Improving Probe Localization for Freehand 3D Ultrasound using Lightweight Cameras

Dianye Huang, Nassir Navab, *Fellow, IEEE*, and Zhongliang Jiang

***Abstract*— Ultrasound (US) probe localization relative to the examined subject is essential for freehand 3D US imaging, which offers significant clinical value due to its affordability and unrestricted field of view. However, existing methods often rely on expensive tracking systems or bulky probes, while recent US image-based deep learning methods suffer from accumulated errors during probe maneuvering. To address these challenges, this study proposes a versatile, cost-effective probe pose localization method for freehand 3D US imaging, utilizing two lightweight cameras. To eliminate accumulated errors during US scans, we introduce PoseNet, which directly predicts the probe's 6D pose relative to a preset world coordinate system based on camera observations. We first jointly train pose and camera image encoders based on pairs of 6D pose and camera observations densely sampled in simulation. This will encourage each pair of probe pose and its corresponding camera observation to share the same representation in latent space. To ensure the two encoders handle unseen images and poses effectively, we incorporate a triplet loss that enforces smaller differences in latent features between nearby poses compared to distant ones. Then, the pose decoder uses the latent representation of the camera images to predict the probe's 6D pose. To bridge the sim-to-real gap, in the real world, we use the trained image encoder and pose decoder for initial predictions, followed by an additional MLP layer to refine the estimated pose, improving accuracy. The results obtained from an arm phantom demonstrate the effectiveness of the proposed method, which notably surpasses state-of-the-art techniques, achieving average positional and rotational errors of 2.03 mm and 0.37°, respectively. Code: `https://github.com/dianyeHuang/FreehandUS_Pose_Estimation`**

I. Introduction

Medical ultrasound (US) is one of the most vital diagnostic tools in modern clinical practice, offering real-time, non-invasive imaging of soft tissues and internal organs. Its versatility has made US imaging a first-line tool across various medical fields, including obstetrics, cardiology, emergency medicine, and radiology. However, interpreting traditional 2D US images can be challenging, as they provide only limited views of internal structures and are often affected by speckles, artifacts, and shadows. To overcome these limitations and provide a more comprehensive view, 3D US imaging has gained significant attention, offering richer contextual information and improving diagnostic accuracy [1], [2].

To achieve 3D US, several methods have been explored, with one of the most intuitive being native 3D US imaging. Hossack *et al.* developed a prototype probe featuring a 2D array of elements capable of directly providing a 3D view of the examined anatomy [3]. This approach was tested on carotid arteries in clinical settings, with results demonstrating its effectiveness. However, native 3D probes are often constrained by a limited field of view and low frame rates due to the drastic increase in computational demands. The cost is also much higher than a 2D probe. Alternatively, the wobbler probe is presented, which mechanically tilts or rotates a 1D array to reconstruct a 3D volume [4]. While this method reduces operator dependency and ensures more consistent data, wobbler probes typically suffer from limited field of view, mechanical artifacts, and bulkier design. These limitations limited their extensive use in clinical practices.

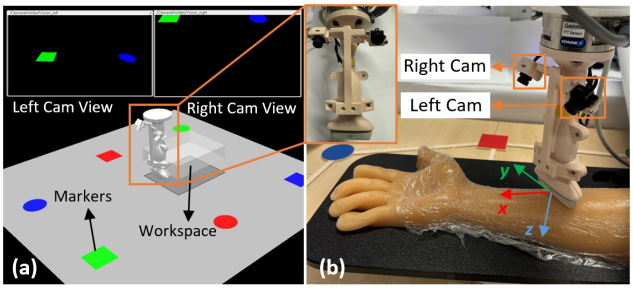

Fig. 1. An illustration of the simulation and real-world settings. *Left*: depicts the virtual environment for simulated data generation. *Right*: shows the corresponding real-world settings and the configuration of two cameras.

To address the issue of limited field of view, external tracking systems, such as electromagnetic (EM) [5] or optical devices [6], have been introduced to monitor the motion of standard 2D ultrasound probes. By acquiring a series of 2D slices while tracking the probe's movement, a complete 3D volume of the scanned area can be reconstructed. To facilitate the rapid prototyping of US-guided intervention systems using various hardware, an open-source software called PLUS was proposed to visualize live reconstruction results based on the streaming of acquired data [7]. This solution aligns well with physicians' workflows; however, potential issues such as occlusion with optical tracking systems or electromagnetic interference with EM tracking devices can limit their practical application. Recently, robotic systems have been employed to provide reliable tracking. Jiang *et al.* mounted the US probe on a robotic arm to compute 3D volume or point clouds [1], [8]–[10]. Similar studies exploring different clinical applications can be found in [11]–[16]. Despite these advancements, the high cost associated with optical tracking and robotic systems restricts their use, particularly for emergency scenarios with limited space and primary examinations in underdeveloped regions.

Thanks to rapid advances in deep learning, image-based freehand 3D US reconstruction has recently garnered increasing attention. Prevost *et al.* utilized a convolutional neural

Dianye Huang, Nassir Navab and Zhongliang Jiang are with the Chair for Computer-Aided Medical Procedures and Augmented Reality, Technical University of Munich, Boltzmannstr. 3, 85748 Garching bei München, Germany; Munich Center for Machine Learning (MCML). (corresponding author: Zhongliang Jiang, zl.jiang@tum.de)

network (CNN) to directly estimate the motion between successive US frames in an end-to-end fashion, incorporating a fully connected layer (FC) to integrate real-time data from an Inertial Measurement Unit (IMU) attached to a 2D probe [17]. This approach achieved average errors of $3.71\ mm$ and $1.88°$ for translation and rotation, respectively. Guo *et al.* presented DCL-Net using multiple US images as input [18]. This method was purely based on US images and explored the contextual features between US frames. To tackle the problem of the ambiguity of moving direction between two frames, Luo *et al.* incorporated prior knowledge and consistency constraints to guide the reconstruction network [19]. Li *et al.* demonstrated the benefits of incorporating long-term dependencies in their extensive experiments to improve 3D reconstruction [20]. Despite these advancements, a common limitation is the accumulation of errors due to the estimation of transformations between neighboring frames. To address this issue, Luo *et al.* proposed a multi-modal online self-supervised framework that uses IMU measurements as weak labels for adaptive optimization, which significantly reduces drift errors [21]. Although this framework showed substantial improvements in reconstruction drift, the reported results—translational and rotational errors of $10.24\ mm$ and $1.55°$, respectively—remain insufficient, which severely hinders their application in real clinical cases. In addition, camera-based methods were also investigated. Sun *et al.* used cameras to estimate probe pose from a skin map, but this approach suffers from accumulated errors, leading to translation errors of up to $10\ mm$ and rotation errors of $6°$ after just $10\ cm$ of probe movement along the abdomen [22].

This paper aims to propose a truly low-cost, robust, and precise solution for freehand 3D US reconstruction. Instead of relying on noisy US images for motion estimation, two lightweight monocular cameras are mounted on the probe (see Fig. 1 (b)) to capture distinct visual changes, enabling accurate probe motion calculation. Unlike the approach in [22], which uses skin maps from individual patients, this study introduces a structured environment with multiple colored circle and rectangle markers distributed around the examined area to enhance tracking accuracy and stability. To tackle the accumulated error for a long sweep, this study directly computes the 6D pose based on the concatenated camera images. The main contributions of this work are summarized below:

1) The cross encoder decoder (CED) module is proposed to effectively learn the relationship between camera pose and camera observations in a structured environment. Since direct pose regression requires a large amount of data densely sampled in the 6D space to ensure accuracy, we first replicate the same environment in a simulator, collecting $200K$ paired pose and images pertaining to CED module.
2) The hybrid loss is proposed based on a cross-reconstruction loss between paired poses and images, and a triplet loss applied to three distinct poses. This design encourages the pose encoder in the CED to benefit from paired image inputs. The inclusion of the triplet loss and cross reconstruction loss is critical, as it helps to manage the inherent redundancy in pose features within the latent space, promoting better alignment between pose and image representations.
3) Inspired by feedback mechanisms from classical control algorithms, the proposed PoseNet for real-world applications adapts certain components from the trained CED and integrates an additional MLP module. This MLP takes the latent representation of real-time camera images and the predicted 6D pose as inputs to compute compensation values for the initial pose predicted by the CED's pose decoder. This compensation step enhances the precision of the final pose estimation, addressing nonideal factors like imperfect camera calibration and bridging the gap between simulation and real-world conditions.

The 3D reconstruction results obtained from experiments on an arm phantom demonstrate that the proposed PoseNet performs robustly in real-world scenarios. PoseNet significantly outperforms existing methods in terms of prediction accuracy (2.03 mm and $0.37°$ for transnational and rotational error), especially during long scans, as it effectively eliminates accumulated errors, a common issue in other approaches. The code will be made public on this website.

II. Preliminaries

This section details the setup. Two low-cost monocular cameras are firmly mounted on the US probe to capture visual changes during scanning. To minimize environmental noise in real-world conditions, a structured environment featuring a variety of basic shapes (circles and squares) is created, providing consistent visual references for accurate probe localization. The same setup is replicated in a simulator to generate a larger dataset for training, facilitating more robust and accurate model development.

A. Lightweight Monocular Camera Configuration

In this study, we utilize two affordable (€15 each) monocular cameras (2303U, China) mounted on the US probe to perceive the structured environment for pose estimation. To ensure a wide field of view (FoV) and capture enough markers during scanning, two cameras are symmetrically positioned on the US probe, both facing toward the scanning direction. To further increase the vertical FoV, one of the cameras is tilted downward at approximately $25°$, as illustrated in Fig. 1 (b). Real-time images are streamed to the main workstation via OpenCV at a rate of 30 frames per second (fps). The original image resolution is 360 (Height) $\times$ 640 (Width), with a horizontal FoV set to $80°$. After mounting the cameras on the probe, classical eye-in-hand calibration was performed to get the transformation from the camera frame to the robotic base frame in the real world. These calibration results are used to configure the virtual camera setting in the simulator.

B. Semi-Structured Environment with Visual Markers

The environment is constructed using six markers in three distinct colors (red, green, and blue) and two basic shapes

(circle and square). Each square marker measures 6 cm in length, while each circular marker has a diameter of 6 cm. The markers are arranged asymmetrically to prevent ambiguity in pose estimation, as depicted in Fig. 1(a). As depicted in Fig. 1(a), the scanning workspace is defined as a rectangular cuboid, with dimensions of 180 mm (longitudinal) $\times$ 150 mm (lateral) $\times$ 40 mm (vertical). At each positional configuration, the probe can rotate within the following ranges: tilting $\pm 20^{\circ}$ around its long axis, swaying $\pm 20^{\circ}$ around its short axis, and rotating $\pm 45^{\circ}$ around its centerline. In such space, 1.56×10^{11} images can be obtained from corresponding grid points with $1\ mm$ translation accuracy and 1° in rotation accuracy. Collecting such a large amount of data is unrealistic in real scenarios.

C. Data Acquisition

Given the impracticality of acquiring a large volume of data in real-world scenarios, we developed a simulation environment using CoppeliaSim[1], also known as V-REP [23]. The simulation setup mirrors the real-world configuration, including the placement of two monocular cameras on the US probe, along with their internal parameters, ensuring consistency between the real and virtual environments. To have the same pose description in real scenarios and simulations, a global reference frame is determined based on the layout of markers, for example, the overall center point of all six markers. In the real world, to transfer the pose description from the robotic base frame to this selected world coordinate system, we need to move the robotic arm (Franka Emika Panda, Franka GmbH) to position the US probe physically to its coordinate origin.

1) Data Acquisition in Simulation: Even in simulation, collecting 1.56×10^{11} pairs of probe poses and camera images is impractical. Additionally, due to camera calibration errors and the sim-to-real gap, densely sampling all images is unnecessary. In this study, we uniformly sampled 2×10^5 pairs of probe poses and corresponding camera images from the simulation. Despite covering a wide range of poses, this sample set remains relatively sparse considering the possible combinations, which will not lead to overfitting the PoseNet in the given environment. The images captured from the two cameras are concatenated into a single image with a resolution of 720×640 pixels. This combined image is then resized to 512×512 pixels, and its intensity values are normalized to a range between zero and one. The corresponding probe poses are directly retrieved from the simulation environment. The dataset is randomly split into training, validation, and testing sets with an 8:1:1 ratio, ensuring balanced and effective model evaluation.

2) Data Acquisition in Real World: In the real world, a linear US probe (12L3, ACUSON Juniper, SIEMENS AG) with a dual camera is mounted to the end-effector of a robotic arm (Franka Emika Panda, Franka GmbH) to scan a human-like arm phantom (BPA304, Blue Phantom GmbH) and the US images are captured via a frame grabber (MAGEWELL). The US images is also streamed in the rate of 30 fps. It is worth noting that the US images are not used to predict the probe pose in this study. The recorded US images are only used to show a reconstructed 3D volume.

To fine-tune the pre-trained model using real-world data, we collect paired camera images and probe poses. The images captured from the two cameras are processed in the same manner as in the simulation, ensuring consistency. The probe pose, measured through robotic kinematics, is mapped to the same world coordinate system used in the simulation, based on the layout of the markers, allowing for seamless integration between the simulated and real-world data. The arm phantom is randomly placed on the table, and we recorded 95 linear trajectories on the phantom. The similarities of these trajectories are computed and ranked. We selected the top 50 dissimilar trajectories and randomly split them into training, validation, and testing sets with numbers of 25 (4.4k tracked images), 10 (1.6k tracked images), and 15 (2.4k tracked images) trajectories, respectively.

III. Probe Pose Localization Network

This section outlines the details of the proposed networks in both simulation and real-world scenarios. The overall architecture is shown in Fig. 2. The Cross Encoder-Decoder (CED) is primarily trained on simulation data, while PoseNet incorporates the pre-trained components from CED and undergoes fine-tuning using data collected in real-world settings for enhanced accuracy. The code will be made public on this website.

A. Probe Pose Estimation in Simulation

To accurately predict the probe pose from camera images, it is essential to link the pose features and camera image features within a shared latent space. To achieve this, we propose a Cross Encoder-Decoder (CED) network, which not only predicts the pose based on camera image features but can also work in reverse, predicting the camera image from the pose information. This bidirectional capability ensures a robust relationship between the pose and image features for more accurate predictions. As shown in Fig. 2 (a), CDE consists of image Encoder & Decoder that follows the ResNet 18 structure, and pose Encoder & Decoder constructed by MLP layers. The core concept is to align the pose latent features with the image latent features, enabling the CED network to effectively perceive the structured environment. To facilitate this alignment, the encoded features must have pathways to exchange information, which influences the encoder parameters. The CED module provides four such cross-information exchange routes: (1) input pose to output reconstructed pose, (2) input image to output reconstructed image, (3) input pose to output reconstructed image, and (4) input image to output reconstructed pose. During inference, the fourth route is utilized for pose estimation, ensuring that image features directly infer accurate pose predictions.

To train the CDE modules, several losses are considered. Firstly, the pose and image should be correctly decoded from their own encoded features, and also from the other features, thus we have the reconstructed loss (MSE loss) as follows:

$$\mathcal{L}_{rec} = \gamma(\mathcal{L}_i^{rec} + \mathcal{L}_i^{crec}) + (\mathcal{L}_p^{rec} + \mathcal{L}_p^{crec}) \tag{1}$$

[1]https://www.coppeliarobotics.com/

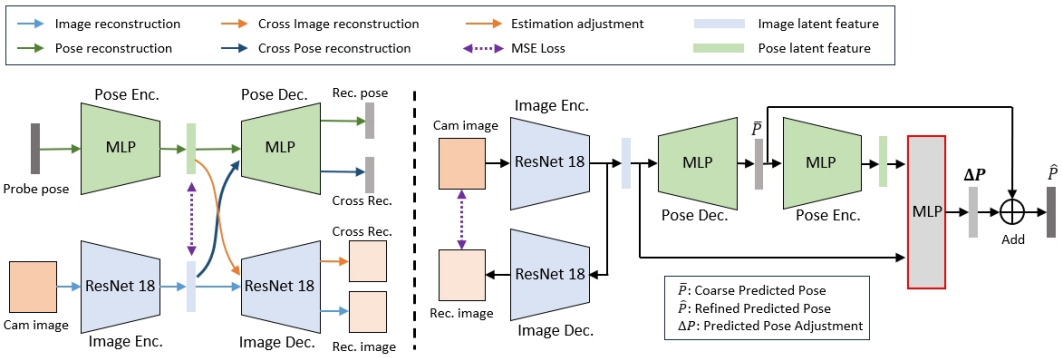

Fig. 2. Structure of the proposed (a) Cross Encoder Decoder (CED) module, and (b) the PoseNet. The CED is trained using simulated data. PoseNet integrates select components from the trained CED along with an additional MLP module to predict the probe pose based on input from two camera images in the real world.

where $\gamma = 2$ is used in this study to assign a relatively higher priority to the image reconstruction task, as the encoded feature of the pose is theoretically redundant and, therefore, is more flexible to be reconstructed.

To make the encoded feature more spatially related, we induce a triplet loss to the encoded pose feature to enforce the CED modules to learn a latent feature that takes the distance of the poses into account. In other words, triplet loss is used to enforce smaller differences in latent features between nearby poses compared to distant ones. This auxiliary learning task helps CED understand the underlying spatial physics, which will help properly encode the unseen poses.

$$\begin{aligned}\mathcal{L}_{tri} = \max(0, d(f_p(P_a), f_p(P_p)) - \\ d(f_p(P_a), f_p(P_n)) + \beta d(P_a, P_n))\end{aligned} \tag{2}$$

where $d(\cdot,\cdot)$ dentoes the norm-2 distance of the inputs, $f_p(\cdot)$ stands for the pose Encoder, and (P_a, P_p, P_n) are the input samples that constructed based on the norm-2 distance. $\beta = 0.1$ to modulate the marginal settings. The most important part of the loss is:

$$\begin{aligned}\mathcal{L}_{latent} = (1-\alpha)\|f_i(I) - sg(f_p(P))\|^2 \\ +\alpha\|sg(f_i(I)) - f_p(P)\|^2\end{aligned} \tag{3}$$

where $sg(\cdot)$, $f_i(\cdot)$, I, P denotes stop gradient, image Encoder, image, and pose, respectively. $\alpha = 0.3$ is to encourage the image feature to be closer to the pose embedding, thus integrating more spatial information.

The overall training loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{tri} + \sigma\mathcal{L}_{latent} \tag{4}$$

where $\sigma = 2.0$ to put more weight on the latent feature alignment.

B. Probe Pose Estimation in Real World

To address the inevitable sim-to-real gaps, we propose that PoseNet utilizes selected trained components from the CED, combined with an additional MLP module. This module takes the latent representations of real-time camera images and the initially predicted 6D pose as inputs, computing compensation values to refine the predicted pose, resulting in more precise estimation outcomes. This design is inspired by the classical feedback control. Then, the PoseNet is fine-tuned using the pairs of probe pose and camera readings in the real setting. The finetuning loss is designed as follows:

$$\mathcal{L} = \eta\mathcal{L}_i^{rec} + \|P - \hat{P}\|^2 \tag{5}$$

where $\eta = 0.1$ is here to balance the importance of pose regression; $\hat{P} = \bar{P} + \Delta P$ is the final prediction results, $\bar{P} = g_p(f_i(I))$ is the inference of the CED module, $\Delta P = MLP(f_i(I), f_p(\bar{P}))$, $g_p(\cdot)$ denotes the pose decoder. With such loss, PoseNet will adjust the image encoder to align with visual perception from the real world; the MLP layer that receives the encoded feature from the image encoder and pose encoder is in charge of adapting to adjust the prediction results from the pre-trained CED modules.

C. Training Details

The proposed networks were trained on a workstation with an NVIDIA GeForce RTX 4070 GPU and an Intel i7-13700KF CPU, using the Adam optimizer and a CosineAnnealingLR scheduler. For the CED module, the training was conducted over 100 epochs with a batch size of 16, and the learning rate ranged from a maximum of 5e-4 to a minimum of 1e-5. For PoseNet, training was performed over 40 epochs with the same batch size and learning rate range. Initially, the encoders and decoders were frozen for the first five epochs to allow the MLP layer to warm up, after which the entire network was fine-tuned for the remaining epochs.

IV. Experimental Results

This section presents the quantitative results of the proposed methods in both simulation and real-world scenarios, alongside a comparison with recent US image-based deep learning techniques and classical ArUco marker-based methods. The reconstruction outcomes are visualized by stacking US images in 3D space based on the computed probe poses, demonstrating performance on a human-like arm phantom.

A. Performance on Simulation Data

In order to use the network to predict the probe pose based on the two lightweight camera observations, we first

TABLE I

RESULTS ON SIMULATION DATA

Methods	Translation Error (mm)	Orientation Error (deg)
ResNet-MLP	1.57±1.21	0.52±0.32
ResNet-MLP w. Att.	1.60±1.21 (2%↑)	0.47±0.28 (9%↓)
CED	**1.35±1.22** (14%↓)	**0.26±0.25** (50%↓)

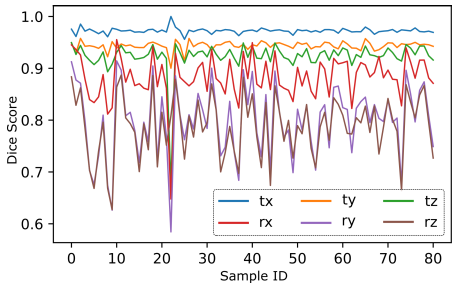

Fig. 3. Sensitivity analysis of the translation and orientation motions.

train the cross-encoder-decoder module using $200K$ pairs of pose and camera images. It is noted that the camera-based method is good for orientation estimation while less sensitive to the change in translations. To investigate this, we sampled 80 images by randomly maneuvering the probe in each translation direction and rotational direction, respectively. The translational and rotational step interval is 3 mm and 1.5°. To demonstrate whether the visual observation from cameras is sensitive to pose variation in each degree of freedom, we compute the dice score [24] between every two consecutive camera images. The result is depicted in Fig. 3.

The figure shows that the Dice score for all three translation directions remains relatively stable, indicating that visual observations are less sensitive to translation, leading to relatively larger errors in translation estimation. In contrast, the Dice score for rotational directions fluctuates significantly (ranging from 0.7 to 0.9), suggesting that even small rotations cause substantial changes in camera images. This indicates that the model is more sensitive to rotational changes than to translational ones. This phenomenon presents a challenge for learning accurate translation regression. To improve overall pose estimation, a deep network is required to effectively encode images and capture subtle changes in translation.

Based on the findings, we adopt the ResNet18 structure [25] instead of a conventional CNN layer to extract deep features from camera observations. Furthermore, to demonstrate the effectiveness of the hybrid loss in the proposed cross encoder-decoder (CED) module [see Fig. 2 (a)], we assess pose estimation performance using simulation data. For comparison, the "ResNet-MLP" model is used as a baseline, where camera images are fed into ResNet18, and the pose is predicted via an MLP with three fully connected layers, mapping latent features to dimensions of 1024, 256, and 6, respectively. Besides, to further refine the spatial feature for pose estimation, we integrated a spatial attention module [26] in ResNet for comparison, which is termed as "ResNet-MLP w. Att." in this study. The results of 2000 random samples in the simulation are summarized in Table I. The proposed CED significantly outperforms other methods, reducing translation and orientation errors by 14% and 50%, respectively. This improvement is due to CED's ability to force the network to learn the implicit relationship between the pose and the perceived image. Additionally, the attention mechanism did not yield better results, probably because the input camera images are binary, making it difficult for attention to further refine the active regions for pose prediction.

B. Performance in Real-World Setting

To evaluate the performance of the PoseNet in real-world applications, we fine-tune the network using 4.4k tracked images with paired poses. The statistical results for each axis and the overall errors are presented in Table II. The ArUco markers were placed on the table to ensure they remained within FoV of the left camera during each scan. We can see that the performance improves with larger ArUco markers. However, the ArUco method exhibits the largest maximum drifts (48.87 mm for the small marker and 38.99 mm for the large marker), indicating poor performance. This significant deviation may be attributed to non-perfect internal and external camera calibration. Additionally, lighting conditions and large angles between the camera and the ArUco marker likely reduce performance as well. However, all these dependencies indicate that the ArUco-based method is hard to apply in clinical practice. In order to provide an overview of the results of the methods that predict the relative poses. In table II, we listed some representative works' results reported in the corresponding cited papers to ensure a relatively fair comparison. Due to the accumulated error nature and large maximum drifts (7.65 mm to 48.87 mm), even in their own applications, such methods are not good enough for practical usage. In contrast, the proposed methods deliver accurate predictions for each axis, especially for orientation accuracy. However, the estimation of y-axis translation exhibits a relatively large deviation (1.62 mm in Table II), which corresponds to the lateral direction of the scan. This inaccuracy can account for the bias of the fine-tuning dataset, which lacks sufficient samples that move along the y-axis.

Fig. 4 provides a detailed example of the trajectory prediction results. In plots (a) and (b), the predicted trajectories from the ArUco method exhibit significant fluctuations around the ground truth for both translation and orientation, which prevents the reconstruction of a plausible 3D volume of the target vessel [mean errors of 20 mm and 18 mm for ArUco small and ArUco large respectively]. The violin plots also indicate high standard deviations. Although the proposed method demonstrates significantly lower prediction errors, we can still see noises around the predicted trajectories. The noise in the prediction results can be attributed to variations in lighting, which affect the segmentation of the marker borders in the real-world environment. Considering the fact that the movement of the probe is continuous, a smoothing filter with a kernel size of 5 is applied on the predicted trajectories by PoseNet. The violin plots on the right-hand side of Fig. 4 show that after smoothing, the mean errors for both translation and orientation are reduced. Nonetheless, by smoothing the predicted trajectory, we reconstruct a plausible 3D volume of the target vessel [see bottom right of Fig. 4], demonstrating the practicality of the proposed method.

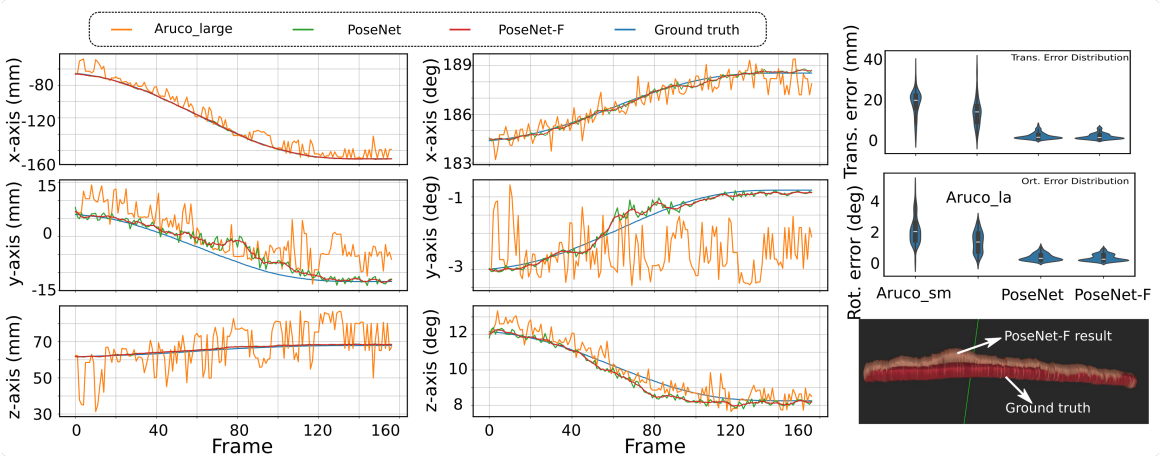

Fig. 4. A detailed performance comparison in real-world scenarios. The *Left* and *Middle* plots display the translation and orientation predictions for each axis. The *Upper Right* shows violin plots of the prediction results, while the *Bottom Right* overlays the 3D reconstruction of PoseNet's filtered prediction trajectory with the ground truth. The 3D reconstruction process was performed offline using *ImFusion Suite* (ImFusion GmbH, Munich, Germany).

TABLE II

PERFORMANCE COMPARISON WITH OTHER METHODS IN REAL APPLICATION

Methods	**Average Absolute Translation Errors (mm)**					**Average Absolute Orientation Errors (°)**			
	t_x	t_y	t_z	t_{all}	Maximum Drift	θ_x	θ_y	θ_z	θ_{all}
CNN-OF [17]*	1.58	2.86	1.47	-	7.65	1.37	0.84	0.98	-
DCL-Net [18]*	-	-	-	10.33	27.03	-	-	-	-
MoNet [21]*	-	-	-	-	10.24	-	-	-	1.55
RecON [27]*	-	-	-	-	18.69	-	-	-	2.26
ArUco small†	10.01	11.56	16.27	25.17	48.87	1.64	1.06	1.31	2.68
ArUco large†	4.38	5.77	12.59	16.21	38.99	0.76	1.54	0.58	1.93
Proposed method	**0.56**	**1.62**	**0.72**	**2.03**	**5.89**	**0.17**	**0.14**	**0.24**	**0.37**

*: The numbers were reported in the original references.
†: The ArUco method is sensitive to experimental conditions, and the results shown are from a random trial.

V. DISCUSSION

In this study, we proposed a low-cost, versatile probe localization method for freehand 3D US imaging using two lightweight cameras. Compared with US image-based solutions that are prone to accumulate relative pose errors, the proposed method focuses on global pose estimation. This can significantly improve reconstruction accuracy overall, particularly for long sweeps. While using ArUco markers can offer global estimations, these approaches suffer from accumulated calibration errors due to the kinematic chain required to compute the final US probe pose. Moreover, factors such as lighting conditions, distance, and marker quality can hinder the accuracy and stability of ArUco detection. In this regard, the proposed method is more adaptable to calibration inaccuracies in both the camera's internal and external parameters.

In addition to the significant improvements in accuracy and stability, there are some limitations to consider. First, for optimal results in real-world scenarios, accurate marker segmentation is essential. In this study, markers were extracted from the background using an HSV threshold. However, future work could explore training a more reliable segmentation network, such as U-Net [28], to detect these structural markers with greater precision. The asynchrony in camera streaming can also result in reduced performance. Additionally, as discussed in Section IV-A, camera-based methods typically suffer from low resolution in translation. While our use of ResNet as an encoder and a reconstruction task mitigates this problem, it does not fully resolve the inherent limitations in translational accuracy. To tackle this problem further in the future, there are several ways to improve translation accuracy, such as replacing the 2D markers with 3D markers to provide more feature information and using a mask autoencoder to complete the occluded markers.

VI. CONCLUSION

This work presents a low-cost, robust, and precise solution for freehand 3D US reconstruction based on two lightweight cameras. In order to have an effective encoder to extract the distinct visual difference from camera observation, the CED module was trained using $200K$ simulated pairs of 6D probe pose and camera images. Then, the proposed PoseNet integrates selected components from the trained CED along with an additional MLP module that was fine-tuned based on the data obtained in real scenarios. The experimental results demonstrate that PoseNet can significantly outperform existing methods in terms of prediction accuracy (2.03 mm and 0.37° for transnational and rotational error), as it effectively eliminates accumulated errors. Overall, the proposed method offers a promising solution for freehand 3D US reconstruction and holds potential for integration with US image-based methods, which could further enhance performance and accuracy in clinical applications. Future studies will need to address practical challenges in US scanning further, such as motion-aware [29], [30] and force-induced deformation-aware [31], [32] 3D US imaging systems.

REFERENCES

[1] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Medical image analysis*, p. 102878, 2023.

[2] Y. Bi, Z. Jiang, F. Duelmer, D. Huang, and N. Navab, "Machine learning in robotic ultrasound imaging: Challenges and perspectives," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7.

[3] J. A. Hossack, T. S. Sumanaweera, S. Napel, and J. S. Ha, "Quantitative 3-d diagnostic ultrasound imaging using a modified transducer array and an automated image tracking technique," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 49, no. 8, pp. 1029–1038, 2002.

[4] P. Chatelain, A. Krupa, and N. Navab, "Confidence-driven control of an ultrasound probe," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1410–1424, 2017.

[5] A. Lang, P. Mousavi, G. Fichtinger, and P. Abolmaesumi, "Fusion of electromagnetic tracking with speckle-tracked 3d freehand ultrasound using an unscented kalman filter," in *Medical Imaging 2009: Ultrasonic Imaging and Signal Processing*, vol. 7265. SPIE, 2009, pp. 399–410.

[6] J. Guerrero, S. E. Salcudean, J. A. McEwen, B. A. Masri, and S. Nicolaou, "Real-time vessel segmentation and tracking for ultrasound imaging applications," *IEEE transactions on medical imaging*, vol. 26, no. 8, pp. 1079–1090, 2007.

[7] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "Plus: open-source toolkit for ultrasound-guided intervention systems," *IEEE transactions on biomedical engineering*, vol. 61, no. 10, pp. 2527–2537, 2014.

[8] Z. Jiang, Y. Gao, L. Xie, and N. Navab, "Towards autonomous atlas-based ultrasound acquisitions in presence of articulated motion," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7423–7430, 2022.

[9] Z. Jiang, Y. Kang, Y. Bi, X. Li, C. Li, and N. Navab, "Class-aware cartilage segmentation for autonomous us-ct registration in robotic intercostal ultrasound imaging," *IEEE Transactions on Automation Science and Engineering*, 2024.

[10] Z. Jiang, X. Li, C. Zhang, Y. Bi, W. Stechele, and N. Navab, "Skeleton graph-based ultrasound-ct non-rigid registration," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4394–4401, 2023.

[11] M. Chen, Y. Huang, J. Chen, T. Zhou, J. Chen, and H. Liu, "Fully robotized 3d ultrasound image acquisition for artery," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2690–2696.

[12] Q. Huang, J. Lan, and X. Li, "Robotic arm based automatic ultrasound scanning for three-dimensional imaging," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1173–1182, 2018.

[13] X. Ma, M. Zeng, J. C. Hill, B. Hoffmann, Z. Zhang, and H. K. Zhang, "Guiding the last centimeter: Novel anatomy-aware probe servoing for standardized imaging plane navigation in robotic lung ultrasound," *arXiv preprint arXiv:2406.11523*, 2024.

[14] Q. Huang, B. Gao, and M. Wang, "Robot-assisted autonomous ultrasound imaging for carotid artery," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[15] J. Tan, J. Li, Y. Li, B. Li, Y. Leng, Y. Rong, and C. Fu, "Autonomous trajectory planning for ultrasound-guided real-time tracking of suspicious breast tumor targets," *IEEE Transactions on Automation Science and Engineering*, 2023.

[16] Z. Jiang, Y. Bi, M. Zhou, Y. Hu, M. Burke, and N. Navab, "Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations," *The International Journal of Robotics Research*, vol. 43, no. 7, pp. 981–1002, 2024.

[17] R. Prevost, M. Salehi, S. Jagoda, N. Kumar, J. Sprung, A. Ladikos, R. Bauer, O. Zettinig, and W. Wein, "3d freehand ultrasound without external tracking using deep learning," *Medical image analysis*, vol. 48, pp. 187–202, 2018.

[18] H. Guo, S. Xu, B. Wood, and P. Yan, "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*. Springer, 2020, pp. 463–472.

[19] M. Luo, X. Yang, X. Huang, Y. Huang, Y. Zou, X. Hu, N. Ravikumar, A. F. Frangi, and D. Ni, "Self context and shape prior for sensorless freehand 3d ultrasound reconstruction," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer, 2021, pp. 201–210.

[20] Q. Li, Z. Shen, Q. Li, D. C. Barratt, T. Dowrick, M. J. Clarkson, T. Vercauteren, and Y. Hu, "Long-term dependency for 3d reconstruction of freehand ultrasound without external tracker," *IEEE Transactions on Biomedical Engineering*, 2023.

[21] M. Luo, X. Yang, H. Wang, L. Du, and D. Ni, "Deep motion network for freehand 3d ultrasound reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 290–299.

[22] S.-Y. Sun, M. Gilbertson, and B. W. Anthony, "Probe localization for freehand 3d ultrasound by tracking skin features," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17*. Springer, 2014, pp. 365–372.

[23] E. Rohmer, S. P. Singh, and M. Freese, "V-rep: A versatile and scalable robot simulation framework," in *2013 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2013, pp. 1321–1326.

[24] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[27] M. Luo, X. Yang, H. Wang, H. Dou, X. Hu, Y. Huang, N. Ravikumar, S. Xu, Y. Zhang, Y. Xiong *et al.*, "Recon: Online learning for sensorless freehand 3d ultrasound reconstruction," *Medical Image Analysis*, vol. 87, p. 102810, 2023.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[29] Z. Jiang, H. Wang, Z. Li, M. Grimm, M. Zhou, U. Eck, S. V. Brecht, T. C. Lueth, T. Wendler, and N. Navab, "Motion-aware robotic 3d ultrasound," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 494–12 500.

[30] Z. Jiang, N. Danis, Y. Bi, M. Zhou, M. Kroenke, T. Wendler, and N. Navab, "Precise repositioning of robotic ultrasound: Improving registration-based motion compensation using ultrasound confidence optimization," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[31] Z. Jiang, Y. Zhou, D. Cao, and N. Navab, "Defcor-net: physics-aware ultrasound deformation correction," *Medical Image Analysis*, vol. 90, p. 102923, 2023.

[32] Z. Jiang, Y. Zhou, Y. Bi, M. Zhou, T. Wendler, and N. Navab, "Deformation-aware robotic 3d ultrasound," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7675–7682, 2021.