wBCAM: Windowed Bi-directional Cross-Attention with Mamba for Enhanced 3D Vascular Reconstruction in Free-Hand Photoacoustic and Ultrasound Imaging

SiYeoul Lee[1], SeongKyu Park[1], MinKyung Seo[1], EonSeung Seong[1], DongEon Lee[1], Doyun Kim[1], Sehoon Kim[1], DoYeon Kim[1], MinJu Jin[1], Hyeyun Lee[2], GyuBin Kim[1], and MinWoo Kim[3,4]

[1] Department of Information Convergence Engineering, College of Information and Biomedical Convergence Engineering, Pusan National University, Yangsan, Korea
[2] Department of Radiology, Pusan National University, Yangsan, Korea
[3] School of Biomedical Convergence Engineering, College of Information and Biomedical Engineering, Pusan National University, Yangsan, Korea
[4] Center for Artificial Intelligence Research, Pusan National University, Busan, Korea
mkim180@pusan.ac.kr

Abstract. Integrated photoacoustic and ultrasound (PAUS) imaging is promising for clinical translation because it enables simultaneous acquisition of conventional ultrasound (US) and photoacoustic (PA) images in a unified setup. However, 3D vascular reconstruction in hand-held PAUS remains challenging without reliable scan trajectory estimation, while external tracking devices increase cost and complexity. We propose a sensor-free free-hand PAUS framework based on windowed Bi-directional Cross-Attention with Mamba (wBCAM) for scan motion estimation and 3D PA vascular reconstruction. wBCAM models long-range scan dynamics by combining windowed bi-directional cross-attention with Mamba-based global memory for temporally consistent motion estimation. The proposed method is validated on dynamic in-vivo human forearm datasets. Quantitative and qualitative results demonstrate improved long-range motion estimation and stable high-resolution 3D vascular reconstruction. The source code is publicly available at (https://github.com/pnu-amilab/wBCAM).

Keywords: Photoacoustic Ultrasound · Free-hand Imaging · 3D Vessel Visualization · Motion Estimation

1 Introduction

Photoacoustic (PA) imaging is a promising non-invasive and safe modality for clinical applications [18, 10]. In PA imaging, pulsed optical energy is absorbed by tissue chromophores and converted into ultrasound (US) waves via thermoelastic expansion. Because hemoglobin is a major endogenous absorber, PA imaging

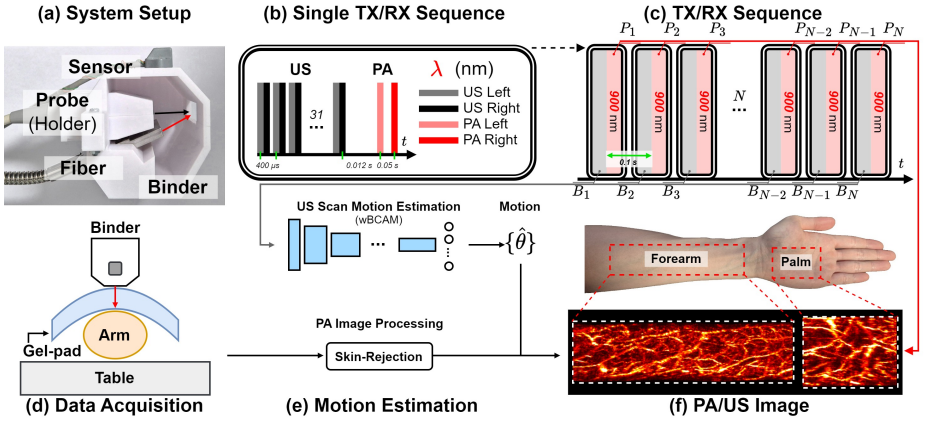

Fig. 1: Overview of the proposed free-hand 3D PAUS framework: (a) PA/US acquisition setup with a custom transducer–fiber holder, (b) TX/RX sequence for a single PA/US frame, (c) TX/RX timing for real-time PA/US imaging, (d) free-hand forearm scanning with a gel pad, (e) sensor-free motion estimation and PA processing for 3D reconstruction, and (f) reconstructed 3D vascular structures.

enables high-contrast visualization of blood microvasculature [21], making it attractive for diagnosing cancer, ischemia, and dermatological diseases [4, 3].

Among PA imaging systems, an integrated photoacoustic and ultrasound (PAUS) framework that combines a conventional clinical US system with a laser system is particularly attractive for clinical translation [12, 11, 8]. PAUS can leverage standard US hardware and workflows while providing complementary PA/US information, and free-hand scanning with a handheld transducer enables flexible real-time examination in routine clinical settings [7].

Despite these advantages, most clinical PAUS implementations remain limited to 2D imaging with a restricted field of view (FoV), which limits comprehensive assessment of vascular morphology and spatial continuity. In contrast, 3D imaging can reduce operator dependency and enable more reliable structural analysis within a region of interest (RoI) [16]. From a translational standpoint, an ideal solution is to obtain 3D images using a general PAUS system without additional tracking hardware. In our preliminary study, we demonstrated the feasibility of free-hand 3D PA imaging by stacking multiple 2D slices acquired during a predominantly elevational sweep, without an external trajectory tracking device [14]. In that framework, transducer motion was estimated solely from US B-mode images and used to reconstruct a 3D PA vascular volume from the corresponding 2D PA images.

However, feasibility alone is not sufficient for clinical deployment. The trajectory estimation accuracy of our preliminary framework remains inadequate for robust clinical use. This issue is particularly critical in PA reconstruction because microvascular structures are highly sensitive to accumulated motion errors; even

small trajectory inconsistencies can disrupt vessel continuity and distort 3D vascular topology. Therefore, improving scan trajectory estimation is a prerequisite for clinically meaningful free-hand 3D PAUS imaging.

Many existing deep learning (DL)-based scan motion estimation methods primarily rely on local frame-to-frame alignment. Although effective for short-term motion cues, such local formulations are vulnerable to cumulative drift over long scan trajectories. For free-hand 3D PAUS reconstruction, preserving structural consistency over extended sequences requires explicit modeling of long-range temporal dependencies.

To address these limitations, we propose a DL framework, windowed Bi-directional Cross-Attention with Mamba (wBCAM), for image-based sensor-free scan motion estimation. wBCAM encodes paired frames into local representations and processes them within overlapping temporal windows. A local-to-global (L2G) cross-attention module writes motion evidence into learnable global tokens that act as structured motion memory, and a Mamba-based state-space module (SSM) [5] updates these tokens across windows to capture long-range temporal dependencies. The refined global context is then fed back to local tokens through global-to-local (G2L) cross-attention for motion-consistent representation learning. By coupling local motion evidence with long-range temporal memory, the proposed framework improves trajectory estimation accuracy and stabilizes 3D PA vascular reconstruction in free-hand PAUS. An overview of the proposed framework is shown in Fig. 1.

2 Methods

2.1 System Setup

We used a programmable US system (Vantage 64LE, Verasonics, USA) with a linear array transducer (L11-5v, Verasonics) and a Q-switched Nd:YAG laser system (Phocus Mobile, OPOTEK, USA) with a linear fiber bundle for real-time PAUS imaging. As shown in Fig. 1(a), the fiber bundle was attached to the transducer using a custom holder, forming a handheld PAUS probe. The illumination angle was empirically optimized to maximize optical fluence coverage over the imaging plane.

The TX/RX sequence for a single PAUS cycle is shown in Fig. 1(b). Because the system supports 64 receive channels, full-aperture receive beamforming required two acquisitions using the left and right halves of the aperture. US B-mode frames $\{\mathbf{B}_i\}_{i=1}^{N}$ were reconstructed using delay-and-sum (DAS) beamforming with 31-angle plane-wave compounding, requiring 62 TX/RX events per frame. After US acquisition, laser pulses at a selected wavelength were fired to acquire PA channel data, which were reconstructed using DAS beamforming into PA monochromatic frames $\{\mathbf{P}_i\}_{i=1}^{N}$ (i.e., initial pressure distribution maps) [13]. The laser system supports a maximum pulse repetition frequency of 20 Hz. With two laser shots per PAUS cycle, the maximum PAUS frame rate was 10 Hz.

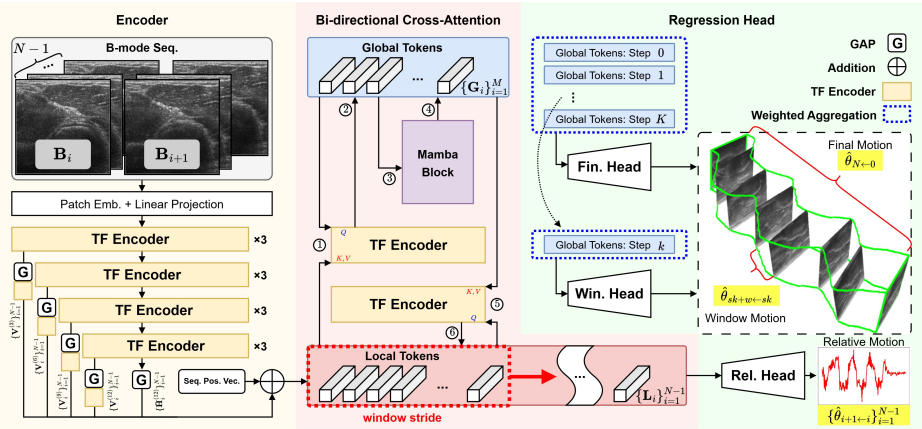

Fig. 2: Architecture of the proposed windowed Bi-directional Cross-Attention with Mamba (wBCAM). Consecutive US B-mode frame pairs are encoded into local tokens, which are refined through windowed local-global bi-directional cross-attention with Mamba-based global memory. Relative-, window-, and sequence-level heads regress 6D scan motions.

2.2 Problem Setup

Let $\{\theta_i = (t_i^x, t_i^y, t_i^z, a_i^x, a_i^y, a_i^z)\}_{i=1}^N$ denote the 6D absolute probe motion vectors, where t and a represent translation and Euler-angle rotation components, respectively. We convert $\{\theta_i\}_{i=1}^N$ into absolute homogeneous transformation matrices $\{\mathbf{T}_i\}_{i=1}^N$, and compute the relative transforms as $\mathbf{T}_{i+1\leftarrow i} = \mathbf{T}_{i+1}\mathbf{T}_i^{-1}$ for $i = 1, \ldots, N-1$. These are then converted into 6D relative motion vectors $\{\theta_{i+1\leftarrow i}\}_{i=1}^{N-1}$. Our goal is to learn a deep regression model $\mathcal{F}(\cdot)$ that predicts the relative motion sequence from the input US B-mode sequence, i.e., $\mathcal{F}(\{\mathbf{B}_i\}_{i=1}^N) \approx \{\theta_{i+1\leftarrow i}\}_{i=1}^{N-1}$.

2.3 Scan Sequence Modeling

The overall architecture of wBCAM is illustrated in Fig. 2. First, consecutive B-mode frame pairs are encoded in parallel using a vision transformer (ViT) encoder [2]. The encoder consists of 12 sequential transformer encoder blocks and captures spatial dependencies through patch-wise self-attention. For the i-th frame pair, the output of the j-th block is denoted by $\mathbf{H}_i^{(j)} \in \mathbb{R}^{(16\times16+1)\times D}$.

To inject temporal context into the encoder in a multi-scale manner, we additionally use skip features from every third transformer block [9, 20]. Specifically, for the j-th transformer block, its compact output $\bar{\mathbf{H}}_i^{(j)} \in \mathbb{R}^D$ is obtained by applying global average pooling (GAP). These compact sequences are processed by each temporal self-attention block to model sequential context, producing temporally enhanced feature token $\mathbf{V}_i^{(j)}$. Finally, the local token is defined as: $\mathbf{L}_i = \bar{\mathbf{H}}_i^{(12)} + \sum_{j\in\{3,6,9,12\}} \mathbf{V}_i^{(j)} + \mathbf{e}^{\text{pos}}$, where $\mathbf{e}^{\text{pos}}$ denotes a positional vector.

We then introduce M learnable global tokens $\{\mathbf{G}_m\}_{m=1}^{M}$ to model long-range scan dynamics via windowed bi-directional cross-attention with Mamba. For each strided temporal window (stride s), local window tokens are first written into the global tokens through a local-to-global (L2G) cross-attention block. The global tokens are then updated by a Mamba block, producing refined global memory tokens at the k-th step. Next, the refined global context is fed back to the local tokens through a global-to-local (G2L) cross-attention block to recalibrate local motion representations. By iterating this process over overlapping windows, wBCAM accumulates long-range motion cues in the global tokens while preserving locally discriminative motion information in the local tokens. Unlike prior separate local/global-stream or dense-feature SSM approaches, wBCAM uses Mamba only to propagate compact global-memory tokens across windows, enabling unified local/global modeling without additional fusion branches.

The recalibrated local tokens are passed to a relative-motion head to estimate relative-level motion $\{\theta_{i+1\leftarrow i}\}_{i=1}^{N-1}$. In parallel, the global tokens at each step are aggregated using a softmax-normalized weighted sum and fed to a global-motion head to estimate the window-level motion $\{\theta_{sk+w\leftarrow sk}\}_{k=1}^{\lfloor (N-w)/s \rfloor+1}$, where w is the window length. Finally, all global tokens are aggregated in the same manner to estimate the sequence-level motion $\theta_{N\leftarrow 1}$. Each prediction head consists of a fully connected layer that maps the corresponding feature representation to a 6D motion vector.

2.4 Loss Functions

To train the model, we use a basic regression loss that combines an ℓ_1 error term with a correlation-aware term [6]. Using cosine similarity $\kappa(\cdot,\cdot)$, the basic loss is defined as

$$\mathcal{L}_{\text{basic}} = \|\hat{\theta} - \theta\|_1 + \big(1 - \kappa(\hat{\theta}, \theta)\big), \tag{1}$$

where $\hat{\theta}$ and θ denote the predicted and ground-truth motion vectors.

During training, we regularize the global tokens to encourage diverse global-context representations. Although using multiple global tokens can improve long-range sequence modeling, they may collapse to similar representations without an explicit diversity constraint. To mitigate this issue, we introduce the following regularization terms:

$$\mathcal{L}_{\text{centering}} = \|\mathbb{E}[\mathbf{G}]\|_2^2, \quad \mathcal{L}_{\text{spreading}} = \frac{1}{|\mathcal{P}|} \sum\nolimits_{(i,j)\in\mathcal{P}} \kappa(\mathbf{G}_i, \mathbf{G}_j), \tag{2}$$

where $\mathbf{G} = \{\mathbf{G}_m\}_{m=1}^{M}$ denotes the set of global tokens and $\mathcal{P}$ is the set of all unordered pairs (i, j) such that $i < j$. The centering term prevents the global-token distribution from drifting toward a biased mean, while the spreading term reduces redundancy among tokens by penalizing pairwise similarity.

The final loss is defined as $\mathcal{L}_{\text{final}} = \lambda_1\mathcal{L}_{\text{basic}} + \lambda_2\mathcal{L}_{\text{centering}} + \lambda_3\mathcal{L}_{\text{spreading}}$, where $\lambda_1 = 1.0$, $\lambda_2 = 10^{-2}$, and $\lambda_3 = 1.0$ are weighting coefficients.

3 Experimental Results

3.1 Data Acquisition

PAUS Scan: We scanned both the right and left forearms of each subject using two linear scans and two S-shaped scans from 19 subjects. The scans were split into training, validation, and test sets with a ratio of 12 : 3 : 4. The split was performed at the subject level, so all scans from each subject were assigned to only one split. In total, 76 scans ($2 \times 2 \times 19$) were used for experiments. The maximum pulsed laser energy was 13.54 mJ/cm^2 at 700 nm, 12.59 mJ/cm^2 at 800 nm, and 9.13 mJ/cm^2 at 900 nm, all of which are below the American National Standards Institute (ANSI) safety limits of 16.83 mJ/cm^2, 25.23 mJ/cm^2, and 25.23 mJ/cm^2 at the corresponding wavelengths [1].

US Scan: To validate motion estimation performance under dynamic scan trajectories, we additionally scanned both the right and left forearms of each subject three times using S-shaped motion from 33 subjects. The scans were split into training, validation, and test sets with a ratio of 20 : 6 : 7, ensuring that the PAUS test subjects and US test subjects did not overlap. This split was also performed at the subject level. In total, 198 scans ($3 \times 2 \times 33$) were used for experiments. These US scans were performed using a holder without a binder. The US-only sequences were acquired at 20 Hz.

For supervised learning, scan motion was recorded using an electromagnetic (EM) sensor (Patriot, Polhemus), which was attached to the binder or holder during PA/US scanning. The recorded trajectories were initialized in the image plane coordinate system. All scans covered the entire forearm (150-200 mm) and were performed in a transverse view. The acquisition and use of human subject data were approved by the local Institutional Review Board (IRB) (PNU IRB/2023 074 HR).

3.2 Implementation Details

For the main comparison, we selected CNN [17], DCL-Net [6], MoGLo-Net [14], DualTrack [19], and FiMA [20], including several state-of-the-art (SOTA) methods. We implemented all comparison models from their public repositories.

All models were implemented using PyTorch and trained using NVIDIA RTX 4090 GPU. Except for DualTrack, models were trained for 4,000 epochs with AdamW (l.r. 1e-4, w.d. 1e-2) and cosine annealing learning rate scheduling. For DualTrack, we followed the training settings provided in its official repository. For wBCAM, we used M=16 global tokens, N=65 frames, w=16, and s=8. During training, random subsequences were sampled from each scan.

3.3 Quantitative & Qualitative Analysis

Table 1 summarizes the quantitative comparison on the US and PAUS datasets. Average error (AE) and cosine similarity (CS) are computed from the predicted

Models	relative Error			absolute Error		
	rAE ↓	rCS ↑	rFE ↓	aAE ↓	aCS ↑	aFE ↓
US Dataset						
CNN [17]	0.05805	0.89263	0.13185	7.56293	0.98507	18.91404
DCL-Net [6]	0.04824	0.93494	0.10007	5.68047	0.99099	14.57179
MoGLo-Net [14]	0.04544	0.93427	0.09574	5.95810	0.98905	12.66530
DualTrack [19]	0.03945	0.95127	0.08375	4.61368	0.99478	12.04894
FiMA [20]	**0.03854**	**0.95132**	**0.08201**	4.40335	0.99483	10.44783
wBCAM (Ours)	0.04124	0.94521	0.09052	**3.93011**	**0.99601**	**9.25238**
PAUS Dataset						
CNN [17]	0.11713	0.78898	0.31618	10.56372	0.97611	28.51080
DCL-Net [6]	0.09366	0.85764	0.26699	6.29142	0.99272	16.59572
MoGLo-Net [14]	0.09120	0.86160	0.25671	4.92504	0.99526	13.24179
DualTrack [19]	0.09273	0.85163	0.26512	4.88800	0.99558	14.85182
FiMA [20]	**0.08543**	**0.88143**	**0.23917**	6.18725	0.99305	17.93673
wBCAM (Ours)	0.09223	0.85283	0.26496	**4.10756**	**0.99600**	**10.99699**
G W R	**Ablation**					
× × ×	0.04276	0.94085	0.09394	4.36441	0.99512	10.42516
✓ × ×	0.04236	0.94203	0.09239	4.04995	0.99587	9.66178
✓ ✓ ×	0.04175	0.94382	0.09065	4.03385	0.99559	9.31113
✓ ✓ ✓	**0.04124**	**0.94521**	**0.09052**	**3.93011**	**0.99601**	**9.25238**

Table 1: Quantitative comparison on US and PAUS datasets and ablations.

6D motion vectors. Frame error (FE) is computed after geometric reconstruction as the ℓ_2 distance between the four corner points of the predicted and ground-truth image planes [15]. Using these metrics, we evaluate both relative and absolute motion estimation performance: relative metrics (rAE, rCS, and rFE) are computed from $\{\theta_{i+1\leftarrow i}\}_{i=1}^{N-1}$, whereas absolute metrics (aAE, aCS, and aFE) are computed from $\{\theta_{i\leftarrow 1}\}_{i=1}^{N}$.

On the US dataset, wBCAM achieves the best performance on all absolute-motion metrics (aAE, aCS, and aFE), while remaining competitive on relative-motion metrics. This trend suggests that the proposed model effectively suppresses error accumulation over long scan sequences, leading to improved long-range trajectory stability.

On the PAUS dataset, the gap between relative and absolute performance becomes more pronounced, likely because the lower frame rate (10 Hz) leads to larger inter-frame motion and weaker short-range temporal continuity. In this setting, methods that strongly exploit local inter-frame correlation (e.g., MoGLo-Net and FiMA) achieve better relative-motion metrics. Nevertheless, wBCAM achieves the best absolute-motion performance (aAE, aCS, and aFE), indicating improved robustness to cumulative drift, which is critical for stable 3D vascular reconstruction.

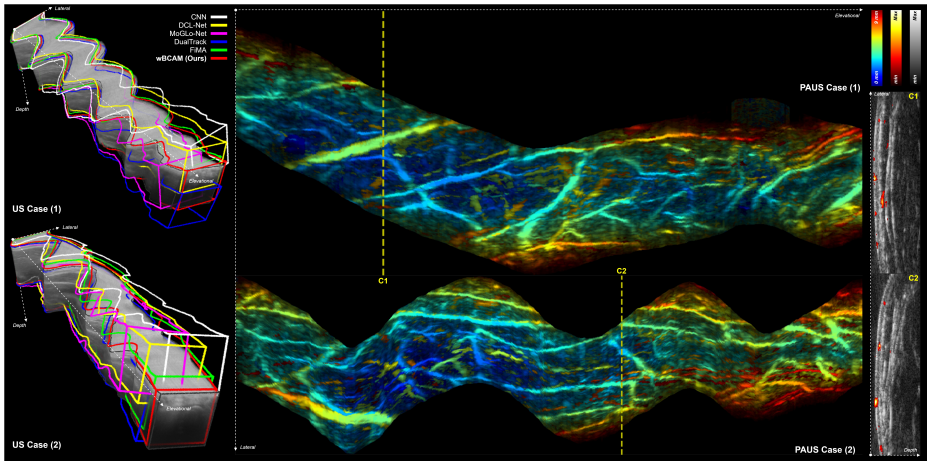

Fig. 3: Qualitative results on the US (left) and PAUS (right) datasets, with two representative cases for each. For the US dataset, reconstructed volumes and estimated scan trajectories are visualized. For the PAUS dataset, wBCAM estimates scan motion from US B-mode frames, which is used to reconstruct 3D vascular structures from the corresponding PA frames after skin-signal suppression and inter-frame consistency enhancement. Reconstructed vessels are rendered using depth color encoding, where color indicates depth and opacity indicates PA signal intensity. Cross-sections of vessels (dotted lines) are overlaid on the corresponding US B-mode frames.

The ablation study further validates each component. Starting from the baseline (encoder + relative head; $\times/\times/\times$), adding global tokens (G) improves both relative and absolute accuracy, confirming the benefit of explicit global context modeling. Adding windowed processing (W) further improves most metrics, and adding global-token regularization (R) yields the best overall performance, showing that the proposed regularization stabilizes long-range scan modeling.

Fig. 3 presents representative qualitative results on the US and PAUS datasets. On the US dataset, wBCAM produces more accurate scan trajectories and more coherent reconstructed volumes than competing methods, indicating reduced cumulative drift. On the PAUS dataset, wBCAM better preserves vascular continuity despite deliberate zig-zag free-hand scanning, resulting in more stable 3D vessel reconstruction.

4 Conclusion

In this study, we achieved sensor-free free-hand 3D PAUS imaging using a DL-based motion estimation framework, wBCAM. The proposed method reconstructs 3D vascular structures from free-hand scans without external tracking sensors, improving practicality while reducing hardware dependency. By

introducing structured global memory through windowed bi-directional cross-attention and a Mamba-based state-space module, wBCAM enables stable long-range motion modeling and suppresses cumulative drift, leading to improved structural consistency in reconstructed 3D volumes.

Extensive experiments on both US and PAUS datasets demonstrate robust accumulated motion estimation and reliable 3D vascular reconstruction. These results suggest that the proposed framework can support clinically practical free-hand PAUS imaging, particularly for applications requiring assessment of superficial vascular structures. In particular, the method may facilitate volumetric evaluation of peripheral perfusion for ischemia-related assessment and may also be useful for low-depth clinical applications, such as thyroid or dermatologic imaging, where PA imaging is well suited.

Acknowledgement. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25491857). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2026-RS-2023-00254177) grant funded by the Korea government(MSIT).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Charschan, S.S., Rockwell, B.A.: Update on ansi z136. 1. In: International Laser Safety Conference. vol. 1999, pp. 48–55. Laser Institute of America (1999)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Ermilov, S.A., Khamapirad, T., Conjusteau, A., Leonard, M.H., Lacewell, R., Mehta, K., Miller, T., Oraevsky, A.A.: Laser optoacoustic imaging system for detection of breast cancer. Journal of biomedical optics **14**(2), 024007–024007 (2009)
4. Favazza, C.P., Jassim, O., Cornelius, L.A., Wang, L.V.: In vivo photoacoustic microscopy of human cutaneous microvasculature and a nevus. Journal of biomedical optics **16**(1), 016015–016015 (2011)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Guo, H., Xu, S., Wood, B., Yan, P.: Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 463–472. Springer (2020)
7. Hajireza, P., Shi, W., Zemp, R.J.: Real-time handheld optical-resolution photoacoustic microscopy. Optics Express **19**(21), 20097–20102 (2011)
8. Harrison, T., Ranasinghesagara, J.C., Lu, H., Mathewson, K., Walsh, A., Zemp, R.J.: Combined photoacoustic and ultrasound biomicroscopy. Optics express **17**(24), 22041–22046 (2009)

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hu, S., Wang, L.V.: Photoacoustic imaging and characterization of the microvasculature. Journal of biomedical optics **15**(1), 011101–011101 (2010)
11. Kim, C., Favazza, C., Wang, L.V.: In vivo photoacoustic tomography of chemicals: high-resolution functional and molecular optical imaging at new depths. Chemical reviews **110**(5), 2756–2782 (2010)
12. Kim, J., Park, S., Jung, Y., Chang, S., Park, J., Zhang, Y., Lovell, J.F., Kim, C.: Programmable real-time clinical photoacoustic and ultrasound imaging system. Scientific reports **6**(1), 35137 (2016)
13. Kim, M., Pelivanov, I., O'Donnell, M.: Review of deep learning approaches for interleaved photoacoustic and ultrasound (paus) imaging. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **70**(12), 1591–1606 (2023)
14. Lee, S., Kim, S., Seo, M., Park, S., Imrus, S., Ashok, K., Lee, D., Park, C., Lee, S., Kim, J., et al.: Enhancing free-hand 3d photoacoustic and ultrasound reconstruction using deep learning. IEEE Transactions on Medical Imaging (2025)
15. Li, Q., Saeed, S.U., Huang, Y., Luo, M., Yan, Z., Chen, J., Yang, X., Ni, D., Winter, N., Nguyen, P., et al.: Tus-rec2024: A challenge to reconstruct 3d freehand ultrasound without external tracker. arXiv preprint arXiv:2506.21765 (2025)
16. Mozaffari, M.H., Lee, W.S.: Freehand 3-d ultrasound imaging: a systematic review. Ultrasound in medicine & biology **43**(10), 2099–2124 (2017)
17. Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., Wein, W.: 3d freehand ultrasound without external tracking using deep learning. Medical image analysis **48**, 187–202 (2018)
18. Steinberg, I., Huland, D.M., Vermesh, O., Frostig, H.E., Tummers, W.S., Gambhir, S.S.: Photoacoustic clinical imaging. Photoacoustics **14**, 77–98 (2019)
19. Wilson, P.F., Ronchetti, M., Göbl, R., Markova, V., Rosenzweig, S., Prevost, R., Mousavi, P., Zettinig, O.: Dualtrack: Sensorless 3d ultrasound needs local and global context. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 3–12. Springer (2025)
20. Yan, Z., Yang, X., Luo, M., Chen, J., Chen, R., Liu, L., Ni, D.: Fine-grained context and multi-modal alignment for freehand 3d ultrasound reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 340–349. Springer (2024)
21. Yao, J., Maslov, K.I., Zhang, Y., Xia, Y., Wang, L.V.: Label-free oxygen-metabolic photoacoustic microscopy in vivo. Journal of biomedical optics **16**(7), 076003–076003 (2011)