Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PA-SfM: Tracker-free freehand 3D photoacoustic imaging via differentiable acoustic radiation

Changhui Li

chli@pku.edu.cn

Peking University https://orcid.org/0000-0002-6179-2684

Yao Yao

Nanjing University https://orcid.org/0000-0001-9866-4291

Shuang Li

Peking University https://orcid.org/0000-0002-6965-5860

Jian Gao

Nanjing University

Chulhong Kim

Pohang University of Science and Technology

Seongwook Choi

Korea University

Hao Huang

Zhejiang Lab

Xuanhao Wang

Zhejiang Lab https://orcid.org/0009-0006-7360-363X

Junhui Shi

Zhejiang Lab https://orcid.org/0000-0002-5741-2781

Qian Chen

Peking University

Yibing Wang

Peking University https://orcid.org/0009-0004-9841-3846

Shuang Wu

Nanjing University

Yu Zhang

Peking University

Yucheng Zhou

Peking University

Article

Keywords: Freehand 3D photoacoustic imaging, pose recovery, differentiable acoustic radiation, tracker-free imaging, structure from motion

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11501955/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

PA-SfM: Tracker-free freehand 3D photoacoustic imaging via differentiable acoustic radiation

Shuang Li[1†], Jian Gao[2†], Chulhong Kim[3], Seongwook Choi[7], Hao Huang[4,6], Xuanhao Wang[4,5], Junhui Shi[4,5,6], Qian Chen[1], Yibing Wang[1], Shuang Wu[2], Yu Zhang[1], Yucheng Zhou[1], Yao Yao[2*], Changhui Li[1,8*]

[1]Department of Biomedical Engineering, College of Future Technology, Peking University, Beijing, China.
[2]School of Intelligence Science and Technology, Nanjing University, Suzhou, China.
[3]Department of Electrical Engineering, Convergence IT Engineering, Mechanical Engineering, and Medical Science and Engineering, Medical Device Innovation Center, Pohang University of Science and Technology, Pohang, Republic of Korea.
[4]Zhejiang Lab, Hangzhou, China.
[5]State Key Laboratory of Ocean Sensing, Hangzhou, China.
[6]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China.
[7]School of Health and Environmental Science, Korea University, Seoul, Republic of Korea.
[8]National Biomedical Imaging Center, Peking University, Beijing, China.

*Corresponding author(s). E-mail(s): yaoyao@nju.edu.cn; chli@pku.edu.cn;
†These authors contributed equally to this work.

Abstract

Three-dimensional photoacoustic imaging (3D PAI) has shown great potential in preclinical studies and clinical applications, but sparse sensor arrays commonly limit both spatial sampling and the field of view (FOV). Freehand scanning could overcome these limitations through anatomy-adaptive scanning that extends spatial coverage and provides complementary views beyond the fixed array aperture, but its flexibility is conventionally constrained by the need for motor feedback or external pose tracking. Here we introduce PA-SfM, a tracker-free structure-from-motion (SfM) framework that recovers multiple poses directly from PA measurements, enabling freehand 3D PAI without external pose measurements. PA-SfM integrates a differentiable acoustic radiation model with hierarchical optimization and rigid array constraints to jointly estimate transformations between views and reconstruct 3D volumes. The performance of PA-SfM was demonstrated in numerical simulations, animal studies and human imaging experiments. Compared with motor encoder-based fusion of the mechanically scanned system, PA-SfM reduced vessel misalignment in expanded FOV reconstructions. Controlled quantitative validations yielded peak signal-to-noise ratio (PSNR) values over 40 dB and structural similarity index (SSIM) values about 0.98 relative to ground truth or known pose reference reconstructions. Freehand experiments further demonstrated robust pose recovery at high scan speeds, with inter-frame displacements of 9 mm at intervals of 0.1 s. These results support PA-SfM as a tracker-free approach to large-scale freehand 3D PAI and provide a basis for further evaluation in clinical settings.

Keywords: Freehand 3D photoacoustic imaging, pose recovery, differentiable acoustic radiation, tracker-free imaging, structure from motion

Introduction

Photoacoustic imaging (PAI) is a hybrid imaging modality that combines optical absorption contrast with ultrasound detection, and is widely explored in both preclinical and clinical settings[1–6]. Technological advances have greatly facilitated three-dimensional (3D) PAI based on two-dimensional matrix arrays, including spherical and planar arrays[7–14], as well as synthetically scanned systems[15,16]. 3D PAI has been used for *in vivo* imaging of peripheral vessels[15,17], breast tissue[18,19], and small animals[20].

Despite these advances, sensor arrays in 3D PAI are often sparse because of physical and engineering constraints. As a result, a single acquisition pose may provide insufficient spatial sampling density or angular coverage for high quality volumetric reconstruction[21,22]. Acquiring data from multiple views is therefore a common strategy to increase both the sensor density and effective acquisition aperture, mitigating reconstruction artifacts. In practice, these additional views are often generated through mechanical scanning consisting of rotation and translation[1,15,23,24]. Multi-view reconstruction requires accurate knowledge of the sensor pose across views. Even

minor positioning errors can lead to structural blurring, vessel discontinuity, and loss of quantitative fidelity in the reconstructed volume.

Various external tracking mechanisms are employed to provide physical pose feedback. In controlled systems, motorized stages and mechanical encoders are used to trace the acquisition trajectory. In freehand settings, external positioning devices such as optical tracking systems, electromagnetic trackers, and inertial measurement units are commonly employed[25,26]. Although these methods can provide geometric information, they also increase system complexity, cost, and calibration burden. Optical tracking requires a clear line of sight that is often difficult to maintain in crowded experimental or clinical environments[25,27]. Electromagnetic tracking is susceptible to field distortion from metallic objects and nearby electronics[26]. Inertial sensing is portable and inexpensive, but its estimates are vulnerable to drift accumulation over time[25,28]. Mechanical scanning systems offer controlled motion, yet the final reconstruction remains limited by encoder precision, assembly tolerance, and motion induced vibration.

These limitations have motivated tracker-free pose estimation[29–31]. Unlike ultrasound, where speckle tracking supports motion estimation[32,33], photoacoustic images may lack sufficient texture for reliable tracking[34]. MoGLo-Net[35] and wBCAM[36] infer probe motion from B-mode ultrasound image sequences to guide freehand PA and ultrasound reconstruction.

Image-based registration[37] enables tracker-free panoramic optoacoustic imaging through vascular centerline registration, intensity-based refinement and spatial compounding. Fourier-domain rotation and translation estimates enabled volumetric compounding[38], as subsequently used in HELIOS[39]. In photoacoustic microscopy, FS-PAM combines vascular feature matching with affine mosaicking[40]. PAATAM combines vascular geometry and intensity in a sliding-window factor graph to recover six-degree-of-freedom motion in optical-resolution photoacoustic angiography[41]. These methods derive pose constraints from reconstructed PA or ultrasound images. However, sparse sampling, image artifacts and reduced overlap can disrupt image correspondence during freehand scanning.

To address these challenges, we present PA-SfM, a tracker-free framework for freehand 3D PAI with joint pose recovery and reconstruction in multi-view, expanded field of view (FOV). PA-SfM estimates relative imaging poses directly from PA signals by combining a differentiable acoustic radiation model with hierarchical optimization under rigid array geometry constraints. Starting from the initial view, PA-SfM localizes the sensor array in subsequent views, enforces rigid array consistency to obtain coarse relative poses, and further refines the global transformations through differentiable pose optimization before joint multi-view or multi pose reconstruction. We demonstrate genuine freehand 3D PAI of human hand vasculature, where arbitrary hand motion provides multi-view measurements for pose recovery and joint reconstruction of a large FOV vascular network based on PA-SfM without external motion tracking. In this freehand experiment, the maximum average probe displacement between adjacent poses exceeded 9 mm. We further validate the method in mechanically scanned 3D PAI acquisitions with translation and rotation, *in vivo* rat experiments with known relative geometry, and numerical simulations. In mechanically rotated imaging, PA-SfM

reduced vessel misregistration compared with reconstruction based on motor encoder feedback. In translational multi pose imaging, PA-SfM enabled expanded FOV PA mapping without using translation stage pose feedback. We also compare PA-SfM with image-based method in measured hand data across array densities and scan speeds.

Results

PA-SfM framework for freehand 3D imaging

PA-SfM recovers array poses directly from PA measurements and combines the recovered views into a joint 3D reconstruction (Fig. 1). A differentiable acoustic radiation model enables pose recovery without external tracking (Supplementary Note 1).

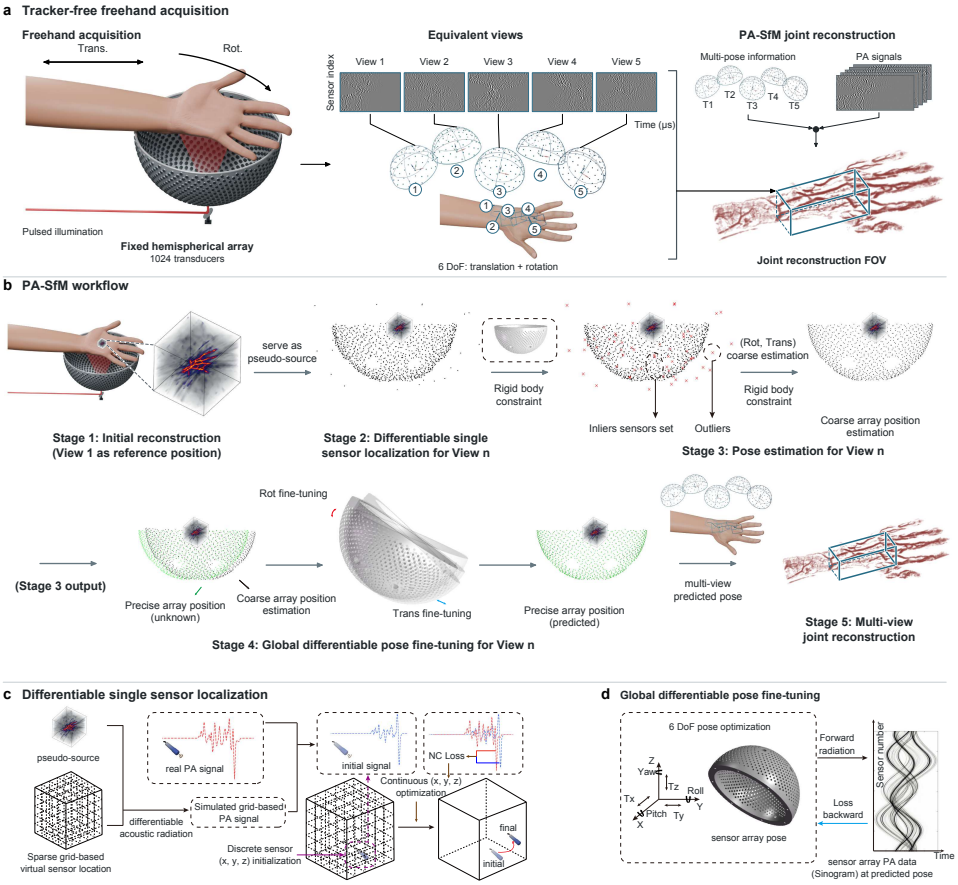


Fig. 1. Overview of the PA-SfM framework for tracker-free freehand 3D PAI. **a** Freehand imaging of the hand with a fixed hemispherical array and multi-view joint reconstruction. **b** The PA-SfM pipeline. **c** Differentiable single sensor localization. **d** Global differentiable pose fine tuning.

PA-SfM encodes the measured photoacoustic signals into a latent pseudo-source representation by optimizing a volumetric source distribution through the differentiable acoustic radiation model. The pseudo-sources provide an acoustic reference for predicting signals at candidate sensor locations and optimizing sensor coordinates against the measurements. Rigid array constraints then yield a common rotation and translation, followed by global differentiable refinement and joint reconstruction. Repeating these steps for successive views registers the measurements in a common coordinate system and extends the field of view. Detailed optimization procedures are given in Methods.

Freehand 3D reconstruction of hand vessels

To evaluate the feasibility of PA-SfM, we performed tracker-free freehand PAI of a 26-year-old male volunteer's hand using the 3D-PanoPACT system [23]. 3D-PanoPACT is a hemispherical array PAI system with 1024 transducer elements and a diameter of 20 cm, providing volumetric acoustic detection over a large angular aperture. The central frequency of each transducer element is 3.16 MHz. More details regarding the system's configuration and imaging performance can be found in the literature [23].

In this freehand experiment, the hand was placed above the imaging aperture and freely moved without external tracking, while the PAI system is fixed, as shown in Figure 2(a). The hand motion included extension, retraction and swinging trajectories, thereby producing a sequence of partially overlapping photoacoustic views with unknown relative poses. PA-SfM was then used to estimate the transformations between views directly from the photoacoustic signals and to recover the absolute coordinates for each individual view, thus integrating into an equivalent large array for joint reconstruction. The results were reconstructed using the classical delay-and-sum (DAS) algorithm, and maximum amplitude projections (MAPs) were used for visualization.

Figure 2 illustrates the overall workflow. Compared with conventional static imaging, freehand acquisition provides multiple observations of the target from different relative viewpoints, but also introduces unknown hand motion and view-dependent misalignment. Fig. 2(b) presents the imaging geometry from top and side views, and Fig. 2(c) is a photography of experimental configuration. The freehand hand imaging experiment procedure can be found in Supplementary Movie 1. Although the volumetric frame rate is 10 Hz, to demonstrate the capability of PA-SfM to recover pose without much overlap in FOV, five single views were selected from over ten candidate views in chronological order as shown in Fig. 2(d), and each individual view captured only partial vascular information and exhibited view-dependent incompleteness. After PA-SfM pose recovery and joint reconstruction, the reconstructed volume reveals a more complete 3D vascular network of the hand with a substantially expanded FOV of $50\,\mathrm{mm} \times 30\,\mathrm{mm} \times 10\,\mathrm{mm}$ and fewer artifacts, as demonstrated by the top-, side- and front view MAPs and the volumetric display in Fig. 2(e). The 3D display of the result can also be found in Supplementary Movie 2. The probe displacement between adjacent poses exceeded 9 mm, while the PA-SfM reconstruction preserved fine vascular details without visible vessel duplication or discontinuity.

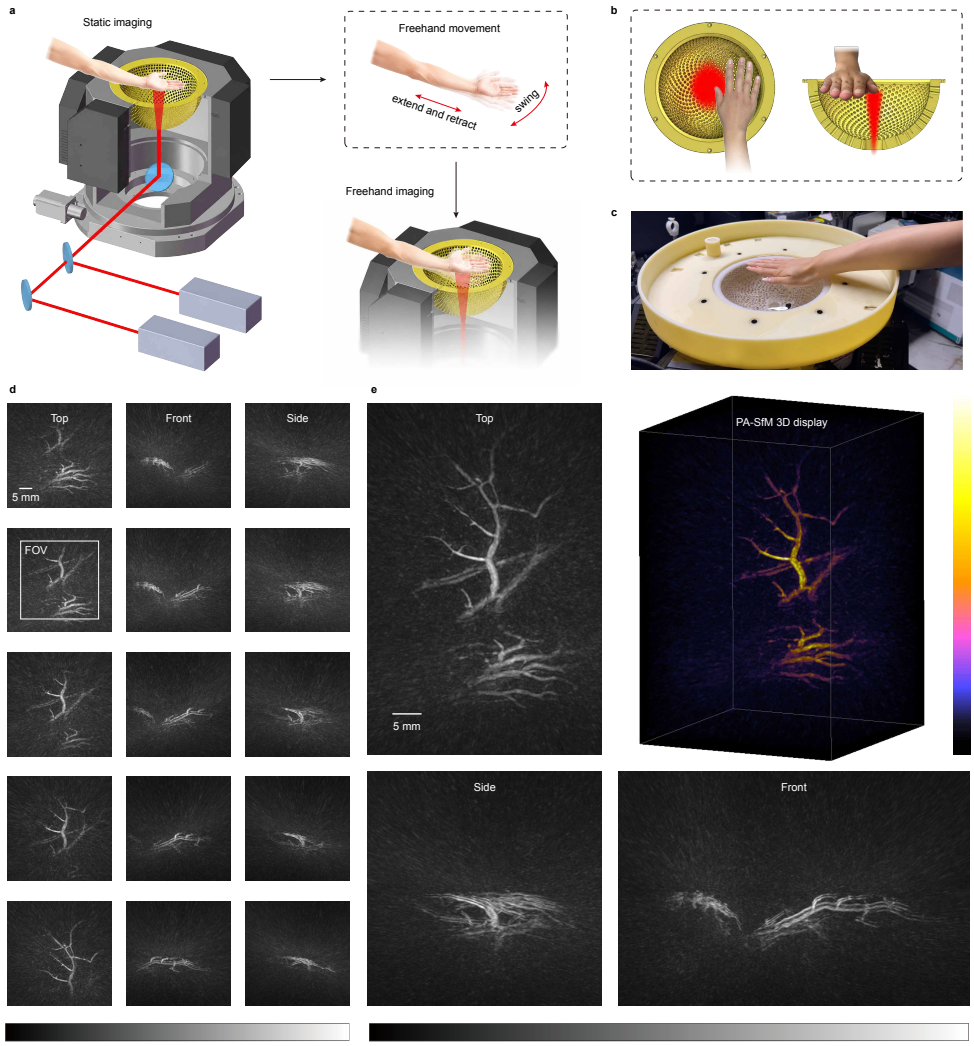


Fig. 2. PA-SfM freehand 3D reconstructions of human hand. a Workflow of freehand photoacoustic imaging of the hand using the 3D-PanoPACT system. During imaging, the hand performed freehand motion, including extension–retraction and swinging trajectories, to provide multi-view measurements for PA-SfM reconstruction. **b** Detailed top and front views of the imaging geometry, illustrating the hand position, detection aperture and optical excitation region. **c** Photograph of the experimental setup. **d** Top-, front-, and side view maximum amplitude projections (MAP) of reconstructions from five sequentially ordered single views (Scale: 5 mm; grayscale; reconstruction with DAS), in which the white box represent the FOV of a single view. **e** Top-, front-, side view MAP and 3D display for PA-SfM freehand 3D reconstruction result (Scale: 5 mm; colormap: gray, magma).

To assess the repeatability of PA-SfM under freehand acquisition, we performed additional independent hand imaging experiments, as shown in Fig. 3. In the first experiment, eight sequentially ordered but non-consecutive frames (fewer overlaps between frames) from imaging area 1 were selected for PA-SfM reconstruction. These frames were from the same freehand data used in Fig. 2, with three additional frames included to increase the spatial coverage and provide more diverse viewing angles. As shown in Fig. 3(a), PA-SfM recovered the vascular structure from these sparsely sampled views.

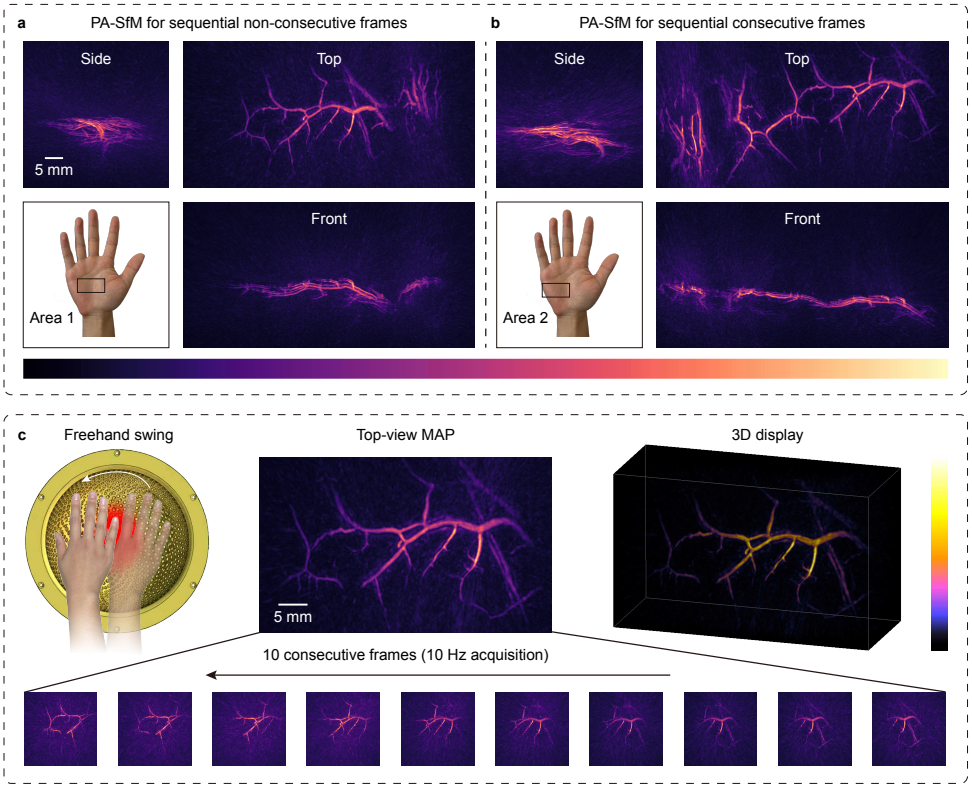


Fig. 3. Repeatability validation of PA-SfM freehand 3D reconstruction of hand vessels using two partially overlapping imaging regions. a PA-SfM reconstruction of imaging area 1 using eight sequentially ordered but non-consecutive frames acquired from the same freehand hand vessel data as in Fig. 2, with three additional frames included (Scale: 5 mm; colormap: magma). **b** PA-SfM reconstruction of another hand vessel data acquired from imaging area 2 using 25 consecutive frames during a freehand swing trajectory (**a** and **b** are independent experiments derived from the same subject). **c** Pose recovery and joint 3D reconstruction visualization using the first 10 consecutive frames from **b** (Scale: 5 mm; colormap: magma).

In the second experiment, an independent freehand acquisition was performed on a different imaging area 2 of the same palm. This data consisted of 25 consecutive frames acquired (10 Hz sampling rate) during a freehand swing trajectory, providing a denser multi-view acquisition of the hand vasculature. PA-SfM reconstruction from these consecutive frames yielded a continuous 3D vascular map with consistent vessel morphology across the reconstructed FOV, as shown in Fig. 3(b). The reconstructions in Fig. 3(a) and Fig. 3(b) show similar vessel morphology despite the different frame selections.

Since two independent imaging experiments have a partially overlapping area, we can validate the repeatability by comparing the reconstruction results in the overlapped area. Fig. 3(c) shows the PA-SfM pose recovery and joint reconstruction using the first 10 consecutive frames from the second experiment, which corresponds to the region overlapping with area 1. The major vascular branches and local vessel patterns reconstructed from the two experiments show consistent morphology in the shared region, despite being acquired from different frame selections and motion trajectories.

Image-based comparison in freehand hand imaging

We also compared PA-SfM with image-based method on freehand imaging sequence of the hand in Fig. 3(c). Figure 4(a,b) shows the reference panorama formed from the preceding frames and the new frame at its initial pose. With the final Image-based pose estimate, several vascular branches remained displaced relative to the reference (Fig. 4(c)). The PA-SfM estimate brought the new frame vessels into closer agreement with the same reference panorama in the three orthogonal projections (Fig. 4(d)).

Expanded FOV reconstruction with mechanical translation

We evaluated PA-SfM under translational multi pose acquisition to extend the FOV. Using the same 3D PAI system, the *in vivo* imaging data is from a mechanical scanning experiment with four poses, as illustrated in Fig. 5(a). Starting from the initial position, the hemispherical array was mechanically translated three times, with a step size of 4 mm. Therefore, the four acquired poses jointly covered a larger anatomical region than a single stationary acquisition. For comparison, we reconstructed each individual pose using DAS and then performed multi pose joint DAS reconstruction using either the hardware-tracked transformations or the transformations estimated by PA-SfM, as shown in Fig. 5(b), which also demonstrates that the single pose reconstructions capture only partial vascular structures.

In the top view MAPs in Fig. 5(c), the hardware-tracked reconstruction contains fragmented, duplicated and broadened vessels. These artifacts are reduced in the PA-SfM reconstruction, particularly in the regions marked by green arrows and dashed boxes. Fig. 5(d) shows the corresponding 3D reconstructions.

Multi-view reconstruction with mechanical rotation

We evaluated the performance of PA-SfM using *in vivo* animal experimental data[23] acquired with the 3D-PanoPACT system. Mechanical rotation of the transducer

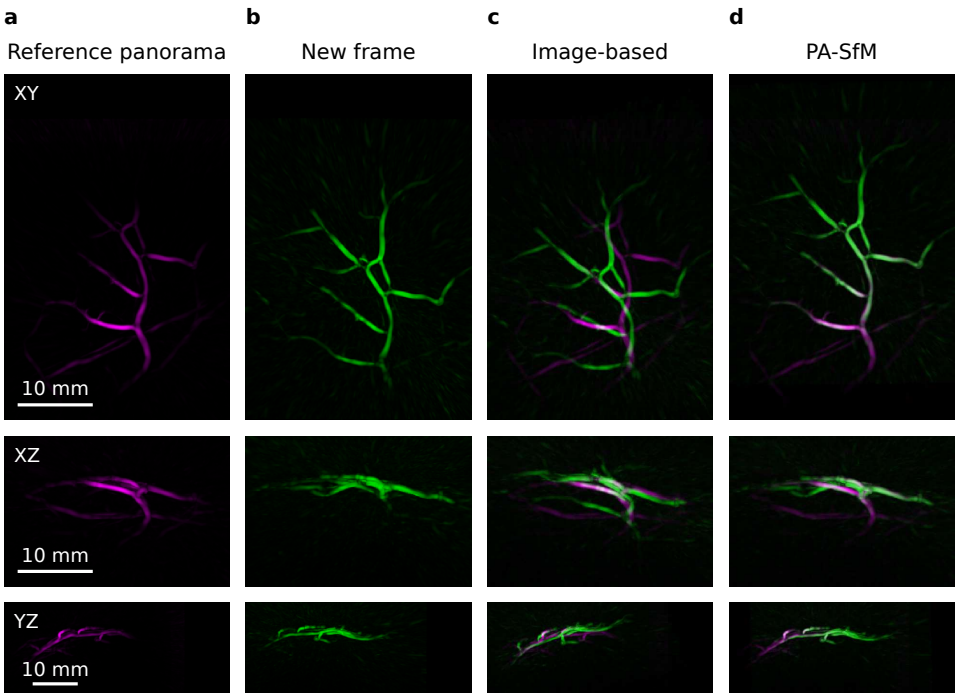


Fig. 4. Pose recovery comparison for a freehand PA imaging of the human hand. **a** Reference panorama reconstructed from the preceding frames. **b** Newly acquired frame at its initial pose. **c,d** Overlays obtained with the final Image-based and PA-SfM pose estimates, respectively. Each column shows the XY, XZ and YZ projections from top to bottom. Magenta and green represent maximum projections of the reference and new frame Frangi vesselness features; white indicates projected overlap. Both pose estimates are displayed against the same reference panorama using a shared linear intensity range. Scale bars in the first column: 10 mm.

array increases the effective spatial sampling density and reduces limited view artifacts. However, the final reconstruction quality is highly sensitive to the accuracy of mechanical positioning feedback. Hardware-induced errors, including encoder inaccuracies, mechanical structural deviations and stochastic vibrations during motor movement, can introduce misalignment between views and thereby degrade multi-view reconstruction.

The *in vivo* animal study is illustrated in Fig. 6(a), where the mouse hepatic vasculature was imaged from two views controlled by a mechanical 8.0° rotation of the hemispherical array around the system's central Z axis. To compare PA-SfM with encoder-based tracking, multi-view joint reconstructions using either hardware feedback or PA-SfM are presented in Fig. 6.

As shown in Fig. 6(b), both the original and rotated single view reconstructions provide the same FOV with different fine vascular structures due to sparse spatial sampling. In the zoomed-in MAPs in Fig. 6(c), the green arrows highlight misregistered

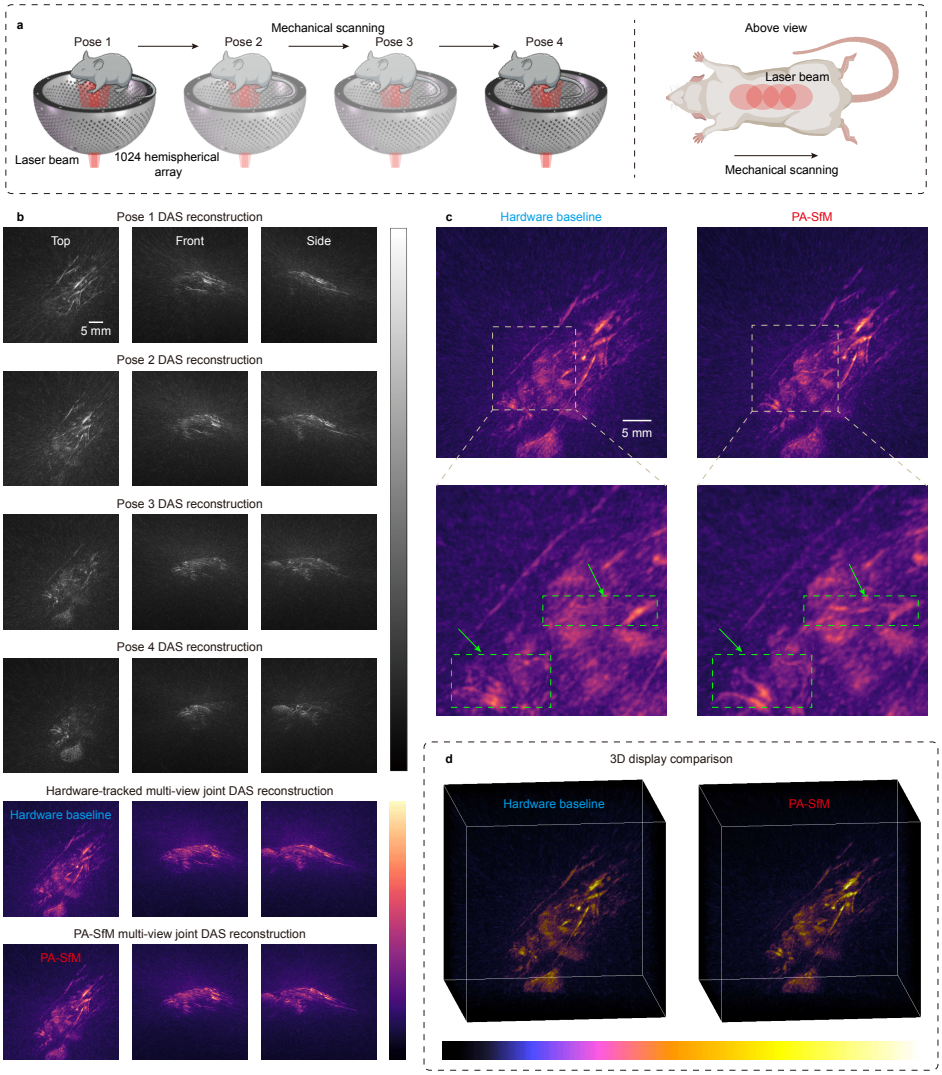


Fig. 5. Expanded field of view multi pose PA mapping using mechanical translation. a Schematic of the mechanical scanning experiment with the 1024-element hemispherical array. The animal was imaged at four poses, including the initial pose and three subsequent mechanical translations. Each translation step was 4 mm along a direction forming a 48° angle with the negative Y-axis of the array. **b** Top-, front-, and side view MAPs of single pose DAS reconstructions and multi pose joint DAS reconstructions using hardware-tracked poses and PA-SfM estimated poses (Scale: 5 mm; colormap: gray, magma). **c** Top view MAP comparison between the hardware-tracked baseline and PA-SfM reconstruction, with local magnifications highlighting vessel continuity and structural sharpness (Scale: 5 mm; colormap: magma). **d** 3D display comparison of the hardware-tracked and PA-SfM multi pose reconstructions.

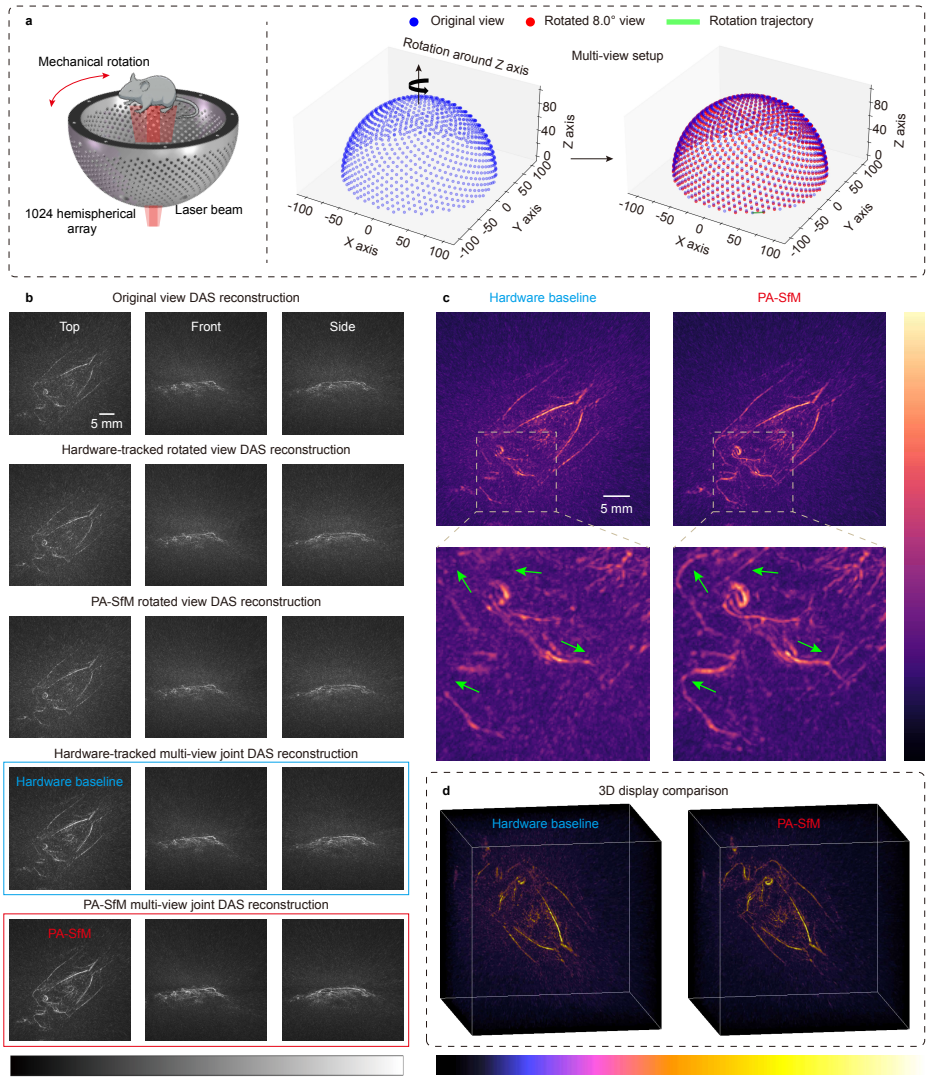


Fig. 6. Multi-view PA reconstructions of mouse hepatic vasculature. a 3D PAI imaging system based on a mechanically rotating hemispherical array. **b** Top-, front-, and side view MAPs of reconstructions from single views (original, hardware-tracked and PA-SfM rotated) and multi-views (hardware-tracked baseline, and PA-SfM result) (Scale: 5 mm; grayscale; reconstruction with delay-and-sum (DAS)). **c** Locally magnified top view MAPs comparing the hardware-tracked baseline and PA-SfM fine tuned reconstruction (Scale: 5 mm; colormap: magma). **d** 3D display comparing hardware-tracked and PA-SfM reconstructions.

and fragmented vessels in the hardware-tracked reconstruction, whereas the corresponding branches remain connected in the PA-SfM reconstruction. Fig. 6(d) shows the corresponding 3D reconstructions.

To validate the repeatability of PA-SfM, we chose another set of data that imaged a different part of the mouse abdomen with a nominal rotation angle of 4.0° around the system's Z axis. Supplementary Fig. 1 compares the PA-SfM and hardware-tracked reconstructions. Supplementary Fig. 1(a) presents two single view reconstructions and the joint reconstruction result. In the regions marked by green arrows in Supplementary Fig. 1(b), vessel branches that appear blurred or interrupted in the hardware-tracked reconstruction are resolved in the PA-SfM result.

In vivo validation with known relative geometry

We next validated PA-SfM on real *in vivo* animal imaging data with a known relative geometry between views. Specifically, we used data from *in vivo* multi-view 3D PAI of rat kidney (Fig. 7) and rat liver (Supplementary Fig. 2), which were acquired by Professor Kim's lab using a hemispherical ultrasound (US) transducer array with 1024 elements and a radius of 60 mm, and the average center frequency of each US transducer element in the array is 2.02 MHz[42]. More details regarding the 3D PAI system and animal experiments can be found in the literature[43–49].

Based on the arrangement pattern of array elements, we deliberately partitioned the full 1024-element spherical array into two 512-element sparse subsets, separated by an exact 90° relative rotation along the azimuthal axis (Fig. 7(a-b)). These two subsets, designated as View A and View B which represent two independent imaging poses, were used to generate the ground truth (GT) multi-view reconstruction for quantitative evaluation. The final *in vivo* 3D reconstruction results are presented in Fig. 7(c-d).

For the rat kidney experiment, Fig. 7(c) compares the single view reconstruction, the PA-SfM multi-view reconstruction, and the GT multi-view reconstruction. MAPs from different viewpoints show the vascular structures, with the top view also displayed in a turbo colormap to show background noise and reconstruction artifacts. As shown in Fig. 7(c-d), the single view reconstruction provides limited structural information and exhibits prominent limited view artifacts. In contrast, the PA-SfM multi-view reconstruction combines View A and View B to recover a more complete vascular network of the *in vivo* rat kidney. The reconstructed vessel branches correspond to those in the GT multi-view reconstruction.

When evaluated against the GT multi-view reconstruction, the single view kidney reconstruction yielded a peak signal-to-noise ratio (PSNR) of 30.20 dB and a structural similarity index measure (SSIM) of 0.7019. By contrast, the PA-SfM multi-view reconstruction achieved a PSNR of 41.42 dB and an SSIM of 0.9864. A further comparison between the PA-SfM recovered pose and the GT pose showed that the mean and minimum localization errors for all sensors were 0.34 mm and 0.027 mm, respectively.

We also evaluated PA-SfM using the *in vivo* rat liver imaging data. Following the same visualization scheme, Supplementary Fig. 2(a-b) presents the liver reconstruction results. The single view liver reconstruction suffers from incomplete vessel connectivity and elevated background noise, which is particularly evident in the turbo colormap.

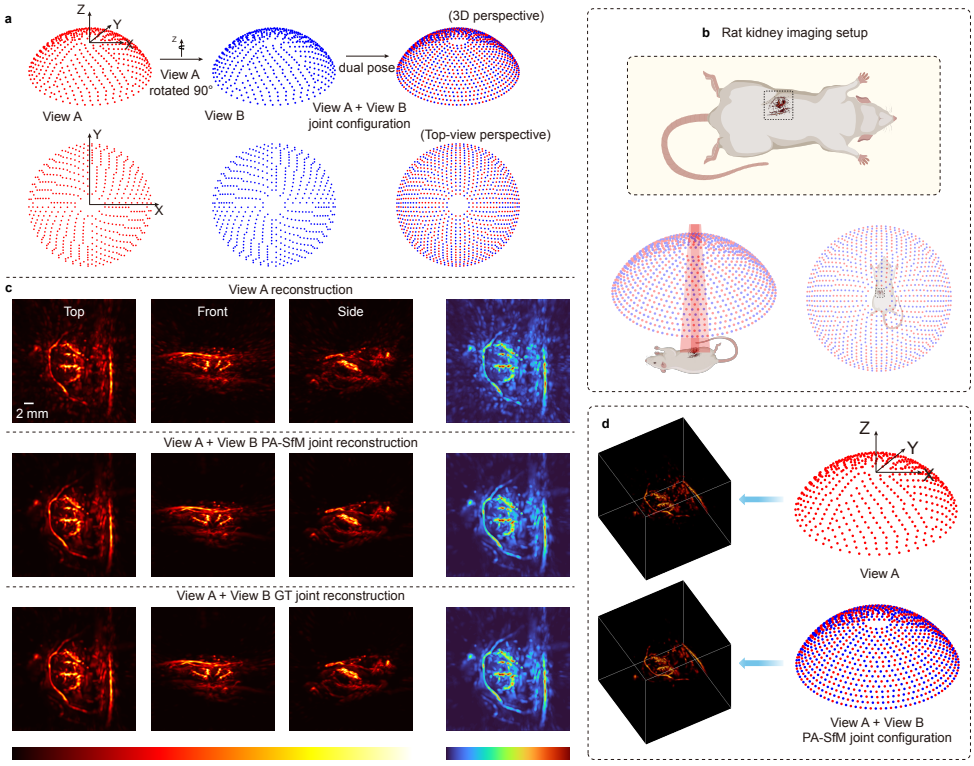


Fig. 7. The reconstruction results of multi-view *in vivo* rat kidney with PA-SfM algorithm. a The dual view 3D PAI for *in vivo* rat kidney imaging. **b** The rat kidney imaging setup. **c** The single view reconstruction, multi-view reconstruction using the PA-SfM pose, and ground truth (GT) of the *in vivo* rat kidney (Scale: 2 mm). **d** The 3D results of single view, and PA-SfM multi-view reconstruction.

In comparison, the PA-SfM multi-view reconstruction resolves vessel branches missing from the single view result and follows the branching pattern in the GT reconstruction.

Compared with the GT multi-view reconstruction, the single view liver MAP yielded a PSNR of 28.01 dB and an SSIM of 0.6947. The PA-SfM multi-view reconstruction yielded a PSNR of 44.31 dB and an SSIM of 0.9961. Under the PA-SfM estimated array poses, the mean and minimum localization errors across all sensors were 0.21 mm and 0.016 mm, respectively. The effect of global differentiable pose fine tuning on these reconstructions is shown in Supplementary Note 2 and Supplementary Fig. 3.

Quantitative validation in sparse array simulations

To quantitatively evaluate pose recovery and image reconstruction under fully controlled conditions, we conducted numerical simulations using a phantom liver model

under a multi-view setup. As shown in Fig. 8(a), the simulation utilized a sparse hemispherical array to acquire data from multiple views. Specifically, the array consisted of only 33 sparse elements distributed across a spherical cap with a radius of 60 mm. View 1 served as the reference array pose, while View 2 and View 3 were generated by translating and rotating the array around the target. The single view and PA-SfM multi-view joint reconstruction results are shown in Fig. 8(b,c).

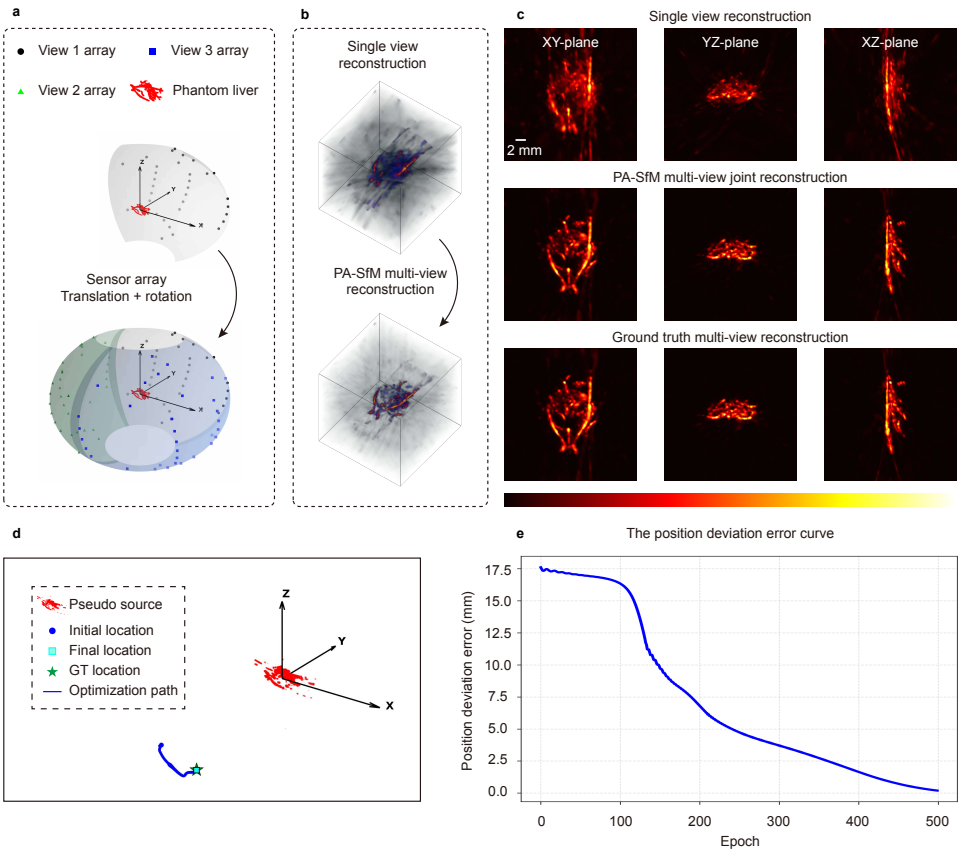


Fig. 8. The reconstruction results of multi-view simulation data with PA-SfM algorithm. a The multi-view setup of the simulation experiment. **b** The 3D results of single view, and PA-SfM multi-view reconstruction. **c** The single view reconstruction, multi-view reconstruction using the PA-SfM pose, and GT of the phantom liver (Scale: 2 mm). **d** Differentiable single sensor localization optimization. **e** The absolute position deviation error curve of the single sensor localization optimization.

To demonstrate the pose optimization mechanism, Fig. 8(d-e) illustrates a representative target sensor during single sensor localization process. Starting from an initial sensor location, the differentiable localization procedure progressively optimized the sensor position by minimizing the discrepancy between the simulated PA signal

and the GT supervision signal. The absolute position deviation of a randomly selected sensor decreased during optimization (Fig. 8(e)). After PA-SfM pose optimization, the mean and minimum Euclidean distance errors of sensor localization were reduced to 0.751 mm and 0.113 mm, respectively.

Based on the recovered PA-SfM poses, we further compared the 3D reconstruction results. As depicted in Fig. 8(b,c), the single view reconstruction suffered from severe limited view artifacts, yielding quantitative metrics of 25.51 dB for the PSNR and 0.6775 for the SSIM in the MAP along the XY plane. In contrast, the PA-SfM multi-view reconstruction expanded the angular coverage and reduced limited view artifacts. The reconstruction yielded a PSNR of 41.14 dB and an SSIM of 0.9799. The contribution of global differentiable pose fine tuning is examined in Supplementary Note 3 and Supplementary Fig. 4.

Pose recovery across array densities and scan speeds

To evaluate pose recovery as spatial sampling and overlap between views decrease, we simulated freehand scanning of three vascular samples with hemispherical arrays containing 32, 64 and 128 detectors. Scan speeds of 20, 40, 60 and 80 mm/s at a volumetric acquisition rate of 10 Hz yielded 18, 10, 7 and 6 views, respectively, giving 36 combinations of sample, array density and scan speed. Figure 9(a,b) shows the array configurations and scanning geometry. The corresponding mean overlap between adjacent illuminated volumes decreased from 86.6% to 59.0% (Fig. 9(c)), with projected acquisition footprints shown in Fig. 9(d). Acquisition details are provided in Supplementary Note 4. The scanning geometry and progressive FOV extension are illustrated in Supplementary Fig. 5 and Supplementary Movie 3.

We compared PA-SfM with Image-based, which combines vessel enhancement, centerline registration, correlation-based refinement and compounding. Both methods used data derived from the same simulated acoustic measurements. Reconstructions from the estimated poses were generated using the universal back-projection (UBP), with the true pose reconstruction (GT pose) as the reference. Pose error was defined as the root-mean-square displacement of 27 fixed control points across all non-reference frames, accounting for both translation and rotation.

Across the 27 conditions at 20–60 mm/s, PA-SfM achieved pose errors of 0.34–1.56 mm, with a mean of 0.73 mm, compared with 9.41 mm for Image-based. At 60 mm/s with 128 detectors, the errors for the three vascular samples were 0.47, 0.38 and 0.47 mm for PA-SfM, versus 15.92, 22.90 and 4.40 mm for Image-based. The three-dimensional pose comparison for Vessel 1 shows close agreement of the PA-SfM positions and orientations with ground truth, whereas the Image-based estimates deviate along the scan (Fig. 9(e,f)). Across the three samples, the branching patterns reconstructed by PA-SfM agreed with those in the GT pose reconstruction (Fig. 10(a)). Local magnifications further show preserved vessel continuity with PA-SfM and displaced or missing branches with Image-based (Fig. 10(b)). The complete comparisons for all three samples are provided in Supplementary Figs. 6–8 and Supplementary Tables 1–3.

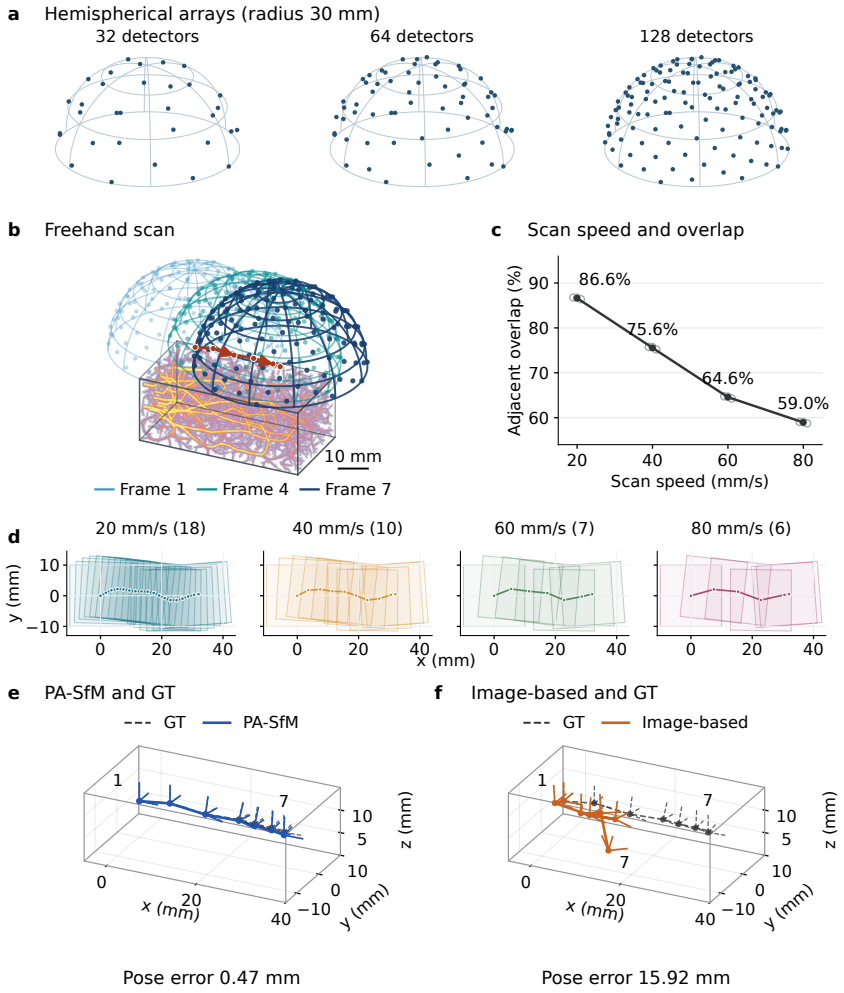


Fig. 9. Expanded FOV simulation setup and three-dimensional pose recovery. a Hemispherical arrays with 32, 64 and 128 detectors (radius 30 mm). **b** Ground truth (GT) vascular volume and three acquisition poses; red line, probe center trajectory. **c** Adjacent volume overlap (illumination ≥ 0.5): three vessel mean (line) and individual vessels (circles). **d** XY projections of local 20×20 mm squares at $z = -12$ mm; the dots mark their centers. **e,f** Estimated probe centers and 4 mm orientation axes against GT, in identical coordinates without further alignment. Numbers identify the first and last frames. Panels **b,d–f** show Vessel 1; **b,e,f** use 128 detectors at 60 mm/s. Pose error is pooled root-mean-square (RMS) displacement of 27 control points over six non-reference frames.

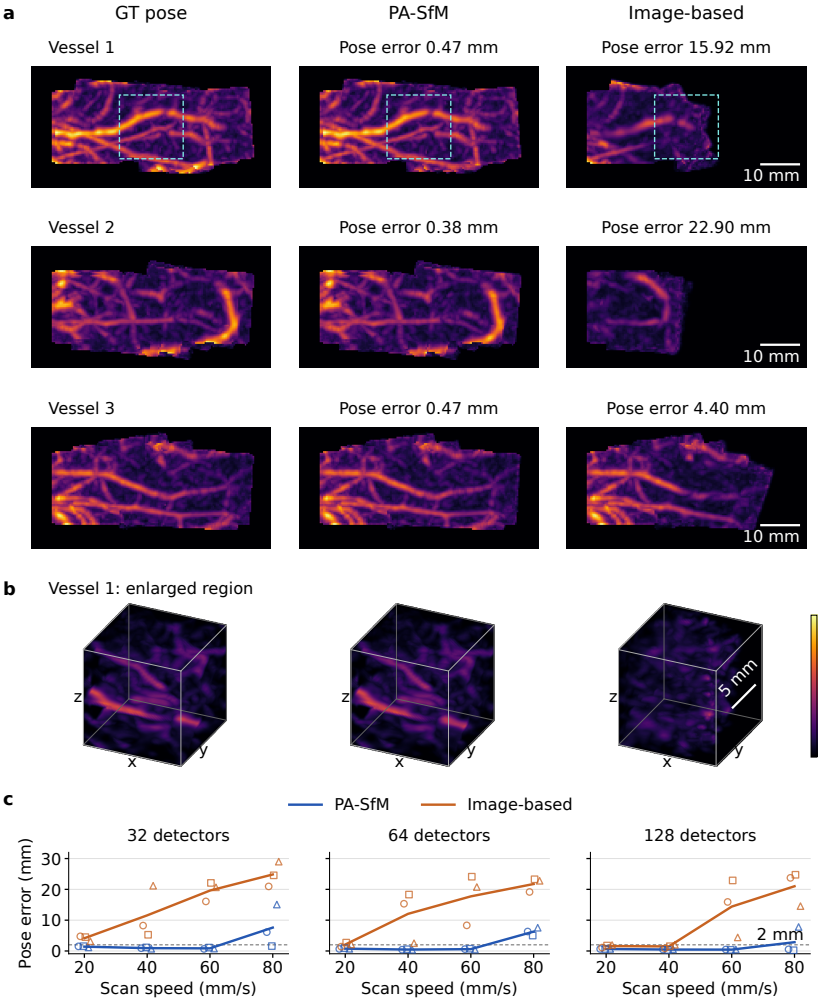


Fig. 10. Reconstruction comparison across three vascular samples. a XY maximum amplitude projections (MAPs) of Vessels 1–3 reconstructed using GT poses, PA-SfM and Image-based at 60 mm/s with 128 detectors. **b** Three-dimensional MAPs of the 16×16×16 mm region outlined in cyan in **a**. For each sample, the three reconstructions share the evaluation mask and linear intensity range; **b** retains the range of Vessel 1 in **a**. **c** Pose errors with 32, 64 and 128 detectors, from left to right. Open circles, squares and triangles denote Vessels 1, 2 and 3; lines show their mean. The dashed line marks 2 mm. GT pose denotes reconstruction using the true acquisition poses. Scale bars: 10 mm in **a** and 5 mm in **b**.

Across array densities, pose recovery became less reliable at 80 mm/s, with errors exceeding 2 mm in six of nine PA-SfM conditions and all nine Image-based conditions (Fig. 10(c)).

Discussion

PA-SfM enables tracker-free freehand 3D PAI by estimating acquisition geometry from acoustic measurements and jointly reconstructing multiple views. We demonstrated this principle in freehand imaging of human hand vasculature, mechanically translated and rotated array acquisitions, *in vivo* rat studies with known relative geometry, and simulations. In freehand imaging, the recovered relative poses allow measurements acquired during uncontrolled motion to contribute to a common 3D volume.

PA-SfM was more robust than image-based registration during rapid freehand scanning in the expanded FOV simulations. Pose errors remained below 1.6 mm at scan speeds up to 60 mm/s, preserving vascular continuity despite reduced overlap between views. Errors exceeded 2 mm only at 80 mm/s, where overlap was lowest.

Compared with mechanical tracking, PA-SfM improves reconstruction quality with more accurate pose estimation for the 3D PAI system used in this study. This improvement arises because the nominal encoder positions could be influenced by calibration errors, backlash, structural deviations or stochastic motion, leading to image blurring for fine structures. Optimizing the imaging geometry against the PA signals reduced these misregistration artifacts. Nevertheless, the method still requires sufficient structural overlap and recognizable PA features across views, and its performance could degrade for sparse absorbers, weak contrast, severe artifacts or limited overlap. In addition, during expanded FOV freehand imaging, the array poses near the periphery of the overall FOV can receive relatively weaker optical illumination than those near the central region, resulting in reduced signal intensity close to edges in the joint reconstruction. We therefore applied edge compensation to the results shown in Fig. 2(e) and Fig. 3(c); further discussion is provided in Supplementary Note 5 and Supplementary Fig. 9. The current formulation also assumes rigid relative geometry and does not explicitly model non-rigid tissue deformation, which should be addressed in future work for applications involving larger tissue motion or deformation.

In the expanded FOV simulations, the mean time to estimate a complete trajectory was 31.74, 15.79, 10.62 and 8.98 s for PA-SfM at 20, 40, 60 and 80 mm/s, respectively, compared with 64.31, 39.68, 30.91 and 29.19 s for Image-based. These times include the internal pose estimation steps and exclude the common input volume reconstruction and final joint reconstruction (Supplementary Table 4). Higher scan speeds sampled the same trajectories with fewer frames. For high-resolution hand imaging, pose recovery and joint reconstruction required 175 s per additional pose on a single NVIDIA A100 SXM4 40 GB GPU, excluding reference view initialization; the complete two-view workflow required 287.8 s (Supplementary Note 6). We have released the complete automated workflow for the continuous 10-frame PA-SfM pose estimation and joint reconstruction example in Fig. 3(c), together with the corresponding 10-frame dataset; detailed instructions are provided in Supplementary Note 7. The complete 10-pose workflow takes less than 28 minutes using a single reconstruction grid

of $400\times400\times400$ voxels, with GPU memory usage of approximately 5.1 GB throughout the process. The current implementation supports offline high-resolution volumetric 3D PAI while jointly recovering unknown poses and reconstructing the volume without external tracking, encoder feedback or predefined motion. Future work will focus on algorithmic acceleration, memory optimization and parallel implementation to enable real-time pose recovery for interactive freehand 3D PAI.

Methods

Differentiable acoustic radiation

The core of the framework is a differentiable acoustic radiation engine that models acoustic wave propagation and computes gradients. To map the 3D PA initial pressure distribution $\mathbf{P}_c \in \mathbb{R}^{N_x\times N_y\times N_z}$ to the time-resolved sensor signals $p(\mathbf{x}_s, t)$ at sensor location $\mathbf{x}_s \in \mathbb{R}^3$, we employ a Gaussian kernel-based propagation model:

$$p(\mathbf{x}_s, t) = \sum_{\mathbf{x}\in\Omega} \mathbf{P}_c(\mathbf{x}) \cdot \mathcal{G}(\|\mathbf{x} - \mathbf{x}_s\|, t; \sigma), \tag{1}$$

where Ω denotes the reconstructed volume, and $\mathcal{G}$ is the spatiotemporal Gaussian radiation kernel parameterized by width σ.

The Gaussian radiation kernel is derived from the analytical solution of the PA wave equation. The pressure field generated by an initial pressure distribution $p_0(\mathbf{r}')$ can be expressed using Green's function as[50,51]

$$p(\mathbf{r}, t) = \frac{1}{4\pi v_s^2}\frac{\partial}{\partial t}\left[\frac{1}{v_s t}\int d\mathbf{r}'\, p_0(\mathbf{r}')\,\delta\left(t - \frac{|\mathbf{r} - \mathbf{r}'|}{v_s}\right)\right], \tag{2}$$

where $\mathbf{r}$ is the observation position, $\mathbf{r}'$ is the source position, v_s is the speed of sound, and $p_0(\mathbf{r}')$ is the initial pressure distribution.

For a compact, spherically symmetric PA source element with radial initial pressure distribution $p_0(r)$, the analytical solution of the PA wave equation can be further written as

$$p(r, t) = \frac{1}{2r}\left[(r + v_s t)p_0(r + v_s t) + (r - v_s t)p_0(|r - v_s t|)\right], \tag{3}$$

where r is the propagation distance (the detailed derivation is provided in Supplementary Note 1). Under the far field condition, where the propagation distance is larger than the characteristic source size, the inward-propagating component becomes negligible and the detected signal is dominated by the outward-propagating component:

$$p(r, t) \approx \frac{1}{2r}(r - v_s t)p_0(|r - v_s t|). \tag{4}$$

In this work, each voxel is modeled as a local Gaussian source element which has been demonstrated to be effective for 3D PA reconstruction[52], $p_0(r) = p_c \exp(-r^2/2\sigma^2)$.

Substituting this profile into the far field expression gives

$$p(r,t) \approx \frac{p_c}{2r}(r - v_s t) \exp\left[-\frac{(r - v_s t)^2}{2\sigma^2}\right], \tag{5}$$

which motivates the Gaussian radiation kernel $\mathcal{G}$ used in Eq. (1). The detailed derivation from the photoacoustic wave equation to the far field spherically symmetric solution, the Gaussian ball source approximation, and the GPU-accelerated far field discrete model used in our engine is provided in Supplementary Note 1.

In the discrete implementation, each sensor's temporal signal is the superposition of contributions from all independent Gaussian radiation sources according to its propagation distance and initial pressure value. This formulation exposes two natural parallel dimensions, sensor index and temporal index, and is therefore well suited for GPU execution. The forward and backward propagation operators are implemented as custom OpenAI Triton kernels and wrapped by `torch.autograd.Function`. As a result, gradients can be propagated not only to the source intensity $\mathbf{P}_c$ for image reconstruction, but also to the continuous sensor coordinates $\mathbf{x}_s$ for differentiable sensor localization and global pose refinement.

The main PA-SfM optimization workflow is summarized in Algorithm 1, with multi-view extension implemented by sequentially repeating the pose recovery and joint reconstruction steps for additional views.

Stage 1: Initial reference reconstruction

The first stage is to establish a structural reference map (coarse model) using data acquired at a known, fixed pose (Pose 1 or View 1). Given the known sensor array coordinates $\mathbf{X}_1$ and the corresponding detected temporal signals $\mathbf{S}_1$, the initial source distribution $\mathbf{P}_c$ is initialized on a voxel grid. The reconstruction is formulated as an inverse optimization problem:

$$\hat{\mathbf{P}}_c = \arg\min_{\mathbf{P}_c} \left(\|\mathcal{F}(\mathbf{P}_c, \mathbf{X}_1) - \mathbf{S}_1\|_2^2 + \lambda \mathcal{L}_{\mathrm{TGV}}(\mathbf{P}_c) \right) \tag{6}$$

where $\mathcal{F}$ represents the forward simulation engine, and $\mathcal{L}_{\mathrm{TGV}}$ is the total generalized variation (TGV) regularization term utilized to preserve structural edges while suppressing background noise. The objective function is iteratively minimized using the Adam optimizer, yielding the initial 3D anatomical structure serving as the reference map.

Stage 2: Differentiable single sensor localization

To estimate the unknown sensor pose (Pose 2 or View 2), we independently localize each array element using the reference map $\hat{\mathbf{P}}_c$ obtained in Stage 1. By fixing $\hat{\mathbf{P}}_c$, the individual sensor coordinates $\mathbf{x}_{s,i}$ become the optimizable parameters. The localization

is driven by a negative correlation (NC) loss:

$$\mathcal{L}_{\mathrm{NC}}(\mathbf{x}_{s,i}) = -\frac{\mathrm{Cov}(\mathcal{F}(\hat{\mathbf{P}}_c, \mathbf{x}_{s,i}), \mathbf{S}_{2,i})}{\sigma_{\mathrm{sim}}\sigma_{\mathrm{meas}}} \tag{7}$$

where σ_{sim} and σ_{meas} denote the standard deviations of the simulated and measured signals, respectively.

To circumvent local minima, a coarse-to-fine search strategy is employed, as demonstrated in Fig. 1(c):

1) **Global coarse search:** The target space is discretized into grids, and candidate locations are ranked based on $\mathcal{L}_{\mathrm{NC}}$ to select the Top-K priors.
2) **Gradient-based refinement:** Starting from the candidate priors, gradient descent is utilized to optimize the (x, y, z) coordinates.
3) **Dynamic smoothing:** The Gaussian kernel width σ in Eq. (1) is dynamically decayed (from large to small) during optimization, progressively reducing spatial smoothing.

This stage outputs a set of independently predicted, albeit noisy, sensor coordinates $\tilde{\mathbf{X}}_2$.

Stage 3: Robust rigid array estimation

The reference map $\mathbf{P}_c$ obtained in Stage 1 contains noise and many artifacts due to sparsity, leading to errors in sensor localization in Stage 2. To suppress these errors, the inherent geometric constraints of the rigid sensor array are leveraged. Given the predefined geometric template of the array $\mathbf{X}_{\mathrm{template}} \in \mathbb{R}^{3\times N_e}$ (where N_e is the number of sensor elements), the task is transformed into finding the optimal global rotation matrix $\mathbf{R} \in SO(3)$ and translation vector $\mathbf{t} \in \mathbb{R}^3$.

A modified random sample consensus (RANSAC)[53] algorithm is implemented. A geometric consistency pre-check reduces the number of candidate subsets: randomly sampled point sets are evaluated based on their pairwise edge distances before performing singular value decomposition (SVD). For valid subsets, the Kabsch algorithm[54] is applied to solve the rigid transformation. This stage yields a corrected global array coordinate set $\mathbf{X}_2^{\mathrm{corr}}$ and identifies a subset of inlier sensors $\mathcal{I}$, whose relative positions satisfy the rigid array constraint within a preset small error tolerance.

Stage 4: Global differentiable fine tuning

To further enhance spatial precision, a `RigidArrayOptimizer` module is constructed. The optimization variables are globally parameterized as Euler angles $\boldsymbol{\theta}$ and the translation vector $\mathbf{t}$ (Fig. 1(d)), ensuring that the array strictly adheres to rigid body kinematics: $\mathbf{X}_2(\boldsymbol{\theta}, \mathbf{t}) = \mathbf{R}(\boldsymbol{\theta})\mathbf{X}_{\mathrm{template}} + \mathbf{t}$.

We introduce an inlier-masked gradient flow mechanism. The global loss is computed exclusively on the signals from the inlier set $\mathcal{I}$ (identified in Stage 3):

$$\mathcal{L}_{\mathrm{global}}(\boldsymbol{\theta}, \mathbf{t}) = \sum_{i\in\mathcal{I}} \mathcal{L}_{\mathrm{NC}}\left(\mathbf{R}(\boldsymbol{\theta})\mathbf{x}_{\mathrm{template},i} + \mathbf{t}\right) \tag{8}$$

Although the loss uses only the inlier signals, its gradients update the global parameters $(\boldsymbol{\theta}, \mathbf{t})$. The resulting rigid transformation is applied to all sensors. The contribution of this stage is evaluated in the simulation and in vivo studies (Figs. 8 and 7; Supplementary Fig. 2). The detailed analysis of the intermediate coarse pose estimation and global differentiable pose fine tuning is provided in Supplementary Note 3 and Supplementary Note 2.

Stage 5: Joint dual view reconstruction

In the final stage, the data from the known Pose 1 $(\mathbf{X}_1, \mathbf{S}_1)$ and the calibrated Pose 2 $(\mathbf{X}_2^{\text{fine}}, \mathbf{S}_2)$ are combined to form an augmented synthetic aperture. The fused dual view data can be jointly reconstructed using either back-projection reconstruction or model-based iterative reconstruction methods.

Stage 6: Joint multi-view reconstruction

For scenarios involving more than two imaging views, the framework adopts a sequential pose chain calibration strategy to recover all sensor poses within a unified global coordinate system. The coordinate system of Pose 1 is defined as the global reference frame. After Pose 2 is localized and refined through Stages 2–4, its estimated rigid transformation is used to map the Pose 2 sensor array into the Pose 1 coordinate frame.

For each subsequent unknown view, the pose is estimated relative to the previously calibrated view rather than independently in isolation. Specifically, the relative rigid transformation between Pose k and Pose $k-1$ is first recovered using the same differentiable localization, robust rigid estimation, and global fine tuning pipeline described in Stages 2–4. The resulting relative transformation is then compounded with the accumulated transformations from previous stages, thereby expressing Pose k in the original Pose 1 coordinate system:

$$\mathbf{T}_{1\leftarrow k} = \mathbf{T}_{1\leftarrow 2}\mathbf{T}_{2\leftarrow 3}\cdots\mathbf{T}_{k-1\leftarrow k}, \tag{9}$$

where $\mathbf{T}_{a\leftarrow b} \in SE(3)$ denotes the rigid transformation that maps coordinates from the Pose b frame to the Pose a frame.

By recursively chaining these relative transformations, all sensor arrays from multiple views are consistently registered into the initial Pose 1 coordinate frame. The calibrated coordinates from all poses are then combined to form a large synthetic sensing aperture:

$$\mathbf{X}_{\text{multi}} = \left[\mathbf{X}_1, \mathbf{T}_{1\leftarrow 2}\mathbf{X}_2, \mathbf{T}_{1\leftarrow 3}\mathbf{X}_3, \ldots, \mathbf{T}_{1\leftarrow M}\mathbf{X}_M\right], \tag{10}$$

where M is the total number of imaging views. Joint reconstruction then uses all calibrated sensor coordinates and their corresponding measurements.

Algorithm 1. Differentiable photoacoustic structure from motion (PA-SfM)

Input: Known Pose 1 data: $\mathbf{X}_1$, $\mathbf{S}_1$; Unknown Pose 2 signals: $\mathbf{S}_2$; Rigid array template: $\mathbf{X}_{\text{template}}$; Simulation engine $\mathcal{F}$ with TGV weight λ.
Output: High-resolution joint 3D volume $\mathbf{P}_{\text{final}}$; Fine tuned Pose 2 array coordinates $\mathbf{X}_2^{\text{fine}}$.

Stage 1: Initial reference reconstruction
Initialize source distribution $\mathbf{P}_c$ on a 3D grid.
while not converged **do**
  $\hat{\mathbf{P}}_c \leftarrow \arg\min_{\mathbf{P}_c} \left(\|\mathcal{F}(\mathbf{P}_c, \mathbf{X}_1) - \mathbf{S}_1\|_2^2 + \lambda \mathcal{L}_{\text{TGV}}(\mathbf{P}_c) \right)$
end while

Stage 2: Differentiable single sensor localization
Initialize empty noisy coordinate set $\tilde{\mathbf{X}}_2 \leftarrow \emptyset$
for each sensor $i = 1, 2, \ldots, N$ in Pose 2 **do**
  $\mathbf{x}_{s,i}^0 \leftarrow$ Top-K candidates from global coarse search using $\mathcal{L}_{\text{NC}}$
  for $\sigma = \sigma_{\max}$ down to $\sigma_{\min}$ **do** ▷ Dynamic Gaussian smoothing
    $\mathbf{x}_{s,i} \leftarrow \mathbf{x}_{s,i} - \eta \cdot \nabla_{\mathbf{x}_{s,i}} \mathcal{L}_{\text{NC}} \left(\mathcal{F}(\hat{\mathbf{P}}_c, \mathbf{x}_{s,i}), \mathbf{S}_{2,i}; \sigma \right)$
  end for
  $\tilde{\mathbf{X}}_2 \leftarrow \tilde{\mathbf{X}}_2 \cup \{\mathbf{x}_{s,i}\}$
end for

Stage 3: Robust rigid array estimation
$(\mathbf{R}_{\text{init}}, \mathbf{t}_{\text{init}}, \mathcal{I}) \leftarrow$ Modified RANSAC$(\tilde{\mathbf{X}}_2, \mathbf{X}_{\text{template}})$
▷ $\mathcal{I}$ is the robust inlier subset passing geometric consistency pre-check

Stage 4: Global differentiable fine tuning
Initialize Euler angles $\boldsymbol{\theta}$ from $\mathbf{R}_{\text{init}}$, and translation $\mathbf{t} \leftarrow \mathbf{t}_{\text{init}}$
while not converged **do**
  $\mathcal{L}_{\text{global}} \leftarrow \sum_{i \in \mathcal{I}} \mathcal{L}_{\text{NC}} \left(\mathbf{R}(\boldsymbol{\theta}) \mathbf{x}_{\text{template},i} + \mathbf{t} \right)$ ▷ Inlier-masked loss
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{global}}$ ▷ Update global rotation
  $\mathbf{t} \leftarrow \mathbf{t} - \alpha \cdot \nabla_{\mathbf{t}} \mathcal{L}_{\text{global}}$ ▷ Update global translation
end while
$\mathbf{X}_2^{\text{fine}} \leftarrow \mathbf{R}(\boldsymbol{\theta}) \mathbf{X}_{\text{template}} + \mathbf{t}$ ▷ Kinematic constraint applied to ALL sensors

Stage 5: Joint multi-view reconstruction
$\mathbf{X}_{\text{joint}} \leftarrow \mathbf{X}_1 \cup \mathbf{X}_2^{\text{fine}}$
$\mathbf{S}_{\text{joint}} \leftarrow \mathbf{S}_1 \cup \mathbf{S}_2$
$\mathbf{P}_{\text{final}} \leftarrow \left(\mathbf{X}_{\text{joint}}, \mathbf{S}_{\text{joint}} \right)$
return $\mathbf{P}_{\text{final}}$, $\mathbf{X}_2^{\text{fine}}$

Expanded FOV simulations and image-based comparison

The expanded FOV simulations used three vascular volumes sampled at 0.4 mm and hemispherical arrays of 30 mm radius. The 32- and 64-detector arrays were spatially distributed subsets of the 128-detector array. For each sample, the same continuous trajectory was sampled at four scan speeds, with translation and rotation in three dimensions. A spatially varying illumination window moved with the array. Acquisition

geometry and illumination are detailed in Supplementary Note 4.

Image-based registration used vesselness filtering and skeletonization, followed by rigid iterative closest point registration and normalized cross-correlation refinement. Each new volume was registered to the preceding panorama, initialized by the previous estimated pose, and incorporated using weights inversely proportional to distance from the volume center. For the final reconstruction comparison, both sets of estimated poses were applied to the same acoustic data with a common back-projection operator, grid and evaluation mask. GT pose reconstructions used the true poses with the same measurements and reconstruction settings.

Pose error evaluation

For a sequence of M frames, pose error was calculated as

$$E_{\text{pose}} = \left[\frac{1}{27(M-1)} \sum_{k=2}^{M} \sum_{j=1}^{27} \left\| \hat{\mathbf{R}}_k \mathbf{q}_j + \hat{\mathbf{t}}_k - \left(\mathbf{R}_k^{\text{GT}} \mathbf{q}_j + \mathbf{t}_k^{\text{GT}} \right) \right\|_2^2 \right]^{1/2}, \tag{11}$$

where $\mathbf{q}_j$ are fixed control points distributed over the local imaging volume, and $(\hat{\mathbf{R}}_k, \hat{\mathbf{t}}_k)$ and $(\mathbf{R}_k^{\text{GT}}, \mathbf{t}_k^{\text{GT}})$ denote the estimated and true poses. The first frame defined the common reference and was excluded from the error calculation. No subsequent alignment to ground truth was performed. Means across conditions were calculated from their individual sequence errors. Reconstruction metrics and the control point coordinates are provided in Supplementary Note 4.

Data availability

The data supporting the findings of this study are provided within the article and its Supplementary Information. The other datasets generated during the study are available for research purposes from the corresponding authors on reasonable request.

Code availability

The source code and the continuous 10-frame dataset for the example in Fig. 3(c) are publicly available at https://github.com/JaegerCQ/PA-SfM; usage instructions are provided in Supplementary Note 7.

References

1. Park, J. *et al.* Clinical translation of photoacoustic imaging. *Nature Reviews Bioengineering* **3**, 193–212 (2025).

2. Wang, L. V. & Hu, S. Photoacoustic tomography: in vivo imaging from organelles to organs. *Science* **335**, 1458–1462 (2012).

3. Assi, H. *et al.* A review of a strategic roadmapping exercise to advance clinical translation of photoacoustic imaging: from current barriers to future adoption. *Photoacoustics* **32**, 100539 (2023).

4. Deán-Ben, X. L., Gottschalk, S., Mc Larney, B., Shoham, S. & Razansky, D. Advanced optoacoustic methods for multiscale imaging of in vivo dynamics. *Chemical Society Reviews* **46**, 2158–2198 (2017).

5. Lin, L. & Wang, L. V. The emerging role of photoacoustic imaging in clinical oncology. *Nature Reviews Clinical Oncology* **19**, 365–384 (2022).

6. Ntziachristos, V. Addressing unmet clinical need with optoacoustic imaging. *Nature Reviews Bioengineering* **3**, 182–184 (2025).

7. Matsumoto, Y. *et al.* Label-free photoacoustic imaging of human palmar vessels: a structural morphological analysis. *Scientific Reports* **8**, 786 (2018).

8. Matsumoto, Y. *et al.* Visualising peripheral arterioles and venules through high-resolution and large-area photoacoustic imaging. *Scientific Reports* **8**, 14930 (2018).

9. Ivankovic, I., Merčep, E., Schmedt, C.-G., Deán-Ben, X. L. & Razansky, D. Real-time volumetric assessment of the human carotid artery: handheld multispectral optoacoustic tomography. *Radiology* **291**, 45–50 (2019).

10. Deán-Ben, X. L. & Razansky, D. Portable spherical array probe for volumetric real-time optoacoustic imaging at centimeter-scale depths. *Optics Express* **21**, 28062–28071 (2013).

11. Nagae, K. *et al.* Real-time 3D photoacoustic visualization system with a wide field of view for imaging human limbs. *F1000Research* **7**, 1813 (2018).

12. Kim, W., Choi, W., Ahn, J., Lee, C. & Kim, C. Wide-field three-dimensional photoacoustic/ultrasound scanner using a two-dimensional matrix transducer array. *Optics Letters* **48**, 343–346 (2023).

13. Piras, D., Xia, W., Steenbergen, W., Van Leeuwen, T. G. & Manohar, S. Photoacoustic imaging of the breast using the Twente photoacoustic mammoscope: present status and future perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* **16**, 730–739 (2010).

14. Heijblom, M. *et al.* Visualizing breast cancer using the Twente photoacoustic mammoscope: what do we learn from twelve new patient measurements? *Optics Express* **20**, 11582–11597 (2012).

15. Li, S. *et al.* Photoacoustic imaging of peripheral vessels in extremities by large-scale synthetic matrix array. *Journal of Biomedical Optics* **29**, S11519 (2024).

16. Wang, Y. & Li, C. Comprehensive framework of GPU-accelerated image reconstruction for photoacoustic computed tomography. *Journal of Biomedical Optics* **29**, 066006 (2024).

17. Wray, P., Lin, L., Hu, P. & Wang, L. V. Photoacoustic computed tomography of human extremities. *Journal of Biomedical Optics* **24**, 026003 (2019).

18. Wang, M. *et al.* Functional photoacoustic/ultrasound imaging for the assessment of breast intraductal lesions: preliminary clinical findings. *Biomedical Optics Express* **12**, 1236–1246 (2021).

19. Han, T. *et al.* A three-dimensional modeling method for quantitative photoacoustic breast imaging with handheld probe. *Photoacoustics* **21**, 100222 (2021).

20. Sun, Y., Wang, Y., Li, W. & Li, C. Real-time dual-modal photoacoustic and fluorescence small animal imaging. *Photoacoustics* **36**, 100593 (2024).

21. Choi, S. & Ferrara, K. W. Deep learning for scaling large-aperture photoacoustic computed tomography: From single fingers to the human hand. *Ultrasonics* **165**, 108081 (2026).

22. Jeong, H. *et al.* A hybrid diffusion model enhances multiparametric 3D photoacoustic computed tomography. *Advanced Science* **13**, e13624 (2026).

23. Wang, X. *et al.* Cross-regional real-time visualization of systemic physiology and dynamics with 3D panoramic photoacoustic computed tomography (3D-PanoPACT). *Nature Communications* **16**, 10077 (2025).

24. Zhang, Y. *et al.* Rotational ultrasound and photoacoustic tomography of the human body. *Nature Biomedical Engineering* (2026). URL https://doi.org/10.1038/s41551-025-01603-5.

25. Dou, Y., Mu, F., Li, Y. & Varghese, T. Sensorless end-to-end freehand three-dimensional ultrasound reconstruction with physics guided deep learning. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **71**, 1514–1525 (2024).

26. Lee, C., Kim, C. & Park, B. Review of three-dimensional handheld photoacoustic and ultrasound imaging systems and their applications. *Sensors* **23**, 8149 (2023).

27. Lasso, A. *et al.* PLUS: open-source toolkit for ultrasound-guided intervention systems. *IEEE Transactions on Biomedical Engineering* **61**, 2527–2537 (2014).

28. Frey, B. N. *et al.* UltraFlex: Iterative model-based ultrasonic flexible-array shape calibration. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **72**, 1462–1475 (2025).

29. El Hadramy, S., Verde, J., Beaudet, K.-P., Padoy, N. & Cotin, S. *Trackerless volume reconstruction from intraoperative ultrasound images. International conference on medical image computing and computer-assisted intervention*, 303–312 (Springer, 2023).

30. Li, Q. *et al. Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5 (IEEE, 2023).

31. He, W. *et al.* Trackerless 3D ultrasound volume reconstruction from 2D freehand scans using a hybrid transformer–CNN framework. *Medical Physics* **52**, e70103 (2025).

32. Blessberger, H. & Binder, T. Two dimensional speckle tracking echocardiography: basic principles. *Heart* **96**, 716–722 (2010).

33. Mondillo, S. *et al.* Speckle-tracking echocardiography: a new technique for assessing myocardial function. *Journal of Ultrasound in Medicine* **30**, 71–83 (2011).

34. Glatz, T., Scherzer, O. & Widlak, T. Texture generation for photoacoustic elastography. *Journal of Mathematical Imaging and Vision* **52**, 369–384 (2015).

35. Lee, S. *et al.* Enhancing free-hand 3-D photoacoustic and ultrasound reconstruction using deep learning. *IEEE Transactions on Medical Imaging* **44**, 4652–4665 (2025).

36. Lee, S. *et al. wBCAM: Windowed bi-directional cross-attention with Mamba for enhanced 3D vascular reconstruction in free-hand photoacoustic and ultrasound imaging. Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, Vol. 16888 of *Lecture Notes in Computer Science* (Springer Nature Switzerland, 2026). URL https://papers.miccai.org/miccai-2026/1152-Paper3737.html.

37. Nitkunanantharajah, S., Hennersperger, C., Deán-Ben, X. L., Razansky, D. & Navab, N. Trackerless panoramic optoacoustic imaging: a first feasibility evaluation. *International Journal of Computer Assisted Radiology and Surgery* **13**, 703–711 (2018).

38. Knauer, N., Deán-Ben, X. L. & Razansky, D. Spatial compounding of volumetric data enables freehand optoacoustic angiography of large-scale vascular networks. *IEEE Transactions on Medical Imaging* **39**, 1160–1169 (2020).

39. Liu, X. *et al.* Handheld embedded laser-diode illumination optoacoustic system for real-time 3D angiography of deep tissues. *Laser & Photonics Reviews* **19**, e01006 (2025).

40. Chen, J., Zhang, Y., Zhu, J., Tang, X. & Wang, L. Freehand scanning photoacoustic microscopy with simultaneous localization and mapping. *Photoacoustics* **28**, 100411 (2022).

41. Xin, H. *et al.* Robust hybrid feature-driven on-the-fly mapping enables freehand 3D panoramic photoacoustic angiography. *Light: Science & Applications* **15**, 328 (2026).

42. Choi, S. *et al.* Deep learning enhances multiparametric dynamic volumetric photoacoustic computed tomography in vivo (DL-PACT). *Advanced Science* **10**, 2202089 (2023).

43. Choi, W. *et al.* Recent advances in contrast-enhanced photoacoustic imaging: overcoming the physical and practical challenges. *Chemical Reviews* **123**, 7379–7419 (2023).

44. Kim, J. *et al.* 3D multiparametric photoacoustic computed tomography of primary and metastatic tumors in living mice. *ACS Nano* **18**, 18176–18190 (2024).

45. Kim, J. *et al.* Deep learning acceleration of multiscale superresolution localization photoacoustic imaging. *Light: Science & Applications* **11**, 131 (2022).

46. Choi, S. *et al.* Photoacoustic computed tomography monitors cerebrospinal fluid dynamics and glymphatic function. *Nature Communications* **17**, 2677 (2026).

47. Park, J. *et al.* Clinical ultrasound, photoacoustic, and fluorescence image-guided lymphovenous anastomosis microsurgery via a transparent ultrasound transducer array. *Nature Communications* **16**, 9853 (2025).

48. Yang, J. *et al.* Multiplane spectroscopic whole-body photoacoustic computed tomography of small animals in vivo. *Laser & Photonics Reviews* **19**, 2400672 (2025).

49. Kim, J. *et al.* Non-invasive photoacoustic cerebrovascular monitoring of early-stage ischemic strokes in vivo. *Advanced Science* **12**, 2409361 (2025).

50. Xu, M. & Wang, L. V. Universal back-projection algorithm for photoacoustic computed tomography. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **71**, 016706 (2005).

51. Xu, M. & Wang, L. V. Photoacoustic imaging in biomedicine. *Review of Scientific Instruments* **77**, 041101 (2006).

52. Li, S. *et al.* SlingBAG: point cloud-based iterative algorithm for large-scale 3D photoacoustic imaging. *Nature Communications* **17**, 128 (2026).

53. Chum, O., Matas, J. & Kittler, J. *Locally optimized RANSAC. Joint pattern recognition symposium*, 236–243 (Springer, 2003).

54. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **32**, 922–923 (1976).

Acknowledgements

This research was supported by the National Key R&D Program of China (No. 2023YFC2411700, No. 2017YFE0104200); the Beijing Natural Science Foundation (No. 7232177); the National Natural Science Foundation of China (62472213); and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03047902).

We thank Zhibo Xiao, Tingting Huang and Boxin Yao for their assistance.

Author contributions

S.L. and J.G. conceived and designed the study and the algorithms. C.K., S.C., H.H., X.W. and J.S. provided the *in vivo* data. Q.C., Y.W., S.W. and Y.Z. contributed to the simulation experiments. Y.Z. contributed to the metrics calculation. C.L. and Y.Y. supervised the study. All of the authors contributed to writing the paper.

Competing interests

All authors declare no competing interests.

Ethics statement

The human study has passed the ethical approval of the Peking University Institutional Review Board.

Animal Ethics declaration: No new animal experiments were conducted for this study. All animal imaging analyses used previously acquired datasets. The mouse imaging data were obtained from the previously published 3D-PanoPACT study by Wang et al.[23], and the rat kidney and liver imaging data were provided by Professor Kim's laboratory[42].

Consent to Participate declaration: The first author, Shuang Li, voluntarily participated in the human hand imaging experiments and provided informed consent before data acquisition, with an understanding of the experimental procedures and associated risks.

Supplementary information

Supplementary Notes 1 to 7
Supplementary Figs. 1 to 9
Supplementary Tables 1 to 4
Supplementary Movies 1 to 3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMovie1.mp4
- SupplementaryMovie2.mp4
- SupplementaryMovie3.mp4
- SupplementaryInformation.docx