

Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames

Qi Li¹, Ziyi Shen¹, Qian Li^{1,2}, Dean C Barratt¹, Thomas Dowrick¹, Matthew J Clarkson¹, Tom Vercauteren³, and Yipeng Hu¹

¹ Centre for Medical Image Computing, Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Department of Medical Physics and Biomedical Engineering, University College London, London, U.K.

² State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin, China

³ School of Biomedical Engineering & Imaging Sciences, King's College London, London, U.K.

Abstract. Three-dimensional (3D) freehand ultrasound (US) reconstruction without a tracker can be advantageous over its two-dimensional or tracked counterparts in many clinical applications. In this paper, we propose to estimate 3D spatial transformation between US frames from both past and future 2D images, using feed-forward and recurrent neural networks (RNNs). With the temporally available frames, a further multi-task learning algorithm is proposed to utilise a large number of auxiliary transformation-predicting tasks between them. Using more than 40,000 US frames acquired from 228 scans on 38 forearms of 19 volunteers in a volunteer study, the hold-out test performance is quantified by frame prediction accuracy, volume reconstruction overlap, accumulated tracking error and final drift, based on ground-truth from an optical tracker. The results show the importance of modelling the temporal-spatially correlated input frames as well as output transformations, with further improvement owing to additional past and/or future frames. The best performing model was associated with predicting transformation between moderately-spaced frames, with an interval of less than ten frames at 20 frames per second (fps). Little benefit was observed by adding frames more than one second away from the predicted transformation, with or without LSTM-based RNNs. Interestingly, with the proposed approach, explicit within-sequence loss that encourages consistency in composing transformations or minimises accumulated error may no longer be required. The implementation code and volunteer data¹ will be made publicly available ensuring reproducibility and further research.

Keywords: 3D freehand US, transformation estimation, multi-task learning, sequence encoding

¹ <https://github.com/ucl-candi/freehand>

1 Introduction

Reconstructing freehand ultrasound in 3D provides spatial information between acquired 2D frames, potentially for a wide range of clinical applications, and has indeed been adopted in areas including surgical and interventional guidance. In these applications, 3D reconstruction of the anatomy and pathology is essential for tasks such as registration to pre-operative imaging [6] and quantifying 3D tissue motion [7]. It provides a low-cost, accessible alternative with larger and more flexible fields-of-view to the other 3D US imaging techniques such as 2D array transducer [15] and motorised probe. Spatial tracking, electromagnetic or optical, is currently considered most robust approaches for 3D freehand US, but poses practical challenges for clinical adoption, due to the additional requirement such as extra equipment, line-of-sight or interference mitigation. Therefore, tracker-free or image-based methods have generated long-lasting research interest, from previous work in exploiting physics-based speckle correlation models between image frames [1,3] to, more recently, machine learning-based methods [17,16], driven by supervising data often from spatial trackers for training.

Prevost *et. al.* [18] proposed a convolutional neural network (CNN) to reconstruct 3D volume by estimating the transformation between two adjacent 2D images. FlowNet and densely connected networks were used in [11,21]. In [12], ResNet and FlowNetS were integrated for a better localization and optical flow estimation, and consistency loss derived from stereo vision was added. Forward consistency loss was then proposed in [14]. In [13], RNN was used to estimate both relative and absolute probe poses. In [2], 3D CNN and Pearson correlation coefficient based case-wise correlation loss was proposed to enable more smooth trajectories. A novel online learning framework with self-supervised learning method and adversarial training was proposed in [9]. The authors then integrated the IMU information both in training and inference time to extract velocity information and reduce drift error [10]. Recently, [16] used ResNet and transformer to extract local and global features of US sequence.

This work builds on the previous effort in this challenging application and formulates the freehand US transformation estimation problem as a multi-task learning problem, not only focusing on the one transformation between a pair of images (main task) but a set of transformations (auxiliary tasks) between frames of the input image sequence. We show that this formulation is effective to capture strong correlation among input frames and that among output between-frame transformations. It is a generalised algorithm that 1) includes future frames in addition to past frames and their potential correlation; and 2) predicts correlated neighbouring transformations in addition to the main task and takes advantage of cyclic and accumulative consistency between them.

Our contributions include: 1) a new design of freehand US sequence encoding in a novel multi-transformation learning algorithm; 2) extensive experimental results to quantify the benefits from the proposed methodological components; and 3) code and volunteer data for public access.

2 Method

For an US scan consisting of a set of 2D image frames \mathcal{S} , image sequences with a length of M can be sampled $S = \{I_m\}, m = 1, 2, \dots, M$, where $S \subseteq \mathcal{S}$ and m denotes consecutively increasing time-steps at which the frames are acquired. For a given sequence, a spatial transformation $T_{j \leftarrow i}, 1 \leq i < j \leq M$ denotes the relative translation and rotation between the i^{th} and j^{th} frames. This section describes our proposed method to predict the spatial transformation $T_{j^* \leftarrow i^*}$ between a pair of frames (i^*, j^*) , with a $j^* - i^*$ interval, $i^* - 1$ past frames and $M - j^*$ future frames.

After models are trained by sequences randomly sampled from training US scans, a test scan can then be reconstructed by consecutively predicting multiple sequences with a predefined M , such that the $(j^*)^{th}$ frame from the previous sequence is the $(i^*)^{th}$ frame in the subsequent sequence, for the entire scan with variable length. All frames after the initial j^* can be spatially localised with respect to their varying starting reference frame. Different values of M are tested to include potentially useful long-term dependency.

2.1 Input sequence encoding

A recurrent neural network f_{rec} with parameters θ takes the image frames in sequence to predict $T_{j^* \leftarrow i^*}$:

$$T_{j^* \leftarrow i^*} = f_{rec}(I_m, h^{(m-1)}; \theta), \text{ for } m = M \quad (1)$$

$$h^{(m)} = f_{rec}(I_m, h^{(m-1)}; \theta), \forall m \leq M - 1 \quad (2)$$

where $h^{(m)}$ is the internal hidden state at time-step m and the transformation $T_{j^* \leftarrow i^*}$ is predicted at the end of each sequence. Here, the future frames are used if $j^* < M$, leading to a time-delayed transformation prediction. Feed-forward CNN f_{fwd} are also tested to model the same image sequence without considering the sequential steps explicitly:

$$T_{j^* \leftarrow i^*} = f_{fwd}(S; \theta). \quad (3)$$

Given a predefined pair indices (i^*, j^*) and the sequence length M , this formulation includes permutations of available neighbouring frames $I_{m \in ([1, i^* - 1] \cup [j^* + 1, M])}$ and their relative positions, as shown in the example in Fig. 1.

In this work, we propose to consider (i^*, j^*) as hyperparameters, tuned on validation set. This together with M is equivalent to a flexible, generalised frame-encoding that provides a conditioning context for spatial transformation prediction. For example, smaller i^* and $M - j^*$ indicate prediction using a shorter history and fewer future frames, respectively; and a single-pair input is represented by $M = 2$, found in several previous studies. The benefits of an effective and efficient context-enabled encoding have been studied in related areas such as n-gram encoding [19].

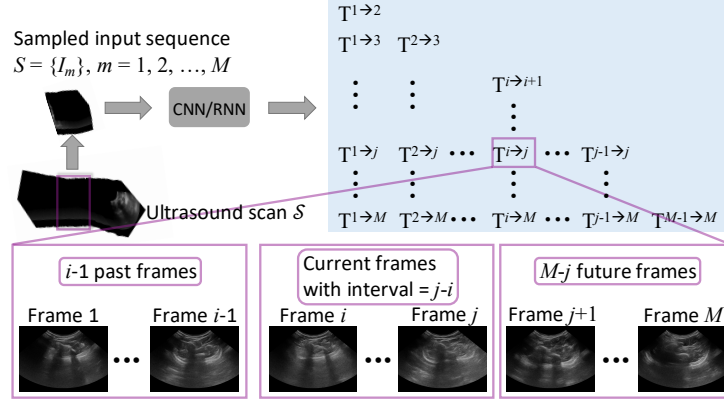


Fig. 1: Illustration of input sequence encoding and multi-task output in the proposed method.

Efficient frame encoding is particularly important in this application due to memory required, for both feed-forward and unfolded recurrent networks, with the high-dimensional image input and potentially long US sequence, and, as shown in this study, may warrant much shorter sequences required for tracking and scan reconstruction. Furthermore, this simultaneously enables a practically data structure for the multi-task learning described in Sec. 2.2.

2.2 Multi-task learning

Whilst predicting $T_{j^* \leftarrow i^*}$ is regarded as the main task, both the recurrent and feed-forward networks can be adapted to predict other transformations $T_{j \leftarrow i}$, $i \neq i^*$ or $j \neq j^*$. This work proposes to predict all these possible $C_M^2 - 1$ transformations as auxiliary tasks, also illustrated in Fig. 1. The differences between the predicted $\hat{T}_{j \leftarrow i}$ and ground-truth $T_{j \leftarrow i}^{(gt)}$ can be averaged as the overall loss for network training. When C_M^2 is very large, randomly selected τ samples of the auxiliary tasks may be used instead, $\tau \leq C_M^2 - 1$.

The proposed multi-task learning for the “neighbouring transformations”, albeit conceptually simple, not only exploits the shared representation from the auxiliary transformation prediction tasks, but also facilitates other losses based on these correlated transformations, such as the consistency loss and the alternative accumulated loss in Sec. 2.3.

2.3 Loss functions

To alleviate empirical tuning between rotational and translational contributions, this work adopts loss functions based on the distance between the prediction-transformed points $\hat{p}_n^{(j)}$, $n = 1, \dots, N$ and the ground-truth-transformed points $p_n^{(j)}$, by the predicted $\hat{T}_{j \leftarrow i}$ and ground-truth $T_{j \leftarrow i}^{(gt)}$, respectively. In this work,

$N = 4$ corner points in j^{th} image are used, in their homogeneous tracking *tool space*. Therefore, the multi-task loss function is the average of the mean-square-errors (MSEs) over the $\tau + 1$ tasks:

$$\mathcal{L}_{multi-task} = \frac{1}{N \cdot (\tau + 1)} \sum_{n=1}^{\tau+1} \sum_{n=1}^N D(p_n^{(j)}, \hat{p}_n^{(j)}) \quad (4)$$

where $D(\cdot)$ denotes MSE between x , y and z coordinates of the two point sets, $\hat{p}_n^{(j)} = \hat{T}_{j \leftarrow i} \cdot T_{(calib)} \cdot p_n^{(i)}$, and $p_n^{(j)} = T_{j \leftarrow i}^{(gt)} \cdot T_{(calib)} \cdot p_n^{(i)}$. $p_n^{(i)}$ represents the same points in the i^{th} image space, and $T_{(calib)}$ is a fixed transformation from image space to tool space, obtained through calibration.

The ground-truth transformation is composed by two tool-to-world transformations, $T_{j \leftarrow i}^{(gt)} = (T_{world \leftarrow j}^{(gt)})^{-1} \cdot T_{world \leftarrow i}^{(gt)}$, at the time-steps i and j , obtained from the optical tracker, thus independent of the *world (camera) space*. A further left-multiplication by $(T_{(calib)})^{-1}$ could compute distance defined in the image space, should it be preferred.

Among the predicted transformations, consistency may be enforced between a direct prediction $\hat{T}_{j \leftarrow i}$ and an indirect prediction $\hat{T}_{j \leftarrow i}^{\oplus} = \hat{T}_{j \leftarrow k} \cdot \hat{T}_{k \leftarrow i}$. Given each time-step k and additional transformations to and from k , a set of consistency losses on the transformed points can be defined for each task:

$$\mathcal{L}_{consistency} = \frac{1}{N} \sum_{n=1}^N D(\hat{p}_n^{(j)}, \hat{p}_n^{(j)\oplus}) \quad (5)$$

where $\hat{p}_n^{(j)\oplus} = \hat{T}_{j \leftarrow i}^{\oplus} \cdot T_{(calib)} \cdot p_n^{(i)}$. This loss function only promotes consistency and does not require ground-truth data. It should be used in conjunction with Eq. 4 (or Eq. 6) to avoid trivial solutions. Importantly, the consistency loss is a form of “teacher forcing” commonly adopted in training sequence models, which makes use of the ground-truth targets, rather than the previous predictions, to supervise the subsequent prediction during training. It has been proven advantageous in sequence-to-sequence models [14].

Alternatively, minimising the difference between $\hat{p}_n^{(j)\oplus}$ and ground-truth $p_n^{(j)}$ forms an accumulated loss:

$$\mathcal{L}_{accumulated} = \frac{1}{N} \sum_{n=1}^N D(p_n^{(j)}, \hat{p}_n^{(j)\oplus}) \quad (6)$$

Although not investigated in this work, finding the optimal relative weighting between these loss terms should further improve the proposed method and be of interest in future studies. The reported results in Sec. 3 used equal weighting to provide a reference performance.

2.4 Evaluation metrics

For each sequence, the Euclidean distance between prediction and ground-truth on four corner points of consecutive frames is defined as *frame prediction accuracy* (ϵ_{frame}), which assesses model generalisation without scan reconstruction.

For each reconstructed scan, two reconstruction errors are reported: 1) an *accumulated tracking error* ($\epsilon_{acc.}$) is the average Euclidean distance over all reconstructed image pixel locations; and 2) a *volume reconstruction overlap* (ϵ_{dice}) measure, Dice between the reconstructed volumes of prediction and ground-truth, where a reconstructed volume is approximated with hexahedrons formed by two adjacent frames. A *final drift* (ϵ_{drift}) is also reported as the Euclidean distance, averaged over the four corners, between the final predicted and ground-truth frames in each scan.

All results are reported on the hold-out test set, unseen to model training and development. These error metrics are designed for a range of freehand US applications that may have different clinical focuses [15].

3 Experiments and results

3.1 Data acquisition

Freehand US scans were acquired on both left and right forearms from 19 volunteers. On each forearm, the US probe was moved, for study purpose, in a straight line, a ‘C’ shape and a ‘S’ shape, in a distal-to-proximal direction. These three scans were repeated, with the curved-linear transducer held (thus the US planes) perpendicular of and parallel to the forearm. After manually cropping the initial and end stages when the probe was largely stationary, between 36 and 430 frames with a size of 480×640 pixels, equivalent to a probe travel distances between 100 and 200 mm, were included. One scan with less than 50 frames was discarded for its uncertain quality. The data was split into train, validation and test sets by a ratio of 3:1:1, without the same forearm in different sets. All US scans were acquired on Ultrasonix machine (BK, Europe) with a curve-linear probe (4DC7-3/40), tracked by an NDI Polaris Vicra (Northern Digital Inc., Canada). B-mode images with median level of speckle reduction were recorded at 20 fps. Spatial (image-to-tool) and temporal differences were calibrated using a pinhead-based method [5] and the Plus Toolkit [8], respectively.

3.2 Network development and implementation

This work aims to provide an established network performance, without focusing on further architecture optimisation. The EfficientNet (b1) [20] was adapted as the feed-forward CNN, with a no-activation output layer to predict $(\tau + 1) \times 6$ dimensional vectors representing the multi-task predictions. The same EfficientNet-based feature encoder followed by a long short-term memory (LSTM) module [4], with a 1024-dimensional hidden feature vector, was used as the recurrent network. A baseline CNN was also trained, with two adjacent frames as input and output transformation between them. A minibatch size of 32 and the Adam optimizer were used to train each model for 50,000 epochs. The best model with the minimum frame prediction accuracy on the validation set was selected, and then report the results on the test set. In addition to the network

Table 1: Reconstruction performance of baseline and proposed method.

Evaluation metrics(mm)	Baseline	$T_{6\leftarrow 1}$	$T_{10\leftarrow 5}$	$T_{9\leftarrow 6}$	$T_{10\leftarrow 6}$	$T_{9\leftarrow 6}^{accumulated}$	$T_{9\leftarrow 6}^{consistency}$
$\epsilon_{frame-cnn}$	0.63 ± 0.54	0.55 ± 0.57	0.53 ± 0.56	0.57 ± 0.57	0.55 ± 0.56	0.58 ± 0.60	0.58 ± 0.59
$\epsilon_{frame-LSTM}$	0.66 ± 0.46	0.53 ± 0.42	0.50 ± 0.41	0.53 ± 0.43	0.51 ± 0.41	0.54 ± 0.44	0.56 ± 0.48
$\epsilon_{acc-cnn}$	24.42 ± 17.17	19.05 ± 13.64	19.09 ± 14.60	19.03 ± 13.68	19.15 ± 14.33	20.94 ± 14.58	20.98 ± 15.18
$\epsilon_{acc-LSTM}$	27.91 ± 15.39	19.18 ± 10.19	18.13 ± 9.49	18.21 ± 9.18	18.56 ± 9.65	20.35 ± 9.82	20.52 ± 13.54
$\epsilon_{dice-cnn}$	0.72 ± 0.22	0.80 ± 0.11	0.81 ± 0.11	0.79 ± 0.12	0.80 ± 0.11	0.76 ± 0.20	0.75 ± 0.22
$\epsilon_{dice-LSTM}$	0.68 ± 0.20	0.76 ± 0.12	0.78 ± 0.12	0.78 ± 0.11	0.78 ± 0.11	0.65 ± 0.48	0.76 ± 0.15
$\epsilon_{drift-cnn}$	46.01 ± 33.34	37.96 ± 27.98	36.82 ± 28.01	37.19 ± 27.37	36.93 ± 27.63	42.33 ± 27.48	40.54 ± 30.64
$\epsilon_{drift-LSTM}$	51.43 ± 30.30	40.56 ± 24.29	36.48 ± 20.77	37.26 ± 20.73	37.36 ± 20.75	40.33 ± 22.09	39.50 ± 27.03

and training options described above and those in Sec. 2, other hyperparameter values including a learning rate of 10^{-4} , tested among $\{10^{-3}, 10^{-4}, 10^{-5}\}$, and a sequence length of 20, tested among $\{10, 20, 30, 40, 49\}$, were selected with $\tau = 79$ based on the validation set performance.

3.3 Comparison to the baseline and ablation study

On the hold-out test set, ablation studies quantify the impact on the performance due to 1) the addition of auxiliary tasks, 2) the number of past frames and future frames included in input sequence; 3) the frame interval $j - i$ between which the transformation is predicted; and 4) the choice between feed-forward CNNs and LSTM-based RNNs.

As shown in Table. 1 and Fig. 2, both ϵ_{frame} and $\epsilon_{acc.}$ were improved after adding auxiliary tasks, regardless their permutations, compared with the baseline ($p \leq 0.001$ for ϵ_{frame} , $\epsilon_{acc.}$, ϵ_{drift} and $p \leq 0.033$ for ϵ_{dice} , paired t-tests at $\alpha = 0.05$), where $\epsilon_{acc.}$ increases with time, whilst ϵ_{frame} is relatively stable between sequence locations in the scan. ϵ_{dice} was computed on the perpendicular scans as an example.

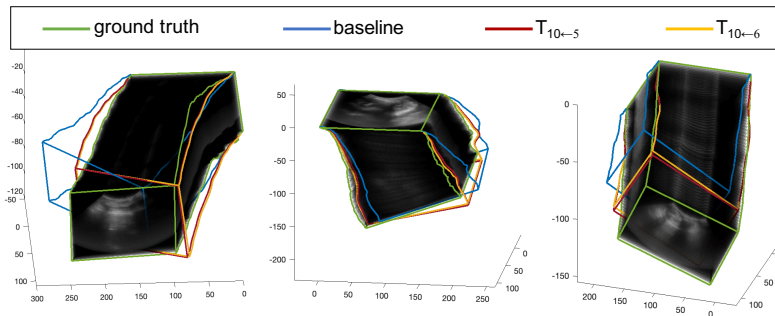


Fig. 2: Reconstruction results of baseline and proposed method (using $T_{10\leftarrow 5}$ and $T_{10\leftarrow 6}$).

Fig. 3 plots the performance in $\epsilon_{acc.}$ versus variable intervals, number of past and future frames, respectively. It shows that a relatively short interval, for both

CNN and LSTM, between 3 and 9, resulted lower errors (e.g. unpaired $p=0.010$, LSTM at interval=9 vs. baseline). The use of past and future frames was clearly beneficial, compared with those without, i.e., $x = 0$ in Fig. 3 b and c. However, an interesting observation is that, performance improved when <5 past frames was added, whilst additional 9-11 future frames values offered lowest $\epsilon_{acc.}$. The need for longer-term dependency was unsubstantiated, for example no significant improvement was found by increasing the sequence length beyond 20 during model development. However, the RNNs yielded consistent lower prediction variance, as shown in Fig. 3, which may indicate a superior within-sequence modelling. ϵ_{dice} and ϵ_{drift} showed consistent conclusions to those based on $\epsilon_{acc.}$, therefore omitted for brevity in the plots.

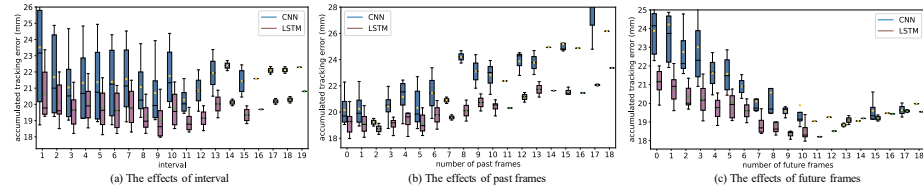


Fig. 3: The performance of accumulated tracking error with various intervals, number of past and future frames.

Fig. 4 plots mean and variance of ϵ_{frame} and $\epsilon_{acc.}$, over all scans in the test set, between baseline and the proposed CNN-based multi-task model. As an example in predicting $T_{10 \leftarrow 6}$, the improvement from the multi-task learning seems increased as the sequences accumulate.

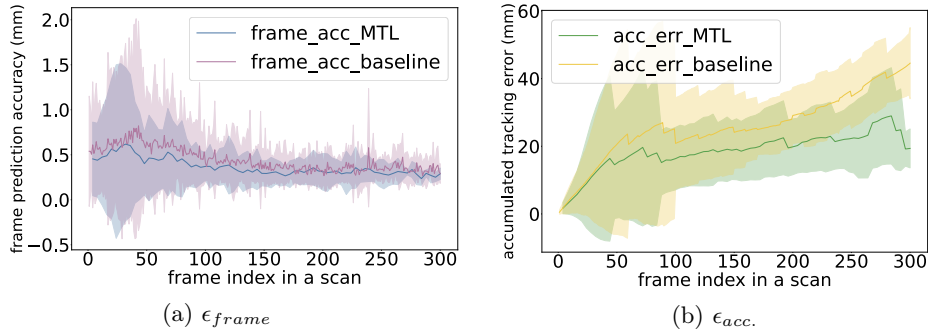


Fig. 4: Comparison of ϵ_{frame} and $\epsilon_{acc.}$ between baseline and proposed method.

In conclusion, the proposed trackerless freehand US improved baseline performance, by utilising sequence modelling and multi-tasking as hyperparameters,

supported by a set of extensive experiments. The published code and data should also be valuable for furthering research in this area.

Compliance with ethical standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of local institution.

Acknowledgement

This work was supported by the EPSRC [EP/T029404/1], a Royal Academy of Engineering / Medtronic Research Chair [RCSRF1819\7\734] (TV), and Wellcome/EPSRC Centre for Interventional and Surgical Sciences [203145Z/16/Z]. Qi Li was supported by the University College London Overseas and Graduate Research Scholarships.

References

1. Chen, J., Fowlkes, J., et al.: 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)
2. Guo, H., Xu, S., et al.: Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In: MICCAI. pp. 463–472. Springer (2020)
3. Hassenpflug, P., Prager, R., et al.: Speckle classification for sensorless freehand 3-d ultrasound. *Ultrasound in medicine & biology* **31**(11), 1499–1508 (2005)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
5. Hu, Y., Gibson, E., et al.: Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks. In: *Molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment*, pp. 105–115. Springer (2017)
6. Hu, Y., Kasivisvanathan, V., et al.: Development and phantom validation of a 3-d-ultrasound-guided system for targeting mri-visible lesions during transrectal prostate biopsy. *IEEE Transactions on Biomedical Engineering* **64**(4), 946–958 (2016)
7. Jiang, Z., Wang, H., et al.: Motion-aware robotic 3d ultrasound. In: ICRA. pp. 12494–12500. IEEE (2021)
8. Lasso, A., Heffter, T., et al.: Plus: open-source toolkit for ultrasound-guided intervention systems. *IEEE transactions on biomedical engineering* **61**(10), 2527–2537 (2014)
9. Luo, M., Yang, X., et al.: Self context and shape prior for sensorless freehand 3d ultrasound reconstruction. In: MICCAI. pp. 201–210. Springer (2021)
10. Luo, M., Yang, X., et al.: Deep motion network for freehand 3d ultrasound reconstruction. In: MICCAI. pp. 290–299. Springer (2022)
11. Mikaeili, M., Bilge, H.: Trajectory estimation of ultrasound images based on convolutional neural network. *Biomedical Signal Processing and Control* **78**, 103965 (2022)

12. Miura, K., Ito, K., et al.: Localizing 2d ultrasound probe from ultrasound image sequences using deep learning for volume reconstruction. In: ASMUS, pp. 97–105. Springer (2020)
13. Miura, K., Ito, K., et al.: Pose estimation of 2d ultrasound probe from ultrasound image sequences using cnn and rnn. In: ASMUS. pp. 96–105. Springer (2021)
14. Miura, K., Ito, K., et al.: Probe localization from ultrasound image sequences using deep learning for volume reconstruction. In: International Forum on Medical Imaging in Asia. vol. 11792, pp. 133–138. SPIE (2021)
15. Mozaffari, M., Lee, W.: Freehand 3-d ultrasound imaging: a systematic review. *Ultrasound in medicine & biology* **43**(10), 2099–2124 (2017)
16. Ning, G., Liang, H., et al.: Spatial position estimation method for 3d ultrasound reconstruction based on hybrid transformers. In: ISBI. pp. 1–5. IEEE (2022)
17. Prevost, R., Salehi, M., et al.: Deep learning for sensorless 3d freehand ultrasound imaging. In: MICCAI. pp. 628–636. Springer (2017)
18. Prevost, R., Salehi, M., et al.: 3d freehand ultrasound without external tracking using deep learning. *Medical image analysis* **48**, 187–202 (2018)
19. Takase, S., Suzuki, J., Nagata, M.: Character n-gram embeddings to improve rnn language models. In: AAAI. vol. 33, pp. 5074–5082 (2019)
20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
21. Xie, Y., Liao, H., et al.: Image-based 3d ultrasound reconstruction with optical flow via pyramid warping network. In: EMBC. pp. 3539–3542. IEEE (2021)