Fine-grained Context and Multi-modal Alignment for Freehand 3D Ultrasound Reconstruction

Zhongnuo Yan[1,2,3⋆], Xin Yang[1,2,3⋆], Mingyuan Luo[1,2,3], Jiongquan Chen[1,2,3], Rusi Chen[1,2,3], Lian Liu[1,2,3,4], and Dong Ni[1,2,3(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, China
nidong@szu.edu.cn
[2] Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University, China
[3] Marshall Laboratory of Biomedical Engineering, Shenzhen University, China
[4] Shenzhen RayShape Medical Technology Inc., Shenzhen, China

Abstract. Fine-grained spatio-temporal learning is crucial for freehand 3D ultrasound reconstruction. Previous works mainly resorted to the coarse-grained spatial features and the separated temporal dependency learning and struggles for fine-grained spatio-temporal learning. Mining spatio-temporal information in fine-grained scales is extremely challenging due to learning difficulties in long-range dependencies. In this context, we propose a novel method to exploit the long-range dependency management capabilities of the state space model (SSM) to address the above challenge. Our contribution is three-fold. First, we propose ReMamba, which mines multi-scale spatio-temporal information by devising a multi-directional SSM. Second, we propose an adaptive fusion strategy that introduces multiple inertial measurement units as auxiliary temporal information to enhance spatio-temporal perception. Last, we design an online alignment strategy that encodes the temporal information as pseudo labels for multi-modal alignment to further improve reconstruction performance. Extensive experimental validations on two large-scale datasets show remarkable improvement from our method over competitors.

Keywords: State Space Model · Multi-modal Alignment · Freehand 3D Ultrasound.

1 Introduction

Freehand 3D ultrasound (US) can provide comprehensive spatial information about the scanned region of interest and has been widely used in clinical diagnosis [9,11]. With the development of deep learning technology, current freehand 3D ultrasound reconstruction is free from dependence on external positioning devices, which were previously routinely utilized. They reconstruct the volume

⋆ Zhongnuo Yan and Xin Yang contribute equally to this work.

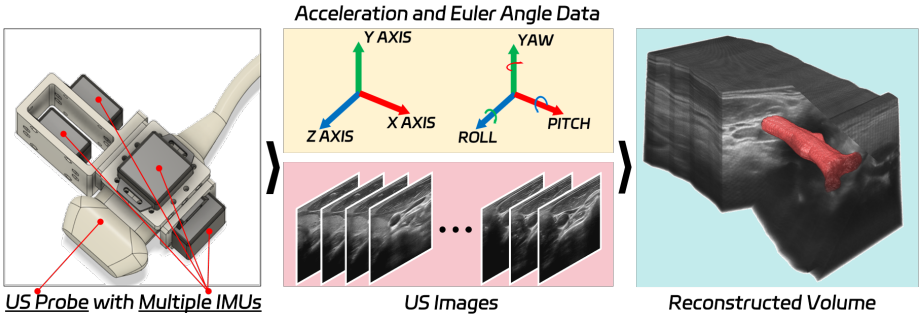

Fig. 1. Pipeline of freehand 3D US reconstruction with multiple IMUs.

by estimating the relative spatial transformations of a series of US images. However, the difficulty in mining spatio-temporal information in fine-grained scales makes it very challenging to accurately infer the relative position.

Recent studies were mainly based on convolutional neural network (CNN) and achieved advanced performance. Prevost et al. [12] introduced an end-to-end method utilizing CNN to estimate the relative motion of US images. Guo et al. [2] proposed a deep contextual-contrastive network (DC^2-Net) and introduced a contrastive learning strategy to enhance reconstruction performance. Li et al. [4] proposed to estimate 3D spatial transformation between US frames using recurrent neural networks (RNNs). Luo et al. [5,6] further improves reconstruction performance by online learning and shape priors. However, we note that the general approach of these studies is to first extract the coarse-grained features of the image and then extract the temporal information contained in these features. This design undoubtedly ignores the fine-grained spatio-temporal information, which is crucial for freehand 3D ultrasound reconstruction, and results in fragmentation between spatial and temporal information.

The lightweight sensor known as the inertial measurement unit (IMU) is an ideal choice for freehand 3D ultrasound reconstruction due to its low cost, low power consumption, and small size, as shown in Figure 1. Prevost et al. [11] have shown that incorporating IMU angles can enhance the accuracy of relative motion estimation. Luo et al. [7,8] have developed two multi-modal networks that leverage the valuable information from acceleration and angle measurements obtained from single or multiple IMUs to improve reconstruction performance. These studies highlight the significant improvement that IMUs can bring to freehand 3D ultrasound reconstruction.

In this study, we propose FiMA (**Fi**ne-grained Context and **M**ulti-modal **A**lignment), which exploits the efficient long-range dependency management capabilities of the state space model (SSM) to mine the spatio-temporal information in fine-grained features. Our contribution primarily revolves around three key aspects. First, we propose ReMamba, which mines multi-scale spatio-temporal information via multi-directional SSM. Additionally, we propose an adaptive fusion strategy that introduces multiple IMUs as additional temporal

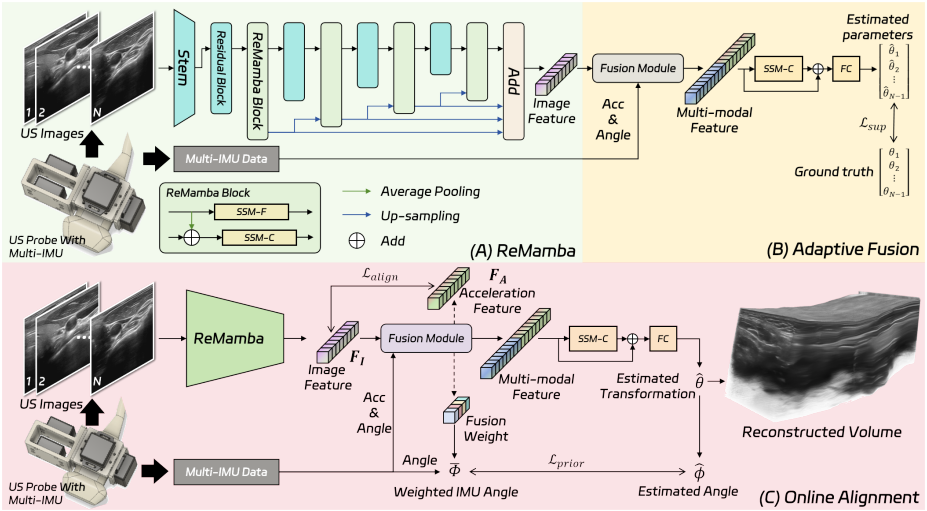

Fig. 2. Overview of the proposed FiMA.

information to enhance spatio-temporal information. Last, we design an online alignment strategy that uses the temporal information of IMUs as pseudo-labels for multi-modal alignment to further improve reconstruction performance.

2 Methods

An overview of our proposed FiMA is shown in Figure 2. It consists of three components: ReMamba for image sequence encoding (Figure 2(A)), adaptive fusion (Figure 2(B)) and online alignment (Figure 2(C)). Given an N-length scanning sequence $I = \{I_i | i = 1, 2, \cdots, N\}$ and corresponding multiple IMU data $U = \{U_i | i = 1, 2, \cdots, N-1\}$, we utilize FiMA to estimate the transformation parameters $\theta = \{\theta_i | i = 1, 2, \cdots, N-1\}$. In this context, θ_i refers to the 3-axis translations $t_i = (t_x, t_y, t_z)_i$ and rotation angles $\phi_i = (\phi_x, \phi_y, \phi_z)_i$ between image I_i and I_{i+1}. There are M independent IMU data $U_i = \{U_i^j | j = 1, 2, \cdots, M\}$. Here, U_i^j consists of 3-axis angles $\Phi_i^j = (\Phi_x, \Phi_y, \Phi_z)_i^j$ and accelerations $A_i^j = (A_x, A_y, A_z)_i^j$. The pre-processing process for Φ_i and A_i follows the method described in [7].

2.1 ReMamba with Multi-directional State Space Model

Fine-grained spatio-temporal information is crucial for accurate reconstruction. Previous methods mainly capture temporal information just in coarse-grained features, which is typically due to mining spatio-temporal information in fine-grained features involves intractable long-range dependencie. Recently, Mamba [1], with a SSM architecture and hardware-aware algorithms, demonstrated excellent

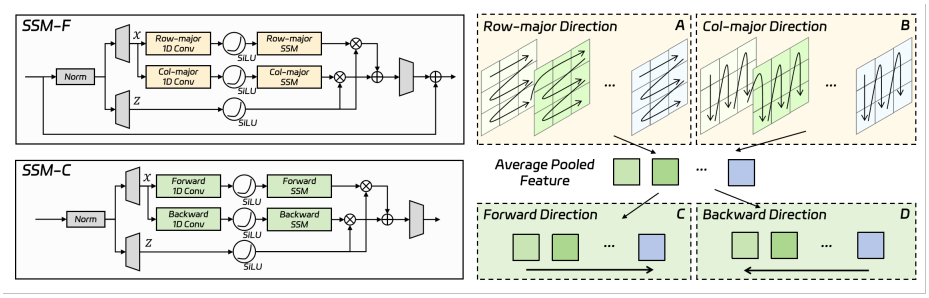

Fig. 3. Detail design of ReMamba Block.

long sequence processing capabilities. Inspired by SSM, we propose ReMamba, which mines multi-scale spatio-temporal information via multi-directional SSM. ReMamba consists of stem, multiple residual blocks and ReMamba blocks with multi-directional SSM. The architecture of ReMamba is shown in Figure 2(A).

Preliminaries. SSM maps a sequence $x(t)$ to another sequence $y(t)$ through a implicit latent state $h(t)$. It contains three learnable matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and satisfies the following system of equation:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{1}$$

The system is continuous, SSM is less effective on discrete data such as image. Mamba is a discrete version of continuous system, it makes SSM parameters a function of the input, which allows the model to selectively propagate or forget information based on current token. This facilitates compression of the context into a small state and better management of long-range dependency.

ReMamba Block. We flatten the three-dimensional image sequence features into one-dimensional sequences in order to model spatial and temporal information in a unified perspective, capturing temporal and spatial information in multi-scale features. Different flatten rules produce different contexts. To capture spatio-temporal information that is suitable for reconstruction, we design two different granularities of SSM within each ReMamba block, fine-grained **SSM-F** and coarse-grained **SSM-C**. Each ReMamba block takes as input two features of different scales and outputs two features of the same shape.

SSM-F is used to capture diverse spatio-temporal information in multi-scale features. SSM-F receives output of previous residual block as input, and after normalised distribution by LayerNorm, it goes through linear projections on the main branch and on the gated branch into high-dimensional space to obtain x, z, respectively. x then passes through the 1D convolution, SiLU and SSM layers in both row-major direction (Figure 3(A)) and col-major direction (Figure 3(B)). z goes through SiLU and products with the outputs of SSM in two directions in the main branch, the products are summed up and finally linearly projected back to the original dimensions. The projection results are summed with the inputs to get fine-grained temporal information as output.

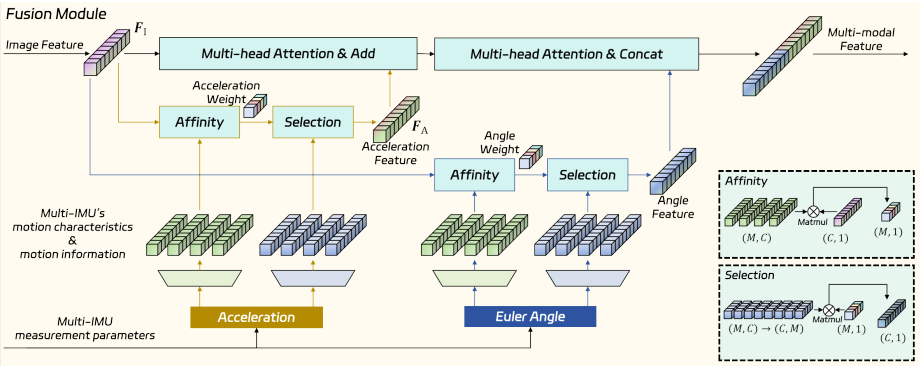

Fig. 4. Details of Fusion Module. Its input are the image features from ReMamba, the acceleration and Euler angles of multiple IMUs. It outputs multi-modal fused feature.

SSM-C is used to perceive overall motion patterns, the average pooling result of the previous residual block output is then summed with the previous SSM-C output, and the result as input to SSM-C, the input does not need to be flattened. Similarly, we use both forward (Figure 3(C)) and backward (Figure 3(D)) direction and finally get the coarse-grained temporal information as output.

2.2 Adaptive Fusion Strategy

IMU can provide motion information beyond the image, recording the state of motion over a period of time. To further enhance the spatio-temporal information, we introduce multiple IMUs, and consider the measurement parameters provided by IMUs as additional temporal information. However, motion of scan sequences is complex and various. In addition, due to each IMU having different characteristics and states and generating different noise, the temporal information provided by multiple IMUs is not always consistent.

We introduce an adaptive fusion strategy (Fusion Module in Figure 2 to address the above problem. Spatio-temporal information extracted from images and temporal information extracted from IMUs are not independent, instead, there should be a correlation between them. By combining the spatio-temporal information extracted from images, we expect the network to be able to judge whether each IMU provides reasonable temporal information at each moment. Specifically, as shown in Figure 4, we map the acceleration/angle of each IMU to the same two spaces as the image, representing the motion characteristics and detailed temporal information of this IMU, respectively. We calculate affinity between spatio-temporal information of images and representation for the motion characteristics of the acceleration/angle of each IMU, the affinity is used as a reference for weighting the detailed temporal information.

For the temporal information obtained from IMU acceleration/angle, we use two multi-head attention to fuse spatio-temporal information of images and temporal information of IMUs, respectively, and get the multi-modal fused features.

Finally, we use a SSM-C module with skip connection and a linear projection to decode the estimated transformation parameters $\hat{\theta}$. In training phase, we optimise the network using mean absolute error and Pearson correlation loss:

$$\mathcal{L}_{\text{sup}} = \|\hat{\theta} - \theta\|_1 + \left(1 - \frac{\mathbf{Cov}(\hat{\theta}, \theta)}{\sigma(\hat{\theta})\sigma(\theta)}\right) \tag{2}$$

where θ denotes the true transformation parameters, **Cov** denotes the covariance calculation, and σ denotes the standard deviation.

2.3 Online Alignment Strategy

To capture appropriate temporal features on unseen data, we propose an online alignment strategy (shown in Figure 2(C)), which takes the multiple IMU information as pseudo-labels and further enhance the reconstruction performance through multi-modal feature alignment in test phase.

For the image feature F_I from ReMamba, and the weighted temporal feature of IMU acceleration F_A from Fusion Module, we maximize the mutual information between them to align their feature spaces, while facilitating their mutual reduction of each other's uncertainty. However, it is difficult to optimize the mutual information directly, and we optimize one of its lower bounds refer to [10]. For an N-length sequence, the alignment loss $\mathcal{L}_{\text{align}}$ is calculated according to the image feature F_{Ii} and acceleration temporal feature F_{Ai} $(0 < i < N)$:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N-1}\sum_{i=1}^{N-1} \log\left(\frac{\exp(F_{Ii} \cdot F_{Ai}/\tau)}{\sum_{j=1}^{N-1}\exp(F_{Ii} \cdot F_{Aj}/\tau)}\right) \tag{3}$$

where τ is the temperature parameter, and we set it to 0.1.

IMUs provide accurate angle measurements in most cases, we use the fusion weights corresponding to IMU angle in the fusion module as prior, and use the prior weight to calculate average angle $\bar{\Phi}$ of these angle measurements, we calculate the loss between the estimated Euler angle $\hat{\phi}$ and weighted IMU angle using Pearson correlation loss:

$$\mathcal{L}_{\text{prior}} = 1 - \frac{\mathbf{Cov}(\hat{\phi}, \bar{\Phi})}{\sigma(\hat{\phi})\sigma(\bar{\Phi})} \tag{4}$$

We optimize FiMA using the sum of the above loss functions.

3 Experiments

Materials and Implementation. We construct two datasets refering to [8], including arm and carotid, from 50 volunteers. The arm dataset contains 583 scans, employing a variety of scanning tactics such as linear, curved, loop, and sector scans. Similarly, the carotid dataset includes 432 scans, utilizing linear,

Table 1. The mean (std) results of different methods on the arm and carotid scans. *indicates that the method does not require a sensor. ReM′ and ReM use Mamba and ReMamba block, respectively. F′ and F represent the direct mapping of acceleration/angle of multiple IMU to high dimension combined and the proposed adaptive fusion strategy, respectively. The best results are shown in blue.

Method	FDR(%)↓	ADR(%)↓	MD(mm)↓	SD(mm)↓	HD(mm)↓	MEA(deg)↓
	Arm dataset					
CNN-OF*	33.03(21.3)	46.80(36.5)	82.94(36.9)	2360.97(1503.3)	74.85(36.4)	4.82(3.0)
ResNet*	21.13(12.8)	31.29(18.0)	55.01(28.2)	1684.58(1795.8)	51.10(26.9)	7.36(4.2)
DC^2-Net*	18.12(12.7)	26.63(15.0)	48.26(30.2)	1399.03(1670.2)	48.15(30.8)	7.02(4.0)
RecON*	15.37(9.7)	22.23(11.7)	34.41(20.1)	1096.15(963.0)	30.97(15.8)	5.23(3.5)
MoNet	14.38(8.7)	21.20(10.4)	32.36(18.7)	1009.60(863.5)	28.96(14.2)	3.70(2.3)
OSCNet	13.06(7.4)	19.90(11.2)	30.81(17.3)	947.06(716.6)	27.69(13.2)	3.45(2.2))
ReM′	17.31(14.7)	26.36(16.6)	45.22(30.9)	1236.33(1135.5)	40.50(28.4)	6.50(3.3)
ReM	13.87(12.5)	20.86(14.6)	35.07(26.0)	969.46(911.6)	32.42(24.9)	5.67(2.9)
ReM+F′	12.87(12.6)	18.98(13.3)	31.88(24.4)	875.87(880.0)	29.78(22.9)	5.17(2.8)
ReM+F	10.85(8.0)	16.86(10.4)	27.38(16.1)	746.32(567.2)	25.60(15.9)	4.58(2.6)
FiMA	9.72(7.1)	15.53(9.6)	24.68(13.6)	677.48(498.4)	23.05(13.6)	3.41(1.8)
	Carotid dataset					
CNN-OF*	28.25(18.3)	42.87(21.6)	45.12(17.5)	1392.58(1057.0)	39.68(16.6)	3.95(2.9)
ResNet*	21.47(13.5)	32.56(13.1)	37.53(16.9)	1157.17(740.4)	33.24(15.6)	5.34(3.2)
DC^2-Net*	19.06(13.0)	30.64(17.1)	33.06(15.2)	1017.02(814.8)	27.99(12.8)	5.43(3.2)
RecON*	15.74(10.5)	26.80(19.0)	24.90(11.2)	800.50(716.9)	22.36(11.3)	4.25(2.8)
MoNet	14.53(9.5)	26.50(19.2)	23.67(10.7)	753.40(593.4)	21.11(11.0)	2.92(1.8)
OSCNet	14.17(9.5)	25.42(19.0)	23.25(10.5)	714.22(526.5)	20.62(10.5)	2.69(1.7)
ReM′	13.13(10.3)	24.07(16.0)	20.79(10.7)	599.77(526.4)	18.21(10.1)	4.60(2.1)
ReM	11.27(7.9)	19.94(10.9)	17.58(8.8)	497.85(389.0)	15.96(8.4)	4.38(2.0)
ReM+F′	10.13(6.6)	18.18(9.1)	16.41(8.4)	459.95(366.2)	14.55(7.6)	3.60(1.8)
ReM+F	9.07(5.9)	17.27(8.7)	15.12(7.5)	421.35(316.3)	13.33(6.9)	3.25(1.6)
FiMA	8.61(5.9)	16.16(8.1)	13.78(6.4)	391.77(298.4)	12.39(6.2)	2.09(1.2)

loop, and sector scan tactics. The average lengths of the arm and carotid scans are 386.39 and 241.25 mm, respectively. The size of scanned images is 248×260 pixels, with an image spacing of 0.15×0.15 mm^2. The positions of IMUs are the same as in Figure 1. The procurement and application of this data received approval from the local Institutional Review Board (IRB), ensuring compliance with ethical guidelines.

The arm and carotid datasets were split into training/validation/test sets in a ratio of 375/104/104 and 276/78/78 scans, respectively, based on volunteer allocation. We performed random augmentations on each scan referring to [8]. We used the Adam optimizer to optimize the model. During the training phase, the epochs and batch size are set to 200 and 1, respectively. To avoid overfitting, we set the initial learning rate to 2×10^{-4} and used a learning rate decay strategy that halves the learning rate every 30 epochs. During the testing phase, the iteration epoch and learning rate are set to 60 and 2×10^{-6}, respectively.

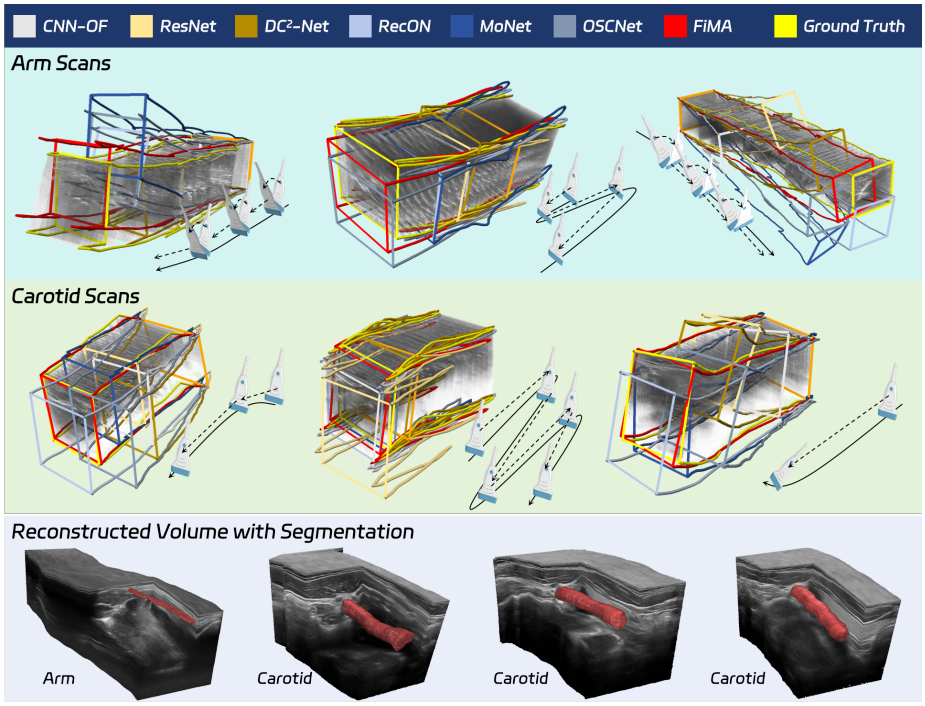

Fig. 5. Reconstruction examples produced by our proposed method. Red surface denotes the vessels reconstructed. Probe trajectory represents the scanning path.

Quantitative and Qualitative Analysis. We use following metrics refering to [8] to quantify the performance of FiMA: final drift rate (FDR), average drift rate (ADR), maximum drift (MD), sum of drift (SD), symmetric Hausdorff distance (HD), and mean error of angle (MEA). We compare FiMA with following methods: CNN-OF [11], ResNet [3], DC^2-Net [2], RecON [6], MoNet [7], OSCNet [8]. All comparison methods were conducted following original experimental settings. The quantitative results are shown in Table 1.

As seen in Table 1, ReMamba outperforms single-IMU-based MoNet in several metrics, even in the absence of an IMU. Moreover, it demonstrates the effectiveness of our proposed multi-directional SSM, adaptive fusion strategy and online alignment strategy. The optimal result achieved further improvements over OSCNet, with 25.57% /21.96% and 39.24%/36.43% improvement in FDR/ADR on the arm and carotid datasets, respectively. FiMA achieves the state-of-the-art performance.

Figure 5 displays representative reconstruction outcomes from all the methods of comparison. It is evident that our FiMA demonstrates superior performance and aligns with the ground truth more closely compared to other methods on both the arm and carotid datasets. Segmentation results on typical recon-

structed volumes show that FiMA can reconstruct blood vessels well, which is expected to provide a reference for 3D analysis of anatomical structures.

4 Conclusion

In this study, we propose FiMA to exploits the efficient long-range dependency management capabilities of SSM. FiMA realises the capture of spatio-temporal information in multi-scale features, including these fine-grained features that are crucial in reconstruction. We innovate an multi-modal fusion strategy to adaptively extracted suitable information from multiple IMUs to guide reconstruction. We propose an online alignment strategy to ensure stable and accurate reconstruction performance of FiMA when inferring on unseen data. The experimental results on the arm and carotid datasets show that above methods have resulted in great performance gains and FiMA achieves state-of-the-art reconstruction performance.

Acknowledgments. This work was supported by the grant from National Natural Science Foundation of China (Nos. 12326619, 62171290, 62101343), Science and Technology Planning Project of Guangdong Province (2023A0505020002), and Shenzhen-Hong Kong Joint Research Program (SGDX20201103095613036).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces (2023)
2. 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)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
4. 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: ISBI. pp. 1–5. IEEE (2023)
5. 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: MICCAI. pp. 201–210. Springer (2021)
6. Luo, M., Yang, X., Wang, H., Dou, H., Hu, X., Huang, Y., Ravikumar, N., Xu, S., Zhang, Y., Xiong, Y., et al.: Recon: Online learning for sensorless freehand 3d ultrasound reconstruction. Medical Image Analysis **87**, 102810 (2023)
7. Luo, M., Yang, X., Wang, H., Du, L., Ni, D.: Deep motion network for freehand 3d ultrasound reconstruction. In: MICCAI. pp. 290–299. Springer (2022)
8. Luo, M., Yang, X., Yan, Z., Li, J., Zhang, Y., Chen, J., Hu, X., Qian, J., Cheng, J., Ni, D.: Multi-imu with online self-consistency for freehand 3d ultrasound reconstruction. In: MICCAI. pp. 342–351. Springer (2023)

9. Mohamed, F., Siang, C.V.: A survey on 3d ultrasound reconstruction techniques. Artificial Intelligence—Applications in Medicine and Biology pp. 73–92 (2019)
10. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019)
11. Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., Wein, W.: 3d freehand ultrasound without external tracking using deep learning. Medical image analysis **48**, 187–202 (2018)
12. Prevost, R., Salehi, M., Sprung, J., Ladikos, A., Bauer, R., Wein, W.: Deep learning for sensorless 3d freehand ultrasound imaging. In: MICCAI. pp. 628–636. Springer (2017)