

# Long-term Dependency for 3D Reconstruction of Freehand Ultrasound Without External Tracker

Qi Li, Ziyi Shen, Qian Li, Dean C. Barratt, Thomas Dowrick, Matthew J. Clarkson, Tom Vercauteren, and Yipeng Hu

**Abstract—Objective:** Reconstructing freehand ultrasound in 3D without any external tracker has been a long-standing challenge in ultrasound-assisted procedures. We aim to define new ways of parameterising long-term dependencies, and evaluate the performance. **Methods:** First, long-term dependency is encoded by transformation positions within a frame sequence. This is achieved by combining a sequence model with a multi-transformation prediction. Second, two dependency factors are proposed, anatomical image content and scanning protocol, for contributing towards accurate reconstruction. Each factor is quantified experimentally by reducing respective training variances. **Results:** 1) The added long-term dependency up to 400 frames at 20 frames per second (fps) indeed improved reconstruction, with an up to 82.4% lowered accumulated error, compared with the baseline performance. The improvement was found to be dependent on sequence length, transformation interval and scanning protocol and, unexpectedly, not on the use of recurrent networks with long-short term modules; 2) Decreasing either anatomical or protocol variance in training led to poorer reconstruction accuracy. Interestingly, greater performance was gained from representative protocol patterns, than from representative anatomical features. **Conclusion:** The proposed algorithm uses hyperparameter tuning to effectively utilise long-term dependency. The proposed dependency factors are of practical significance in collecting diverse training data, regulating scanning protocols and developing efficient networks. **Significance:** The proposed new methodology with publicly available volunteer data and code<sup>1</sup> for parameterising the long-term dependency, experimentally shown to be valid sources of performance improvement, which could potentially lead to better model development and practical optimisation of the reconstruction application.

**Index Terms—**Freehand ultrasound reconstruction, long-term dependency, multi-task learning, sequence modeling

Qi Li, Ziyi Shen, Dean C. Barratt, Thomas Dowrick, Matthew J. Clarkson and Yipeng Hu are with the UCL Centre for Medical Image Computing, and the Wellcome/EPSCRC Centre for Interventional and Surgical Sciences, Department of Medical Physics and Biomedical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: qi.li.21@ucl.ac.uk; ziyi-shen@ucl.ac.uk; d.barratt@ucl.ac.uk; t.dowrick@ucl.ac.uk; m.clarkson@ucl.ac.uk; yipeng.hu@ucl.ac.uk).

Qian Li is with State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin, China, and also with the UCL Centre for Medical Image Computing, and the Wellcome/EPSCRC Centre for Interventional and Surgical Sciences, Department of Medical Physics and Biomedical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: qian-li@ucl.ac.uk).

Tom Vercauteren is with School of Biomedical Engineering & Imaging Sciences, King's College London, London WC2R 2LS, U.K. (e-mail: tom.vercauteren@kcl.ac.uk).

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

## I. INTRODUCTION

**D**ETERMINING the relative 3D spatial positions between ultrasound (US) images can recover 3D anatomy and pathology in these images. External spatial trackers such as those based on mechanical, optical and electromagnetic principles, enabled many clinical ultrasound applications. Removing the need for such external devices has attracted decades of research interest, in order to devise a more portable, accessible and cost-effective freehand ultrasound system, without being constrained by line-of-sight [1] or magnetic field interference [2], particularly preferable in surgical and interventional applications.

Speckle-induced correlation between near-overlapping images has been studied to align spatially close US frames [3]. Statistical image correlation was also investigated [4], under the same assumption. The image slice separation is limited in this kind of approach, usually chosen to be 0.2 mm. These methods have focused on inherent correlation between images, independent of clinical applications with specific protocols and their intended anatomical content, for estimating spatial transformation from images.

To improve upon these general approaches, recent data-driven deep learning-based methods are capable of learning “global” correlations for determining relative image locations. Indeed, recurrent neural networks (RNNs) [5], [6], [7], and transformers [8] have been proposed to model US frames as sequential data, with reported better spatial localisation. For example, Luo et al [7] tested ConvLSTM on sequences with 90-120 frames and Miura et al [9] used ConvLSTM with sequences of 180 frames. This suggests that, even though these methods lack a physical basis, there is still an advantage to be gained from image frames that are spatially and temporally distant, i.e. further than a few neighbouring frames which may contain shared content or signals. This is referred to as the long-term dependency in this work<sup>2</sup>. However, most of the aforementioned studies have incorporated other innovative contributions such as novel network training strategies [8], [10], and added prior knowledge [7]. It is therefore unclear whether and how much the reported improvement originated from this long-term dependency.

This work describes a new freehand US sequence encoding together with a multiple transformation prediction algorithm.

<sup>2</sup>In other fields such as natural language processing, long-term dependency may refer to those with much longer distances, therefore is considered an application-dependent term.

The correlations within the input US sequence, those between a large number of output transformations, and the output dependency on the input sequence can be readily modelled as hyperparameters of RNNs or, perhaps more interestingly, feed-forward convolutional neural networks (CNNs). We show that a large margin of improvement reducing up to 74.5% in final drift, due to including distant (up to 20s) past frames, was possible for specific applications.

To investigate factors that resulted in this improvement, two types of application-specific long-term dependencies are hypothesized, *anatomical dependency* and *protocol dependency*. That is, predicting spatial frame locations is considered dependent on *a)* anatomical/pathological content in acquired images and *b)* on pre-defined scanning paths, probe movements or orientation patterns during image acquisition by trained operators. This work utilises the proposed method to quantify the performance change due to altering these two factors independently.

We argue in this paper that a better understanding of the benefits of long-term dependency, by quantifying reconstruction accuracy as a function of the factorised dependency, is not only an interesting research topic but also practically important. Dependency (hyper-)parameters, defined in Section III-C, are useful for choosing effective models, whilst identifying performance-gaining dependency factors may help generalise these gains, for example by optimising protocols, training cohort of data and the trade-off between computation cost and memory requirement from dedicated hardware.

The sequence modelling method was summarised in a conference paper [11]. This work has a different focus on further developing the methodology specifically for modelling and assessing long-term dependency, with different experiments using more than 5 times longer sequences. The added contributions include: 1) A detailed description and motivation of the proposed input encoding and multiple transformation output method; 2) Presenting extensive experimental results for demonstrating the improvement from long-term dependency; 3) Based on analysis of the long-term dependency, in terms of the defined hyperparameters and proposed dependency factors, a number of interesting conclusions are summarised, some of which are addressed with quantitative evidence for the first time; and 4) The code and volunteer data are made available for public access to ensure study reproducibility and further research.

## II. RELATED WORK

3D US reconstruction, a promising technique for ultrasound examination and ultrasound-guided intervention, has advantages over its 2D counterpart in many clinical scenarios, such as multi-modal registration [12], musculoskeletal assessments [13], [14], volume visualisation and measurement [15]. A large number of approaches has been proposed for 3D US reconstruction, which in this paper are studied using three categories: 1) scanning with 2D-array US probe, which can directly acquire 3D US volume [16]; 2) mechanical scanning, which can efficiently reconstruct the 3D US volume by using motorized mechanical motor to move the US transducer along

predefined trajectories [17]; 3) freehand 3D US scanning, which can reconstruct 3D US using spatial-temporal information of probe obtained by tracker or trackerless methods.

Despite the perceived flexibility and accessibility associated with freehand 3D US scanning, a spatial tracker (often external) is required which adds on cost and other logistic challenges, such as maintaining line-of-sight for optical trackers and avoiding interference for electromagnetic trackers. It has therefore been a strong research interest in developing trackerless freehand 3D US reconstruction, which may historically be further classified into non-deep-learning and deep-learning based methods.

Many popular non-deep-learning based freehand 3D US systems are based on utilising speckle patterns in US images [3]. Although some consider speckle impacts the quality of 3D US reconstruction with studies trying to suppress speckle to enhance tissue contrast [18], the correlation of speckle could be analysed, using statistical or machine learning approaches, to indicate the likelihood of relative positioning of nearby US frames, therefore to achieve tracking without external trackers. Gao et al [19] proposed a wireless and sensorless 3D US imaging system that relied on adaptive speckle decorrelation curve to measure the motion of US probe along a single direction. This study has demonstrated the feasibility of image based US probe tracking method on phantom and real-tissue data, although more work remains to be done for allowing much less unconstrained scanning protocol so they can be clinically useful.

Deep-learning based approaches, featured with helpful representation ability, have been utilized for 3D US reconstruction [20]. For instance, Guo et al [21] proposed a deep contextual learning network (DCL-Net), a sequence modelling method with 3D convolutions, attention module, and a novel case-wise correlation loss, for 3D US reconstruction. Luo et al [22] exploited acceleration and orientation data measured by inertial measurement unit (IMU) to extract velocity information that could help estimate elevational displacements better. They also proposed an online self-supervised strategy for adaptive optimization of the model to reduce the drift. Based on this work, Luo et al afterwards proposed a multi-IMU-based network to reduce noise in IMU data [23], in which a modal-level self-supervised strategy for IMU information fusion and a sequence-level self-consistency strategy for estimation stability enhancement were presented for performance improvement, demonstrated by extensive ablation experiments. A self-supervised learning and adversarial learning based online learning strategy was presented in [24], along with a motion-weighted training strategy, for case-wise adaption to unseen dataset with diverse scanning velocities and poses. Instead of formulating a transformation into rotation and translation components separately, Hou et al [25] trained a pose estimation CNN on manifold  $SE(3)$ , with a left-invariant Riemannian metric. This proposed loss, computed on Riemannian geodesic space, could couple the translation and rotation components, taking into account the structure of Lie group  $SE(3)$ . Yeung et al proposed a pipeline for mapping 2D US images into 3D space with a pairwise comparison module and attention mechanism [26]. Inspired by [27], Yeung et al parameterised

the 3D reconstruction with implicit neural representation, jointly refining the initial pose estimation. A regression CNN was used in [28] with the continuous rotation representation [29], demonstrated on both phantom and real fetal data. Wein et al [30] proposed a pipeline for 3D thyroid assessment, consisting of tracking estimation, joint co-registration, and thyroid segmentation.

In summary, deep-learning-based trackerless freehand 3D US reconstruction seems a promising alternative to previous approaches without using deep neural networks. Existing methods estimated probe positions from US sequence that contained more than two frames, with the assumption that long-term dependency across the sequence can benefit estimation of current probe positions. However, to our best knowledge, no existing methods quantified such long-term dependency and analysed its contributing factors, the two aims of this study. The hypothesis of long-term dependency will be examined and defined in the following sections.

### III. METHOD

Assume a sequence of 2D US frames  $S = \{I_m\}, m = 1, 2, \dots, M$ , with a sequence length  $M$ , and denote the spatial transformation between  $i^{th}$  and  $j^{th}$  frames as  $T_{j \leftarrow i}, 1 \leq i < j \leq M$ . In this work,  $T_{j \leftarrow i}$  is represented by homogeneous matrices describing the relative translation and rotation, such that points  $p^{(i)}$  in  $i^{th}$  image coordinate system, in  $[x, y, z, 1]$  homogeneous coordinates, can be transformed to  $j^{th}$  image coordinate system,  $p^{(j)} = T_{j \leftarrow i} \cdot p^{(i)}$ , thus describing the relative positions between the two frames.

#### A. Input Ultrasound Sequence Modelling

Recurrent models, such as RNNs with long short-term memory (LSTM) modules [31] and transformers [32] can be used to model the input sequential US frames. In this work, we assume a single intended output transformation<sup>3</sup>  $T_{j^* \leftarrow i^*}$  between two predefined frames  $i^*$  and  $j^*$ , and a RNN model  $f_{rnn}$ , with network parameters  $\theta$ , is a function of individual frames  $I_m$  at time step  $m$  and the internal hidden state  $h^{(m-1)}$  from the previous time step.

$$\begin{aligned} T_{j^* \leftarrow i^*} &= f_{rnn}(I_m, h^{(m-1)}; \theta), \text{ for } m = M \\ h^{(m)} &= f_{rnn}(I_m, h^{(m-1)}; \theta), \forall m \leq M - 1 \end{aligned} \quad (1)$$

This many-to-one mapping model enables the use of past frames  $\{I_m\}_{m \in [1, i^* - 1]}$  and future frames  $\{I_m\}_{m \in [j^* + 1, M]}$ , the latter of which necessitates a temporal delay for a real-time system.

It is worth noting that the system or GPU memory required for training an unrolled  $f_{rnn}$  is a function of the entire sequence, rather than individual frames, using back-propagation through time (BPTT) [33]. Algorithms that are less dependent on sequence length, such as truncated BPTT or alternatives [34], have seldom been seen in medical imaging applications,

<sup>3</sup>This assumption of single fixed output is made to enable the hyperparameters described in Section III-C and for investigating image-to-transformation distance without predictions at multiple time points, further discussed in Section III-D.

perhaps due to the potentially excessive computation for training high dimensional image input.

The recurrent models, with the single output at the end of a sequence, when unrolled, are conceptually equivalent to feed-forward models with the same output and the entire sequence as the input. This motivates us to test a CNN  $f_{cnn}$  for this sequence modelling:

$$T_{j^* \leftarrow i^*} = f_{cnn}(S; \theta) \quad (2)$$

#### B. Output Multiple Transformation Prediction

Although the sequence modelling described above only predicts a single transformation at the end of a sequence, supervision, i.e. ground-truth target transformations, at the previous time steps are available and were shown to accelerate training [35], also known as ‘‘teacher forcing’’. In this section, a multi-transformation prediction is proposed to use these additional data.

In addition to the intended  $T_{j^* \leftarrow i^*}$ , both the CNNs and RNNs can be adapted to output other  $M(M - 1)/2 - 1^4$  transformations  $\{T_{j \leftarrow i}\}, i \neq i^* \text{ or } j \neq j^*$ . Based on points  $p_n^{(i)}$  sampled from  $i^{th}$  frame in image coordinates, the proposed overall multi-transformation loss becomes:

$$\mathcal{L}_{MTL} = \frac{1}{N \cdot M(M - 1)/2} \sum_{n=1}^{M(M-1)/2} \sum_{i=1}^N D_{mse}(p_n^{(j)}, \hat{p}_n^{(j)}) \quad (3)$$

where  $p_n^{(j)}$  and  $\hat{p}_n^{(j)}$  are the same points transformed from the  $i^{th}$  to  $j^{th}$  image coordinate systems,  $p_n^{(j)} = T_{j \leftarrow i}^{(gt)} \cdot p_n^{(i)}$  and  $\hat{p}_n^{(j)} = \hat{T}_{j \leftarrow i} \cdot p_n^{(i)}$ , using ground-truth  $T_{j \leftarrow i}^{(gt)}$  and prediction  $\hat{T}_{j \leftarrow i}$ , respectively. Four image corner points were used in this work, i.e.  $N = 4$ . Mean-square-error (MSE) was used as the distance function  $D_{mse}(\cdot)$ , between  $x, y$  and  $z$  coordinates of the two points.

Optimising different transformations to the same image coordinate systems, as in Eq. 3, encourages consistency and minimises accumulated error, as previously proposed [11]. Using a third  $k^{th}$  frame for example, when  $D_{mse}(T_{j \leftarrow i}^{(gt)} \cdot p_n^{(i)}, \hat{T}_{j \leftarrow i} \cdot p_n^{(i)})$  is minimised simultaneously with  $D_{mse}(T_{k \leftarrow i}^{(gt)} \cdot p_n^{(i)}, \hat{T}_{k \leftarrow i} \cdot p_n^{(i)})$  and  $D_{mse}(T_{j \leftarrow k}^{(gt)} \cdot p_n^{(k)}, \hat{T}_{j \leftarrow k} \cdot p_n^{(k)})$ , the difference between  $\hat{T}_{j \leftarrow k} \cdot \hat{T}_{k \leftarrow i}$  and  $T_{j \leftarrow i}^{(gt)}$  is thus implicitly minimised - a form of accumulated error, so is the difference between  $\hat{T}_{j \leftarrow k} \cdot \hat{T}_{k \leftarrow i}$  and  $\hat{T}_{j \leftarrow i}$ , with equal ground-truth  $T_{j \leftarrow k}^{(gt)} \cdot T_{k \leftarrow i}^{(gt)} = T_{j \leftarrow i}^{(gt)}$  - optimising a measure of consistency between sequential predictions. As these multiple output transformations share information and impose regulating constraints on each other, each of them can be regarded as one task in a multi-task learning framework. The multi tasks consist of one main task  $T_{j^* \leftarrow i^*}$  and other  $M(M - 1)/2 - 1$  auxiliary tasks. This multi-task learning framework takes advantage of the shared information among different tasks, which may result in improved performance over single-task learning framework. In addition, this multi-task learning framework can predict various transformations with various intervals, past and future

<sup>4</sup>For an US sequence with length  $M$ , the possible number of output transformations can be  $M$ -combinations of a 2-set  $C_M^2 = M(M - 1)/2$ .

frames using a common set of layers, making it possible to compare the performance of different hyperparameters in a single training run. In practice, when  $M(M-1)/2$  is large,  $\tau + 1$  transformation tasks, including various transformation intervals and number of past and future frames, are sampled due to memory limit, where  $\tau \leq M(M-1)/2 - 1$ .

### C. Parametric Dependency as Hyperparameters

As formulated in Sections III-A and III-B, the dependency of transformation prediction can be quantified and illustrated in Fig. 1. Past- and future- dependencies are represented by the number of the respective frames,  $i^* - 1$  and  $M - j^*$ . We propose to use  $i^*$ ,  $j^*$  and  $M$  as hyperparameters of the models in Eqs. 1 and 2, which are general enough to represent many scenarios to test the dependency of the predicted transformation on frames outside of the transformation. For example, larger  $i^*$  and  $M - j^*$  increase the lengths of past and future dependencies, respectively, whilst a high  $M$  value can test both. It is noteworthy that tuning these hyperparameters may therefore aim for an optimum  $T_{j^* \leftarrow i^*}$  on the validation set, rather than the overall loss in Eq. 3. An extension to this work may investigate cases that predict a future or past transformation using acquired frames before or after the input sequence.

Sequence length  $M$  determines the largest possible number of past and future frames that can be used for predicting transformations. For example, more past and future frames can be used with a larger  $M$  and a smaller transformation interval.  $M$  can be selected based on specific applications. For this study, the relationship between sampled and tested sequence lengths and reconstruction performance is reported in Section V-A.

### D. Sequence Sampling and 3D Scan Reconstruction

From available US scans with variable lengths, sequences  $S = \{I_m\}$  with the predefined  $M$  are randomly sampled, for training models in Eqs. 1 and 2. The ground-truth is used to transform points in the  $i^{\text{th}}$  image coordinates to the  $j^{\text{th}}$  image coordinates,  $T_{j \leftarrow i}^{(gt)} = T_{(calib)}^{-1} \cdot (T_{world \leftarrow j}^{(gt)})^{-1} \cdot T_{world \leftarrow i}^{(gt)} \cdot T_{(calib)}$ , where  $T_{world \leftarrow j}^{(gt)}$  and  $T_{world \leftarrow i}^{(gt)}$  are  $j^{\text{th}}$ -tool-to-world and  $i^{\text{th}}$ -tool-to-world transformations, at the time steps  $j$  and  $i$ , obtained from the optical tracker. Thus, the transformation is independent of the world coordinate system.  $T_{(calib)}$  is a fixed transformation from image to tool coordinate systems, obtained through spatial calibration. In practice, the left-multiplying inverse calibration matrix is not used, to which the loss is invariant to, due to the distance preservation property of orthogonal matrix. Thus, the loss is computed in the  $j^{\text{th}}$  tracking tool coordinate system with a unit of millimeter (mm):  $T_{j \leftarrow i}^{(gt)} = (T_{world \leftarrow j}^{(gt)})^{-1} \cdot T_{world \leftarrow i}^{(gt)} \cdot T_{(calib)}$ .

During the test, a scan can be reconstructed by predicting the optimum  $T_{j^* \leftarrow i^*}$  from consecutive sequences, such that the  $(j^*)^{\text{th}}$  frame from the previous sequence is the  $(i^*)^{\text{th}}$  frame in the subsequent sequence. Depending on the application where localising every possible adjacent frames is required, varying starting reference frames can be used and the relative locations

between them may be determined by the auxiliary tasks, an independent initialisation method or potentially fixed with a predefined protocol. Furthermore, advanced hyperparameter selection methods, adaptive at different time points, and model ensembles to combine different predictions (therefore multiple main tasks) at the same time point, may further optimise the 3D reconstruction. These remain research interests for future study and are not considered in this work.

### E. Evaluation Metrics

As the training loss is computed by using prediction error on each frame, one direct model generalisation metric *frame prediction accuracy* ( $\epsilon_{frame}$ ) is computed, as the difference between prediction  $\hat{p}_n^{(j)}$  and ground-truth points  $p_n^{(j)}$  on the  $j^{\text{th}}$  frame, both transformed from the  $i^{\text{th}}$  frame,  $\epsilon_{frame} = \frac{1}{N} \sum_{n=1}^N D_{dist}(p_n^{(j)}, \hat{p}_n^{(j)})$ , where  $D_{dist}(\cdot)$  denotes the Euclidean distance between two points and  $N = 4$  on four corner points. This metric is useful for monitoring training and model development, but may not be indicative of the performance in predicting  $T_{j^* \leftarrow i^*}$  or scan reconstruction.

For each test scan<sup>5</sup>, three metrics are reported to assess the reconstructed frames: 1) *Accumulated tracking error* ( $\epsilon_{acc}$ ) is the average Euclidean distance of all reconstructed image pixels between prediction and ground-truth,  $\epsilon_{acc} = \frac{1}{\mathcal{J} \cdot N} \sum_{j=1}^{\mathcal{J}} \sum_{n=1}^N D_{dist}(p_n^{(j)}, \hat{p}_n^{(j)})$ , where  $N$  is the number of pixels in an image and  $\mathcal{J}$  is the number of reconstructed frames using  $\hat{T}_{j \leftarrow i}$ ; 2) *Volume reconstruction overlap* ( $\epsilon_{dice}$ ) is a Dice score, computed as the overlap between the ground-truth- and prediction- reconstructed scan volumes; and 3) *Final drift* ( $\epsilon_{drift}$ ) measures as the average Euclidean distance, over the four corner points on the last frame of the scan, between ground-truth and prediction.

### F. Factorised Dependency and Reduced Variance Analysis

Anatomy dependency refers to the long-term dependency contributed by anatomical self-correlation with respect to anatomical variance. The common spatial movement pattern inherent in scanning protocols can also contribute to the long-term dependency, defined as protocol dependency. The abundance of anatomy and scanning protocol within the training dataset determines how much long-term dependency exists and can be learned. As US scan acquired from the same subject have the same anatomical content, anatomy dependency can be investigated by altering the included number of subjects. In the volunteer study, the dataset mentioned in Section IV are acquired by three types of scanning protocols - straight line shape, ‘‘C’’ shape, and ‘‘S’’ shape, as variance of scanning protocol is determined by number of scanning protocols involved. For quantifying anatomical dependency, the original training set can be re-sampled, at the subject level, such that a percentage of  $v_a = 25\%$ ,  $50\%$ ,  $75\%$  of the

<sup>5</sup>This study was performed in accordance with the ethical standards in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Approval was granted by the Ethics Committee of local institution (UCL Department of Medical Physics and Biomedical Engineering) on 20<sup>th</sup> Jan. 2023 [24055/001].

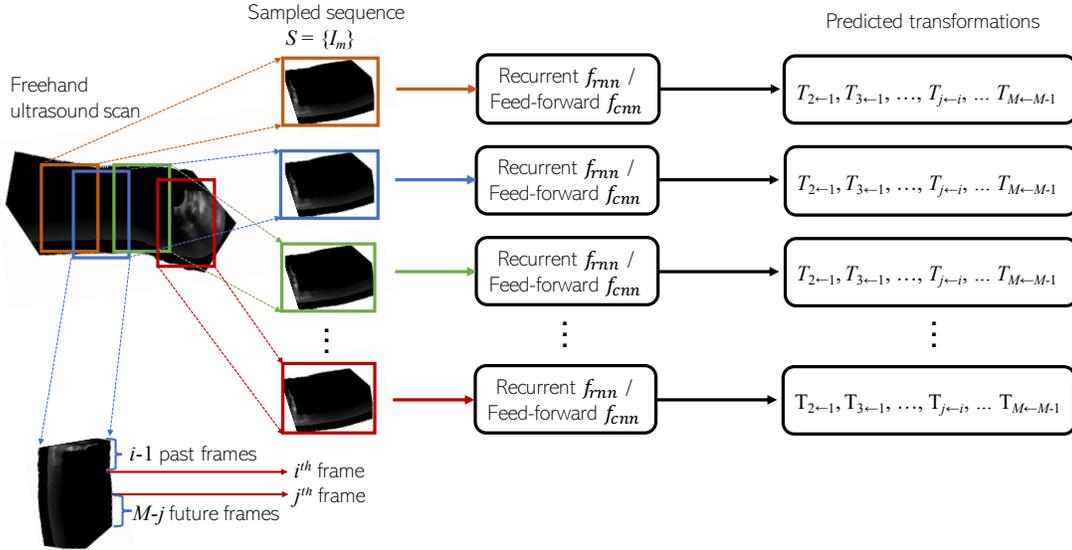


Fig. 1: Illustration of the proposed transformation-prediction algorithm.

subjects are randomly removed, on which the same networks are trained and subsequently tested on the same test data. The difference in performance is quantified with respect to the reduced variance. For investigating the performance changes due to reduced protocol dependency, two sets of additional models are trained, ‘straight’ only and ‘c-shape and s-shape’, using one and two from the three different types of scans. Together with the models trained on all three types of scans, they represent three levels of protocol variance,  $v_p = 1, 2, 3$ . Anatomical and protocol variances may both relate to various percentage of frames in a scan, denoted as  $v_l$ ,  $v_l = 50\%, 75\%$ , and can be tested using training scans that are cropped to 50% and 75% of their original lengths.

## IV. EXPERIMENTS

### A. Data Acquisition

US data were acquired on an Ultrasonix machine (BK, Europe) with a curvilinear probe (4DC7-3/40), from 19 volunteers on both their left and right forearms. Three trajectories, straight, c-shape and s-shape, in a distal-to-proximal direction, were acquired for each forearm. For each trajectory, two scans were obtained by keeping the US probe approximately perpendicular of and parallel to the scanning direction, as illustrated in Fig. 2. Thus, six US scans were acquired on each forearm, each scan containing 36–430 frames (100–200 mm). The dataset contains 228 scans in total, with statistics summarised in Fig. 3, and was split into train, validation and test sets by a ratio of 3:1:1 on the scan level. Images with a size of 480×640 were recorded at 20 fps. The frequency was fixed at 6 MHz with a depth of 9 cm, a dynamic range of 83 dB, an overall gain of 48%, and the speckle reduction was set at median level and the persistence at 3. Spatial calibration from image to tool coordinates was based on a pinhead-based method [36], and the temporal difference between the optical

tracker (NDI Polaris Vicra, Northern Digital Inc., Canada) and imaging was calibrated using the Plus Toolkit [37].

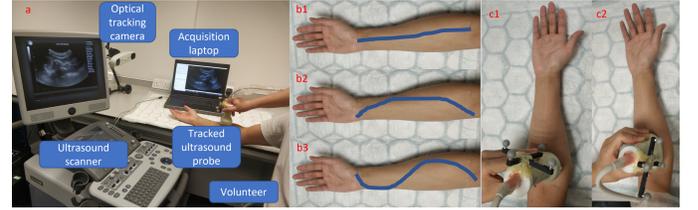
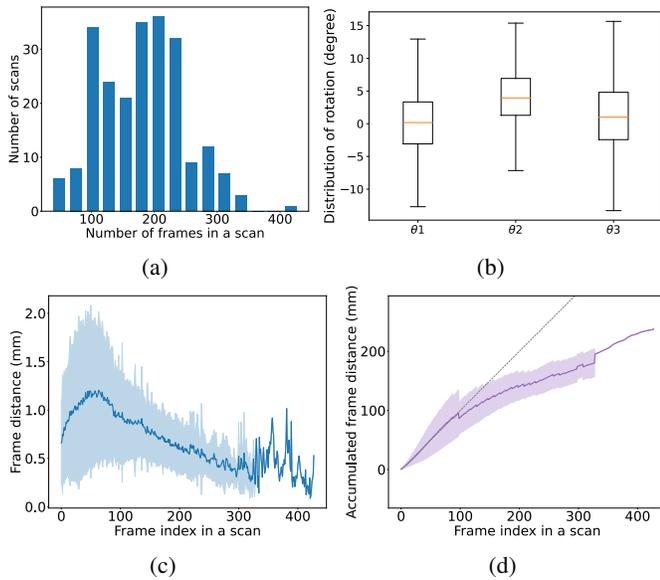


Fig. 2: Photographs of a) the US data acquisition system, b) various US probe trajectories, and c) various US plane orientations.

### B. Network Development and Implementation

Both CNN- and RNN- based networks were trained using randomly-selected  $(\tau + 1)$  tasks with variable  $M$  sequence length, in order to quantify the impact on performance due to varying long-term dependency, as discussed in Section III-C. The commonly used and well-established CNN- and RNN- based networks were adapted in this paper, without excessive fine-tuning, to benchmark the results. The EfficientNet (b1) [38] was adapted as a CNN, outputting  $(\tau + 1) \times 6$  transformation parameters using fully connected layer, where  $\tau$  denotes the number of auxiliary tasks. EfficientNet has the advantage of being smaller and more efficient than other CNN networks while still preserving state-of-the-art performance. The RNN architecture we used is LSTM with 1024-dimensional hidden states, for its capacity to capture long-term dependency, utilizing the same EfficientNet (b1) network as feature encoder (1000-dimensional feature vectors).

In this work, models were trained with  $M = 10, 20, 30, 40, 49, 60, 75, 100$  separately, with  $\tau + 1 =$



**Fig. 3:** Statistics of the dataset: (a) The histogram of scan length. (b) The distribution of three rotation parameters, between two adjacent frames. (c) Mean and standard deviation of distances between consecutive frames. (d) Mean and standard deviation of accumulated consecutive frame distances.

45, 80, 124, 157, 165, 177, 197, 218 sampled tasks<sup>6</sup>. This results in 16 RNN/CNN networks. To study the effect of long-term dependency on reconstruction accuracy, the same CNN and RNN models were adapted with input sequence length  $M = 2$  (i.e. only using two adjacent US frames as input and outputting the transformation between them), used as the baseline. The motivation is to investigate the effect of long-term dependency, with and without further input frames. A recent method for freehand 3D US reconstruction, DCL-Net [21] has also been implemented for comparison. In addition, 4 models with re-sampled  $M = 20, 49, 75, 100$  were trained using the same strategy to quantify the performance change due to altered dependency, i.e. reduced anatomical and protocol variance, detailed in Section III-F.

For all networks, a minibatch size of 32, Adam optimizer, and learning rate  $10^{-4}$  were used. The minibatch size and optimizer were empirically selected based on validation set performance, and the learning rate  $10^{-4}$  were tested among  $\{10^{-3}, 10^{-4}, 10^{-5}\}$ . To test the dependency hyperparameters, results from varying  $M$ ,  $i^*$  and  $j^*$  were computed. Each network was trained for at least 20,000 epochs until convergence, for up to 9 and 4 days on Ubuntu 18.04.6 LTS with a single NVIDIA Quadro P5000 GPU card, for RNNs and CNNs, respectively. All the results are reported on the test set unless otherwise specified.

The optimum predicted transformations, evaluated on the validation set, will be regarded as the main task. However, which one is optimum is unknown before training the model.

<sup>6</sup>As the number of tasks is one of the major sources of computational complexity and memory consumption, we have selected and tested a subset of all possible transformation predictions, with various transformation intervals, and past and future frame numbers.

Therefore, equal weights are given to different tasks to ensure a fair opportunity for each potential main task. In addition, tuning the explicit weighting between tasks may be partially redundant given the large number of auxiliary tasks and different configurations when varying the three hyperparameters,  $M$ ,  $i^*$  and  $j^*$ , some of which are equivalent to weighting the tasks differently. For example, less auxiliary tasks correspond to a bigger weight on the main task, and vice versa.

## V. RESULTS

### A. Results with Varying Dependency Hyperparameters<sup>7</sup>

Fig. 4 summarises the performance of  $\epsilon_{acc.}$  with respect to past- and future- frames, from all the models, with all available training data, described in Section IV. The reconstruction error decreases when more past frames are used. For example, using 74 past frames, the CNN and RNN achieved  $\epsilon_{acc.}$  of  $9.44 \pm 0.50$  mm and  $10.04 \pm 0.56$  mm, respectively. Both represent statistically significant improvement ( $p$ -value  $< 0.001$ ), based on unpaired t-test at a significance level at  $\alpha = 0.05$ , compared with that from the baseline (i.e.,  $M = 2$ ,  $\epsilon_{acc.} = 22.75$  mm) and DCL-Net ( $\epsilon_{acc.} = 22.15$  mm). The influence of past frames is further illustrated in Figs. S-3 to S-6 in Supplementary Material, showing that  $\epsilon_{acc.}$  decreased with more past frames, when the number of future frames is fixed. Fig. 5 also illustrates better reconstruction from longer sequence length.

Other interesting observations include: 1) When using fewer than 25 added past frames, the improvement is not obvious. For example, with 20 past frames, no statistically significant difference was found between the baseline and either CNN ( $p$ -value=0.762) or RNN ( $p$ -value=0.815); 2) There was no statistically significant difference found between CNN and RNN, which may suggest that feed-forward models are equally competent in modelling US sequences with limited length, compared to the more “specialised” RNNs; 3) The same trend was not found when future frames increased. This was first suspected to be caused by non-constant scanning speed, as illustrated in Fig. 3 (d). Additional experiments are illustrated in Fig. 6 to investigate the relationship between the reconstruction performance and scanning speed. The two models are trained using re-sampled train data with relatively constant speed and a reversed frame order. No evidence shows the correlation between  $\epsilon_{acc.}$  and the scanning speed.

Table I shows the effect of sequence length on reconstruction performance, using the best CNN models among sampled tasks, evaluated by four evaluation metrics. It can be concluded that larger  $M$  generally results in a smaller reconstruction error, due to the utilization of a relatively larger long-term dependency, i.e. more past and future frames. On the other hand, larger  $M$  corresponds to higher computational complexity, reflected in speed of forward/backward process and model convergence.

Fig. 7 shows the train and validation losses, trained with the varying train sets. A relatively larger number (20,000) was

<sup>7</sup>Other reconstruction metrics yielded the same conclusions as summarised above, and these detailed results are provided in Supplementary Material for brevity.

empirically selected as the maximum training epoch to train the model until convergence. Although validation loss begins to increase after a relatively rapid decrease, the best model used during inference stage are selected based on the performance on the validation set. Although the aim of this work is primarily to analyse the hyperparameters, consistent results on the validation set were also obtained, shown as in Fig. 7. This suggested the feasibility to tune these hyperparameters for optimum reconstruction for specific applications, on available validation sets.

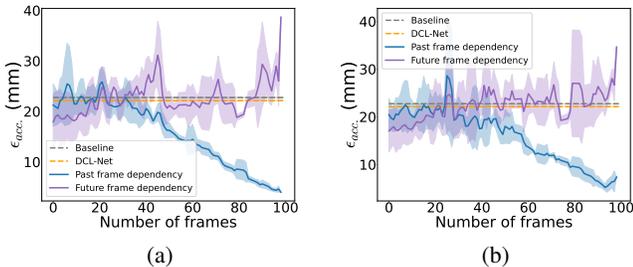


Fig. 4:  $\epsilon_{acc}$ . with respect to dependency, from CNN (a) and RNN (b). The means and standard deviations of  $\epsilon_{acc}$ . are plotted over all test scans, from models with  $M = 20, 49, 75, 100$ .

TABLE I: Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on all train data. Note:  $\epsilon_{dice}$  is computed on the perpendicular scans as an example.

$M$	$\epsilon_{frame}$	$\epsilon_{acc}$ .	$\epsilon_{dice}$	$\epsilon_{drift}$
2	$0.53 \pm 0.46$	$22.75 \pm 17.51$	$0.50 \pm 0.29$	$29.59 \pm 19.53$
10	$0.46 \pm 0.56$	$19.28 \pm 13.25$	$0.54 \pm 0.30$	$26.55 \pm 13.29$
20	$0.39 \pm 0.36$	$16.59 \pm 11.45$	$0.58 \pm 0.27$	$22.92 \pm 11.56$
30	$0.38 \pm 0.35$	$19.37 \pm 12.86$	$0.60 \pm 0.25$	$27.70 \pm 15.70$
40	$0.33 \pm 0.35$	$16.50 \pm 9.21$	$0.63 \pm 0.26$	$23.83 \pm 14.95$
49	$0.32 \pm 0.22$	$18.69 \pm 10.44$	$0.56 \pm 0.27$	$28.62 \pm 17.62$
60	$0.25 \pm 0.10$	$14.08 \pm 8.37$	$0.64 \pm 0.26$	$22.78 \pm 16.73$
75	$0.24 \pm 0.09$	$11.12 \pm 6.60$	$0.75 \pm 0.23$	$18.20 \pm 14.17$
100	$0.19 \pm 0.08$	$4.01 \pm 4.01$	$0.77 \pm 0.17$	$7.24 \pm 8.33$

### B. Ablation Study with Reduced Dependency Factors

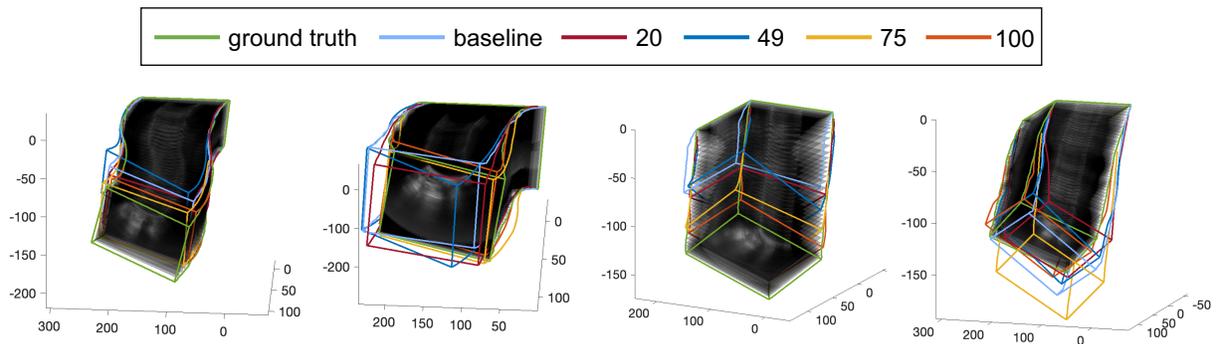
The reconstruction performance for different models trained with various variance-reduced training sets (Section III-F) is shown in Fig. 8, evaluated using  $\epsilon_{acc}$ . and  $\epsilon_{frame}$ , with increasing past frames. In practice, the reduction in either anatomical or protocol dependencies generally led to expected poorer performance. For instance, the model trained with only ‘straight’ scans yielded highest reconstruction errors in both  $\epsilon_{acc}$ . and  $\epsilon_{frame}$ , worse than baseline model ( $M = 2$ ) even when the past frame number is high. This suggested that the improved reconstruction accuracy contributed by the long-term dependency (e.g., seen with more than 70 past frames), was considerably reduced by mismatched protocol variances between train and test sets. To a lesser extent, removing 75% training subjects resulted in similar performance reduction. The other variance reduction models (‘c-shape and s-shape’, 75% training subjects or 75% scan length) yielded much less substantial performance losses, suggesting the current

levels of anatomy or protocol variance still include long-term dependency and benefit the reconstruction. The same conclusion can be drawn when only perpendicular or parallel scans are sampled for testing (as shown in Fig. 9).

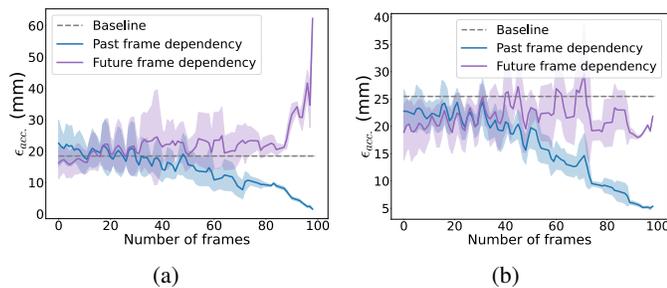
## VI. DISCUSSION

Many previous studies have focused on improving the networks and their training strategies [8], [10], including those with prior knowledge [7] and additional sensors, such as IMU [22]. These developments are not necessarily specific to utilising long-term dependency, and may be applied in addition to the proposed input encoding and multi-transformation prediction to further advance the performance. The presented work adopted established CNNs and RNNs for providing the uncomplicated results to quantify the advantageous long-term dependency, as the first step towards maximising its utility.

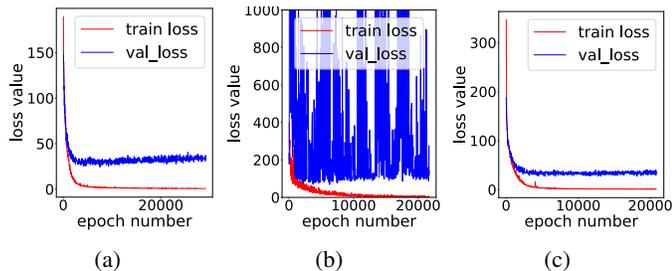
These results suggested that 1) Both the hypothesized anatomical and protocol dependency are likely to be factors for long-term dependency-improved reconstruction, shown in Section V-B; 2) Between the two, the protocol dependency is more likely to be the predominant source for the benefits from including long-term dependency; and 3) The interesting difference consistently observed between the past and future frames, in contributing to reconstruction accuracy, remains unexplained and a subject of investigation in our ongoing study. It is important to emphasize the application-specific nature of the second conclusion. The forearm dataset used in this paper may be considered specific in a number of aspects, including anatomical content richness, compared with other anatomical targets. As a result, conclusions drawn in this study may need further evaluation, when a different data set is considered. The proposed methodology however shall still be applicable and fit-for-purpose for different clinical indications. Future work should thus aim to access tracked ultrasound data from different clinical usages. Furthermore, 3D visualisation in addition to its geometric reconstruction of US volume can be crucial for many applications, which should be investigated further for its clinical applicability. It is worth noting that the influence of scanning speed, image quality, and overlap between adjacent frames remains open research questions for freehand US reconstruction, and may affect the robustness and generalisation of the tested reconstruction method, in turn may be specific to our conclusion on long-term dependency. In addition, definitions for these factors also remain an open research question and multiple potential solutions have been proposed, such as protocol-adaptation [39] and application-specific quality-control [40]. The generalisability of our conclusion should be interpreted with respect to individual applications and reconstruction algorithms, although the methodology of factorising the long-term dependency may nevertheless be useful. Besides, quantifying the difference between the anatomy and protocol dependency can also be application-dependent, with varying practical costs for changing their complexity. Further experimental results for these applications are mandatory. What is more, US images acquired using different ultrasound scanner/probe, by different researchers, may be different due to the specific parameter configurations and the intra- and inter-variability of the acquired



**Fig. 5:** Reconstruction from baseline and various  $M = 20, 49, 75, 100$ , with a selected transformation of  $T_{18\leftarrow 8}$ ,  $T_{32\leftarrow 30}$ ,  $T_{69\leftarrow 64}$ , and  $T_{94\leftarrow 92}$ , respectively. The trajectories are ‘S’ shape, ‘C’ shape, straight line shape, and straight line shape, from left to right, respectively.



**Fig. 6:**  $\epsilon_{acc}$ . with respect to dependency, trained using re-sampled train data with relatively constant speed (a) and a reversed frame order (b), with a CNN model.  $\epsilon_{acc}$ . is shown as the mean and standard deviation over all scans in the test set, computed using models with  $M = 20, 49, 75, 100$ .



**Fig. 7:** Train and validation loss of models trained with all data in train set (a), straight data in train set (b), and 25% subject reduction of the train set (c).

US images. Therefore, acquiring various types of dataset by using various kinds of ultrasound scanner/probe, by a number of researchers, is our future research focus, which can be used to test the generalization of the conclusion.

## VII. CONCLUSION

This work proposed a new parametric dependency based on frame encoding and multi-tasking transformation, to quantify dependency factors originated from anatomical and protocol characteristics. The experiments showed that long-term dependency based on either recurrent or feed-forward models can significantly improve reconstruction, and the improvement

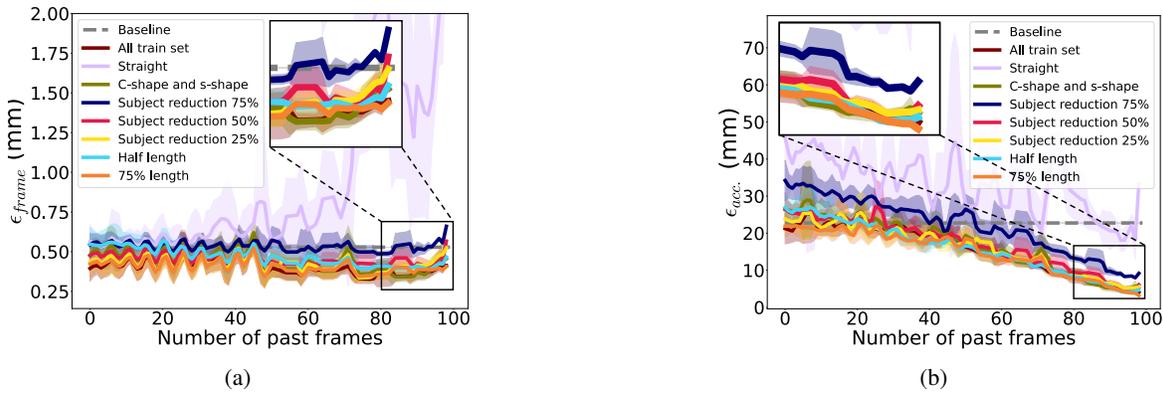
was dependent on the frame-to-transformation distance and transformation intervals. It was also found that the scanning protocol and, to a lesser degree, the anatomical content are both important in utilising the long-term dependency. As the proposed approach is based on the multi-task learning framework, negative transfer among tasks could lead to inferior performance. The ongoing work investigates methods such as task grouping [41] to further improve the main task and for potentially utilising multiple main tasks.

## ACKNOWLEDGMENT

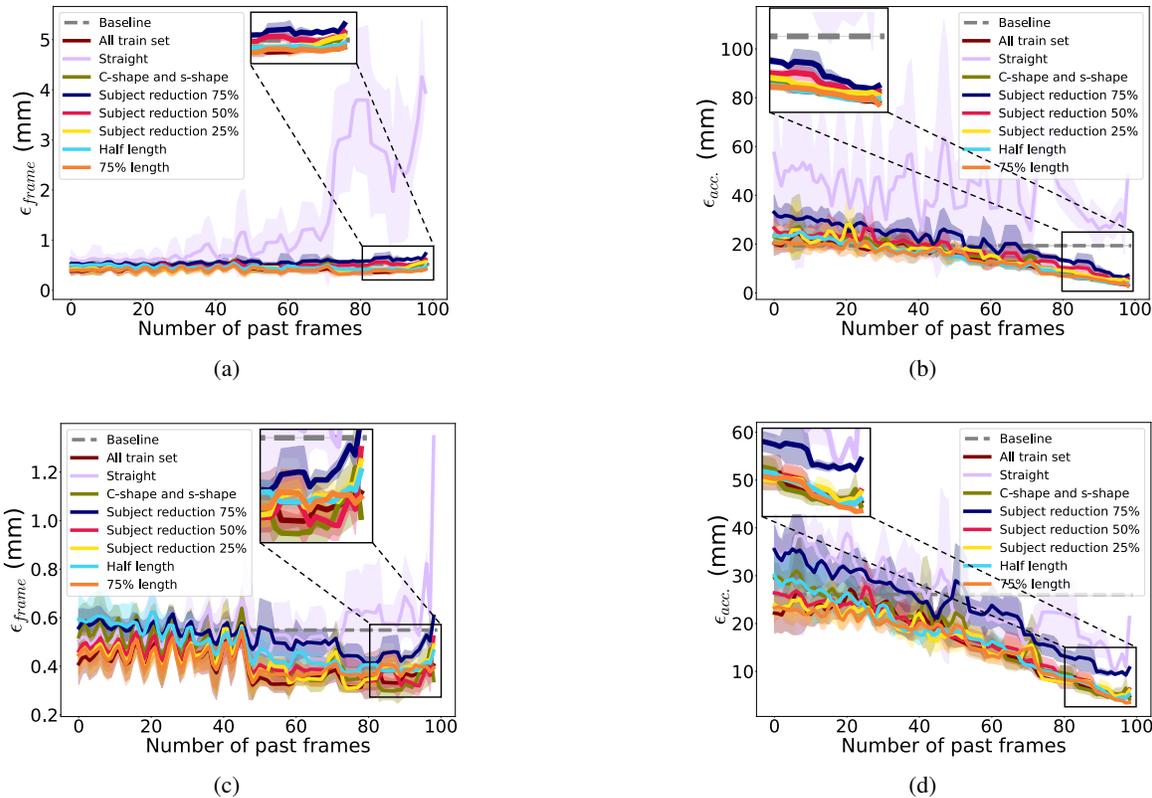
This work was supported by the EPSRC [EP/T029404/1], a Royal Academy of Engineering / Medtronic Research Chair [RCSRF1819\7\734] (TV), Wellcome/EPSRC Centre for Interventional and Surgical Sciences [203145Z/16/Z], and the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK [C28070/A30912; C73666/A31378], Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester. TV is co-founder and shareholder of Hypervision Surgical. Qi Li was supported by the University College London Overseas and Graduate Research Scholarships. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## REFERENCES

- [1] A. Benjamin et al., “Renal volume estimation using freehand ultrasound scans: an ex vivo demonstration,” *Ultrasound Med. Biol.*, vol. 46, no. 7, pp. 1769–1782, 2020.
- [2] S. Chung et al., “Freehand three-dimensional ultrasound imaging of carotid artery using motion tracking technology,” *Ultrasonics*, vol. 74, pp. 11–20, 2017.
- [3] J. Chen et al., “Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test,” *International Journal of Imaging Systems and Technology*, vol. 8, no. 1, pp. 38–44, 1997.
- [4] R. W. Prager et al., “Sensorless freehand 3-d ultrasound using regression of the echo intensity,” *Ultrasound Med. Biol.*, vol. 29, no. 3, pp. 437–446, 2003.
- [5] K. Miura et al., “Localizing 2d ultrasound probe from ultrasound image sequences using deep learning for volume reconstruction,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent. Workshop: ASMUS*. Springer, 2020, pp. 97–105.



**Fig. 8:** The reconstruction performance with regards to the past dependency. The performance is shown as the means and standard deviations of  $\epsilon_{frame}$  and  $\epsilon_{acc.}$  over all test scans, from models with  $M = 20, 49, 75, 100$ . All models trained with different variance-reduced data are tested on the same original test set.



**Fig. 9:** The reconstruction performance with regards to the past long-term dependency. The performance is shown as the means and standard deviations of  $\epsilon_{frame}$  and  $\epsilon_{acc.}$ , from models with  $M = 20, 49, 75, 100$ , trained with different variance-reduced data, tested on parallel or perpendicular scans in the original test set: (a) Performance of  $\epsilon_{frame}$  on parallel scans. (b) Performance of  $\epsilon_{acc.}$  on parallel scans. (c) Performance of  $\epsilon_{frame}$  on perpendicular scans. (d) Performance of  $\epsilon_{acc.}$  on perpendicular scans.

- [6] K. Miura et al., “Probe localization from ultrasound image sequences using deep learning for volume reconstruction,” in *SPIE*, 2021, vol. 11792, pp. 133–138.
- [7] M. Luo et al., “Self context and shape prior for sensorless freehand 3d ultrasound reconstruction,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent.* Springer, 2021, pp. 201–210.
- [8] G. Ning et al., “Spatial position estimation method for 3d ultrasound reconstruction based on hybrid transformers,” in *Int. Symposium on Biomed. Imaging*. IEEE, 2022, pp. 1–5.
- [9] K. Miura et al., “Pose estimation of 2d ultrasound probe from ultrasound image sequences using cnn and rnn,” in *Proc. Int. Conf. Med. Image*

- Comput. Computer-Assisted Intervent. Workshop: ASMUS*. Springer, 2021, pp. 96–105.
- [10] Y. Xie et al., “Image-based 3d ultrasound reconstruction with optical flow via pyramid warping network,” in *EMBC*. IEEE, 2021, pp. 3539–3542.
- [11] Q. Li et al., “Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames,” in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.
- [12] A. Lang et al., “Multi-modal registration of speckle-tracked freehand 3d ultrasound to ct in the lumbar spine,” *Med. Image Anal.*, vol. 16, no.

- 3, pp. 675–686, 2012.
- [13] Q. Huang et al., “Correspondence-3-d ultrasonic strain imaging based on a linear scanning system,” *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 62, no. 2, pp. 392–400, 2015.
- [14] Z. Chen et al., “Development of a wireless and near real-time 3d ultrasound strain imaging system,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 2, pp. 394–403, 2015.
- [15] H. Guo et al., “Ultrasound volume reconstruction from freehand scans without tracking,” *IEEE Trans. Biomed. Eng.*, vol. 70, no. 3, pp. 970–979, 2022.
- [16] E. D. Light et al., “Progress in two-dimensional arrays for real-time volumetric imaging,” *Ultrasonic imaging*, vol. 20, no. 1, pp. 1–15, 1998.
- [17] L. Mercier et al., “A review of calibration techniques for freehand 3-d ultrasound systems,” *Ultrasound in medicine & biology*, vol. 31, no. 2, pp. 143–165, 2005.
- [18] Q. Huang et al., “Speckle suppression and contrast enhancement in reconstruction of freehand 3d ultrasound images using an adaptive distance-weighted method,” *Applied Acoustics*, vol. 70, no. 1, pp. 21–30, 2009.
- [19] H. Gao et al., “Wireless and sensorless 3d ultrasound imaging,” *Neurocomputing*, vol. 195, pp. 159–171, 2016.
- [20] R. Prevost et al., “3d freehand ultrasound without external tracking using deep learning,” *Med. Image Anal.*, vol. 48, pp. 187–202, 2018.
- [21] H. Guo et al., “Sensorless freehand 3d ultrasound reconstruction via deep contextual learning,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent.* Springer, 2020, pp. 463–472.
- [22] M. Luo et al., “Deep motion network for freehand 3d ultrasound reconstruction,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent.* Springer, 2022, pp. 290–299.
- [23] M. Luo et al., “Multi-imu with online self-consistency for freehand 3d ultrasound reconstruction,” *arXiv preprint arXiv:2306.16197*, 2023.
- [24] M. Luo et al., “Recon: Online learning for sensorless freehand 3d ultrasound reconstruction,” *Med. Image Anal.*, vol. 87, pp. 102810, 2023.
- [25] B. Hou et al., “Computing cnn loss and gradients for pose estimation with riemannian geometry,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent.* Springer, 2018, pp. 756–764.
- [26] P. Yeung et al., “Learning to map 2d ultrasound images into 3d space with minimal human annotation,” *Med. Image Anal.*, vol. 70, pp. 101998, 2021.
- [27] Z. Wang et al., “Nerf-: Neural radiance fields without known camera parameters,” *arXiv preprint arXiv:2102.07064*, 2021.
- [28] C. Di Vece et al., “Deep learning-based plane pose regression in obstetric ultrasound,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 5, pp. 833–839, 2022.
- [29] Y. Zhou et al., “On the continuity of rotation representations in neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.
- [30] W. Wein et al., “Three-dimensional thyroid assessment from untracked 2d ultrasound clips,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent.* Springer, 2020, pp. 514–523.
- [31] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [32] A. Vaswani et al., “Attention is all you need,” *NeurIPS*, vol. 30, 2017.
- [33] R. Liao et al., “Reviving and improving recurrent back-propagation,” in *Proc. Int. Conf. Machine Learning*. PMLR, 2018, pp. 3082–3091.
- [34] A. Gruslys et al., “Memory-efficient backpropagation through time,” *NeurIPS*, vol. 29, 2016.
- [35] M. Sangiorgio and F. Dercole, “Robustness of lstm neural networks for multi-step forecasting of chaotic time series,” *Chaos, Solitons & Fractals*, vol. 139, pp. 110045, 2020.
- [36] Y. Hu 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.
- [37] A. Lasso et al., “Plus: open-source toolkit for ultrasound-guided intervention systems,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2527–2537, 2014.
- [38] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Machine Learning*. PMLR, 2019, pp. 6105–6114.
- [39] Q. Chen et al., “Cross-device cross-anatomy adaptation network for ultrasound video analysis,” in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent. Workshop: ASMUS*. Springer, 2020, pp. 42–51.
- [40] S. Saeed et al., “Image quality assessment for machine learning tasks using meta-reinforcement learning,” *Med. Image Anal.*, vol. 78, pp. 102427, 2022.
- [41] C. Fifty et al., “Efficiently identifying task groupings for multi-task learning,” *NeurIPS*, vol. 34, pp. 27503–27516, 2021.

## SUPPLEMENTARY MATERIALS

**TABLE S-I:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on all train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
2	0.53 ± 0.00	22.75 ± 0.00	0.50 ± 0.00	29.59 ± 0.00
10	0.51 ± 0.03	20.29 ± 1.48	0.46 ± 0.07	28.15 ± 1.92
20	0.45 ± 0.04	18.91 ± 2.10	0.46 ± 0.08	26.79 ± 3.33
30	0.47 ± 0.06	21.37 ± 1.32	0.41 ± 0.10	30.56 ± 1.96
40	0.46 ± 0.06	21.13 ± 2.20	0.43 ± 0.10	30.80 ± 3.00
49	0.46 ± 0.08	23.89 ± 3.88	0.37 ± 0.10	35.04 ± 5.13
60	0.39 ± 0.09	21.16 ± 5.20	0.41 ± 0.12	32.48 ± 7.57
75	0.37 ± 0.07	21.29 ± 4.91	0.34 ± 0.16	29.17 ± 6.43
100	0.32 ± 0.06	16.35 ± 6.33	0.23 ± 0.15	20.26 ± 7.19

**TABLE S-II:** Mean and standard deviation of best performance of four metrics, over all sampled tasks, with regards to various  $M$  by using RNN-based model, trained on all train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
2	0.57 ± 0.44	26.33 ± 15.99	0.41 ± 0.33	34.54 ± 18.10
10	0.37 ± 0.30	17.29 ± 11.13	0.62 ± 0.24	24.35 ± 13.57
20	0.37 ± 0.29	16.29 ± 9.82	0.61 ± 0.26	24.98 ± 13.11
30	0.37 ± 0.31	19.36 ± 11.53	0.59 ± 0.26	28.53 ± 15.85
40	0.30 ± 0.15	17.07 ± 8.97	0.61 ± 0.26	26.96 ± 15.65
49	0.27 ± 0.16	15.32 ± 7.89	0.65 ± 0.22	24.21 ± 15.84
60	0.26 ± 0.10	14.53 ± 7.64	0.66 ± 0.21	22.58 ± 14.15
75	0.23 ± 0.10	10.64 ± 6.09	0.77 ± 0.12	17.09 ± 11.24
100	0.20 ± 0.07	4.27 ± 3.66	0.73 ± 0.23	6.97 ± 6.79

**TABLE S-III:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using RNN-based model, trained on all train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
2	0.57 ± 0.00	26.33 ± 0.00	0.41 ± 0.00	34.54 ± 0.00
10	0.42 ± 0.03	18.53 ± 0.82	0.50 ± 0.09	25.74 ± 1.25
20	0.45 ± 0.05	18.14 ± 1.07	0.48 ± 0.08	27.36 ± 1.51
30	0.48 ± 0.05	21.44 ± 1.11	0.43 ± 0.10	31.24 ± 1.74
40	0.43 ± 0.08	21.24 ± 3.33	0.42 ± 0.11	31.91 ± 4.75
49	0.45 ± 0.11	21.38 ± 5.31	0.43 ± 0.11	32.59 ± 7.79
60	0.39 ± 0.08	20.84 ± 3.78	0.41 ± 0.13	31.23 ± 5.61
75	0.38 ± 0.09	20.75 ± 4.97	0.40 ± 0.13	30.44 ± 7.03
100	0.34 ± 0.09	16.80 ± 6.78	0.22 ± 0.13	20.87 ± 7.80

**TABLE S-IV:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on straight train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.51 ± 0.31	31.84 ± 17.16	0.45 ± 0.29	39.88 ± 19.33
49	0.43 ± 0.26	23.83 ± 14.05	0.57 ± 0.25	35.22 ± 25.10
75	0.40 ± 0.24	13.06 ± 7.09	0.66 ± 0.22	22.77 ± 17.92
100	0.48 ± 0.25	13.00 ± 23.79	0.64 ± 0.26	22.30 ± 41.10

**TABLE S-V:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on straight train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.57 ± 0.04	39.85 ± 2.03	0.25 ± 0.10	48.20 ± 3.42
49	0.55 ± 0.07	31.93 ± 4.02	0.29 ± 0.11	42.33 ± 4.86
75	0.48 ± 0.07	30.28 ± 9.17	0.31 ± 0.16	38.83 ± 10.42
100	0.92 ± 0.52	58.46 ± 22.81	0.13 ± 0.10	59.04 ± 24.95

**TABLE S-VI:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on c-shape and s-shape train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.48 ± 0.65	23.17 ± 19.62	0.51 ± 0.32	30.29 ± 18.02
49	0.40 ± 0.40	17.62 ± 8.39	0.60 ± 0.28	28.17 ± 20.22
75	0.25 ± 0.13	10.04 ± 6.78	0.78 ± 0.14	17.10 ± 13.16
100	0.24 ± 0.13	4.00 ± 2.72	0.80 ± 0.13	6.74 ± 7.19

**TABLE S-VII:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on c-shape and s-shape train data.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.55 ± 0.04	25.02 ± 1.84	0.38 ± 0.08	33.06 ± 3.10
49	0.54 ± 0.08	24.71 ± 4.92	0.35 ± 0.10	35.96 ± 6.67
75	0.39 ± 0.08	22.68 ± 4.86	0.34 ± 0.14	32.30 ± 6.38
100	0.33 ± 0.07	17.38 ± 7.54	0.25 ± 0.14	21.34 ± 8.36

**TABLE S-VIII:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with subjects reduction rate of 25%.

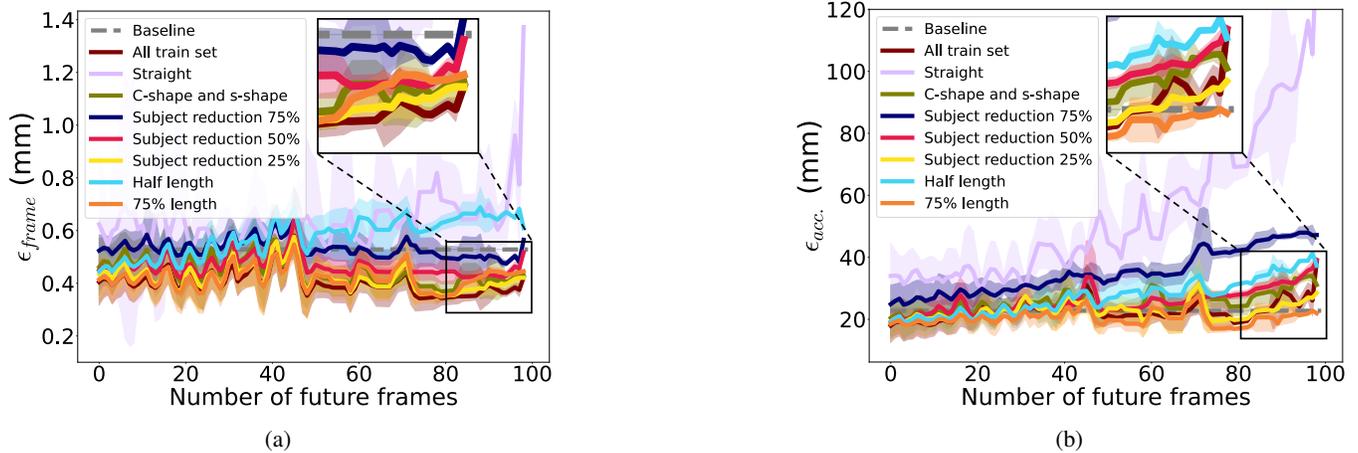
$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.44 ± 0.40	19.64 ± 9.04	0.53 ± 0.30	27.42 ± 10.79
49	0.32 ± 0.32	15.27 ± 7.48	0.63 ± 0.28	24.88 ± 15.80
75	0.24 ± 0.12	10.35 ± 6.27	0.73 ± 0.13	17.31 ± 11.62
100	0.20 ± 0.07	4.49 ± 4.21	0.76 ± 0.20	8.40 ± 9.08

**TABLE S-IX:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with subjects reduction rate of 25%.

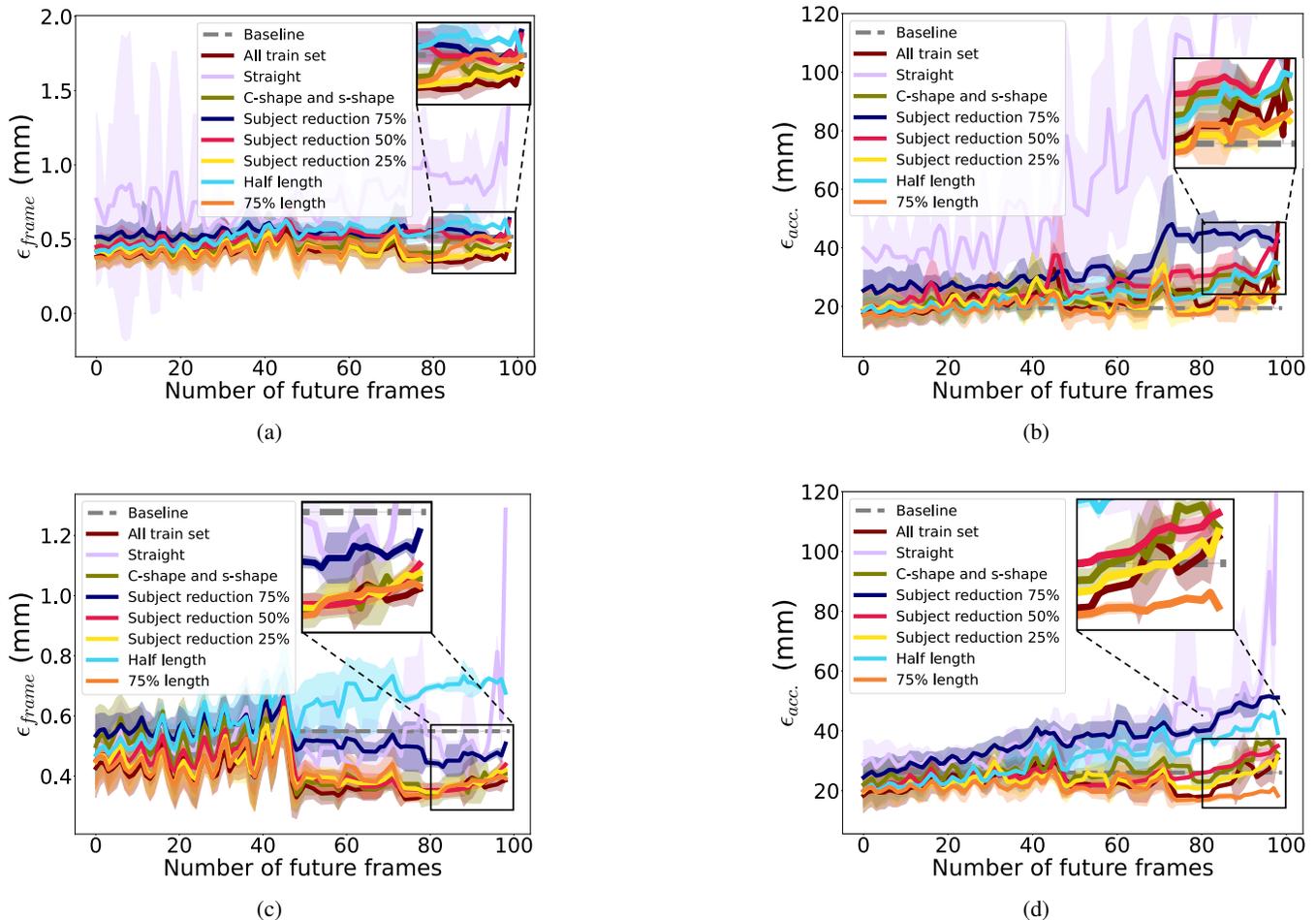
$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.50 ± 0.04	22.45 ± 2.64	0.39 ± 0.08	31.87 ± 3.94
49	0.46 ± 0.09	22.89 ± 4.64	0.40 ± 0.10	34.06 ± 6.47
75	0.40 ± 0.08	22.37 ± 5.34	0.34 ± 0.15	31.87 ± 7.30
100	0.32 ± 0.06	15.49 ± 5.15	0.24 ± 0.15	19.74 ± 5.96

**TABLE S-X:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with subjects reduction rate of 50%.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.47 ± 0.42	20.56 ± 9.62	0.49 ± 0.29	30.02 ± 15.26
49	0.34 ± 0.19	16.07 ± 9.01	0.57 ± 0.22	24.28 ± 17.53
75	0.27 ± 0.11	10.22 ± 6.42	0.71 ± 0.16	19.22 ± 13.09
100	0.29 ± 0.20	4.02 ± 3.82	0.83 ± 0.08	7.61 ± 9.38



**Fig. S-1:** The reconstruction performance with regards to future long-term dependency. The performance is shown as the mean and standard deviation of  $\epsilon_{frame}$  and  $\epsilon_{acc.}$  over all scans in the test set, from models with  $M = 20, 49, 75, 100$ . All models trained with different variance-reduced data are tested on the same original test set. (a) Performance of  $\epsilon_{frame}$  with regards to the number of future frames. (b) Performance of  $\epsilon_{acc.}$  with regards to the number of future frames.



**Fig. S-2:** The reconstruction performance with regards to the future long-term dependency. The performance is shown as the mean and standard deviation of  $\epsilon_{frame}$  and  $\epsilon_{acc.}$ , from models with  $M = 20, 49, 75, 100$ , trained with different variance-reduced data, tested on parallel or perpendicular scans in the original test set: (a) Performance of  $\epsilon_{frame}$  on parallel scans. (b) Performance of  $\epsilon_{acc.}$  on parallel scans. (c) Performance of  $\epsilon_{frame}$  on perpendicular scans. (d) Performance of  $\epsilon_{acc.}$  on perpendicular scans.

**TABLE S-XI:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with subjects reduction rate of 50%.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.53 ± 0.05	24.46 ± 4.46	0.38 ± 0.07	34.87 ± 6.09
49	0.50 ± 0.08	24.35 ± 4.91	0.34 ± 0.10	35.12 ± 7.16
75	0.41 ± 0.07	22.15 ± 4.73	0.33 ± 0.15	31.47 ± 5.84
100	0.40 ± 0.06	19.19 ± 7.27	0.24 ± 0.16	22.73 ± 7.80

**TABLE S-XII:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with subjects reduction rate of 75%.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.55 ± 0.49	26.62 ± 14.07	0.50 ± 0.25	37.02 ± 16.19
49	0.42 ± 0.34	18.51 ± 9.72	0.56 ± 0.30	29.50 ± 19.37
75	0.32 ± 0.21	12.01 ± 7.46	0.62 ± 0.28	20.93 ± 13.01
100	0.41 ± 0.24	7.83 ± 9.08	0.75 ± 0.17	13.66 ± 15.94

**TABLE S-XIII:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with subjects reduction rate of 75%.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.58 ± 0.02	30.03 ± 1.37	0.36 ± 0.08	41.24 ± 2.45
49	0.56 ± 0.06	28.53 ± 4.23	0.30 ± 0.10	39.93 ± 5.13
75	0.50 ± 0.09	28.75 ± 7.59	0.24 ± 0.13	39.82 ± 10.05
100	0.48 ± 0.06	28.38 ± 10.89	0.15 ± 0.15	32.40 ± 10.87

**TABLE S-XIV:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with half length.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.45 ± 0.40	22.58 ± 12.88	0.55 ± 0.25	31.77 ± 20.22
49	0.36 ± 0.31	15.72 ± 7.17	0.63 ± 0.21	24.83 ± 15.46
75	0.28 ± 0.09	10.79 ± 6.19	0.70 ± 0.22	19.06 ± 14.47
100	0.21 ± 0.08	3.80 ± 3.42	0.72 ± 0.27	6.54 ± 6.81

**TABLE S-XV:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with half length.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.53 ± 0.05	23.89 ± 1.32	0.40 ± 0.08	34.67 ± 2.29
49	0.53 ± 0.08	22.84 ± 3.59	0.35 ± 0.11	34.70 ± 5.04
75	0.51 ± 0.11	23.13 ± 6.55	0.28 ± 0.16	33.85 ± 8.66
100	0.43 ± 0.12	18.11 ± 8.75	0.15 ± 0.15	22.78 ± 10.07

**TABLE S-XVI:** Mean and standard deviation of best performance of four metrics, among all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with 75% length.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.41 ± 0.37	18.76 ± 11.11	0.53 ± 0.25	27.33 ± 13.54
49	0.31 ± 0.27	17.15 ± 9.87	0.58 ± 0.28	26.48 ± 19.58
75	0.23 ± 0.10	12.16 ± 8.40	0.73 ± 0.15	20.24 ± 15.60
100	0.19 ± 0.06	3.19 ± 2.36	0.78 ± 0.13	6.46 ± 5.39

**TABLE S-XVII:** Mean and standard deviation of average performance of four metrics, over all sampled tasks, with regards to various  $M$  by using CNN-based model, trained on train data with 75% length.

$M$	$\epsilon_{frame}$	$\epsilon_{acc.}$	$\epsilon_{dice}$	$\epsilon_{drift}$
20	0.47 ± 0.04	20.96 ± 1.35	0.39 ± 0.09	30.03 ± 2.07
49	0.45 ± 0.08	21.77 ± 2.90	0.40 ± 0.11	32.41 ± 3.83
75	0.39 ± 0.08	20.07 ± 3.81	0.39 ± 0.13	29.34 ± 5.32
100	0.32 ± 0.07	13.97 ± 4.88	0.24 ± 0.13	18.17 ± 6.01

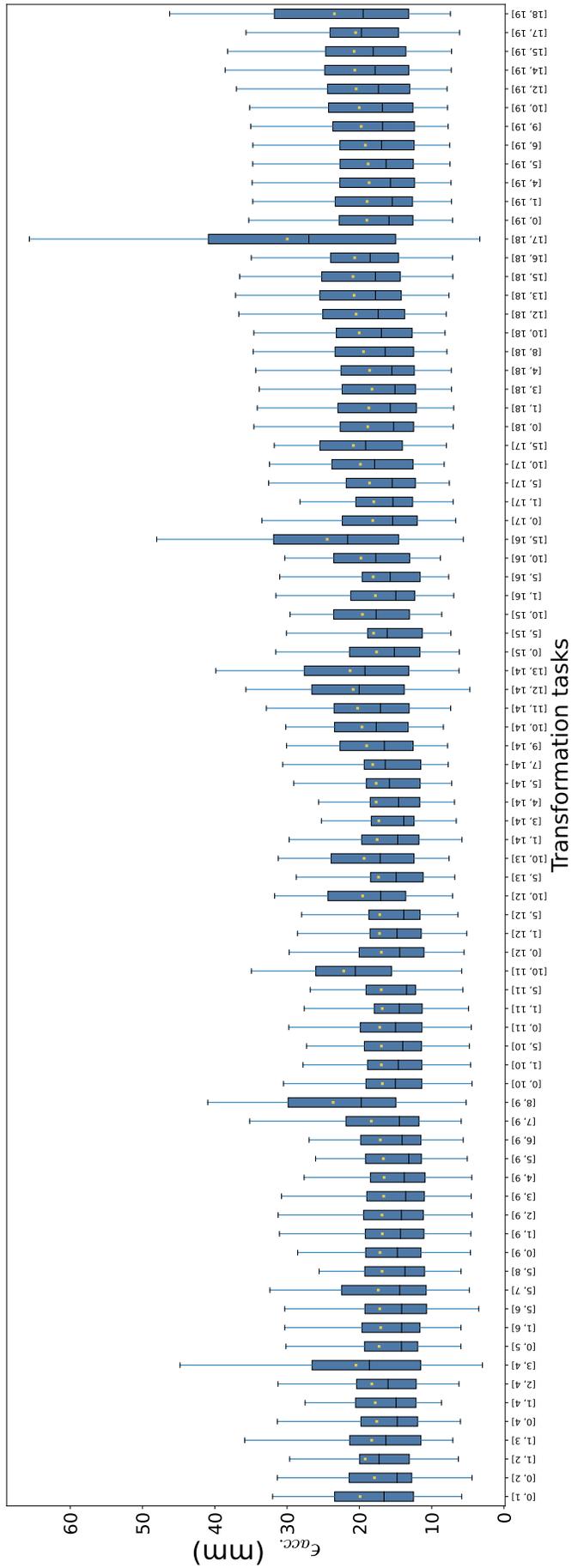


Fig. S-3: Performance of various transformation tasks in terms of  $\epsilon_{acc}$ , when sequence length  $M = 20$ . Each bar denotes the distribution of  $\epsilon_{acc}$ , using the specific transformation  $T_{j \leftarrow i}$  over all scans in the test set.

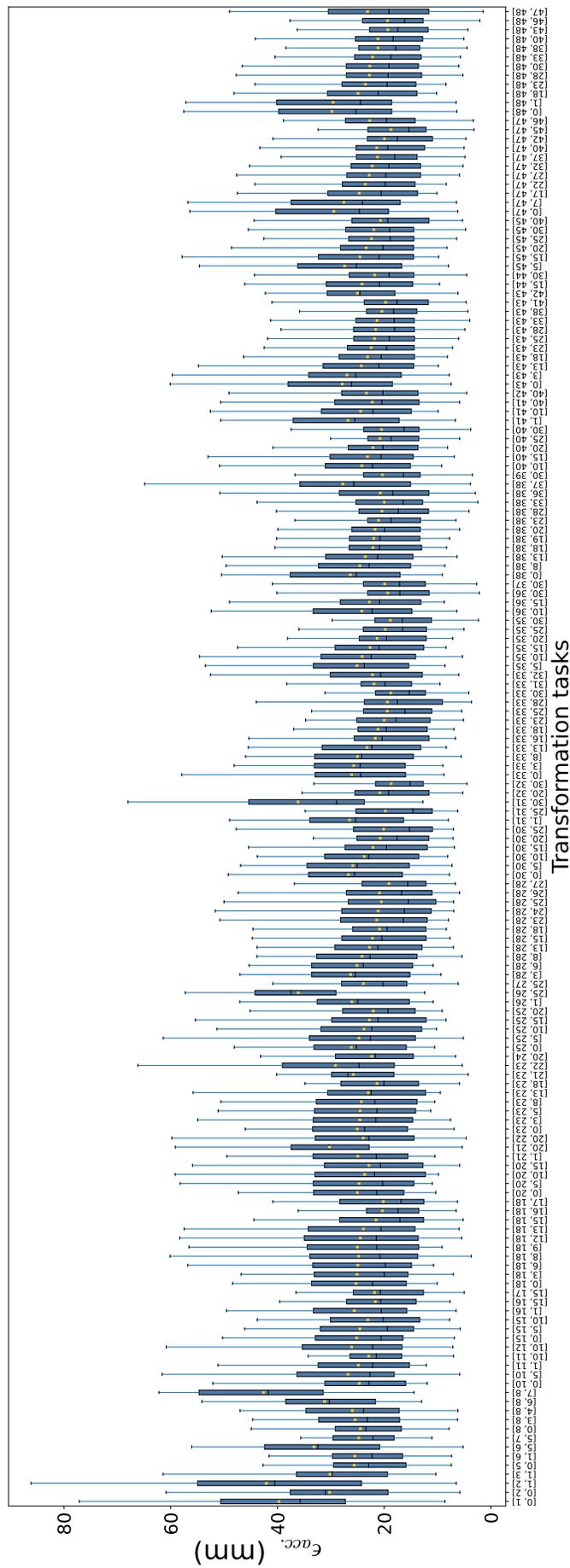
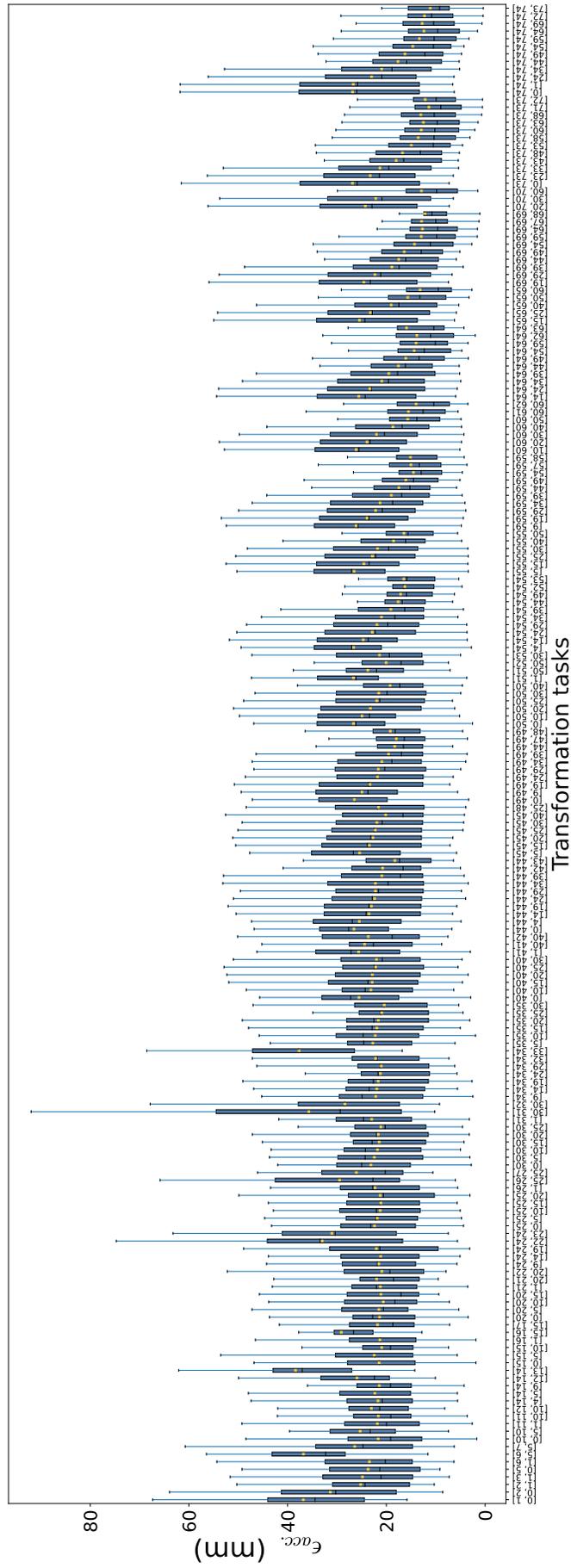
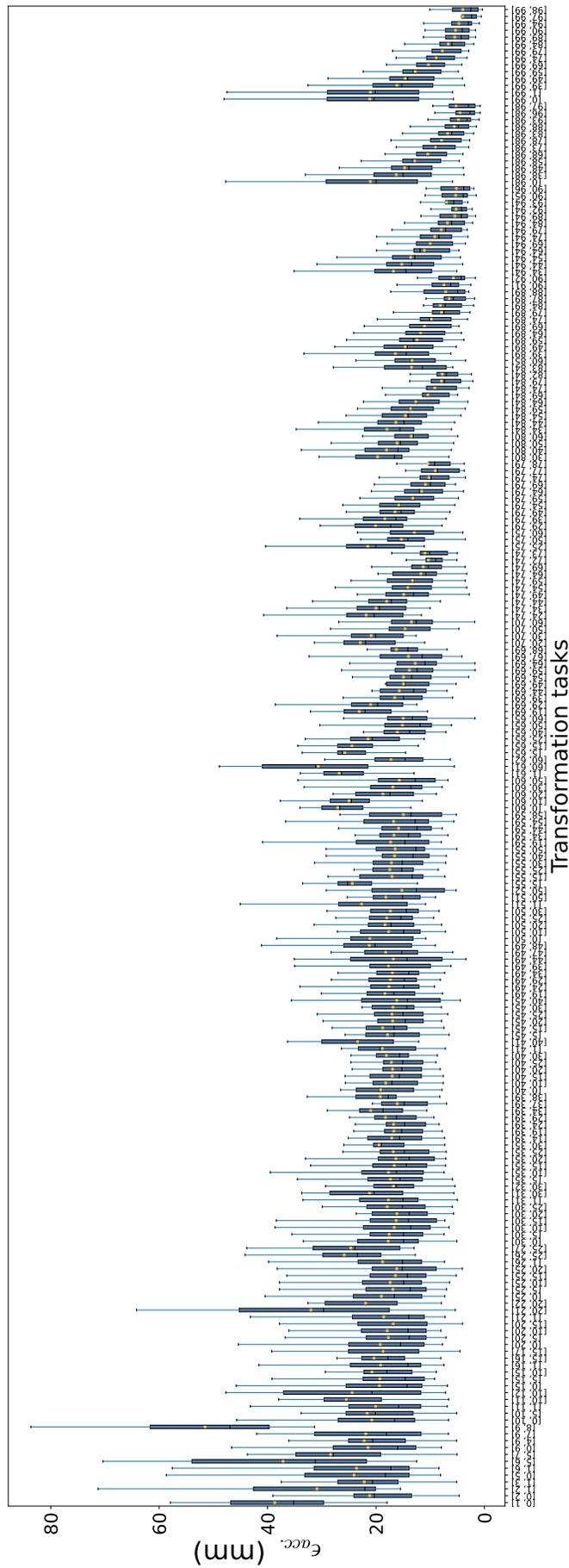


Fig. S-4: Performance of various transformation tasks in terms of  $\epsilon_{acc}$ , when sequence length  $M = 49$ . Each bar denotes the distribution of  $\epsilon_{acc}$ , using the specific transformation  $T_{j \leftarrow i}$  over all scans in the test set.



**Fig. S-5:** Performance of various transformation tasks in terms of  $\epsilon_{acc}$ , when sequence length  $M = 75$ . Each bar denotes the distribution of  $\epsilon_{acc}$ , using the specific transformation  $T_{j \leftarrow i}$  over all scans in the test set.



**Fig. S-6:** Performance of various transformation tasks in terms of  $\epsilon_{acc}$ , when sequence length  $M = 100$ . Each bar denotes the distribution of  $\epsilon_{acc}$ , using the specific transformation  $T_{j \leftarrow i}$  over all scans in the test set.