arXiv:2509.09530v2 [cs.CV] 17 Apr 2026

DualTrack: Sensorless 3D Ultrasound needs Local and Global Context

Paul F. R. Wilson[1], Matteo Ronchetti[2], Rüdiger Göbl[2], Viktoria Markova[2], Sebastian Rosenzweig[2], Raphael Prevost[2], Parvin Mousavi[1], and Oliver Zettinig[2]

[1] Queen's University, Kingston, Canada
[2] ImFusion GmbH, Munich, Germany
paul.wilson@queensu.ca

Abstract. Three-dimensional ultrasound (US) offers many clinical advantages over conventional 2D imaging, yet its widespread adoption is limited by the cost and complexity of traditional 3D systems. Sensorless 3D US, using deep learning to estimate a 3D probe trajectory from a sequence of 2D US images, is a promising alternative. *Local features* such as speckle patterns can help predict frame-to-frame motions, while *global features*, such as coarse shapes and anatomical structures, can situate the scan relative to anatomy and help predict its general shape. In prior approaches, global features are either ignored or tightly coupled with local feature extraction, restricting the ability to robustly model these two complementary aspects. We propose DualTrack, a novel dual encoder architecture leveraging decoupled *local* and *global* encoders specializing in their respective scale of feature extraction. The local encoder uses dense spatiotemporal convolutions to capture fine-grained features, while the global encoder utilizes an image backbone such as a 2D CNN or foundation model and temporal attention layers to embed high-level anatomical features and long-range dependencies. A lightweight fusion module then combines these features to estimate trajectory. Experimental results on a large public benchmark show that DualTrack achieves state-of-the-art accuracy and globally consistent 3D reconstructions, outperforming previous methods and yielding an average reconstruction error below 5 mm.[3]

Keywords: 3D Ultrasound · Dual Encoder · Deep Learning

1 Introduction

3D ultrasound (US) provides significant advantages over conventional 2D imaging, including enhanced visualization, improved anatomical navigation, and accurate volumetric measurements. 3D US systems that use matrix probes or external tracking for volume construction can be cost prohibitive and complex, or

[3] Code and model weights are provided at `https://github.com/ImFusionGmbH/DualTrack`.

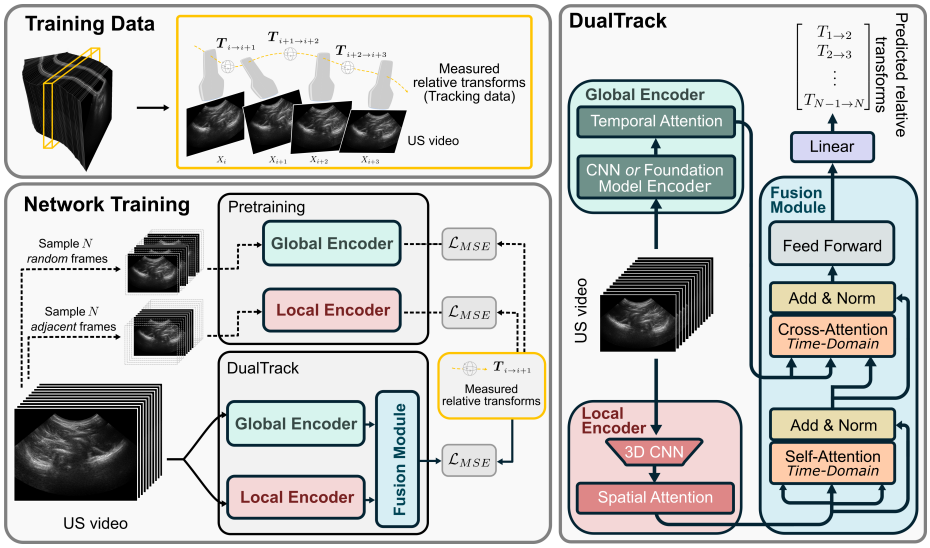


Fig. 1. DualTrack enables sensorless 3D ultrasound by estimating the probe trajectory for a sequence of ultrasound images (Top Left). Specialized local and global encoders are designed to extract low- and high-level features–their outputs are combined using self- and cross-attention through time (Right). The encoders are pretrained at their respective timescales before being finetuned with the fusion module (Bottom Left).

impose physical constraints that are impractical in routine clinical workflows. Sensorless 3D US, approaches that use deep learning to directly estimate the probe's 3D trajectory from 2D image sequences, are a promising alternative.

A sequence of 2D US images contains rich spatiotemporal information that can, in principle, be leveraged to infer the probe's 3D position. *Local features*, such as speckle patterns and short-term frame-to-frame motion cues, capture fine-grained displacements between adjacent frames. Meanwhile, *global features*, derived from identifying coarse anatomical patterns in the image sequence, help situate the probe with respect to that anatomy. For example, identifying structures such as bones, vessels, muscles, and organs in a 2D image sequence, and deriving their 3D properties such as continuity, shape, and landmarks can help inform a realistic 3D reconstruction that is faithful to the original structure. Therefore, an ideal trajectory estimation model should combine local *and* global features to ensure both short-term accuracy and global consistency. In practice, robustly modeling these complementary aspects remains a significant challenge.

Early advances in sensorless 3D US using convolutional neural networks (CNNs) were driven primarily by improvements in *local* feature extraction. These approaches successfully capture low-level image details such as speckle patterns and motion cues [4,3,10,17], but are inherently blind to global features. Recognizing this shortcoming, hybrid architectures have been proposed that append a sequence model (e.g. long short-term memory (LSTM) or transformer) to a

CNN, enabling global feature extraction by capturing long-term temporal dependencies [13,14,16]. Loss functions aimed at enforcing global consistency have also been used [10,14]. The limitation of these approaches is their strong coupling: the same basic encoder is forced to model local and global features, even though the properties of an ideal local feature extractor (locality, texture bias) do not match those of a global feature extractor (shape, long-range dependency).

In this work, we propose a novel approach for sensorless 3D US that maximizes the learning of complementary local and global information from 2D US sequences. Our approach is inspired by the *dual encoder* paradigm, i.e. a network architecture comprised of two separate branches, designed to model different features of an image–e.g., spatial vs. frequency features [1] or texture vs. shape features [2,19,5]. This separation allows for complementary and specialized encoders for each type of feature. We extend the dual encoder paradigm to 3D US by designing specialized spatiotemporal encoders that capture local and global properties. The decoupled nature of our architecture allows us to explore specialized design choices in each encoder, such as the use of CNNs for local feature modeling and pretraining or transfer learning (e.g. self-supervised learning, foundation models) to improve global feature modeling. After an initial independent training phase, these encoders are combined using a fusion module that integrates the outputs of each encoder to predict trajectory.

Our approach, DualTrack, is the first (to our knowledge) dual encoder network for sensorless 3D US. Our main contributions are:

1. We propose a dual encoder architecture that separately models local features (e.g. speckle-based cues) and global features (e.g. high-level anatomical landmarks), before combining them using a fusion module and prediction head to generate accurate and globally consistent trajectory estimates.
2. We design specialized architectures and pretraining objectives for each encoder to maximize its performance in its designated role. In particular, we investigate the potential of medical foundation models in sensorless 3D US by incorporating them as part of the global encoder.
3. Using a large public benchmark dataset, we empirically validate our method. DualTrack significantly outperforms previous approaches, achieving an average global reconstruction error of less than 5 mm.

2 Methods

An overview of our method is shown in Figure 1. We train a network to estimate probe trajectory from a sequence of 2D US images. Our network consists of a local and a global encoder, designed to extract low- and high-level features, respectively. The networks are pre-trained at their respective time scales. Finally, we train a fusion module to combine embeddings from the local and global encoders and produce accurate, globally consistent trajectory estimates.

Problem setup: Given an ordered sequence of 2D US frames $(X_i)_{i=1}^{n}$, the objective is to reconstruct the trajectory of the US probe by estimating its pose at each timepoint i. The pose at timepoint i can be represented by a homogeneous

4×4 matrix T_i or the corresponding six degree of freedom (three translations and three Euler angles) parameterization $\mathbf{p}_i \in \mathbb{R}^6$. Following prior approaches [17,10], we set $T_0 = \mathrm{Id}$ without loss of generality and focus on estimating the relative transformations $T_{j\leftarrow i} = T_i^{-1}T_j$, represented by 6-dimensional parameter vectors $\mathbf{p}_{j\leftarrow i}$. This reduces to a deep regression problem which we solve using a parameterized network $f(\cdot\,;\Theta)$ such that $f\big((X_i)_{i=1}^n;\Theta\big) \approx (\mathbf{p}_{j\leftarrow i})_{i=1}^{n-1}$ by optimizing the mean-squared error loss:

$$\min_{\Theta} \mathbb{E}_{(X_i)\sim\mathrm{Dataset}}\left[\mathcal{L}_{\mathrm{MSE}}\left(f\big((X_i)_{i=1}^n;\Theta\big), (\mathbf{p}_{j\leftarrow i})_{i=1}^{n-1}\right)\right] \tag{1}$$

At inference time, we let (X_i) be the full US scan and reconstruct the tracking sequence (T_i) as $T_i = T_0\prod_{k=0}^{i-1} T_{k+1\leftarrow k}$, with $T_{k+1\leftarrow k}$ converted from $\mathbf{p}_{k+1\leftarrow k}$ estimated by the network.

Local Features Modeling: Our local encoder is designed to capture local spatiotemporal features (e.g., speckle) from the ultrasound sequence. Following the literature, (e.g. [17,4,16,10]), it consists of a local spatiotemporal convolutional network followed by a spatial attention-pooling mechanism. Let N, H, and W denote the number of timesteps, height and width of an ultrasound subsequence. Let C and C' denote hidden feature dimensions. First, a 3D CNN $g_{\mathrm{loc}} : \mathbb{R}^{N\times H\times W} \to \mathbb{R}^{N\times \frac{H}{16}\times\frac{W}{16}\times C}$ extracts a sequence of feature maps from the sequence. Next, the spatial pooling module $h_{\mathrm{loc}} : \mathbb{R}^{N\times \frac{H}{16}\times\frac{W}{16}\times C} \to \mathbb{R}^{N\times C'}$ extracts an embedding vector for each feature map in the sequence. The CNN is a 3D (two spatial and one time dimension) ResNet18 modified to have small temporal windows, ensuring the network captures only local temporal dependencies. The pooling module is a lightweight transformer [18]. We train the network by attaching a linear layer and optimizing the tracking estimation objective (Eq. 1). We explicitly force locality by (i) using short contiguous subsequences as input; and (ii) keeping the image data at full resolution to preserve speckle features.

Global Features Modeling: Our global encoder is designed to extract coarse anatomical features from the ultrasound sequence. Its first stage is an image encoder $g_{\mathrm{glob}} : \mathbb{R}^{N\times H\times W} \to \mathbb{R}^{N\times C'}$ which extracts a single C'-dimensional feature vector separately for each image in the input sequence, focusing on high-level features. We chose a simple 2D ResNet18 which worked well on this dataset but also explored transfer using pretrained backbones including the USFM [6] encoder, the MedSAM [15] encoder, and an iBOT [20]-pretrained vision transformer. The second stage is a temporal self-attention module which aggregates information across time, $h_{\mathrm{glob}} : \mathbb{R}^{N\times C'} \to \mathbb{R}^{N\times C'}$. This module is based on the transformer [18] architecture, whose powerful ability to capture long-term and global dependencies in sequences makes it a suitable choice for the global encoder. To learn the global features relevant to tracking estimation, we attach a linear layer to the global encoder and train it using the tracking estimation objective (Eq. 1). We explicitly encourage learning global context features by (i) using randomly sampled non-contiguous subsequences as inputs to the model, preventing it from leveraging short-term speckle motion and forcing it instead to rely on global positional awareness; (ii) feeding it with images downsampled

to 224×224 to erase local textural details and allowing the model to focus on higher level semantics.

Local-Global Fusion: The fusion module combines embeddings from the local and global encoders. Following common multimodal late fusion approaches (e.g., [12,7]), we implement this fusion with a transformer decoder-based architecture using self-attention and cross-attention mechanisms. Specifically, local embeddings serve as the input states, while global embeddings act as the key-value states in the cross-attention layers. This design allows the transformer to refine local representations by attending to relevant global context, ensuring that short-term motion cues are interpreted within the broader anatomical and spatial structure of the scan. The fusion module is trained using the tracking estimation objective of Eq. 1 with the full US sequence as input to the local encoder and an evenly spaced subsample[4] to the global encoder.

3 Experiments

To evaluate the DualTrack approach, we performed the following experiments:

1. *Comparison to prior art:* To determine the efficacy of our approach, we compare it to 2-Frame CNN [17], MoNet [14], DCL-net [3] and Hybrid Transformer [16]. These works were chosen on the basis of (i) breadth of coverage of different methodologies in the literature; (ii) suitability of the method for our problem setting (i.e, no external sensor requirement); and (iii) success of reproduction. The methods were reproduced with reference to their manuscripts and, where available, source code. Modifications to the networks were made to account for differences in the dataset and hyperparameter values were re-tuned to allow fair comparison with our approach.
2. *Ablation Studies:* To assess the contribution of each component, we conduct ablation studies by removing key components of our architecture. We test the performance of "local only", where we simply used our local encoder as the predictor, "coupled local-global modeling" where we add a temporal attention transformer on top of our local encoder (no decoupling), in comparison to our dual encoder decoupled approach. Additionally, we evaluate multiple image encoder backbones for the global context encoder.
3. *Qualitative Analysis:* To further assess DualTrack's performance, we use it to reconstruct and visualize trajectories and 3D US based on the model's outputs (Figure 2A). We specifically investigate the types of errors which occur in local predictions but are corrected by the global model.

Dataset: We use the public TUS-REC challenge dataset for this study [8,9]. It consists of 1,248 freehand 2D US scans of forearms from 53 volunteers, with multiple scan shapes, scan directions and probe orientations. Scan shapes included linear scans as well as scans with wavy, non-linear trajectories. The average

[4] Subsampling improves inference speed and memory cost when using a large global encoder, at no observed performance cost. We typically subsample every 8 frames.

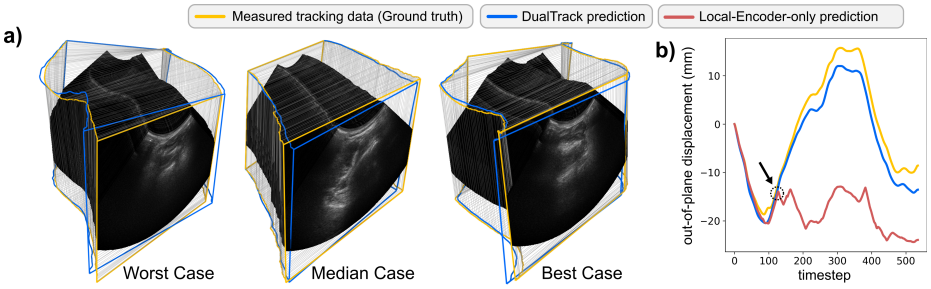


Fig. 2. (a) Trajectories and 3D US images generated with DualTrack predictions vs. ground truth, for three sweeps in the test set with worst/median/best GPE, respectively. (b) Comparison of the out-of-plane displacement prediction for the "local only" module and the full DualTrack model. The local model fails to disambiguate the out-of-plane direction and loses track after the first turn (see indicator); DualTrack corrects this error using contextual cues from global features.

translational distance from start to end of a scan is 167mm; the median scan length is 546 time steps (27 seconds). The total dataset size is approximately 226 GB. The public dataset is divided into training (first 50 subjects, 1176 scans) and test (last 3 subjects) splits. We further subdivided the training data into train (40 subjects) and validation (10 subjects) splits.

Evaluation strategy: We evaluate our models by measuring reconstruction errors from predicted trajectories compared to ground truth: final drift rate (FDR) is the translational distance between the estimated and true position at the final timestep, normalized by the translational distance from start to end of a scan. Maximum drift is the largest absolute translational distance between the predicted and true position at any timestep during a scan. Following [11], we also use the *point displacement errors* obtained by averaging the difference between the spatial positions of the five points (center and four corners) of an image when applying the true vs. predicted transform. The local point error (LPE) measures the average point error between two adjacent frames in the scan, while the global point error (GPE) measures the average point error between the true position and the fully reconstructed trajectory estimate at each timestep.

Implementation Details: Hyperparameters of our method were hand-tuned to minimize validation error. *Training:* Models were implemented using PyTorch and trained using a Nvidia A40 GPU. Local encoder pretraining took approximately 7 days, divided into (i) 4000 epochs of training with the CNN encoder (learning rate (l.r.) $1e-4$, weight decay (w.d.) $1e-3$); and (ii) 500 epochs of training with the spatial attention and frozen CNN module (l.r. $1e-4$, w.d. 0). Each training epoch for the local encoder sampled one subsequence of length 16 from each scan. Global encoder pretraining took 24 hours (800 epochs), and fusion module training 48 hours for 500 epochs (both with l.r. $1e-4$, w.d. 0). Baselines training took an average of 7 days to converge. All training phases used the AdamW optimizer and cosine annealed learning rate schedule. *Archi-*

Method			GPE (mm)	LPE (μm)	FDR (%)	Max. Drift (mm)
Baseline Comparison						
DCL-net [4]			10.77	134.21	11.68	18.75
2-Frame CNN [17]			8.90	140.21	8.93	14.49
MoNet [14]			9.19	127.97	8.87	14.46
Hybrid Transformer [16]			6.01	122.49	6.22	10.00
DualTrack (ours)			**4.93**	**122.01**	**5.10**	**8.33**
Ablation Results						
Local	Global	Dual Encoder				
✗	✓	✗	11.26	757.64	12.48	19.73
✓	✗	✗	7.36	140.90	6.96	11.70
✓	✓	✗	5.92	135.32	6.17	9.77
✓	✓	✓	**4.93**	**122.01**	**5.10**	**8.33**

Table 1. Averaged reconstruction metrics on 72 public test scans. For all metrics, lower is better with best results in bold. DualTrack (our method) outperforms the literature across all metrics, while ablation studies highlight the importance of each component.

tecture: Transformer modules in the global encoder and fusion modules used a hidden size of 512 and intermediate size of 1024, 8 layers, and 8 attention heads. We found the inclusion of an additional small transformer (hidden size 64, intermediate size 32, 4 layers and attention heads) between the local encoder and fusion modules improved performance and training stability. *Inference:* On a single RTX Quatro 6000 GPU, DualTrack reconstructs a full sweep of 546 frames in 0.73 s, versus 0.52 s for Hybrid-Transformer and 0.46 s for MoNet, with a comparable memory footprint ($<$ 7 GB). Given the median acquisition duration of 27 seconds, the extra 0.21 s inference time of our method is negligible. Most of the memory and time cost of inference is due to the local encoder CNN module, which could be replaced with a lighter-weight CNN if time compute resources were constrained.

4 Results and Discussion

DualTrack outperforms prior methods across all metrics. Table 1 (top) reports averaged error metrics on the TUS-REC test set. Compared to Hybrid Transformer, its leading competitor, DualTrack achieves lower errors (-1.08 mm GPE, -0.48 μm LPE, -1.12% FDR, -1.67 mm Max. Drift). Figure 3a shows error distributions across scans, with DualTrack exhibiting consistently lower values for key global metrics (GPE, LPE, FDR). These improvements were statistically significant (Wilcoxon signed-rank, $p \leq 0.005$). Models relying solely on local features (Rows 1-2) were outperformed by those using global features (Rows 3-5). The additional performance gains of DualTrack over other global-aware models (Rows

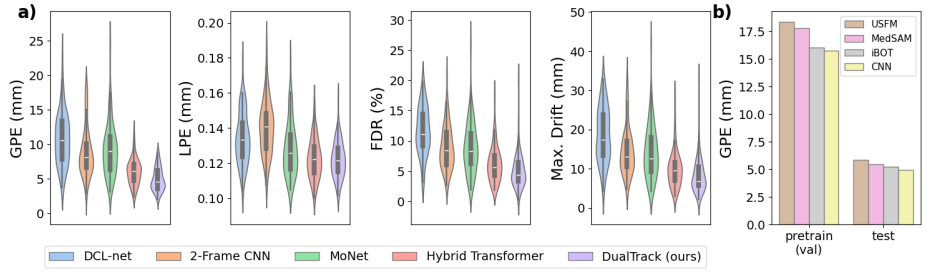


Fig. 3. (a) Distribution of tracking reconstruction errors across the 72 test scans for DualTrack and competing methods. Our method significantly ($p \leq 0.005$) outperforms its leading competitor on all metrics. (b) DualTrack can be successfully used with a variety of image backbones, giving the flexibility to leverage transfer learning strategies.

3-4) likely stem from its dual encoder architecture, which explicitly separates local and global modeling for more robust feature extraction and integration.

Results of our ablation study (Table 1, bottom) show the importance of both global feature extraction (improvements from Row 2 to 3), and the dual encoder setup (improvements from Row 3 to 4), further highlighting the strength of DualTrack's design. Figure 3b shows the impact of different global encoder backbones on GPE. In general, all backbones performed well, indicating the flexibility of the dual encoder. A simple 2D CNN gave the best performance on this dataset, followed by the iBOT self-supervised model. We hypothesize that the best choice of backbone is likely to be dataset-specific with (i) smaller datasets, (ii) datasets featuring more recognizable anatomical structures than the forearm, or (iii) datasets with organs that were represented in a foundation model's training corpus. Further studies are needed to test this hypothesis.

Visualization of trajectories and 3D US reconstructions generated by DualTrack show notable quality. We highlight visualizations on three test cases (Figure 2a) which had the worst, median, and best GPE respectively. The model's prediction aligns with ground truth. Even the worst case followed the general trajectory closely with a slight angular offset at the end of the scan, with a GPE of only 8 mm–the best case had a GPE less than 2 mm. A fourth case (Figure 2B) was chosen to highlight the positive impact of global modeling by comparing the out-of-plane displacement prediction for the "local only" model vs. the full DualTrack model. Unlike DualTrack, the local model fails to disambiguate the out-of-plane direction and loses track after the first turn (see indicator).

5 Conclusion

In this work, we addressed the need for robust local and global feature modeling in sensorless 3D US reconstruction. We introduced DualTrack, a dual encoder architecture that explicitly separates fine-grained features from high-level anatomical cues. Experimental results demonstrate that this decoupled design

significantly reduces reconstruction errors compared to existing state-of-the-art. Given its low error rate, DualTrack is likely capable of generating 3D ultrasound volumes that meet the accuracy requirements of clinical applications, especially those involving short and steady scans such as carotid plaque quantification: further investigation into these applications should be the subject of future work. Furthermore, future work should investigate the efficacy of DualTrack for complex freehand scans of anatomically variable regions such as the liver or heart.

Acknowledgments. This study was partially funded by the Accelerate program of Mitacs Inc., Canada.

References

1. Farshad, A., Yeganeh, Y., Gehlbach, P., Navab, N.: Y-net: A spatiospectral dual-encoder network for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 582–592. Springer (2022)
2. Fu, Z., Li, J., Hua, Z.: Deau-net: Attention networks based on dual encoder for medical image segmentation. Computers in Biology and Medicine **150**, 106197 (2022)
3. Guo, H., Chao, H., Xu, S., Wood, B.J., Wang, J., Yan, P.: Ultrasound volume reconstruction from freehand scans without tracking. IEEE Transactions on Biomedical Engineering **70**(3), 970–979 (2023). https://doi.org/10.1109/TBME.2022.3206596
4. Guo, H., Xu, S., Wood, B., Yan, P.: Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 463–472. Springer (2020)
5. Hong, Z., Chen, M., Hu, W., Yan, S., Qu, A., Chen, L., Chen, J.: Dual encoder network with transformer-cnn for multi-organ segmentation. Medical & biological engineering & computing **61**(3), 661–671 (2023)
6. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
7. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
8. Li, Q., Saeed, S.U., Barratt, D.C., Clarkson, M.J., Vercauteren, T., Hu, Y.: Trackerless 3d freehand ultrasound reconstruction challenge. In: 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024). Zenodo (2024). https://doi.org/10.5281/zenodo.10991501, https://doi.org/10.5281/zenodo.10991501
9. Li, Q., Saeed, S.U., Huang, Y., Luo, M., Yan, Z., Chen, J., Yang, X., Ni, D., Winter, N., Nguyen, P., et al.: Tus-rec2024: A challenge to reconstruct 3d freehand ultrasound without external tracker. arXiv preprint arXiv:2506.21765 (2025)
10. Li, Q., Shen, Z., Li, Q., Barratt, D.C., Dowrick, T., Clarkson, M.J., Vercauteren, T., Hu, Y.: Trackerless freehand ultrasound with sequence modelling and auxiliary

transformation over past and future frames. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
11. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 689–699. Springer (2024)
12. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
13. Luo, M., Yang, X., Huang, X., Huang, Y., Zou, Y., Hu, X., Ravikumar, N., Frangi, A.F., Ni, D.: 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 (2021). https://doi.org/10.1007/978-3-030-87231-1_20, https://doi.org/10.1007/978-3-030-87231-1_20
14. Luo, M., Yang, X., Wang, H., Du, L., Ni, D.: Deep motion network for freehand 3d ultrasound reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 290–299. Springer (2022)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. Ning, G., Liang, H., Zhou, L., Zhang, X., Liao, H.: Spatial position estimation method for 3d ultrasound reconstruction based on hybrid transfomers. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
17. Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., Wein, W.: 3d freehand ultrasound without external tracking using deep learning. Medical image analysis **48**, 187–202 (2018)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
19. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24. pp. 14–24. Springer (2021)
20. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)