Trackerless Ultrasound Pose Estimation via Correspondence-aware Tokenization and Flow-Matching Transformers

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

[1] Electrical Engineering Department, KAIST, 34141 Daejeon, South Korea
hmbae@kaist.ac.kr
[2] Barreleye Inc., 06211 Seoul, South Korea
[3] Department of Emergency Medicine, SNUBH, 13620 Seong-nam, South Korea

Abstract. Trackerless freehand 3D ultrasound reconstructs volumes from 2D B-mode sequences by estimating probe motion without external tracking. In B-mode ultrasound, accurate motion estimation hinges on recognizing subtle inter-frame appearance changes, yet repetitive speckle and weak visual anchors make correspondences ambiguous and drift accumulates when composing relative transforms. We propose a correspondence-amplifying conditioning pipeline with a flow-matching transformer for pose-delta sequence generation. For each adjacent frame pair, we build a bidirectionally supported token correspondence distribution from ViT features, use it to align current-frame evidence onto the previous token grid, and derive a correspondence-induced lattice displacement cue. We further apply entropy-based confidence weighting to suppress ambiguous regions and aggregate high-dimensional correspondence-aware tokens into compact context tokens via a Q-Former. Conditioned on this context, a flow-matching DiT generates coherent pose-delta sequences through ODE sampling. Experiments on tracked freehand ultrasound sequences show improved step accuracy and substantially reduced long-horizon drift.

Keywords: Trackerless ultrasound · Freehand 3D ultrasound · Flow matching · Diffusion Transformer

1 Introduction

Three-dimensional ultrasound (3D-US) is increasingly used for image-guided diagnosis and intervention because it provides a volumetric context that is difficult to infer from isolated 2D B-mode frames. In routine practice, however, reliable 3D-US acquisition remains challenging, since dedicated 3D probes and external tracking systems are cost-prohibitive and constrain standard scanning workflows. These limitations have motivated learning-based trackerless freehand 3D-US, where probe motion is inferred directly from the ultrasound image stream.

A key challenge is that B-mode ultrasound image provides ambiguous motion signals at the frame-to-frame level. Due to speckle patterns, shadowing, and the

limited stable visual anchors, it is difficult to capture reliable inter-frame feature changes. Consequently, inferring accurate inter-frame pose deltas is challenging. As a result, the estimation is prone to small per-step pose errors, which can accumulate into substantial drift when composing relative transforms for 3D reconstruction.

Prior trackerless freehand 3D-US methods have demonstrated deep contextual learning for inter-frame transformation prediction, often augmented with attention mechanisms, correlation-style objectives [3,4,14,16], and sequence modeling [9,10] to better exploit temporal structure. Nevertheless, the correspondence signal in ultrasound is often implicit. Previous pipelines rely on a network to internally discover subtle patch-level alignments from nearly repetitive appearance, which can lead to spurious associations and systematic biases that compound over time.

In this work, we address this challenge by explicitly constructing and amplifying inter-frame correspondences before pose estimation. We introduce a correspondence-aware tokenization that forms a mutual correspondence distribution between patch tokens extracted through ViT [1] and aligns current-frame evidence onto the previous token grid, revealing subtle matched appearance changes together with a correspondence-induced displacement cue. We further apply entropy-based confidence weighting to suppress ambiguous speckle regions, and distill patch-level correspondence evidence into compact conditioning tokens via a Q-Former [7]. Finally, a context-conditioned Transformer [15] trained with flow matching [12] generates pose-delta sequences through ODE sampling, improving stability and reducing long-horizon drift.

2 Method

2.1 Pose-Delta Labels and Representation

Given N ultrasound frames $\{I_0, \ldots, I_{N-1}\}$, we supervise the model with relative probe motion between consecutive calibrated ultrasound image planes. Let $T_k \in SE(3)$ denote the calibrated image-plane pose in the tracker coordinate system at frame k. We define

$$\Delta T_k = T_{k-1}^{-1} T_k, \quad k = 1, \ldots, N-1, \qquad \Delta T_0 = \mathbf{I}_4, \tag{1}$$

Thus, each pose delta is a relative rigid transform, and the target sequence has length N with a pivot at $k = 0$. Each ΔT_k is represented as a 9D vector

$$y_k = [r_k^{(6)}; t_k] \in \mathbb{R}^9, \tag{2}$$

where $r_k^{(6)}$ is the 6D continuous rotation representation proposed in [17] and $t_k \in \mathbb{R}^3$.

2.2 Correspondence-Amplifying Conditioning

Fig. 1 introduces the overview of the proposed framework. The key idea is to make motion evidence explicit before pose generation by constructing correspondence-aware tokens from consecutive frames.

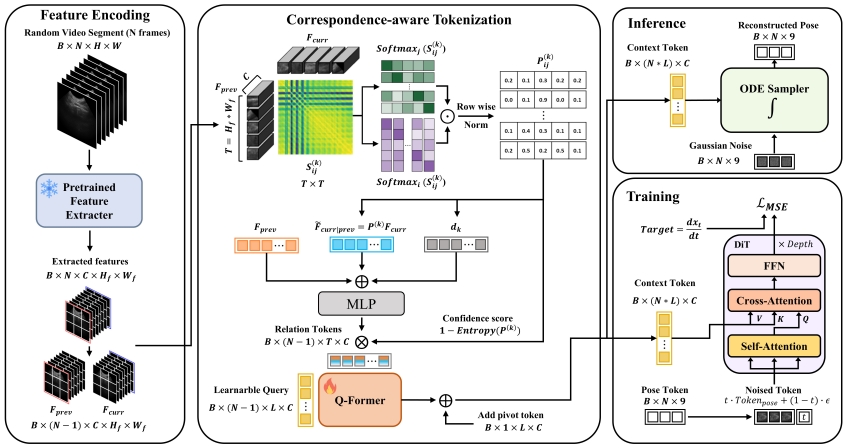

Fig. 1. Overview of the proposed framework. A frozen ultrasound-pretrained ViT encodes B-mode frames into token grids. CAT builds confidence-weighted correspondence-aware tokens from adjacent frames. A Q-Former compresses these tokens, and a flow-matching DiT generates pose deltas through ODE sampling.

To effectively capture ultrasound-specific appearance cues, we leverage a large-scale ultrasound-pretrained feature encoder. Specifically, we used a frozen ViT-B [1] backbone pretrained on B-mode images via masked autoencoding (MAE) [5]. Given an input frame I_k, the encoder extracts a patch-token grid $F_k = [f_{k,1}; \ldots; f_{k,T}] \in \mathbb{R}^{T \times C}$, where H_f and W_f denote the token-grid height and width, and $T = H_f W_f$. We form $(N-1)$ adjacent frame pairs $(k-1, k)$ for $k = 1, \ldots, N-1$. For the k-th pair, we denote $F^{(k)}_{\text{prev}} = F_{k-1}$ and $F^{(k)}_{\text{curr}} = F_k$ for notational convenience.

Next, we construct inter-frame token correspondences by computing a cosine similarity between ℓ_2-normalized tokens:

$$S^{(k)}_{ij} = \langle \hat{f}_{k-1,i},\ \hat{f}_{k,j} \rangle, \tag{3}$$

where $S^{(k)}_{ij}$ is a similarity score between token i in frame $(k-1)$ and token j in frame k.

In ultrasound, repetitive speckle can make many locations appear similarly plausible, producing asymmetric matches where a token strongly prefers a candidate in one direction, but the reverse association is weak. To suppress such asymmetric correspondences, we compute matching distributions in both directions:

$$A^{(k)}_{ij} = \text{softmax}_j(S^{(k)}_{ij}), \qquad B^{(k)}_{ij} = \text{softmax}_i(S^{(k)}_{ij}), \tag{4}$$

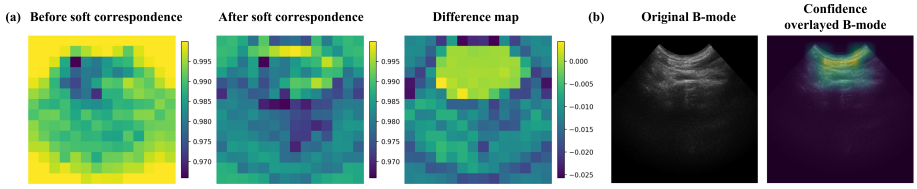

Fig. 2. Visual intuition for correspondence-aware tokenization. (a) Soft matching aligns current evidence to the previous token grid, revealing localized changes that are obscured without explicit correspondence alignment. (b) Entropy-based confidence weighting emphasizes matchable anatomy and suppresses ambiguous speckle/background regions.

and keep only mutually supported matches via the elementwise product

$$M_{ij}^{(k)} = A_{ij}^{(k)} B_{ij}^{(k)}. \tag{5}$$

We finally normalize each row of $M^{(k)}$ to obtain a row-stochastic correspondence distribution:

$$P_{ij}^{(k)} = \frac{M_{ij}^{(k)}}{\sum_{j'} M_{ij'}^{(k)} + \epsilon}, \qquad \sum_{j} P_{ij}^{(k)} = 1. \tag{6}$$

Each row $P_{i\cdot}$ can be interpreted as a probability distribution over candidate locations in the current frame for token i in the previous frame: it becomes peaked when a clear match exists, and remains spread when several candidates are similarly plausible.

Fig. 2(a) illustrates why explicit correspondence is helpful. In B-mode image sequences, many regions exhibit similar speckle statistics and lack distinctive visual anchors, so multiple candidate locations can obtain comparable similarity scores. As a result, the similarity landscape can be low-contrast and motion cues remain implicit. By constructing $P^{(k)}$, we explicitly re-weight current-frame evidence for each previous-frame token, making subtle motion-related changes easier to expose after alignment:

$$\tilde{f}_{k|k-1,i} = \sum_{j=1}^{T} P_{ij}^{(k)} f_{k,j} \in \mathbb{R}^C. \tag{7}$$

This alignment is crucial because $f_{k,j}$ and $\tilde{f}_{k|k-1,i}$ now refer to the corresponding locations in expectation, so direct comparisons become meaningful descriptors of motion-induced change rather than artifacts of fixed indexing.

We also encode $P^{(k)}$ as a compact geometric cue on the token lattice. Let $p_i \in [-1, 1]^2$ denote the fixed coordinate of the token i on the $H_f \times W_f$ grid. We compute the expected matched coordinate and its displacement:

$$\tilde{p}_i^{(k)} = \sum_{j=1}^{T} P_{ij}^{(k)} p_j, \qquad d_i^{(k)} = \tilde{p}_i^{(k)} - p_i \in \mathbb{R}^2. \tag{8}$$

where $d_i^{(k)}$ represents the expected shift induced by correspondence on the token grid: it is stable when $P_{i\cdot}^{(k)}$ is confident and naturally degenerates toward an uninformative average when $P_{i\cdot}^{(k)}$ is less concentrated.

We construct a compact relation descriptor from the aligned feature pair and an explicit correspondence shift. Since the alignment in Eq. (7) already brings current evidence to the previous token grid, motion signals can be read directly from the pair $(f_{k-1,i},\, \tilde{f}_{k|k-1,i})$. We also add a geometric cue via the correspondence-induced displacement in Eq. (8). Specifically, we embed the displacement with the MLP $\psi(\cdot)$ and concatenate

$$u_i^{(k)} = \Big[f_{k-1,i},\ \tilde{f}_{k|k-1,i},\ \psi(d_i^{(k)})\Big], \qquad r_i^{(k)} = \phi(u_i^{(k)}) \in \mathbb{R}^C. \tag{9}$$

followed by a projection head $\phi(\cdot)$ to obtain the relation token $r_i^{(k)} = \phi(u_i^{(k)}) \in \mathbb{R}^C$. The aligned feature pair encodes the degree of appearance consistency at the matched location, while $\psi(d_i)$ encodes how the correspondence shifts on the grid, helping to disambiguate repetitive speckle.

In homogeneous speckle or low-texture regions, $P_{i\cdot}$ tends to spread out and the soft alignment in Eq. (7) becomes less informative. We address this by encoding uncertainty with the entropy of each assignment row:

$$H_i^{(k)} = -\sum_{j=1}^{T} P_{ij}^{(k)} \log\big(P_{ij}^{(k)} + \epsilon\big), \qquad c_i^{(k)} = 1 - \frac{H_i^{(k)}}{\log T}, \tag{10}$$

and weight correspondence-aware tokens as $\bar{r}_i^{(k)} = c_i^{(k)}\, r_i^{(k)}$. Here, $c_i^{(k)}$ is close to 1 when $P_{i\cdot}^{(k)}$ is confident and approaches 0 when the match is ambiguous. As visualized in Fig. 2(b), high-confidence regions concentrate on stable anatomical structures, while background and repetitive speckle are downweighted. This prevents unreliable correspondences from dominating the conditioning signal, improving robustness of per-step motion estimation, and reducing long-horizon drift when composing pose deltas.

Q-Former Context Compression Correspondence-aware tokens are high-dimensional ($T = H_f W_f$), so directly conditioning the pose generator on all tokens is costly. We adopt a Q-Former [7] as a learnable query bottleneck, where a small set of queries cross-attends to the full 2D token grid and adaptively aggregates salient spatial evidence into a fixed-length context. We compress confidence-weighted tokens $\bar{R}_k \in \mathbb{R}^{T\times C}$ for each pair $(k-1, k)$ into L learned queries:

$$Z_k = \mathrm{QFormer}(Q, \bar{R}_k) \in \mathbb{R}^{L\times C}, \qquad k = 1, \ldots, N-1, \tag{11}$$

where $Q \in \mathbb{R}^{L\times C}$ are learnable query tokens. We set the pivot context to zeros ($Z_0 = \mathbf{0}$), yielding $Z \in \mathbb{R}^{N\times L\times C}$, which is flattened into $B \times (NL) \times C$ context tokens for cross-attention.

2.3 Flow-Matching DiT for Pose-Delta Generation

To obtain accurate and stable pose-delta sequence prediction, we formulate pose generation as conditional flow matching and learn a velocity field that maps Gaussian noise to the target pose-delta sequence. Given ground truth $y \in \mathbb{R}^{N\times 9}$ and noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, we sample $t \in [0, 1]$ and define

$$x_t = (1-t)\epsilon + ty, \qquad v^\star = \frac{dx_t}{dt} = y - \epsilon. \tag{12}$$

A context-conditioned DiT [15] predicts $v_\theta(x_t, t\,;\, Z) \in \mathbb{R}^{N\times 9}$ using self-attention to pose tokens and cross-attention to context tokens. We train with a flow-matching [12] objective:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\big[\|v_\theta(x_t, t\,;\, Z) - v^\star\|_2^2\big]. \tag{13}$$

2.4 ODE Sampling and Trajectory Composition

At inference, we sample $z_0 \sim \mathcal{N}(0, \sigma^2 I)$ and solve the ordinary differential equation (ODE):

$$\frac{dz_t}{dt} = v_\theta(z_t, t\,;\, Z), \quad t \in [0, 1], \tag{14}$$

with an Euler solver for 50 steps, yielding predicted deltas $\hat{y} = z_1$. Anchored poses for reconstruction are obtained by composing $\widehat{\Delta T}_k$:

$$\widehat{T}_0 = \mathbf{I}_4, \qquad \widehat{T}_k = \widehat{T}_{k-1}\widehat{\Delta T}_k. \tag{15}$$

3 Experiments

3.1 Dataset and Splits

We evaluated on the public TUS-REC challenge dataset [8], which contains 1,272 tracked freehand 2D B-mode forearm ultrasound scans from 53 volunteers across diverse probe orientations, scan directions, and L-, C-, and S-shaped trajectories. We follow the public subject-disjoint protocol of the challenge, using 1,080 scans from 45 subjects for training, 120 scans from 5 subjects for model selection, and 72 scans from 3 subjects for the final validation in Table 1. These 72 scans are the official public validation split and are never used for training.

3.2 Implementation Details

All frames are resized to 224×224 and normalized to $[0, 1]$. During training, we sample subsequences of length $N = 32$. We used a frozen ViT-B [1] encoder pretrained with masked autoencoding (MAE) to extract dense features ($C = 768$, $H_f \times W_f = 14 \times 14$, $T = 196$ tokens per frame). For each adjacent pair, we compute a mutual, row-stochastic correspondence matrix P to align current-frame features with the previous frame features and estimate a correspondence-induced

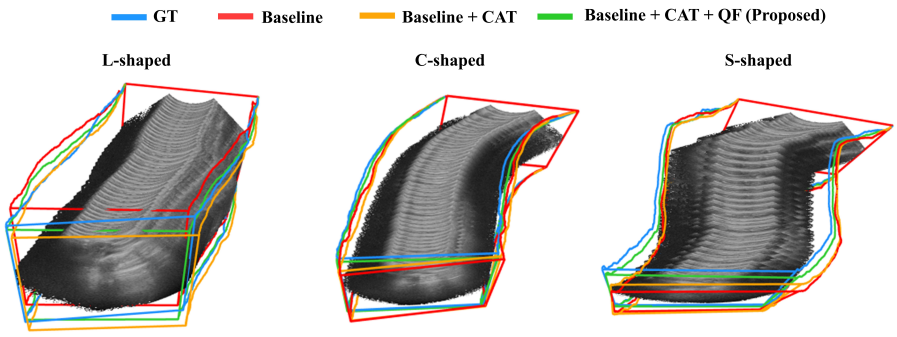

Fig. 3. Qualitative trajectory overlays for representative scan shapes. The proposed correspondence-aware tokenization with Q-Former conditioning more closely matches the ground-truth trajectory and reduces long-horizon drift compared to feature concatenation and linear-only conditioning.

displacement on the token lattice. Correspondence-aware tokens are confidence-weighted using row-wise entropy. We compress dense tokens into $L = 16$ query tokens per timestep using a Q-Former [7], and the resulting context tokens condition a flow-matching DiT [12, 15] (hidden size 768, 8 blocks, 12 heads). We train with Adam (lr 10^{-4}, batch size 32) and exponential moving average (EMA). At inference, pose deltas are generated via Euler ODE sampling with 50 steps.

3.3 Qualitative Assessment

We qualitatively assess reconstruction fidelity by overlaying the composed probe trajectory with the ground-truth trajectory on representative scan patterns, as shown in Fig. 3.

Baseline conditions the model using simple feature concatenation from two adjacent frames, followed by a linear layer that compresses the concatenated features into a fixed-size conditioning vector. **Baseline + CAT** replaces this naive concatenation with our correspondence-aware tokenization (CAT). **Proposed** method further applies Q-Former (QF) compression, distilling dense correspondence-aware tokens into a compact set of query tokens per timestep.

Across all scan shapes in Fig. 3, the Baseline trajectory exhibits noticeable drift and shape distortion. Adding CAT improves alignment to the ground truth by making subtle inter-frame motion evidence more explicit, reducing accumulated deviation along the trajectory. The proposed method achieves the closest agreement with the ground truth, with visibly reduced long-horizon drift and a more stable reconstruction of the overall scan geometry.

Table 1. Quantitative comparison on the official public TUS-REC evaluation split.

Method	GPE (mm)↓	LPE (mm)↓	FDR (%)↓
DCL-Net [4]	11.77	0.134	17.68
PLPPI [2]	9.80	0.133	13.24
NR-Rec-FUS [11]	9.12	0.128	12.11
MoGLo [6]	8.09	0.124	10.45
Baseline	12.58	0.145	17.89
Baseline + CAT	5.84	0.123	9.91
Baseline + CAT + QF (Proposed)	**5.31**	**0.121**	**8.56**

3.4 Quantitative Assessment

We evaluate pose accuracy using the point-wise distance error introduced in [8] with four image-corner points, and report the mean Euclidean distance for both the local and global reconstructions.

Local point error (LPE) reflects single-step motion accuracy, while the use of composed transforms anchored at the first frame yields the **global point error (GPE)**, which captures long-horizon drift accumulation. In addition, we report the **final drift rate (FDR)** introduced in [13], which measures endpoint drift normalized by the accumulated scan length. We compare against reproducible methods reported in [8] benchmark and recent trackerless networks that operate in an offline, feed-forward manner without external tracking hardware. Table 1 summarizes the results.

Table 1 shows that the Baseline suffers from severe drift, as it conditions the pose generator using unaligned features from two adjacent frames and compresses them with a single linear projection, leaving the model to infer inter-frame correspondences implicitly. Replacing this step with CAT yields a large improvement in both local and global accuracy which indicates that explicitly forming correspondence-aware relation tokens provides a substantially more reliable conditioning signal in the presence of repetitive speckle and weak anchors. Replacing the simple linear projection with Q-Former yields further consistent improvements, suggesting that query-based attention pooling better preserves task-relevant correspondence cues for conditioning than a single linear compression.

4 Conclusion

We proposed a correspondence-amplifying conditioning framework for trackerless freehand 3D ultrasound, targeting the core difficulty of B-mode imaging where inter-frame motion cues are subtle and correspondences are often ambiguous due to repetitive speckle and weak visual anchors. Our key idea is to explicitly construct a token-level correspondence distribution and use it to align current-frame evidence onto the previous token grid, exposing matched appearance changes

together with an expected lattice-displacement cue. We further incorporate entropy-based confidence weighting to suppress unreliable regions and adopt a Q-Former to distill patch-level correspondence evidence into compact context tokens for efficient conditioning. Conditioned on this context, a flow-matching DiT generates coherent pose-delta sequences via ODE sampling, improving robustness to accumulated drift. Experiments on the TUS-REC dataset demonstrate consistent gains in step accuracy and substantial reductions in long-horizon drift over feature-concatenation baselines and competitive trackerless methods, leading to precise trajectory reconstruction across diverse scan patterns.

Acknowledgments. This work was supported by Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00444862), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2024-00349441), and BK21 FOUR (Connected AI Education & Research Program for Industry and Society Innovation, KAIST EE, No. 4120200113769).

Disclosure of Interests. S-Y. Kim, Y-M. Kim, G. Jung and H. Lee are under contracts with Ministry of Korea National Defense. S-Y. Kim and H. Lee are under scholarship from the Korea Government Scholarship Program.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
2. Dou, Y., Mu, F., Li, Y., Varghese, T.: Sensorless end-to-end freehand 3-d ultrasound reconstruction with physics-guided deep learning. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **71**(11), 1514–1525 (2024)
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 (2022)
4. Guo, H., Xu, S., Wood, B., Yan, P.: Sensorless freehand 3d ultrasound reconstruction via deep contextual learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 463–472. Springer (2020)
5. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
6. Lee, S., Kim, S., Seo, M., Park, S., Imrus, S., Ashok, K., Lee, D., Park, C., Lee, S., Kim, J., et al.: Enhancing free-hand 3d photoacoustic and ultrasound reconstruction using deep learning. IEEE Transactions on Medical Imaging (2025)
7. 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., 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)

9. Li, Q., Shen, Z., Li, Q., Barratt, D.C., Dowrick, T., Clarkson, M.J., Vercauteren, T., Hu, Y.: Long-term dependency for 3d reconstruction of freehand ultrasound without external tracker. IEEE Transactions on Biomedical Engineering **71**(3), 1033–1042 (2023)
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. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 201–210. Springer (2021)
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. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
16. Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., Wein, W.: Deep learning-based 3d freehand ultrasound reconstruction with inertial measurement units (2018)
17. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019)