Real-Time 6D Ultrasound Probe Pose Tracking in Freehand Scanning

Jungjae Son[1], Guil Jung[1], Seok-Hwan Oh[2], Hyeonjik Lee[1], Young-Min Kim[2], Sang-Yun Kim[1], Myeong-Gee Kim[2], HyukSool Kwon[3], and Hyeon-Min Bae[1]

[1] Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
{jjson0509,hmbae}@kaist.ac.kr
[2] Barreleye Inc., Seoul, South Korea
[3] Seoul National University Bundang Hospital (SNUBH), Seongnam, South Korea

Abstract. The quality of ultrasound images strongly depends on the operator's experience and expertise. To mitigate this operator dependence, recent studies have investigated methods to support sonographer training, skill assessment, and probe guidance systems. These applications require accurate and robust estimation of the 6D ultrasound probe pose during freehand scanning. In this study, we propose a real-time 6D ultrasound probe pose tracking framework using a single RGB-D camera. This framework is designed to operate under real-world scanning conditions, where hand-induced occlusion and dynamic motion frequently occur. The proposed framework consists of two stages: probe keypoint detection and keypoint-guided pose estimation. The probe keypoint detection is introduced to enhance tracking robustness even under hand occlusion and rapid motion. Guided by the detected keypoints, the pose estimation finally determines the 6D pose of the probe. Furthermore, we develop an automated self-training strategy which significantly reduces manual annotation effort utilizing a pose estimation foundation model. As a result, the proposed framework achieves state-of-the-art pose accuracy and tracking robustness simultaneously compared to existing methods, while maintaining real-time performance above 25 fps.

Keywords: Ultrasound probe pose estimation · Real-time ultrasound probe pose tracking · Probe localization

1 Introduction

Ultrasound imaging is widely used due to its real-time capability, cost-effectiveness, and non-invasive nature. However, the ultrasound scanning process is highly dependent on operator, requiring substantial training to ensure diagnostic-quality imaging. Consequently, recent studies have explored sonographer training, skill assessment, and probe guidance systems to mitigate the operator dependence and improve scan quality [26].

Accurate 6D pose estimation of the ultrasound probe is fundamental to these systems. Traditional approaches rely on external devices such as electromagnetic

(EM), optical, or inertial measurement unit (IMU)-based tracking systems [2, 11, 13, 16, 15]. Although these devices provide high accuracy, they increase system cost and hardware complexity and limit clinical usability. Therefore, there remains a need for trackerless and markerless approaches to improve usability and enable unconstrained freehand scanning.

Object 6D pose estimation has been extensively studied in generic object settings [5, 7, 10, 21–23]. Notably, pose estimation foundation models such as FoundationPose [23] demonstrates that real-time pose tracking of a novel object can be achieved using only appearance cues. However, in ultrasound probe scanning scenarios, additional challenges arise due to frequent hand occlusion and dynamic motion, which cause tracking instability and limit the direct use of existing methods. Although recent studies have investigated probe pose estimation in hand–object interaction scenarios [1], the validation has been confined to controlled settings, with limited evaluation under real-world conditions.

In this paper, we propose a real-time 6D ultrasound probe pose tracking framework, UPTrack (Ultrasound Probe Tracking), using a single RGB-D camera. Our markerless and trackerless approach is specifically designed to handle realistic scanning scenarios involving severe hand occlusion and dynamic motion. The framework leverages a pose estimation foundation model integrated with novel keypoint-based guidance to enhance tracking robustness. Furthermore, we introduce an automated training pipeline for the keypoint detection network, enabling self-training without manual annotation. As a result, the proposed framework achieves state-of-the-art tracking accuracy and robustness while maintaining real-time performance.

Our contributions can be summarized as follows:

- A real-time 6D ultrasound probe pose tracking framework for freehand scanning using a single RGB-D camera
- A novel keypoint-guided tracking architecture that achieves robust performance under hand occlusion and dynamic probe motion
- An automated self-training pipeline for probe keypoint detection that significantly reduces manual annotation effort
- Comprehensive validation demonstrating state-of-the-art accuracy and real-time stability under realistic freehand scanning conditions

2 Methods

UPTrack, the proposed framework, performs real-time tracking of the 6D ultrasound probe pose from an RGB-D input sequence as shown in Fig. 1. The tracking pipeline consists of two processes: Initial Frame Pose Registration which estimates the initial pose by selecting the pose candidate that best explains the observed frame among multiple hypotheses, and Next Frame Pose Tracking which updates the current pose by comparing the rendered frame from the previous pose with the observed frame. Both processes share two core stages: Probe Keypoint Detection (Section 2.1) and Keypoint-Guided Probe Pose Estimation (Section 2.2).

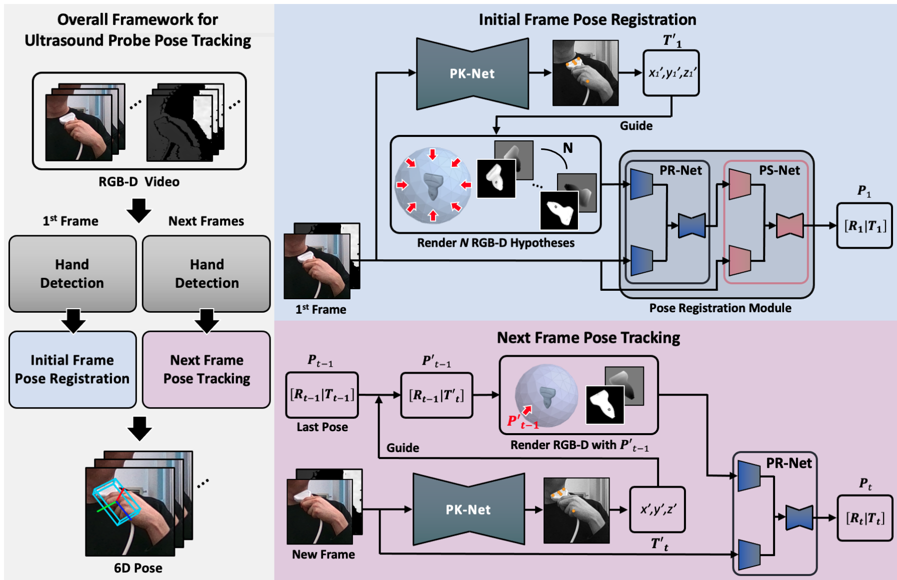

Fig. 1. Overview of the real-time 6D ultrasound probe pose tracking framework.

2.1 Probe Keypoint Detection

A Probe Keypoint Net (PK-Net) is designed to predict a predefined set of landmarks on the ultrasound probe from a single RGB image. The five landmarks are selected from regions with distinctive structural and textural features to ensure stable detection. Specifically, they include: (1,2) the two side endpoints of the transducer, (3) the center of the transducer, (4) the top-center of the probe body, and (5) the tail of the probe.

PK-Net is designed as a lightweight architecture consisting of a ResNeXt-based encoder [25] and a deconvolutional decoder [24]. This design enables real-time performance within the overall framework. The network outputs five probability heatmaps corresponding to predefined probe keypoints, from which 2D keypoint locations are obtained via peak extraction during inference. During training, instead of directly regressing keypoint coordinates, PK-Net is supervised to predict a probability heatmap for each keypoint. Each heatmap is generated as a 2D Gaussian centered at the annotated keypoint location. This supervision mitigates label noise and stabilizes learning, particularly for occluded keypoints whose precise locations are difficult to annotate.

To reduce computational cost, keypoint detection is performed only within a region of interest (ROI) around the hand holding the probe. This strategy avoids unnecessary background computation and stabilizes probe keypoint prediction. The ROI is determined using a pretrained hand detection model [27].

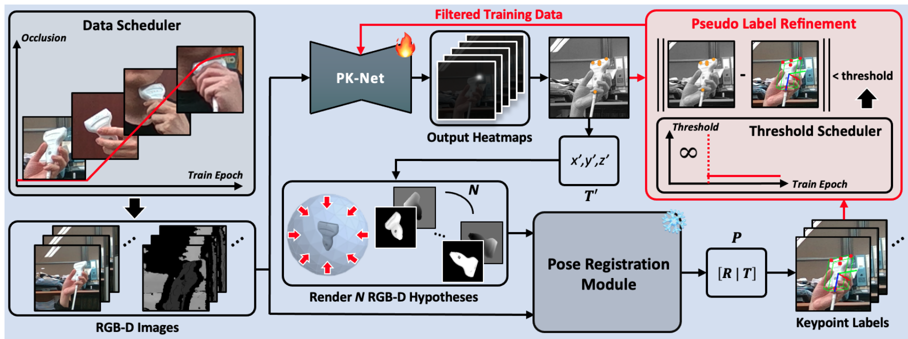

Fig. 2. Overview of the automated self-training strategy for PK-Net.

Automated Self-Training Strategy To achieve robust performance across diverse operators and probe grasping styles, training PK-Net on large-scale data collected under varied conditions is essential. However, manual annotation of dataset of this scale requires substantial effort. To address this challenge, we propose an automated training pipeline as illustrated in Fig. 2.

With this training strategy, PK-Net progressively improves its robustness to occlusion via self-training without manual annotation. We utilize a Pose Registration Module composed of the pretrained Pose Refinement Network (PR-Net) and Pose Selection Network (PS-Net) from FoundationPose [23]. This module selects and outputs the pose that best matches the current frame among N pose hypotheses. As the probe is a rigid object, keypoint pseudo labels can be automatically generated from the estimated pose based on its geometric dimensions.

However, without a sufficiently trained PK-Net, the Pose Registration Module is vulnerable to occlusion. Therefore, we adopt a data scheduler that structures the training data from minimally occluded samples to increasingly occluded samples as PK-Net learns. Furthermore, a refinement step is applied before they are fed to PK-Net for training since incorrect pose estimation can generate noisy pseudo labels. Samples are filtered out if the average pixel L2 distance between pseudo keypoints and PK-Net predictions exceeds a threshold, which is set to infinity for the first 10 epochs and limited to 6 pixels from epochs 10 to 25.

2.2 Keypoint-Guided Probe Pose Estimation

In this stage, the final 6D pose of the ultrasound probe is determined based on the keypoints obtained from the PK-Net. For a given RGB-D video sequence, the first frame undergoes Initial Frame Pose Registration, while all subsequent frames are sequentially processed via Next Frame Pose Tracking.

In Initial Frame Pose Registration, pose prediction is performed using keypoint-based guidance together with PR-Net and PS-Net. In Next Frame Pose Tracking, keypoint-based guidance and PR-Net are used to perform pose updates.

Keypoint-Based Guidance After probe keypoint detection, the 2D coordinates of the five predefined keypoints are obtained. The 2D probe center (u_0, v_0) is approximated as the midpoint between the transducer center keypoint and the probe tail keypoint. The probe depth is estimated from the detected keypoints and the corresponding depth map captured by the RGB-D camera.

Given the 2D pixel coordinates (u, v) and the corresponding depth value d, the 3D coordinates of each keypoint in the camera coordinate system are computed as:

$$\mathbf{X}_{cam} = d\,\mathbf{K}^{-1}\tilde{\mathbf{u}}, \quad \tilde{\mathbf{u}} = \begin{bmatrix} u \; v \; 1 \end{bmatrix}^{\top}, \tag{1}$$

where $\mathbf{X}_{cam} = [x, y, z]^{\top}$ denotes the 3D coordinates in the camera coordinate system, and $\mathbf{K}$ is the camera intrinsic matrix. Using (u_0, v_0) and Eq. (1), an estimated translation of the 3D probe center $\mathbf{T}' = [x', y', z']^{\top}$ is obtained. This is used to guide the translation component of the 6D pose $\mathbf{P} = [\mathbf{R} \mid \mathbf{T}]$, where $\mathbf{R} \in SO(3)$ denotes rotation and $\mathbf{T} \in \mathbf{R}^3$ denotes translation.

Initial Frame Pose Registration Through keypoint-based guidance, a coarse translation of the probe in the initial frame is obtained as $\mathbf{T}'_1 = [x'_1,\, y'_1,\, z'_1]^{\top}$. Subsequently, N rotation hypotheses are generated, producing multiple pose candidates. Specifically, we sampled 42 viewpoints on a sphere using an icosphere-based sampling and applied 6 in-plane rotations per viewpoint, yielding $N = 252$ pose hypotheses. For each pose candidate, a rendered RGB-D result is obtained.

PR-Net takes two inputs: the rendered RGB-D frame corresponding to a pose hypothesis and the actual RGB-D frame. It predicts the relative pose offsets, $\Delta\mathbf{R}$ and $\Delta\mathbf{T}$, required to better align the rendered object with the observed appearance in the captured real RGB-D frame. By updating each pose hypothesis with the predicted pose offset, all pose candidates are refined toward the true pose. Given the relative pose offsets $\Delta\mathbf{R}$ and $\Delta\mathbf{T}$ predicted by PR-Net, the pose is updated as

$$\mathbf{R}_{new} = \Delta\mathbf{R}\mathbf{R}_{old}, \quad \mathbf{T}_{new} = \mathbf{T}_{old} + \Delta\mathbf{T}. \tag{2}$$

After PR-Net refines the pose candidates, PS-Net takes the rendered RGB-D image of each refined candidate and the actual RGB-D frame as input. It outputs the scores indicating how well each candidate explains the observed appearance, and the highest-scoring candidate is selected as the final pose.

Next Frame Pose Tracking In the Next Frame Pose Tracking process, the pose from the previous frame $\mathbf{P}_{t-1} = [\mathbf{R}_{t-1} \mid \mathbf{T}_{t-1}]$ is updated to obtain the current pose $\mathbf{P}_t = [\mathbf{R}_t \mid \mathbf{T}_t]$. This process consists of two sequential updates: first through keypoint-based guidance and second through PR-Net.

First, the keypoint-based guided translation $\mathbf{T}' = [x', y', z']^{\top}$ estimated from the detected keypoints replaces the translation component of $\mathbf{P}_{t-1}$, yielding $\mathbf{P}'_{t-1} = [\mathbf{R}_{t-1} \mid \mathbf{T}']$. Second, a rendered RGB-D frame based on $\mathbf{P}'_{t-1}$ and the captured RGB-D frame at time t are fed into PR-Net. PR-Net predicts $\Delta\mathbf{R}$ and $\Delta\mathbf{T}$, and the pose is updated according to Eq. (2) to obtain $\mathbf{P}_t$.

Table 1. Keypoint localization accuracy of the proposed PK-Net.

Keypoint	Mean L2 [px]	PCK @2px	PCK @4px	PCK @8px
Transducer Sides	2.251	0.458	0.857	0.998
Transducer Center	1.954	0.511	0.918	1.000
Probe Top-Center	1.782	0.546	0.953	1.000
Probe Tail	4.973	0.145	0.413	0.829
Total	2.698	0.425	0.786	0.959

3 Experiments

To validate the proposed framework under real-world scanning conditions, we conducted three experiments: (1) keypoint detection accuracy, (2) pose tracking accuracy, and (3) pose estimation robustness through an ablation study.

Data were acquired using an Intel RealSense D455 camera [4, 18, 19] at a resolution of 1280×720 for RGB-D capture and GE Linear Probe L3-12. Ground truth poses were acquired using an NDI Polaris Vega optical tracker [9, 14, 8, 6], which provides sub-millimeter volumetric accuracy (IQR $<$ 0.1 mm) [3]. All experiments were performed on a system with an NVIDIA RTX 4090 GPU.

Probe scanning was conducted at distances of 0.5–1.5 m from the RGB-D camera. This range was chosen considering the depth reliability of the RealSense D455 and realistic clinical scanning environments.

For quantitative pose evaluation, the optical tracker marker was attached to the probe via a long, thin, and rigid 3D-printed bridge structure. This setup prevented the marker from interfering with natural probe grasping and minimized its influence on vision-based inference. For the qualitative results in Fig. 3, we used a marker-free setup to demonstrate pose inference under practical real-world conditions without markers.

3.1 Keypoint Detection Accuracy

We evaluated the proposed PK-Net by comparing the predicted 2D keypoint locations with ground truth. Localization accuracy was measured using mean L2 error (in pixel, px) and Percentage of Correct Keypoints (PCK).

Table 1 provides the results. The network achieved an overall mean L2 error of 2.698 pixels. For the transducer side keypoints, symmetry was considered during evaluation. The probe tail showed larger error due to frequent hand occlusion.

3.2 Pose Tracking Accuracy

We evaluated the 6D probe pose tracking accuracy on 35 RGB-D test videos using Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) [20, 12]. The test videos were captured under unconstrained freehand scanning by multiple subjects. To exclude sequences in which all methods completely fail

Table 2. Quantitative evaluation of the proposed framework, including comparisons with existing 6D pose estimation methods in probe pose accuracy (ATE/RPE: t = translation [mm], r = rotation [°]) and runtime performance (FPS). Runtime denotes end-to-end pose tracking speed, excluding GUI rendering overhead.

Method	Seg. Mask	Occlusion	$\textbf{ATE}_t$ mae↓	$\textbf{ATE}_t$ rmse↓	$\textbf{RPE}_t$ mae↓	$\textbf{RPE}_t$ rmse↓	$\textbf{RPE}_r$ mae↓	$\textbf{RPE}_r$ rmse↓	FPS↑
Any6D	SAM2	No-occ	27.72	30.42	24.38	51.14	63.47	88.63	<1
		Hand-occ	41.23	46.02	40.62	32.00	77.20	93.20	
SAM6D	SAM2	No-occ	10.99	17.40	12.58	23.82	9.16	15.02	13.70
		Hand-occ	34.50	50.00	35.09	60.41	18.00	25.54	
BundleSDF	SAM2	No-occ	2.42	2.67	1.86	2.11	3.17	3.62	13.33
		Hand-occ	17.46	19.11	7.59	8.35	6.41	7.13	
Ours	Not required	No-occ	**2.17**	**2.44**	**1.77**	**2.01**	**2.73**	**3.09**	**25.62**
		Hand-occ	**13.04**	**15.00**	**7.50**	**8.31**	**4.95**	**5.58**	

tracking, severe occlusion and rapid motion were not included. For methods that use segmentation masks, probe masks were generated using SAM2 [17].

Table 2 shows quantitative tracking accuracy results and presents a comparative study against representative pose estimation models [7, 10, 22]. Overall, the proposed framework achieved the highest tracking accuracy and FPS among the compared methods, while operating without requiring a precomputed segmentation mask.

3.3 Ablation Study on Pose Estimation Robustness

We conducted an ablation study on the proposed keypoint-based guidance using PK-Net to analyze its contribution to tracking robustness. Robustness is evaluated separately for initial frame pose registration and next frame pose tracking. The evaluation is performed on RGB-D videos and quantified by the success rate, defined as the absence of pose registration failure or pose tracking failure within each sequence. A total of 110 test videos were collected from multiple subjects under diverse real-world environments, including severe hand occlusion and dynamic motion. The subjects were instructed to grasp the probe in the manner typically used by sonographers during freehand scanning. Tracking failure is defined with respect to optical ground truth, and a frame is considered failed if the translation error exceeds 20 mm or the rotation error exceeds 20°.

Initial Frame Pose Registration As shown in Fig. 3 (Top), the pose estimation success rate on the first frame of each test video is evaluated. Under severe hand occlusion, the baseline achieves a success rate of 54%, whereas the proposed framework achieves 91%. The segmentation-based baseline more frequently fails to estimate probe translation, leading to registration errors. In contrast, the proposed keypoint-guided approach successfully localizes probe keypoints, including those in occluded regions, resulting in more robust pose registration.

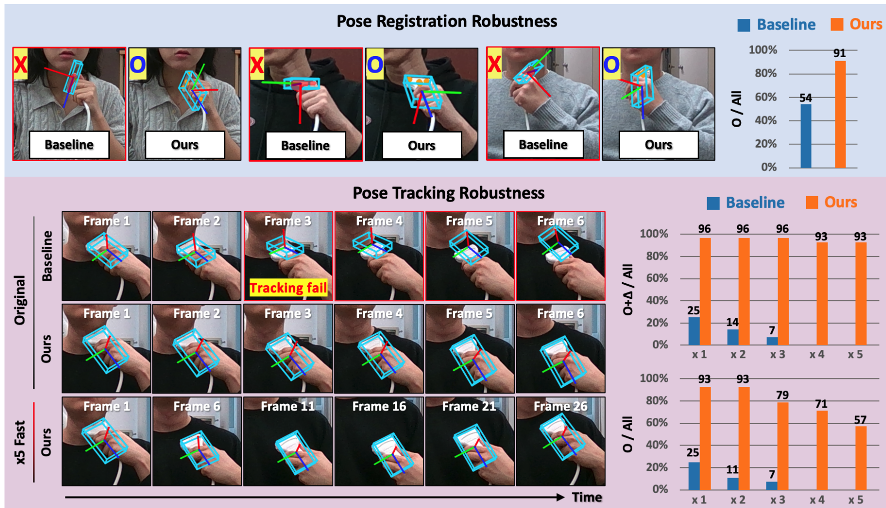

Fig. 3. (Top) Pose Registration Robustness: qualitative results and pose estimation success rate comparison. (Bottom) Pose Tracking Robustness: qualitative results and tracking success rate comparison under varying motion speeds.

Next Frame Pose Tracking Tracking robustness is presented in Fig. 3 (Bottom). A sequence is considered strictly successful (O) if no tracking failure occurs throughout the sequence. Under the relaxed criterion (△), a sequence is also considered successful provided that tracking failures do not occur for three or more consecutive frames.

Furthermore, to evaluate robustness under dynamic motion, frame dropping is applied to emulate fast scanning motion, where 0 drop denotes the original video and 1–4 drops correspond to $\times 2$–$\times 5$ motion speeds, respectively. The baseline achieves 25% robustness on the original videos and fails on all sequences at $\times 4$ speed. In contrast, the proposed framework achieves 93% robustness (O) and 96% robustness (O+△) on the original videos, and maintains stable tracking even at $\times 5$ motion speed, achieving 57% (O) and 93% (O+△). Additionally,

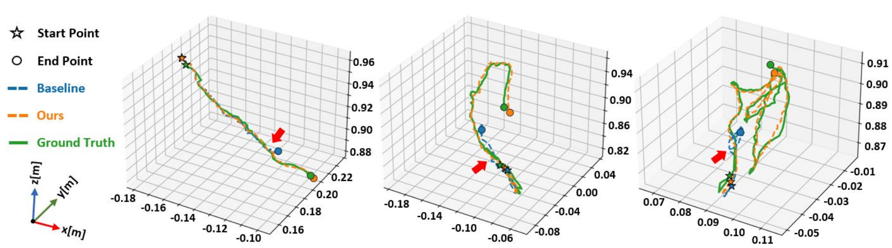

Fig. 4. Trajectory comparison of the proposed method and baseline, with red arrows indicating baseline tracking failure points.

Fig. 4 visualizes representative trajectory comparisons between the baseline and the proposed method, where the baseline fails to maintain tracking while the proposed method remains stable.

4 Conclusion

We propose UPTrack, a real-time 6D ultrasound probe pose tracking framework for freehand scanning using a single RGB-D camera, eliminating the need for markers or external tracking devices. By integrating probe keypoint detection with keypoint-guided pose estimation, the framework achieves robust performance under hand occlusion and dynamic probe motion. Additionally, by leveraging a pose estimation foundation model with automated self-training, it enables adaptation with minimal annotation effort. Experiments demonstrate state-of-the-art accuracy and tracking robustness compared to existing pose estimation methods. The proposed framework can be further extended to downstream applications, such as sonographer training, skill assessment, and probe guidance systems, for validation in practical clinical scenarios.

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

References

1. Birlo, M., Caramalau, R., Edwards, P.J. Dromey, B., Clarkson, M.J., Stoyanov, D.: HUP-3D: A 3D multi-view synthetic dataset for assisted-egocentric hand-ultrasound-probe pose estimation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15001. Springer Nature Switzerland (October 2024)
2. Cai, Q., Peng, C., Prieto, J.C., Rosenbaum, A.J., Stringer, J.S., Jiang, X.: A low-cost camera-based ultrasound probe tracking system: Design and prototype. In: 2019 IEEE International Ultrasonics Symposium (IUS). pp. 997–999. IEEE (2019)
3. Fattori, G., Lomax, A.J., Weber, D.C., Safai, S.: Technical assessment of the ndi polaris vega optical tracking system. Radiation oncology **16**(1), 87 (2021)
4. Intel Corporation: Intel RealSense D400 Series Datasheet (Revision 022, Dec 2025). Intel Corporation (2025), includes specifications for RealSense D455
5. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: Megapose: 6d pose estimation of novel objects via render & compare. arXiv preprint arXiv:2212.06870 (2022)
6. 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)
7. Lee, T., Wen, B., Kang, M., Kang, G., Kweon, I.S., Yoon, K.J.: Any6d: Model-free 6d pose estimation of novel objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11633–11643 (2025)
8. 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)

9. 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: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15004. Springer Nature Switzerland (October 2024)
10. Lin, J., Liu, L., Lu, D., Jia, K.: Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27906–27916 (June 2024)
11. 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: International conference on medical image computing and computer-assisted intervention. pp. 342–351. Springer (2023)
12. Meng, X., Yang, J., Chen, M., Yan, C., Shi, Y., Ding, W., Kneip, L.: Dynopets: A versatile benchmark for dynamic object pose estimation and tracking in moving camera scenarios. IEEE Robotics and Automation Letters (2025)
13. Mozaffari, M.H., Lee, W.S.: Freehand 3-d ultrasound imaging: a systematic review. Ultrasound in medicine & biology **43**(10), 2099–2124 (2017)
14. Naheem, M., Shyam, A., Dumpuri, P., Lakshmanan, M., Sivaprakasam, M., et al.: Optical tracker assessment for image guided surgical interventions. In: 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA). pp. 1–6. IEEE (2022)
15. Paralikar, A., Mantripragada, P., Nguyen, T., Arjoune, Y., Shekhar, R., Monfaredi, R.: Robot-assisted ultrasound probe calibration for image-guided interventions. International Journal of Computer Assisted Radiology and Surgery **20**(5), 859–868 (2025)
16. Peng, C., Cai, Q., Chen, M., Jiang, X.: Recent advances in tracking devices for biomedical ultrasound imaging applications. Micromachines **13**(11), 1855 (2022)
17. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollar, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=Ha6RTeWMd0
18. Servi, M., Mussi, E., Profili, A., Furferi, R., Volpe, Y., Governi, L., Buonamici, F.: Metrological characterization and comparison of d415, d455, l515 realsense devices in the close range. Sensors **21**(22), 7770 (2021)
19. Servi, M., Profili, A., Furferi, R., Volpe, Y.: Comparative evaluation of intel realsense d415, d435i, d455 and microsoft azure kinect dk sensors for 3d vision applications. IEEE Access (2024)
20. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012)
21. Wen, B., Bekris, K.: Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8067–8074. IEEE (2021)
22. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–617 (2023)
23. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17868–17879 (2024)

24. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 466–481 (2018)
25. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
26. Yeo, L., Romero, R.: Optical ultrasound simulation-based training in obstetric sonography. The Journal of Maternal-Fetal & Neonatal Medicine **35**(13), 2469–2484 (2022)
27. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214 (2020)