.

To demonstrate the versatility of our proposed method, we also acquired a few additional test samples using the

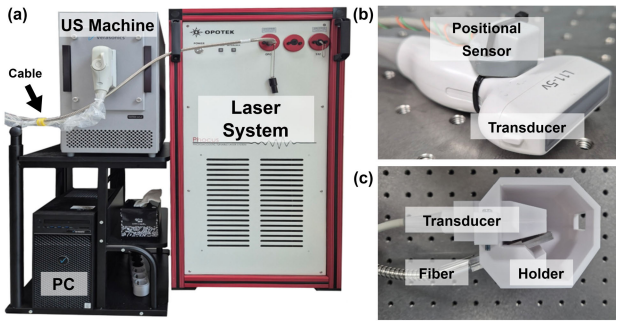

Fig. 4. Experimental setup for PAUS data acquisition and visualized results. (a) US machine connected to the laser system. (b) and (c) Transducer setups.

same ultrasound system for the purpose of 3D vascular reconstruction. In these acquisitions, each B-mode frame, generated using 31 plane waves, was immediately followed by a PD acquisition comprising 20 plane waves transmitted at a pulse repetition frequency (PRF) of 3,000 kHz. These data samples were used exclusively for inference (testing) and were not included in the training or validation.

2) Dataset 2 (In-House): To further demonstrate the versatility of our proposed method, we additionally acquired PA data to construct 3D vascular structures using the same programmable ultrasound system described in Dataset 1. As shown in Fig. 4 (a), a Q-switched Nd:YAG laser system (OPOTEK Phocus Mobile) was integrated with the US machine to enable real-time PAUS imaging. Specifically, the US system received a flash lamp trigger signal from the laser system, waited for a specified delay time (optical build-up time for the laser), and then commenced data acquisition precisely when the laser fired. The laser was delivered via fibers arranged on one side of the linear array US transducer (Fig. 4 (c)). Subsequent to the 31 plane waves used for a B-mode image frame, a laser with a wavelength of 750 nm was employed to insonify for a PA image frame. PA images were reconstructed using a standard delay-and-sum (DAS) method.

Although the laser operated at 20 Hz, the ultrasound system was constrained to receive data from only 64 half-channels at a time. As two laser firings were required to generate a single PA frame, the resulting PAUS acquisition rate was 10 Hz. Furthermore, to achieve sufficient out-of-plane resolution for microvessel imaging, the scan speed was slower than that used in Dataset 1. To improve image contrast for microvessels located near the skin, a small coupling gap was maintained between the transducer and the skin surface. This configuration enhanced optical fluence at the target region but also introduced a non-tissue region at the top of the image. As a result, the region of interest (ROI) was defined as a subregion taken from the bottom portion of each image to exclude this coupling space. The dataset was randomly divided into training, validation, and testing subsets with a ratio of 5:1:4.

3) Dataset 3 (Public): Our dataset may contain inherent biases that could positively influence our model's performance. To validate the model's performance in different environments and mitigate the effects of such biases, we utilized a publicly available dataset [15]. This dataset comprises transverse and longitudinal scans of the forearm with diverse-shaped trajectories, collected from 19 subjects, resulting in a total of 19 × 12 scans. For each subject, both forearms were scanned using three distinct trajectory shapes (S, C, and L) in two orientations: parallel and perpendicular to the forearm, yielding 12 scans per subject.

TABLE I
SUMMARY OF THE DATASETS USED IN THIS STUDY

Attribute	Dataset 1 (in-house)	Dataset 2 (in-house)	Dataset 3 (public)
Transducer Type	Linear	Linear+Holder	Convex
# of Subjects	9	10	19
# of Scans	54	60	228
# of Frames	51,300	28,445	30,552
FPS (Hz)	20	10	20
Travel Length (mm)	199.27±23.81	153.12±13.20	144.51±14.26
Scan Speed (mm/s)	8.4704±4.20	2.9986±1.46	23.8798±12.89
Image Size	256×256	256×256	480×640
Split Ratio	5:1:3	5:1:4	10:4:5

The ultrasound images were acquired using an Ultrasonix machine (BK, Europe) operating at 20 Hz with a convex transducer (4DC7-3/40). The probe motion was tracked by an NDI Polaris Vicra (Northern Digital Inc., Canada). The B-mode images were processed with a moderate level of speckle reduction. Each scan consists of 36 to 430 frames with a resolution of 480 × 640 pixels, and the probe travel distance ranges approximately between 100 and 200 mm. The dataset was randomly divided into training, validation, and testing subsets based on subjects, with a ratio of 10:4:5.

Since images captured using the convex probe have a sector shape with background regions outside the sector, we extracted a square RoI (320 × 320 pixels) from the central foreground region to focus on relevant areas. To handle the increased input size, the batch size was reduced to 12.

A detailed summary of all datasets used in this study is provided in Table. I

B. Implementation Details

For the comparative experiments, we selected models currently recongnized for motion estimation, including CNN (Prevost et al.) [9], DC^2-Net (Guo et al.) [19], Ning et al. [12], Dou et al. [26], Yan et al. [24], Li et al. [15], and Luo et al. [14]. We used publicly available code when provided by the original authors. For models without released implementations, we re-implemented the architectures based on detailed descriptions in the corresponding papers. Minor modifications were made, primarily to match the input tensor sizes used in our datasets, including adjustments to the dimensions of fully connected layers when required. During the training phase, sequences of length s were sampled from each subject in every epoch. This method generally requires a relatively large number of epochs (20,000) to ensure model convergence.

The Adam optimizer was employed with an initial learning rate of 1e-5, which was reduced by a factor of 0.8 every 100 epochs to facilitate the convergence process. Training was performed on an NVIDIA RTX 4090 GPU (24 GB) with a batch size of 14. Under these settings,

TABLE II
MODEL PERFORMANCE - DATASET 1

Models	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
	Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
CNN [9]	0.2435±0.108	0.1458±0.118	19.5551±12.21	16.3043±13.25	0.5045±0.074	40.5307±14.06	0.5282±0.232	32.7253±15.21
DCL [10]	0.1715±0.074	0.0830±0.039	19.3659±11.06	9.8433±7.61	0.3463±0.060	39.4931±13.18	0.8464±0.103	29.5583±9.75
DC^2 [19]	0.1783±0.075	0.0938±0.056	13.9982±10.55	8.4472±6.34	0.3584±0.070	29.8875±13.98	0.8775±0.098	21.3799±11.15
Dou et al. [26]	0.1794±0.076	0.1002±0.060	16.6838±12.17	8.4188±4.77	0.3641±0.055	34.1346±15.24	0.7688±0.127	29.3495±12.20
Ning et al. [12]	0.2434±0.113	0.1486±0.126	13.5194±7.29	13.2908±7.85	0.5052±0.076	29.4476±7.95	0.7288±0.113	17.0679±10.18
Yan et al. [24]	0.1724±0.078	0.0932±0.052	13.2459±9.32	9.7304±7.09	0.3581±0.058	27.9322±12.48	0.7545±0.162	21.6784±11.06
Li et al. [15]	0.1587±0.070	0.0768±0.029	13.0227±6.81	7.9726±4.05	0.3214±0.062	25.4893±9.76	0.7999±0.097	21.8983±8.86
Luo et al. [14]	0.1584±0.072	0.0734±0.027	10.6084±7.29	7.0597±5.29	0.3212±0.059	21.9479±9.13	0.8826±0.069	15.5101±7.18
MoGLo	**0.1430**±0.072	**0.0663**±0.023	**7.9097**±4.89	**5.9412**±3.83	**0.2917**±0.054	**16.1944**±6.09	**0.9217**±0.048	**12.0577**±6.45

the proposed MoGlo-Net required 4.7 hours to train and achieved a processing speed of 95 frames per second (FPS) with a batch size of 32 during the inference phase. In terms of architectural complexity, MoGLo-Net consists of approximately 27.91 million parameters and requires 0.0774 TFLOPs per forward pass (input size: 1 × S=5 × C=1 × H=256 × W=256). Additionally, reconstructing a 3D US volume consisting of 950 frames took 1.8076 seconds.

All experiments were conducted under identical conditions, including sequence length, dataset splitting, and training parameters, to ensure a fair comparison among the models.

C. Evaluation Metrics

For the 6-dimensional output, we utilized the relative Average Error (**rAE**; mm, °) and accumulated Average Error (**aAE**; mm, °) to evaluate $\Delta\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_i$. The rAE measures the relative motion estimation accuracy, while the aAE quantifies the accumulated motion error over the scan trajectory. For the reconstructed 3D ultrasound volumes, we measured the Euclidean distance between the grid points of the true and predicted frames. The relative Frame Error (**rFE**; mm) reflects the relative displacements between frames, whereas the accumulated Frame Error (**aFE**; mm) accounts for the cumulative displacements along the scan path [19]. Additionally, we used Correlation (**Corr**), defined as cosine similarity, to measure the underlying trends of reconstructed trajectories in 3D Euclidean space. The Final Drift (**FD**; mm) [9], which represents the aFE of the final frame, increases proportionally with the scan length. To normalize FD, we employed the Final Drift Rate (**FDR**; %) [11], calculated by dividing FD by the total length of the scan trajectory. FDR serves as a subsidiary metric, as FD typically exhibits relatively large deviations in S-shaped scans.

D. Results Using Dataset 1

1) Quantitative Results: As shown in Table II, DCL [10] improves upon the baseline CNN [9], with additional performance gains achieved through the incorporation of contrastive learning in DC^2-Net [19]. The correlation-based approach proposed by Dou et al. [26] and the transformer-based method introduced by Ning et al. [12] yield moderate improvements by enhancing inter-frame motion representation. The Mamba-based architecture presented by Yan et al. [24] captures fine-grained temporal dependencies and achieves competitive results in several metrics. The regression-based method proposed by Li et al. [15] demonstrates improved accuracy by leveraging sequential context, while the method described by Luo et al. [14] exhibits robust performance by incorporating motion-based constraints and contextual consistency. Our proposed method, MoGLo-Net, achieves the highest overall performance across all evaluation metrics. These results highlight the effectiveness of our attention-enhanced architecture, correlation-guided motion modeling, and supervision strategy in delivering both accurate motion estimation and high-fidelity 3D ultrasound volume reconstruction.

2) Qualitative Results: Consequently, the final 3D US volume can be reconstructed using the estimated motion and B-mode images. As illustrated in Fig. 5, CNN exhibits pronounced distortion, while DC^2 and Li et al. show noticeable improvements over CNN. However, both models display considerable FD due to their limited generalization capabilities. Among the comparison groups, Luo et al. demonstrate the most reliable results in terms of both trajectory similarity and FD. Models that utilize correlation volume and global-local attention tend to further improve the quality of 3D US volumes, with the proposed MoGLo-Net showcasing the most accurate results, closely resembling the ground truth.

3) Ablation Study 1 (Model Components): We created a minor version of MoGLo-Net by removing the global-local attention module (G), correlation operation (C), or motion-based MAE (M) to further validate the contribution of each component in MoGLo-Net. As shown in Table III, the global-local attention significantly improved overall metrics, particularly in accumulated errors, as it effectively highlights critical regions for motion estimation. Employing the correlation operation resulted in partial metric improvements, as it provides comprehensive but unrefined correlation information between frames. Finally, the inclusion of MMAE led to improvements, especially in accumulated errors, by emphasizing fast motion, which is sparsely distributed. By leveraging all components, MoGLo-Net exhibits significantly less distortion compared to its minor version.

4) Ablation Study 2 (Input Features): We also examined the effect of input data type by leveraging the programmable ultrasound machine. We designed three settings, which are summarized in Table IV: 1) Log: inputting raw B-mode images, referring to the images before log-compression for contrast enhancement; 2) IQ: inputting in-phase and

TABLE III
ABLATION STUDY 1 - MODEL COMPONENTS

G	C	M	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
			Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
×	×	×	0.1539±0.068	0.0746±0.028	9.6156±4.99	7.6975±5.75	0.3127±0.053	20.1755±5.37	0.8723±0.050	17.4204±6.22
✓	×	×	0.1479±0.070	0.0704±0.027	9.0556±6.32	7.8395±4.69	0.3002±0.053	18.8602±7.46	0.9008±0.066	14.6913±5.85
×	✓	×	0.1503±0.067	0.0716±0.026	9.6346±6.15	7.3151±4.99	0.3074±0.048	20.3687±7.90	0.8806±0.087	14.8185±7.05
×	×	✓	0.1512±0.068	0.0735±0.029	9.9514±4.92	6.6290±5.01	0.3081±0.052	19.8897±6.55	0.8900±0.063	15.7800±7.08
✓	✓	×	0.1488±0.072	0.0700±0.026	8.1816±4.94	6.2784±3.98	0.3018±0.053	16.7405±6.12	0.9060±0.067	14.1837±5.66
✓	×	✓	0.1475±0.067	0.0691±0.025	8.1578±5.70	**5.9117**±3.99	0.2981±0.053	17.0640±6.56	0.9002±0.078	13.5590±6.53
×	✓	✓	0.1486±0.067	0.0688±0.025	9.4741±5.76	6.6669±3.94	0.3019±0.049	19.1706±7.20	0.8998±0.064	15.1482±7.68
✓	✓	✓	**0.1430**±0.072	**0.0663**±0.023	**7.9097**±4.89	5.9412±3.83	**0.2917**±0.054	**16.1944**±6.09	**0.9217**±0.048	**12.0577**±6.45

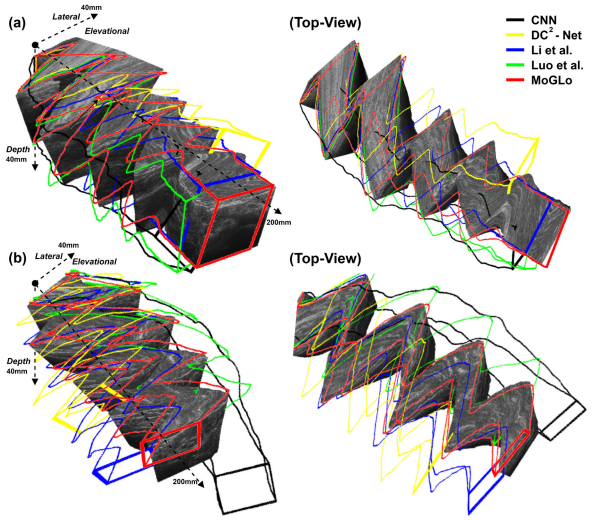

Fig. 5. Two 3D reconstruction cases using publicly open dataset. The 3D ground-truth image is constructed by stacking 2D B-mode images using ground-truth positions. The outlined 3D figures in different colors (unfilled) are constructed using estimated positions from various deep learning models to compare their trajectories with the ground truth.

quadrature (IQ) data as well as B-mode images. To utilize the IQ data, we applied log compression to its amplitude and concatenated it with the B-mode; and 3) SP: inputting B-mode images after applying noise reduction to mitigate speckle noise using the Lee Filter [34].

As shown in Table IV, when raw B-mode images were used, overall performance decreased due to poor contrast in dark regions, which occupy the majority of the image. Adding IQ data did not yield improvement, as it increased the complexity of the input features without providing additional cues for motion estimation. Lastly, using denoised B-mode images reduced overall performance, as speckle patterns are strongly correlated with motion vectors.

5) Ablation Study 3 (Sequence Length): Lastly, we conducted an ablation study on the effect of sequence length ($s + 1$, as shown in Fig. 1) by varying it from 2 to 50. The upper bound of 50 was determined by the memory limitations of our GPU. As presented in Table V, both extremely short sequences (e.g., 2) and overly long sequences led to degraded performance. In particular, the case with sequence length 2 showed a significant drop in accuracy, indicating that insufficient temporal context negatively impacts motion estimation. On the other hand, increasing the sequence length beyond a certain point did not yield further improvements, but it incurred additional computational cost. We attribute this saturation to the limited shared physical content among distant frames, which provides little additional information for estimating motion. Although this observation contrasts with the findings in [15], similar performance trends have been reported in [10], supporting the idea that an optimal range of sequence length exists for motion modeling.

6) US 3D Vascular Imaging: The benefit of motion tracking extends clinical applications beyond the compilation of B-mode images. In this study, we adapted the acquisition sequence in the US system to obtain not only B-mode images but also PD-mode images, which are specialized for visualizing vessels. The B-mode images were utilized for motion estimation through MoGLo-Net, and both the motion estimates and PD images were employed to construct 3D vessels. For PD images, singular value thresholding (SVT) filtering [35] was applied to suppress clutter and noise signals, and top-hat filtering [36] was used to enhance vessel contrast. The overall quality of the reconstructed volumes can be indirectly evaluated from Table II, as the B-mode data were acquired under the same conditions and protocols. Due to the limited lateral field of view (38 mm) of the transducer, forearm scans were performed with a relatively narrow lateral range to capture the full radial artery along the elevational axis. Fig. 6 illustrates examples of reconstructed 3D vessels (radial artery) in a forearm. Notably, each vessel appears almost straight and closely resembles the natural anatomical shape, even though the scan motion trajectory was wavy.

E. Results Using Dataset 2

Furthermore, we set up the PAUS system to acquire both B-mode and PA images, enabling the reconstruction of finer vessels (in a forearm) with high contrast. Similar to the PD scan setting, B-mode images are used to estimate the scan motion, while PA images are utilized to reconstruct vascular structures based on the estimated trajectories. Fig. 7 illustrates the PA imaging results using maximum amplitude projection (mAP) according to the depth.

As described in Section IV-A and summarized in Table I, although the acquisition system remained the same, the imaging conditions for PAUS data (Dataset 2) differ significantly from those of Dataset 1. As a result, applying a model trained solely on Dataset 1 to the PAUS data led to noticeably

TABLE IV
ABLATION STUDY 2 - INPUT FEATURES

Models	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
	Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
Log	0.1527±0.074	0.0726±0.029	12.5220±8.45	6.9884±5.01	0.3105±0.056	25.5322±10.56	0.8790±0.087	18.9047±9.16
IQ	0.1549±0.071	0.0681±0.020	9.1483±5.98	**5.4822**±4.14	0.3157±0.056	18.8984±7.05	0.8925±0.077	13.5999±4.75
SP	0.1474±0.067	0.0701±0.026	9.1138±5.99	6.5060±3.47	0.2988±0.051	18.9144±7.79	0.9021±0.063	14.8614±7.53
MoGLo	**0.1430**±0.072	**0.0663**±0.023	**7.9097**±4.89	5.9412±3.83	**0.2917**±0.054	**16.1944**±6.09	**0.9217**±0.048	**12.0577**±6.45

TABLE V
ABLATION STUDY 3 - SEQUENCE LENGTH

Models	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
	Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
seq=2	0.1484±0.073	0.0707±0.025	10.9096±7.14	7.1582±4.94	0.3021±0.054	22.6424±6.33	0.8653±0.110	17.3014±5.80
seq=5 (**MoGLo**)	**0.1430**±0.070	**0.0663**±0.023	**7.9097**±4.88	5.9412±3.83	**0.2917**±0.054	**16.1944**±6.09	**0.9217**±0.048	**12.0577**±6.45
seq=10	0.1512±0.073	0.0690±0.025	8.7425±6.39	6.7344±5.41	0.3073±0.057	18.1948±6.85	0.9062±0.054	12.5697±5.53
seq=15	0.1494±0.071	0.0694±0.025	8.1883±5.33	6.6791±4.99	0.3035±0.056	17.3325±6.64	0.9123±0.053	14.2481±6.91
seq=20	0.1516±0.075	0.0693±0.026	8.8489±5.99	6.2355±4.50	0.3082±0.058	18.6632±6.23	0.9078±0.063	14.0908±6.22
seq=25	0.1483±0.072	0.0691±0.025	8.7174±5.99	**5.9275**±3.89	0.3016±0.056	18.2472±6.92	0.9043±0.067	13.4284±5.79
seq=30	0.1633±0.075	0.0773±0.031	17.7583±13.17	7.6148±6.54	0.3310±0.052	36.8998±13.53	0.8523±0.071	26.8630±9.01
seq=35	0.1522±0.070	0.0722±0.029	12.2959±7.83	6.3925±4.12	0.3082±0.056	25.3265±8.01	0.8863±0.082	19.3284±7.01
seq=40	0.1540±0.071	0.0863±0.040	14.6164±10.68	8.7355±7.45	0.3149±0.052	30.5636±15.55	0.8379±0.084	27.1425±14.03
seq=45	0.1532±0.073	0.0719±0.027	17.2188±11.53	7.9110±5.65	0.3120±0.051	35.4784±13.03	0.8500±0.084	25.7042±9.80
seq=50	0.1499±0.071	0.0710±0.026	14.6400±10.29	6.7093±5.27	0.3042±0.055	28.9903±9.97	0.8491±0.082	22.1961±6.13

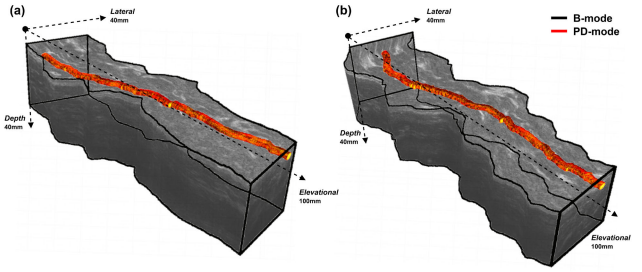

Fig. 6. Two 3D vascular (radial artery in a forearm) reconstruction cases using US PD acquisitions. Each 3D vessel is visualized by superimposing B-mode imaging to verify the scan motion trajectory.

degraded performance. To address this, we conducted transfer learning by fine-tuning the pre-trained model on a subset of the PAUS dataset to improve performance and demonstrate the generalization (cross-domain adaptability) capability of our framework. Additionally, we evaluated a fully supervised scenario, where the model was trained from scratch using the PAUS dataset only.

The results are summarized in Table VI. Here, 1→2 (Trf) refers to the model pre-trained on Dataset 1 and directly tested on Dataset 2 without fine-tuning. 1→2 (Frz) indicates a fine-tuned model in which only the estimator is updated while the encoder remains frozen. 1→2 (All) represents full fine-tuning on Dataset 2 using a model pre-trained on Dataset 1. Both the Frz and All scenarios used only 10% of the full training schedule (2,000 epochs), highlighting the efficiency of transfer learning. Lastly, 2→2 (Scratch) refers to a model trained entirely from scratch using Dataset 2.

As expected, the Trf model showed significantly lower performance compared to other settings. However, both fine-tuning approaches (Frz and All) achieved substantial improvements, even with limited training. Notably, the All setting produced results comparable to the fully supervised Scratch scenario, demonstrating the strong adaptability of our pre-trained model. Furthermore, the Frz setting, which only updates the estimator, achieved competitive results while offering faster training and reduced computational cost.

F. Results Using Dataset 3

Dataset 3 is an open-source dataset acquired using a system and scanning protocol entirely different from those used in our in-house Dataset 1 (see Table I). For this experiment, our model and all comparative models were trained from scratch using a portion of Dataset 3 for training and evaluated on a separate, independent testing subset.

As shown in Table VII, although MoGLo-Net does not achieve the best score on every individual metric, the differences in angular errors are marginal and unlikely to affect practical outcomes. More importantly, MoGLo-Net achieves the best performance on clinically meaningful metrics such as Frame Error (FE), Correlation (Corr), and Final Drift Rate (FDR), which are directly related to the quality of 3D volume reconstruction. The consistent ranking of models across both the in-house and open datasets further supports the reliability of the experimental findings, indicating that the observed trends are not dataset-specific but rather reflect the inherent characteristics of each model. Fig. 8 illustrates the 3D trajectories obtained from ground-truth and estimated positions.

To further evaluate the generalizability of the proposed model, we conducted cross-domain experiments between Dataset 1 and Dataset 3. The results are presented in Table VIII. Specifically, the format $a \rightarrow b$ denotes that the model was trained on Dataset a and tested on Dataset b. For instance, 1→3 (Trf) and 3→1 (Trf) refer to models pre-trained on Dataset 1 and Dataset 3, respectively, and directly evaluated

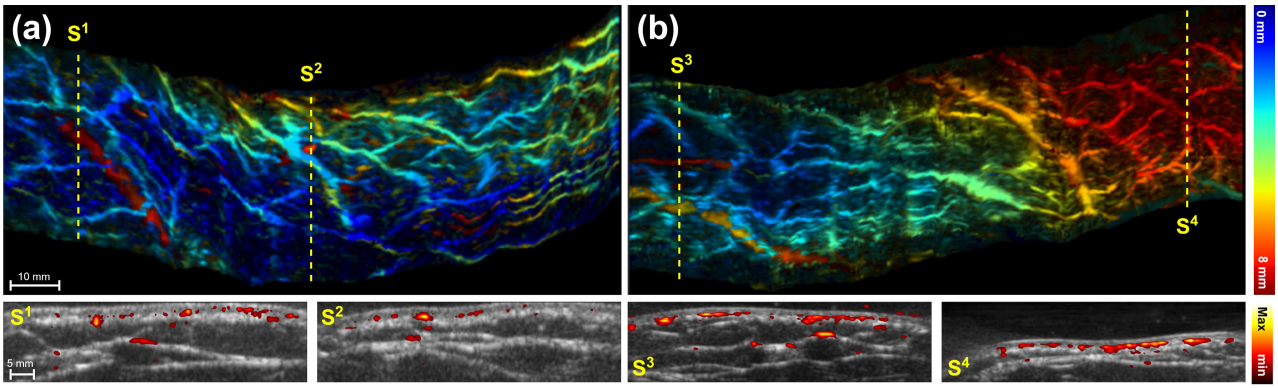

Fig. 7. Two 3D vascular reconstruction cases using PAUS acquisitions. Each case is visualized using maximum amplitude projection (mAP) according to the depth, alongside corresponding cross-sectional B-mode images.

TABLE VI
MODEL PERFORMANCE - DATASET 2

Models	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
	Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
1 → 2 (Trf)	0.1120±0.061	0.0374±0.011	19.0656±14.59	6.8638±4.49	0.2199±0.074	38.5333±11.23	0.6432±0.151	50.7000±12.64
1 → 2 (Frz)	0.0608±0.029	**0.0216**±0.008	5.4837±4.53	2.3272±1.60	0.1255±0.044	11.5556±4.75	0.9122±0.055	13.5616±4.36
1 → 2 (All)	0.0600±0.026	0.0220±0.008	4.6134±3.15	**2.1876**±1.75	0.1229±0.042	9.2950±3.75	0.8964±0.057	11.2285±3.66
2 → 2 (Scratch)	**0.0597**±0.026	0.0221±0.008	**4.3204**±3.31	2.2230±1.88	**0.1225**±0.042	**8.9360**±3.99	**0.9165**±0.059	**10.6663**±3.87

TABLE VII
MODEL PERFORMANCE - DATASET 3

Models	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
	Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
CNN [9]	0.2801±0.155	0.0782±0.026	14.2539±10.51	2.5688±1.72	0.5820±0.173	28.6686±12.52	0.8132±0.189	34.1899±17.46
DCL [10]	0.2472±0.164	0.0752±0.027	12.9903±10.19	2.5297±1.60	0.5225±0.201	26.9548±10.99	0.8376±0.171	33.2981±15.10
DC^2 [19]	0.2405±0.158	0.0745±0.027	12.7417±10.08	2.4095±1.56	0.5084±0.193	26.6126±10.67	0.8389±0.181	33.0940±14.30
Dou et al. [26]	0.2403±0.137	0.0804±0.026	10.8019±7.98	2.5380±1.66	0.5005±0.164	21.9254±9.31	0.8650±0.160	28.2128±13.24
Ning et al. [12]	0.2110±0.132	0.0758±0.026	8.9332±6.48	2.3811±1.63	0.4389±0.167	18.1773±7.16	0.9052±0.123	22.5305±9.88
Yan et al. [24]	0.2294±0.145	**0.0733**±0.029	10.9620±8.35	**2.3309**±1.53	0.4795±0.188	22.2426±10.08	0.8609±0.159	25.3002±12.03
Li et al. [15]	0.2138±0.116	0.0784±0.028	8.7304±5.91	2.8622±1.90	0.4414±0.152	17.6650±6.86	0.8900±0.124	22.0754±10.73
Luo et al. [14]	0.2105±0.131	0.0762±0.026	8.8976±6.56	2.4126±1.69	0.4372±0.166	18.0273±7.72	0.8945±0.121	22.0038±9.46
MoGLo	**0.1979**±0.124	0.0755±0.026	**8.0867**±6.12	2.3956±1.57	**0.4113**±0.159	**16.4062**±7.33	**0.9164**±0.105	**19.8096**±10.30

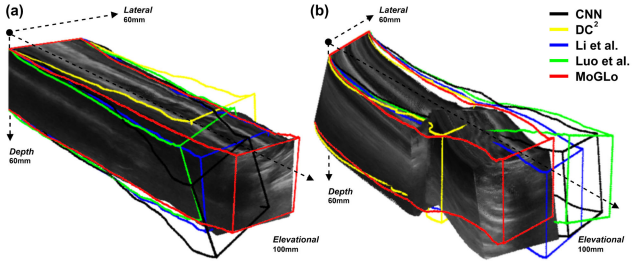

Fig. 8. Two 3D reconstruction cases using publicly open dataset. The 3D ground-truth image is constructed by stacking 2D B-mode images using ground-truth positions. The outlined 3D figures in different colors (unfilled) are constructed using estimated positions from various deep learning models to compare their trajectories with the ground truth.

on the other dataset without any fine-tuning. 1→3 (Frz) and 3→1 (Frz) indicate that only the estimator module was fine-tuned on the target dataset, while the encoder was kept frozen. In contrast, 1→3 (All) and 3→1 (All) represent full fine-tuning of the entire model on the target dataset, updating both encoder and estimator components. For both Frz and All configurations, we used only 10% of the full training schedule (2,000 epochs). Finally, 1→1 (Scratch) and 3→3 (Scratch) refer to models trained and evaluated solely on Dataset 1 and Dataset 3, respectively, without any transfer from other domains.

In both transfer directions (1→3 and 3→1), models evaluated without any adaptation (Trf) showed a marked drop in performance. This decline is primarily due to domain discrepancies such as variations in probe hardware, anatomical features, and scanning conditions. In contrast, fine-tuning approaches yielded notable performance recovery, even with just 10% of the total training epochs. Among these, the All configuration achieved performance levels comparable to those of models trained from scratch on the target dataset, demonstrating strong generalization capability and transferability of the learned representations. Meanwhile, the Frz configuration, which updates only the estimator while keeping the encoder fixed, also delivered competitive results with significantly lower training time and computational burden.

TABLE VIII
CROSS-DOMAIN VALIDATION - DATASET 1 ↔ DATASET 3

Models	rAE		aAE		rFE (mm)	aFE (mm)	Corr	FDR (%)
	Shift (mm)	Angle (°)	Shift (mm)	Angle (°)				
1 → 3 (Trf)	0.4216±0.323	0.0985±0.032	31.1916±27.67	5.4970±3.16	0.9217±0.270	72.8266±10.55	0.6577±0.147	95.0757±6.88
1 → 3 (Frz)	0.2189±0.136	**0.0716**±0.028	10.3878±8.17	**2.2630**±1.47	0.4585±0.169	21.5564±8.57	0.8233±0.199	24.1403±10.43
1 → 3 (All)	0.2032±0.128	0.0720±0.027	9.5677±7.04	2.5524±1.73	0.4240±0.154	19.5941±7.45	0.8885±0.133	23.1133±9.48
3 → 3 (Scratch)	**0.1979**±0.124	0.0755±0.026	**8.0867**±6.12	2.3956±1.57	**0.4113**±0.159	**16.4062**±7.33	**0.9164**±0.105	**19.8096**±10.30
3 → 1 (Trf)	0.2304±0.069	0.1650±0.120	36.8319±21.93	50.9955±20.04	0.4822±0.071	83.8407±9.187	0.3192±0.129	75.8134±7.7
3 → 1 (Frz)	0.1708±0.076	0.0809±0.037	12.6097±7.92	5.3763±2.72	0.3501±0.058	25.1374±9.678	0.8461±0.087	18.7620±7.71
3 → 1 (All)	0.1565±0.073	0.0705±0.025	9.6867±6.01	7.2888±5.74	0.3198±0.050	19.7797±7.372	0.8996±0.070	14.8677±7.10
1 → 1 (Scratch)	**0.1430**±0.070	**0.0663**±0.023	**7.9097**±4.88	**5.9412**±3.83	**0.2917**±0.054	**16.1944**±6.093	**0.9217**±0.048	**12.0577**±6.45

V. DISCUSSION AND CONCLUSION

The primary advantage of handheld PAUS systems is the flexibility they offer during scanning. However, this comes at the cost of requiring skilled and experienced operators with a deep understanding of anatomical 3D structures, as standard 1D array transducers provide only 2D images with a restricted FoV. These challenges can be addressed by reconstructing 3D volumes, which allow the visualization of complex 3D structures and provide arbitrary cross-sections of the RoI. Historically, there have been various hardware-based approaches [37], [38] to obtain 3D volumes in both PA and US fields.

The ideal solution, however, lies in software-only approaches, which are more cost-effective. The freehand 3D approach is an extension of the panoramic image mode, which has already been commercialized in many ultrasound products. Nevertheless, one of the key challenges remains addressing elevational motion (out-of-plane motion), where two adjacent frames have relatively low correlation. Some groups have introduced the use of IMU sensors attached to the transducer as minimal hardware support to aid motion estimation in software. However, bulkier transducers hinder the operator's ability to scan, particularly for PAUS systems, which are already large due to the inclusion of laser fibers.

Our proposed model, MoGLo-Net, is optimized for tracking motion without the need for additional sensors. The model employs correlation operations between feature maps from adjacent B-mode images, explicitly utilizing their closeness. Since the correlation tensor captures all correlations between subspaces across sequential images, the model can track not only in-plane motion but also out-of-plane motion.

In additional experiments, we observed that the 3D volume generated from encoded feature maps ($\mathbf{E}_i^1$ and $\mathbf{E}_{i+1}^1$) is more effective than the one generated directly from B-mode images ($\mathbf{B}_i$ and $\mathbf{B}_{i+1}$). Fig. 9 (a) shows the 2D map obtained by averaging the correlation volume (from encoded feature map) over the z-axis in Fig. 2, while Fig. 9 (b) shows the corresponding 2D map obtained in the same manner from B-mode images. The feature correlation exhibits better contrast than B-mode correlation, displaying clear patterns. We expect that the relative spatial variations in the correlation map contribute to in-plane motion estimation, while the overall mean value across the space contributes to out-of-plane motion estimation.

Additionally, the global-local attention module generates local representations of images by learning attention weights that emphasize significant sub-regions for motion features. This allows the model to effectively highlight sub-regions, such as fully-developed speckle areas, which are critical for motion estimation. As illustrated in Fig. 1, the final motion parameters are derived from both the global features—condensed from deeper layers covering broader spatial regions—and the local features—extracted from shallower layers and locally enhanced via attention.

In extra experiments, we visualized the patch-wise attention scores in the module (Fig. 3), as shown in Fig. 10. Under conditions of slow elevational motion, adjacent frames share significant overlapping regions, preserving speckle patterns. As a result, the model focused on uniform areas, such as fully-developed speckle regions (Fig. 10(b)). In contrast, under conditions of fast elevational motion, speckle patterns become disrupted due to the larger gap between adjacent frames. Despite this, high-echogenic tissues often maintain consistency along the elevational axis. In such cases, the model highlighted these regions, as illustrated in Fig. 10(e). This dynamic attention adjustment contrasts with conventional speckle tracking methods, which consistently emphasize high-echogenic boundaries (Fig. 10(c, f)), regardless of motion speed, due to their persistent correlation profiles.

We also observed that the accuracy of motion estimation is tightly coupled to the degree of overlap between adjacent frames. In Dataset 1, we found that when the scan speed exceeded approximately 0.3 mm/frame, the scale of speckle granularity, the model shifted from emphasizing speckle regions to high-echogenic structures. Our frame-subsampling experiments confirmed that estimation errors increased exponentially with reduced temporal resolution. These findings underscore the sensitivity of motion estimation to frame overlap and motivate future simulation-based studies using tools like Field II [39] or k-Wave [40] for further validation under controlled conditions.

Finally, the model utilizes both global and local representations to leverage complementary mutual information, while the LSTM processes features from the B-mode sequences to capture temporal motion patterns. The MoGLo-Net demonstrated consistent improvements across all metrics over the state-of-the-art (SOTA) model [14], producing more realistic 3D US volumes. Moreover, we validated the effect of each

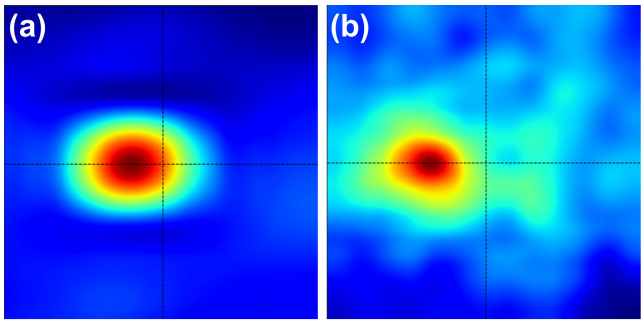

Fig. 9. (a) 2D map obtained by averaging the correlation volume from encoded feature maps over the z-axis in Fig. 2, and (b) corresponding 2D map obtained in the same manner from B-mode images.

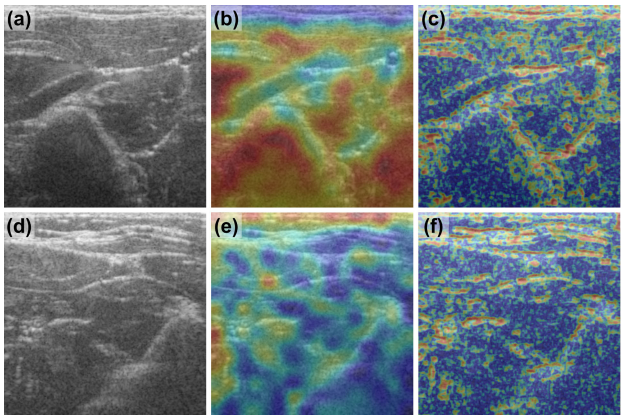

Fig. 10. Qualitative comparison of attention maps under different elevational motion conditions. (a–c) illustrate a slow-motion scenario, while (d–f) depict a fast-motion scenario. (a, d) show the original B-mode ultrasound images. (b, e) present the attention maps generated by the proposed global-local attention module. (c, f) show conventional speckle-based correlation maps.

component of MoGLo-Net through an ablation experiment by omitting key parts, such as the attention module, correlation volume, and motion-based MAE. In addition, we examined the impact of potential information sources, such as IQ data, raw B-mode images, and speckle patterns. As a result, we found that speckle patterns, which are highly related to scan motion, serve as critical cues for scan motion estimation.

While the model architecture draws inspiration from prior work on correlation and attention mechanisms, its originality lies in how these components are restructured and tailored for ultrasound-specific motion estimation. Our use of sparse correlation volumes, dual regression streams, and task-specific attention formulations contributes to a unique and effective solution for sensorless motion estimation in freehand 3D imaging.

Currently, the reconstruction pipeline uses a pixel-based method based on estimated 6-DoF probe motion. Although this approach enables accurate compounding, speckle-related artifacts may still degrade the continuity of reconstructed volumes. In future work, we plan to incorporate implicit neural representations (INRs) [41], [42], [43], [44] and view-consistent generative modeling to improve spatial smoothness, particularly for thin vascular structures. Such techniques are particularly promising when applied to Doppler or photoacoustic data, where high-contrast vascular signals reduce the confounding effects of speckle.

We also acknowledge that advanced reconstruction frameworks, such as geometry-guided or anatomy-aware models (e.g., RecON [14]), could be integrated with our motion estimation pipeline to further mitigate tracking errors and improve final volume quality.

To the best of our knowledge, this is the first attempt at free-hand 3D vascular imaging using ultrasound Doppler acquisition data or photoacoustic data. In this study, we specifically targeted small vessels to emphasize their shape and structure. For ultrasound imaging, we chose power-Doppler over color-Doppler image processing due to its superior sensitivity in detecting weak blood flow signals in smaller vessels. For photoacoustic (PA) imaging, we selected an optimal laser wavelength to maximize blood signal detection and reconstructed the images from raw data using appropriate hyperparameters in the DAS and filtering procedures.

For small vessels, quantitative parameters such as blood volume and flow velocity tend to remain relatively stable over time. In contrast, larger vessels exhibit significant fluctuations in these values due to pulsatile flow. Thus, for accurate quantitative imaging, the ideal target and conditions are those where the values exhibit minimal temporal variation but may vary spatially. This ensures that spatial changes in vascular structure are captured effectively without being obscured by temporal variations in blood flow during at least scanning.

The free-hand 3D PAUS technique holds immense clinical potential in diagnostic imaging and related interventions. For example, in the assessment of thyroid nodules, hands-free 3D PA imaging could provide detailed visualization of the nodule's vascularity and tissue composition in real time. PA imaging has the unique ability to highlight abnormal blood vessel growth, which is often associated with malignant nodules. By integrating 3D reconstructions, clinicians could more accurately differentiate between benign and malignant thyroid nodules, improving the precision of fine-needle aspiration biopsies and reducing the number of unnecessary procedures. These applications are the focus of our ongoing studies.

However, in clinical settings, even minor errors can have significant implications, emphasizing the need for prediction models with enhanced precision and accuracy. Furthermore, the methodology proposed in this study exhibits generalizability primarily on forearm scans, necessitating an expansion of the dataset to encompass diverse body parts.

In conclusion, we introduced MoGLo-Net, a novel approach for estimating PAUS scan motion without the need for an additional positional sensor. The scan data encompassed complex motions, including dynamic in-plane movements and unidirectional out-of-plane shifts. These scans often trace relatively long trajectories exceeding 200 mm, which complicates the task as accumulated errors tend to magnify in proportion to the scan length. Thus, achieving generalization performance requires not only minimizing frame-wise errors but also reducing bias and final drift. By leveraging specialized supervision methods and deep learning architectures, we have attained superior performance, particularly in terms of accumulated errors. For future work, we aim to enhance this study by integrating anatomical information of tissues or embedding

the order of scan sequences to estimate arbitrary motions more accurately. This development promises to further refine motion estimation in ultrasound imaging, potentially leading to more precise and reliable diagnostic tools.

REFERENCES

[1] Y.-J. Ho, C. Huang, C. Fan, H. Liu, and C. Yeh, "Ultrasonic technologies in imaging and drug delivery," *Cellular Mol. Life Sci.*, vol. 78, nos. 17–18, pp. 6119–6141, Jul. 2021.

[2] D. H. Evans, J. A. Jensen, and M. B. Nielsen, "Ultrasonic colour Doppler imaging," *Interface Focus*, vol. 1, no. 4, pp. 490–502, Aug. 2011.

[3] C. Lee et al., "Panoramic volumetric clinical handheld photoacoustic and ultrasound imaging," *Photoacoustics*, vol. 31, Jun. 2023, Art. no. 100512.

[4] M. H. Mozaffari and W.-S. Lee, "Freehand 3-D ultrasound imaging: A systematic review," *Ultrasound Med. Biol.*, vol. 43, no. 10, pp. 2099–2124, 2017.

[5] A. Sorriento et al., "Optical and electromagnetic tracking systems for biomedical applications: A critical review on potentialities and limitations," *IEEE Rev. Biomed. Eng.*, vol. 13, pp. 212–232, 2020.

[6] J.-F. Chen, J. B. Fowlkes, P. L. Carson, and J. M. Rubin, "Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test," *Int. J. Imag. Syst. Technol.*, vol. 8, no. 1, pp. 38–44, 1997.

[7] 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.

[8] 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.

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

[10] H. Guo, S. Xu, B. J. Wood, and P. Yan, "Sensorless freehand 3D ultrasound reconstruction via deep contextual learning," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Lima, Peru. Cham, Switzerland: Springer, Jan. 2020, pp. 463–472.

[11] M. Luo et al., "Self context and shape prior for sensorless freehand 3D ultrasound reconstruction," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Strasbourg, France. Cham, Switzerland: Springer, Jan. 2021, pp. 201–210.

[12] G. Ning, H. Liang, L. Zhou, X. Zhang, and H. Liao, "Spatial position estimation method for 3D ultrasound reconstruction based on hybrid transfomers," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[13] 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. 2nd Int. Workshop Simplifying Med. Ultrasound*, Strasbourg, France. Cham, Switzerland: Springer, Sep. 2021, pp. 96–105.

[14] M. Luo et al., "RecON: Online learning for sensorless freehand 3D ultrasound reconstruction," *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102810.

[15] 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.

[16] G. E. Trahey, S. W. Smith, and O. T. V. Ramm, "Speckle pattern correlation with lateral aperture translation: Experimental results and implications for spatial compounding," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. UFFC-33, no. 3, pp. 257–264, May 1986.

[17] R. 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.

[18] 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.

[19] 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.

[20] M. Luo, X. Yang, H. Wang, L. Du, and D. Ni, "Deep motion network for freehand 3D ultrasound reconstruction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2022, pp. 290–299.

[21] M. Luo et al., "Multi-IMU with online self-consistency for freehand 3D ultrasound reconstruction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 342–351.

[22] A. K. Z. Tehrani and H. Rivaz, "Displacement estimation in ultrasound elastography using pyramidal convolutional neural network," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2629–2639, Dec. 2020.

[23] 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.

[24] Z. Yan et al., "Fine-grained context and multi-modal alignment for freehand 3D ultrasound reconstruction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2024, pp. 340–349.

[25] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[26] Y. Dou, F. Mu, Y. Li, and T. Varghese, "Sensorless end-to-end freehand 3-D ultrasound reconstruction with physics-guided deep learning," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 71, no. 11, pp. 1514–1525, Nov. 2024.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 402–419.

[29] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3922–3931.

[30] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: Global–;ocal attention for image description," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 726–737, Mar. 2018.

[31] N. Le, K. Nguyen, A. Nguyen, and B. Le, "Global–local attention for emotion recognition," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21625–21639, Dec. 2022.

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[34] J.-S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 2, pp. 165–168, Mar. 1980.

[35] C. Demené et al., "Spatiotemporal clutter filtering of ultrafast ultrasound data highly increases Doppler and fUltrasound sensitivity," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2271–2285, Nov. 2015.

[36] P. Soille, *Morphological Image Analysis: Principles and Applications*, vol. 2. Cham, Switzerland: Springer, 1999.

[37] J. T. Yen, J. P. Steinberg, and S. W. Smith, "Sparse 2-D array design for real time rectilinear volumetric imaging," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 47, no. 1, pp. 93–110, Jan. 2000.

[38] D. B. Downey and A. Fenster, "Vascular imaging with a three-dimensional power Doppler system," *Amer. J. Roentgenol.*, vol. 165, no. 3, pp. 665–668, 1995.

[39] J. A. Jensen, "Field: A program for simulating ultrasound systems," *Med. Biol. Eng. Comput.*, vol. 34, no. 1, pp. 351–353, 1997.

[40] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Proc. SPIE*, vol. 15, no. 2, 2010, Art. no. 021314.

[41] H. Chen et al., "Neural implicit surface reconstruction of freehand 3D ultrasound volume with geometric constraints," *Med. Image Anal.*, vol. 98, Dec. 2024, Art. no. 103305.

[42] S. Zhang et al., "Ultra-malin: Robotic ultrasound mapping and localization via implicit neural representation," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–12, 2025.

[43] Y. Zou, Y. Lin, and Q. Zhu, "PA-NeRF, a neural radiance field model for 3D photoacoustic tomography reconstruction from limited bscan data," *Biomed. Opt. Exp.*, vol. 15, no. 3, p. 1651, 2024.

[44] Y. Xiao et al., "Limited-view photoacoustic imaging reconstruction via high-quality self-supervised neural representation," *Photoacoustics*, vol. 42, Apr. 2025, Art. no. 100685.