This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Structure-Aware Sensorless 3D Ultrasound and Photoacoustic Reconstruction

Yi Huang[1,2], Wenxin Gong[1,2], Yongjian Zhao[4], Li Shen[3], Hengrong Lan[1,2⋆], and Fei Gao[1,2,5*]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, PR China
[2] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, PR China
[3] School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, 211189, China
[4] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, 999770, China
[5] Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, PR China

Abstract. Sensorless freehand 3D ultrasound and photoacoustic reconstruction enables flexible clinical diagnosis, yet conventional visual tracking suffers from severe long-range drift induced by speckle decorrelation and acoustic clutter. To overcome these physical degradations, we propose SCC-Net, a structure-aware channel cross-attention network for stable spatiotemporal tracking. The architecture utilizes a channel decoupling module driven by global spatial statistics for adaptive noise filtering, alongside a cross-attention mechanism capturing adjacent-frame geometric topologies. These components are jointly optimized via a progressive training strategy incorporating a motion-modulated mean absolute error to prioritize highly dynamic regions. Extensive independent experiments on a large-scale ultrasound forearm dataset and a complex photoacoustic palm dataset demonstrate the exceptional generalizability of SCC-Net in suppressing cumulative errors. Notably, compared to previous baselines, the method reduces the final drift rate by 13.72% and absolute trajectory error by 14.29% on the ultrasound dataset. Simultaneously, despite the severe acoustic clutter and intricate reciprocating trajectories of the photoacoustic dataset, the network consistently achieves state-of-the-art trajectory estimation. This enables high-fidelity three-dimensional functional visualization without external hardware, demonstrating exceptional adaptability to highly noisy environments.

Keywords: Structure-aware decoupling · Cross-attention · Sensorless reconstruction · Spatiotemporal tracking.

⋆ Hengrong Lan and Fei Gao are the Co-corresponding Authors, E-mail: lanhengrong@163.com, fgao@ustc.edu.cn.

1 Introduction

Ultrasound (US) and photoacoustic (PA) imaging serve as essential clinical tools because of their real-time, non-invasive characteristics and biochemical mapping capabilities [4, 22]. However, conventional two-dimensional (2D) slices lack the spatial context required for complex surgical guidance and precise lesion assessment. To eliminate the need for expensive and interference-prone external tracking devices, sensorless freehand three-dimensional (3D) reconstruction has become a primary research focus [19]. This technique infers probe poses directly from image sequences, building on the theoretical foundation of classic speckle decorrelation algorithms [3, 21].

Recent progress in deep learning has notably advanced these techniques. Early architectures such as DCL-Net [8] and DC^2-Net [7] use 3D contextual information and contrastive learning to improve the efficiency of feature representation . Subsequent long-range sequence modeling shows substantial promise, as demonstrated by the FiMA network used multi-directional state-space models [17], and the MoGLo-Net relied on global-local self-attention[11]. The latter marks the initial application of 3D reconstruction to the PA domain. Additionally, multimodal approaches suppress cumulative noise by integrating dynamic data from lightweight inertial measurement units (IMUs) to mitigate the scarcity of local features in pure visual tracking [26]. Despite the accuracy improvements provided by multimodal integration, pure visual solutions remain highly desirable because they require no additional hardware and offer broad clinical compatibility.

Nevertheless, existing algorithms continue to face physical degradation issues in practical clinical applications [18]. Current methods typically rely on coarse-grained macro features, overlooking the fine-grained acoustic fingerprints—such as US speckle patterns and PA clutter—that contain critical microscopic motion information. During freehand scanning, complex out-of-plane movements and mechanical compression from the probe induce intense US speckle decorrelation and dynamic PA clutter variations [5, 15]. This physical degradation forces deep feature matching to converge to local optima, which leads to severe trajectory drift over extended sequences [1]. Consequently, physically decoupling these unstable acoustic patterns within deep network layers and maintaining robust feature tracking under dynamic compression represent the primary challenges in sensorless 3D reconstruction [25]. To resolve these issues, we introduce a structure-aware channel cross-attention network that overcomes the limitations of traditional implicit regression by integrating high-level semantic information with stable low-level structural features.

The contributions of our work center on the proposed SCC-Net, a structure-aware reconstruction architecture equipped with a bidirectional channel and cross-attention (BCCA) module. Specifically, it integrates a structure-aware channel decoupling (SACD) module to adaptively filter modality-specific noise, namely decorrelated speckles and acoustic clutter, alongside a cross-attention mechanism that establishes robust spatial correspondences across adjacent frames to suppress motion-induced artifacts. Furthermore, this study demonstrates sen-

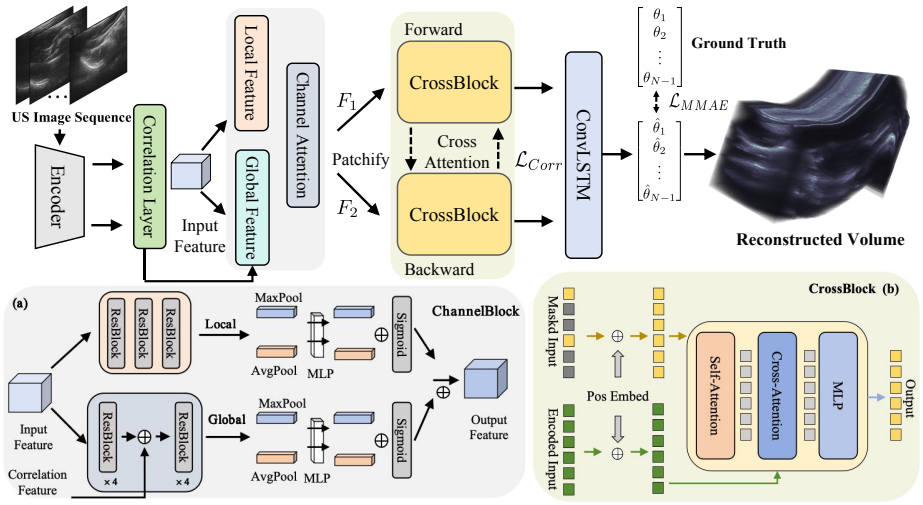

Fig. 1. Overview of the proposed SCC-Net architecture. (a) The ChannelBlock module adaptively filters unstable speckle noise or artifacts through global spatial statistics. (b) The CrossBlock module establishes robust spatio-temporal correspondences to maintain tracking stability under dynamic compression.

sorless 3D reconstruction in the pure PA domain, achieving high-fidelity 3D visualization of large-scale vascular networks without external tracking assistance. Additionally, we develop a sequential self-supervised learning and temporal regularization strategy. This strategy applies the spatio-temporal consistency of bidirectional sequences as a prior constraint to overcome directional biases and improve the robustness of long-range reconstruction.

2 Methods

Assume a scan sequence comprises N US or PA images denoted as $B = \{B_i | i = 1, 2, ..., N\}$. SCC-Net estimates the displacement parameters $\theta = \{\theta_i | i = 1, 2, ...N-1\}$, where each θ_i is decomposed into translation $t_i = (t_x, t_y, t_z)_i^T$ and rotation $\varphi_i = (\varphi_x, \varphi_y, \varphi_z)_i^T$ between consecutive frames B_i and B_{i+1}. Once all transformations between each frame and a reference frame are calculated, scans containing multiple sequences can be reconstructed into a three-dimensional volume. The reference frame can be any image within the scan. Fig. 1 illustrates the proposed framework.

2.1 Structure-Aware Channel Decoupling

During freehand scanning, out-of-plane motion and probe compression alter the spatial distribution of microscopic scatterers in ultrasound and induce acoustic

clutter in PA imaging. If the network uniformly weighs all extracted features when estimating the displacement θ_i, it becomes susceptible to these modality-specific artifacts. This inevitably introduces local matching errors that subsequently accumulate into severe long-range trajectory drift.

To address this, we propose a structure-aware channel decoupling (SACD) module. For intermediate features $F \in \mathbb{R}^{C\times H\times W}$ extracted from the 336×336 input, individual feature channels $F_c \in \mathbb{R}^{H\times W}$ function as implicit acoustic filters. Inspired by channel attention [10, 23], this module utilizes global spatial statistics to adaptively suppress channels capturing unstable artifacts, such as US speckle decorrelation or PA clutter, while preserving stable structural geometries.

We use average pooling F^c_{avg} and max pooling F^c_{max} of the feature tensor F as statistical descriptors.This module employs a shared network to learn physically meaningful rescaling weights $W_{attn} \in \mathbb{R}^{C\times 1\times 1}$, calculated as follows:

$$W_{attn} = \sigma\left(\mathbf{W}_2\delta(\mathbf{W}_1 F_{avg}) + \mathbf{W}_2\delta(\mathbf{W}_1 F_{max})\right) \tag{1}$$

where σ and δ denote Sigmoid and nonlinear activation layers, and $\mathbf{W}_1, \mathbf{W}_2$ are learnable weights. Element-wise modulation yields the decoupled robust features $\tilde{F} = W_{attn} \otimes F$. During optimization, channels representing unstable patterns receive lower gating weights due to abnormal temporal variance from phase decorrelation or clutter fluctuations. Simultaneously, structural channels anchored to stable landmarks, such as fascial interfaces in US or vascular networks in PA imaging, receive higher weights to ensure robust tracking.

2.2 Masked Feature Modeling via Cross-Attention

To accurately capture the invariant global geometric topology associated with probe motion and enhance feature generalization, it is essential to prevent the tracking network from overfitting to localized, domain-specific acoustic textures. Inspired by recent paradigms in asymmetric masked representation learning [9] and cross-view feature completion [24], we introduce a low-ratio masking strategy within the latent feature space, followed by a CrossBlock structure. This design constructs a spatio-temporal feature perception module that forces the network to reconstruct corrupted representations by aggregating cross-frame context, thereby learning highly generalizable spatial correspondences.

Let $F_t \in \mathbb{R}^{L\times D}$ and $F_{ref} \in \mathbb{R}^{L\times D}$ denote the channel-decoupled feature sequences of the current frame and the temporally adjacent reference frame, respectively. We apply a random masking strategy with a low masking ratio $\rho \leq 10\%$. Defining $\mathcal{M} \subset \{1, 2, \ldots, L\}$ as the set of masked indices with $|\mathcal{M}| = \rho L$, the corrupted input features $\tilde{F}_t$ are constructed by retaining the original features F^i_t for $i \notin \mathcal{M}$, and replacing the elements corresponding to $i \in \mathcal{M}$ with a learnable shared mask token $E_{mask} \in \mathbb{R}^D$. This masking mechanism intentionally disrupts trivial local feature smoothing during pose fitting. Consequently, it compels the model to seek semantic support from the global spatial context and establish robust temporal correspondences with F_{ref} through subsequent cascaded self-attention and cross-attention modules [24].

During the auxiliary feature learning phase, the objective is to ensure that the reconstructed features in the masked regions maintain semantic consistency with the original uncorrupted representations. To achieve this, a cosine similarity loss is employed to measure the alignment of the semantic distributions:

$$\mathcal{L}_{Corr} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(1 - \frac{Y_t^i \cdot F_t^i}{\|Y_t^i\|_2 \|F_t^i\|_2}\right) \tag{2}$$

where Y_t^i denotes the predicted feature output by the attention modules, and F_t^i represents the original feature of the current frame prior to masking. Serving as a strict structural regularization, this auxiliary reconstruction task prevents the network from memorizing domain-specific noise, thereby significantly enhancing the generalization capacity of the learned representations against complex motion artifacts.

2.3 Joint Optimization Strategy

We propose a progressive joint optimization strategy that decomposes the network optimization process into two stages: the construction of basic temporal metrics and the alignment of deep motion semantics. This approach achieves stable end-to-end convergence through a weighted loss function.

In the first stage, the objective is to learn a latent feature manifold characterized by temporal smoothness and physical awareness. Inspired by time-contrastive learning frameworks [20], we introduce a temporal triplet loss to constrain the feature space of adjacent frames:

$$\mathcal{L}_{triplet} = \max(0, \|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2 + m) \tag{3}$$

where f_a, f_p, and f_n represent the latent features of the anchor, temporally positive, and temporally negative frames, respectively, and m is the margin.

In the second stage, we design a Motion Modulated Mean Absolute Error (MMAE) to enable the network to adaptively focus on regions with intense motion and large displacements. Drawing inspiration from adaptive weighting mechanisms that prioritize hard examples during training [16], MMAE introduces dynamic weight rescaling based on motion intensity to overcome the limitations of the standard MAE. Given the predicted value $\hat{y}$, the ground truth y, and a local motion weight prior w extracted from the sequence, the loss is defined as:

$$\mathcal{L}_{MMAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \odot \left(|w_i| + s\right)^{\beta} \tag{4}$$

where s is a smoothing factor for numerical stability, β is a hyperparameter controlling sensitivity to the motion prior, and $\odot$ denotes element-wise multiplication. This mechanism applies nonlinear weighting to the residuals via $(|w_i|+s)^{\beta}$, which assigns larger optimization gradients to the network during complex out-of-plane motions and improves trajectory estimation accuracy in highly dynamic scenarios.

Table 1. Quantitative comparison of 3D trajectory reconstruction across Forearm and Palm datasets. In the fifth column, Final Drift (F-Drift) is reported for the Forearm dataset, while Fréchet Distance (F-Dist) is used for the Palm dataset. Note that the final drift rate (FDR) is consistently calculated across both datasets by dividing the Final Drift by the total trajectory length. **Bold** values indicate the best results.

Method	RTE (mm)↓	RRE (deg)↓	ATE (mm)↓	F-Drift/F-Dist (mm)↓	FDR (%)↓
Forearm dataset					
DCL-Net	0.311(0.049)	0.127(0.031)	35.73(8.93)	57.99(14.76)	31.77(7.93)
DC^2-Net	0.302(0.053)	0.118(0.029)	33.52(10.06)	55.37(16.62)	29.77(8.93)
PLPPI	0.210(0.039)	0.085(0.011)	26.54(10.85)	40.78(16.33)	21.47(8.52)
LongTerm	0.203(0.035)	0.080(0.012)	25.84(9.12)	38.92(13.10)	20.61(6.94)
MoGLo	0.198(0.037)	0.075(0.011)	25.12(10.74)	36.13(15.45)	18.88(8.07)
SCC-Net	**0.179**(0.036)	**0.067**(0.012)	**21.53**(8.52)	**31.49**(12.46)	**16.29**(6.64)
Palm dataset					
DCL-Net	0.240(0.080)	0.133(0.062)	38.40(23.10)	64.51(31.51)	44.52(21.37)
DC^2-Net	0.235(0.061)	0.093(0.051)	36.42(21.10)	61.04(30.63)	41.40(22.48)
PLPPI	0.253(0.070)	0.086(0.049)	28.02(13.02)	49.94(24.15)	35.12(19.12)
LongTerm	0.281(0.082)	0.142(0.070)	40.70(24.89)	71.73(33.15)	51.02(19.34)
MoGLo	0.201(0.043)	0.079(0.054)	24.35(14.35)	43.52(25.61)	28.31(17.40)
SCC-Net	**0.197**(0.055)	**0.073**(0.052)	**22.25**(11.34)	**35.51**(21.91)	**25.75**(16.85)

3 Experiments

Data and Implementation. The experimental evaluation utilizes two datasets: public ultrasound (US) forearm data [12, 14] and photoacoustic (PA) palm data, denoted hereafter as the Forearm and Palm datasets, respectively.The US dataset contains transverse and longitudinal scans from 50 subjects, where images are reconstructed at a resolution of 480×640 and a frame rate of 20 frames per second. For network input, these images are cropped to a resolution of 336×336. These scans follow diverse trajectory shapes, specifically straight, C-shape and S-shape, on both forearms of each subject. The data acquisition involves scanning directions both parallel and perpendicular to the forearm, as well as forward and backward movements, resulting in 24 scans per subject and a total of 1200 scan sequences. Due to the unconstrained nature of freehand scanning, the length of the collected sequences varies. On average, each sequence consists of approximately 400 frames, covering a physical trajectory of roughly 180 to 200 mm. The PA dataset comprises scans from the palms of 18 subjects, yielding a total of 249 trajectories recorded at a resolution of 420×420. Due to the reciprocating nature of the scanning process, data for each subject is partitioned into 13 to 14 trajectories for training, and these trajectories are stitched based on their absolute positions during reconstruction. The US and PA datasets are divided into training, validation, and testing sets based on subject identity using ratios of 3:1:1 and 10:4:4, respectively.

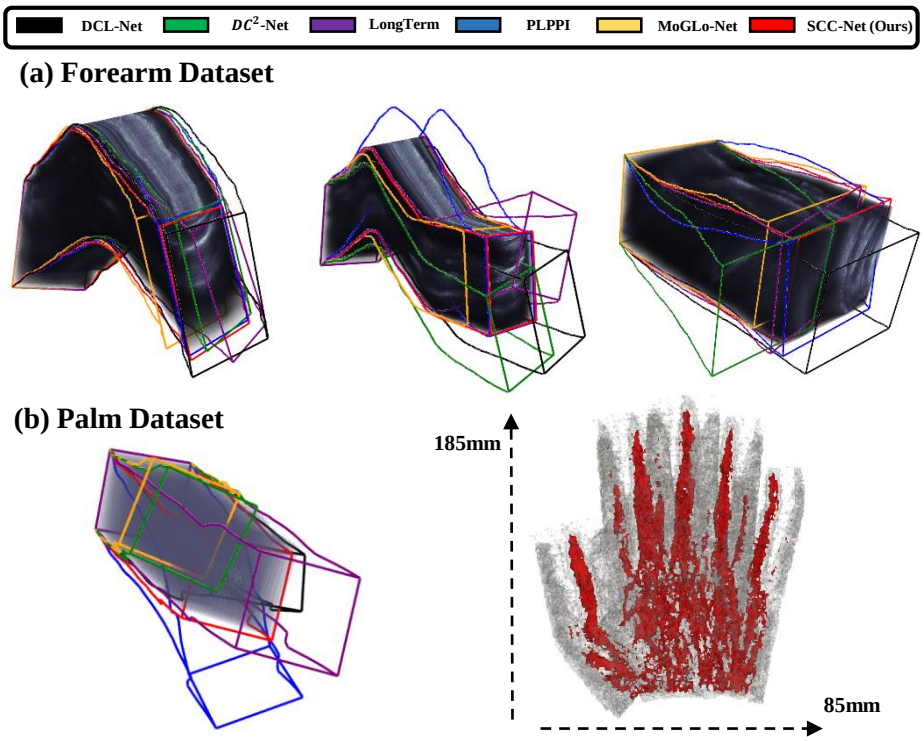

Fig. 2. Qualitative evaluation of 3D reconstruction. (a) Comparison of ultrasound scan volumes reconstructed by the proposed method and baselines against the ground truth under straight, C-shape, and S-shape trajectories. (b) Photoacoustic reconstruction results, featuring a method comparison under a straight trajectory alongside the extracted 3D vascular network of a human palm. Red surfaces denote the reconstructed vessels, and solid lines represent the estimated scanning paths.

To prevent overfitting and enhance robustness, sequence inversion and interval sampling are applied to each scan. The input sequence length is set to 5 frames to ensure baseline performance. During training, each training scan is randomly expanded to 50 sequences, while each test scan is expanded to 10 sequences to simulate complex real-world scenarios. Optimization is performed using the AdamW optimizer with an initial learning rate of 1×10^{-4}, which is reduced by a factor of 0.8 every 50 epochs over a total of 500 epochs. In the curriculum learning phase, the iteration epochs and learning rate are set to 1000 and 1×10^{-5}, respectively. Training was performed on an NVIDIA RTX 4090 GPU with a batch size of 8.

Quantitative and Qualitative Analysis. Performance on the datasets is evaluated using relative translation error (RTE), relative rotation error (RRE),

absolute trajectory error (ATE), and final drift rate (FDR). The FDR is consistently calculated across both datasets by dividing the final drift by the total trajectory length. Additionally, to further evaluate spatial accuracy, dataset-specific metrics are reported. For the Forearm dataset, we report the final drift (F-Drift). Due to the complex reciprocating scanning paths in the Palm dataset, the Fr'echet distance (F-Dist) [2] is employed as a separate metric. This provides a more accurate assessment of spatial curve similarity independent of scanning speed or local pacing.The proposed model is compared against several advanced methods, including DCL-Net [8], DC^2-Net [7], PLPPI [6], LongTerm [13], and MoGLo [11], under identical experimental settings.

Table 1 demonstrates that the proposed SCC-Net achieves the highest accuracy across most evaluation metrics. On the US dataset, the proposed approach yields relative improvements of 13.72% in FDR and 14.29% in ATE over the strong MoGLo baseline. This quantitative superiority is visually corroborated in Fig. 2(a). Under complex scanning paths, particularly the C-shape and S-shape trajectories characterized by intense out-of-plane motions, baseline predictions progressively deviate from the ground truth due to severe speckle decorrelation. In contrast, SCC-Net maintains highly stable spatial representations. By dynamically filtering unstable acoustic features, the network ensures that the predicted paths closely align with the ground truth and effectively suppresses long-range cumulative drift.

For the PA dataset, intricate reciprocating trajectories and severe dynamic clutter increase overall error magnitudes. As Table 1 indicates, conventional architectures struggle under non-linear tissue contrast: DCL-Net and DC^2-Net fail to extract robust spatial features, while LongTerm degrades because temporal correlations cannot resolve rapid transient clutter. In contrast, SCC-Net isolates continuous vascular geometry to achieve state-of-the-art translation accuracy. Although the relative rotation error (RRE) remains comparable to the MoGLo baseline, this reflects the physical sparsity of PA vascular cross-sections, which lack the continuous directional cues required for global orientation tracking. Nevertheless, Fig. 2(b) shows that SCC-Net successfully reconstructs a 3D vascular network of the human palm, demonstrating its robust adaptability to noisy domains without external tracking.

4 Conclusion

In this study, we propose SCC-Net for stable sensorless 3D reconstruction in ultrasound and PA imaging. We introduce a SACD module to adaptively filter unstable acoustic features using global spatial statistics. To establish robust spatio-temporal correspondences, we develop a bidirectional cross-attention mechanism that captures global geometric topologies across adjacent frames. We further implement a self-supervised learning strategy using bidirectional consistency as a prior to enhance long-range reconstruction robustness. Experimental results on extensive ultrasound and complex PA datasets demonstrate that SCC-Net achieves state-of-the-art performance. This work facilitates 3D functional visu-

alization in the pure PA domain without external hardware, while highlighting rotation tracking in sparse vascular networks as a remaining challenge for future research.

Acknowledgements. This work was supported in whole, or in part, by the Suzhou Innovation and Entrepreneurship Leading Talent Program (ZXL2025307) and the National Natural Science Foundation of China (62401323).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adriaans, C.A., Wijkhuizen, M., van Karnenbeek, L.M., Geldof, F., Dashtbozorg, B.: Trackerless 3d freehand ultrasound reconstruction: a review. Applied Sciences **14**(17), 7991 (2024)
2. Alt, H., Godau, M.: Computing the fréchet distance between two polygonal curves. International Journal of Computational Geometry & Applications **5**(01n02), 75–91 (1995)
3. Chen, J.F., Fowlkes, J.B., Carson, P.L., Rubin, J.M.: Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test. International Journal of Imaging Systems and Technology **8**(1), 38–44 (1997)
4. Cleary, K., Peters, T.M.: Image-guided interventions: technology review and clinical applications. Annual Review of Biomedical Engineering **12**(1), 119–142 (2010)
5. Doktofsky, D., Rosenfeld, M., Katz, O.: Acousto optic imaging beyond the acoustic diffraction limit using speckle decorrelation. Communications Physics **3**(1), 5 (2020)
6. Dou, Y., Mu, F., Li, Y., Varghese, T.: Sensorless end-to-end freehand 3-d ultrasound reconstruction with physics-guided deep learning. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **71**(11), 1514–1525 (2024)
7. Guo, H., Chao, H., Xu, S., Wood, B.J., Wang, J., Yan, P.: Ultrasound volume reconstruction from freehand scans without tracking. IEEE Transactions on Biomedical Engineering **70**(3), 970–979 (2022)
8. Guo, H., Xu, S., Wood, B.J., Yan, P.: Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol. 12263, pp. 463–472. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_44
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
11. 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)

12. 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)
13. Li, Q., Shen, Z., Li, Q., Barratt, D.C., Dowrick, T., Clarkson, M.J., Vercauteren, T., Hu, Y.: Long-term dependency for 3d reconstruction of freehand ultrasound without external tracker. IEEE Transactions on Biomedical Engineering **71**(3), 1033–1042 (2023)
14. Li, Q., Shen, Z., Yang, Q., Barratt, D.C., Clarkson, M.J., Vercauteren, T., Hu, Y.: Nonrigid reconstruction of freehand ultrasound without a tracker. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15004, pp. 689–699. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72083-3_64
15. Liang, T., Yung, L., Yu, W.: On feature motion decorrelation in ultrasound speckle tracking. IEEE Transactions on Medical Imaging **32**(2), 435–448 (2012)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
17. Luo, M., Yang, X., Yan, Z., Li, J., Zhang, Y., Chen, J., Hu, X., Qian, J., Cheng, J., Ni, D.: Multi-imu with online self-consistency for freehand 3d ultrasound reconstruction. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14220, pp. 342–351. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43907-0_33
18. Mohamed, F., Siang, C.V.: A survey on 3d ultrasound reconstruction techniques. Artificial Intelligence—Applications in Medicine and Biology **1**, 73–92 (2019)
19. Mozaffari, M.H., Lee, W.S.: Freehand 3-d ultrasound imaging: a systematic review. Ultrasound in Medicine & biology **43**(10), 2099–2124 (2017)
20. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1134–1141. IEEE (2018)
21. Tuthill, T.A., Krücker, J., Fowlkes, J.B., Carson, P.L.: Automated three-dimensional us frame positioning computed from elevational speckle decorrelation. Radiology **209**(2), 575–582 (1998)
22. Wang, L.V., Hu, S.: Photoacoustic tomography: in vivo imaging from organelles to organs. Science **335**(6075), 1458–1462 (2012)
23. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11534–11542 (2020)
24. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. Advances in Neural Information Processing Systems **35**, 3502–3516 (2022)
25. Yan, J., Zhang, T., Broughton-Venner, J., Huang, P., Tang, M.X.: Super-resolution ultrasound through sparsity-based deconvolution and multi-feature tracking. IEEE Transactions on Medical Imaging **41**(8), 1938–1947 (2022)
26. 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: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15007, pp. 340–349. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72104-5_33