MambaXCTrack: Mamba-Based Tracker With SSM Cross-Correlation and Motion Prompt for Ultrasound Needle Tracking

Yuelin Zhang, Long Lei, Wanquan Yan, Tianyi Zhang, Raymond Shing-Yan Tang, and Shing Shin Cheng

***Abstract*—Ultrasound (US)-guided needle insertion is widely employed in percutaneous interventions. However, providing feedback on the needle tip position via US imaging presents challenges due to noise, artifacts, and the thin imaging plane of US, which degrades needle features and leads to intermittent tip visibility. In this letter, a Mamba-based US needle tracker MambaXCTrack utilizing structured state space models cross-correlation (SSMX-Corr) and implicit motion prompt is proposed, which is the first application of Mamba in US needle tracking. The SSMX-Corr enhances cross-correlation by long-range modeling and global searching of distant semantic features between template and search maps, benefiting the tracking under noise and artifacts by implicitly learning potential distant semantic cues. By combining with cross-map interleaved scan (CIS), local pixel-wise interaction with positional inductive bias can also be introduced to SSMX-Corr. The implicit low-level motion descriptor is proposed as a non-visual prompt to enhance tracking robustness, addressing the intermittent tip visibility problem. Extensive experiments on a dataset with motorized needle insertion in both phantom and tissue samples demonstrate that the proposed tracker outperforms other state-of-the-art trackers while ablation studies further highlight the effectiveness of each proposed tracking module.**

***Index Terms*—AI-based methods, deep learning methods, computer vision for medical robotics, visual tracking.**

Received 12 November 2024; accepted 21 March 2025. Date of publication 7 April 2025; date of current version 15 April 2025. This article was recommended for publication by Associate Editor A. Kuntz and Editor J. Burgner-Kahrs upon evaluation of the reviewers' comments. This work was supported in part by the Research Grants Council (RGC) of Hong Kong under Grant T45-401/22-N, Grant CUHK 14217822, and Grant CUHK 14207823, in part by the Innovation and Technology Commission of Hong Kong under Grant ITS/234/21, Grant ITS/233/21, and Grant ITS/235/22, and in part by Multi-scale Medical Robotics Center. *(Corresponding author: Shing Shin Cheng.)*

Yuelin Zhang is with the Department of Mechanical and Automation Engineering and T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong (e-mail: ylzhang@mae.cuhk.edu.hk).

Long Lei is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: longlei@cuhk.edu.hk).

Wanquan Yan was with the Department of Mechanical and Automation Engineering and T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong. He is now with the State Grid Xin Jiang Electric Power Company, Ltd. Supervoltage Branch, 830000, China (e-mail: wqyan@link.cuhk.edu.hk).

Tianyi Zhang is with the Medical Physics Graduate Program, Duke Kunshan University Suzhou 215316, China (e-mail: tz137@duke.edu).

Raymond Shing-Yan Tang is with the Department of Medicine and Therapeutics and Institute of Digestive Disease, The Chinese University of Hong Kong, Hong Kong (e-mail: raymondtang@cuhk.edu.hk).

Shing Shin Cheng is with the Department of Mechanical and Automation Engineering, T Stone Robotics Institute, Shun Hing Institute of Advanced Engineering, Multi-Scale Medical Robotics Center, The Chinese University of Hong Kong, Hong Kong, and also with the Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Hong Kong (e-mail: sscheng@cuhk.edu.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3558377, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3558377

I. Introduction

IN VARIOUS percutaneous intervention procedures, ultrasound (US)-guided needle insertion is commonly adopted in minimally invasive interventions, such as tissue biopsy, tumor ablation, regional anesthesia [1], etc. As a non-invasive, portable, safe, and cost-effective imaging modality [2], US provides real-time intraoperative imaging of the needle and tissue, thereby minimizing the risk of accidental injury to vessels or critical organs. Despite these advantages, US imaging has an inherent limitation of susceptibility to **noise and artifacts** [1], which can obscure or distort the needle tip position. Furthermore, under the narrow US imaging plane, small structures, such as the needle tip, can **disappear intermittently** when they are obscured by anatomical structures or not co-planar with the US imaging plane [2]. These challenges underscore the necessity for robust and accurate needle tracking to ensure successful insertion procedures under challenging environments.

Prior to the widespread adoption of learning-based needle trackers, traditional methods, such as the statistical filter [3] and Gabor filter [4], achieved somewhat satisfactory performance but were hindered by complex workflows and sensitivity to hyper-parameters, failing to address challenging environmental factors in US imaging. A method based on a discriminative correlation filter (DCF) has been proposed in [5], but correlation filter-based methods can be susceptible to background distraction and image distortion [6]. As a result, this DCF method has also been integrated with an optical tracking system for higher accuracy [7], but the deployment of an optical tracking system is a cumbersome constraint. Recently, deep learning methods based on convolutional neural networks (CNN) and transformers [8], [9] have gained popularity in tracking tasks [10], [11], [12], including US needle tracking [13], [14]. Mwikirize et al. proposed a US needle tracker with a two-step structure based on a fully convolutional network and a region-based CNN [15]. A paradigm utilizing digital subtraction is proposed in [16], which augments tip features to enhance visibility prior to tracking. However, these two methods have non-end-to-end structures

that require multiple steps to localize the needle tip, potentially affecting robustness and accuracy. In addition, since the needle tip is often obscured or distorted under artifacts and noise, some other methods perform needle shaft segmentation before localizing the needle tip [17], [18]. While a segmentation mask of the needle shaft can offer useful cues for determining the axial position of the needle tip, this two-stage workflow may inadvertently introduce accumulative errors and discrepancies. Extra data of segmentation mask is also required to train segmentation models, leading to additional obstacles for model deployment.

As an effective structure, cross-correlation (X-Corr) has been widely adopted in end-to-end learning-based trackers. It measures the similarity between a reference template and a search region by performing convolution with sliding windows. The target location is then obtained from the induced similarity score. Following this diagram, many trackers based on X-Corr have been proposed [10], [19], [20], [21]. However, the existing convolutional X-Corr has a limited modeling range constrained by the kernel size. It cannot learn distant semantic features (e.g. needle shaft) which are important for US needle tracking, since local information can unpredictably become unreliable due to degradation by noise and artifacts.

In addition to the challenging environment with noise and artifacts, the intermittently visible needle tip poses another problem that hinders accurate needle tracking. This issue can be caused by deviation of the US imaging plane or obstruction by anatomical structures [22]. Mwikirize et al. proposed a single-shot needle tracker by integrating historical frames to enhance needle tip features, addressing scenarios where the tip is imperceptible or the shaft is invisible [23]. Although it integrates historical information, its feature pre-enhancement can unexpectedly shift the original latent features, causing potential inherent information loss. Integrating historical needle motion into visual tracking presents another approach, since motion information can serve as an effective non-visual prompt to prevent tracking failure when the needle tip is invisible. A motion prediction module is integrated with a visual tracker in [6], yet it is constrained by its explicit motion prediction that poses challenges on generalizability when encountering motion from unseen domains.

Mamba [24] has recently drawn considerable attention and is being applied in real-world tracking tasks [25], [26], [27]. Based on the structured state space models (SSMs) [28], Mamba has a computationally efficient long-range lossless modeling capability with its selective scan mechanism. Leveraging this, Mamba-based trackers can efficiently model long-range information, such as aggregating a video-level template set [26] or integrating historical frames [25]. It should be noted that a transformer-based tracker can hardly achieve long-range modeling under similar model size and complexity since the self-attention mechanism in transformers has quadratic time complexity and memory requirement with respect to the sequence length [8]. It also requires positional encoding that may not effectively capture lossless long-range dependencies compared to Mamba, which is designed specifically for such tasks. Thus, Mamba usually outperforms transformer-based methods under similar model size and complexity [24]. Since no Mamba-based tracker has yet been proposed for US needle tracking, developing one remains an open research area.

To address the aforementioned challenges of noise, artifacts, and intermittent visibility in US needle tracking, in this work, **MambaXCTrack**, a Mamba-based US needle tracker utilizing SSM cross-correlation (**SSMX-Corr**) and an **implicit motion prompt**, is proposed. To the best of our knowledge, it is the first time a Mamba-based tracker has been adopted in US needle tracking. It is also the first time that cross-correlation is implemented with SSM. Leveraging the long-range modeling capability of SSMs, SSMX-Corr enables global search and long-range modeling of distant semantic features. When the needle tip feature is degraded by noise and artifacts, SSMX-Corr avoids tracking failure by implicitly learning distant semantic features that potentially come from visual cues like the needle shaft, rather than by explicitly performing segmentation on the needle shaft like existing methods [17], [18]. By further integrating the proposed cross-map interleaved scan (CIS), SSMX-Corr enjoys global search without losing local pixel-wise interaction between search and template maps to keep positional inductive bias, while existing convolutional X-Corr models [10], [29] only consider local modeling. To address the intermittent visibility of the needle tip, an implicit low-level motion descriptor is introduced as a non-visual prompt in addition to visual features that can unpredictably become unreliable. Different from [6] that trains an external motion predictor to explicitly predict the future motion, the proposed workflow preprocesses motion to obtain the low-level motion descriptor, which is then implicitly integrated with visual features. This implicit low-level motion integration introduces image-agnostic raw motion and ensures an end-to-end network to enhance training stability and tracking robustness. Extensive evaluations on a dataset of motorized needle insertions in both phantom and animal tissue demonstrate MambaXCTrack's superior performance compared to state-of-the-art methods. The main contributions are fourfold:

- SSMX-Corr improves existing cross-correlation with SSM by globally searching and modeling long-range distant semantic features, thus learning potential distant visual cues. This represents the first effort to adopt Mamba effectively for US needle tracking.
- CIS is proposed to provide SSMX-Corr with local pixel-wise interaction and positional inductive bias in addition to global search to enhance overall tracking performance.
- An implicit low-level motion descriptor is adopted to provide a non-visual prompt, thus leveraging motion information to address the challenge of intermittent needle visibility to achieve robust and consistent tracking.
- The proposed tracker achieves state-of-the-art (SOTA) performance on both phantom and tissue experiments. Further ablation studies show the effectiveness of the proposed modules.

II. Methodology

A. Mambaxctrack

The overview of MambaXCTrack is shown in Fig. 1. A ResNet-50 network [30] with 50 layers is adopted as the backbone. The template map Z and the search map X first go through the backbone, from which the embedded features $z \in \mathbb{R}^{H_z \times W_z \times C}$ and $x \in \mathbb{R}^{H_x \times W_x \times C}$ ($C = 128$) are obtained

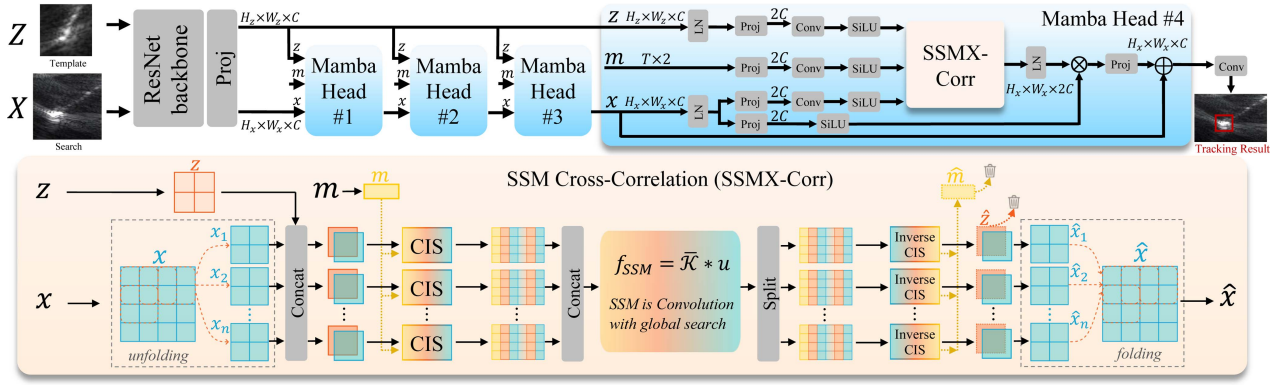

Fig. 1. Structure overview of the proposed MambaXCTrack. The ResNet backbone is cascaded with four Mamba heads. Each Mamba head has the same structure. z and x are the embedding features of template Z and search X.

through a cascaded linear projection. In the following Mamba head, after going through cascaded layer normalization, linear projection (in the dimension of $2C$), convolution (kernel size 3×3, stride 1), and SiLU activation [31], x, z, and the low-level motion descriptor m are transmitted to SSMX-Corr. For each Mamba head, x is serially passed to the next head for a better feature representation, yet z and m are concurrently transmitted by each head to avoid error accumulation on the original template and raw motion.

In SSMX-Corr, x is first *unfolded* with a stride $(\frac{H_z}{2}, \frac{W_z}{2})$ to obtain submaps x_i ($i \in \{1, 2, \ldots, n\}$) in the same size $H_z \times W_z$ with z. The submaps x_i, z, and m are scanned respectively by the proposed CIS, then modeled by SSM after being concatenated. After SSM modeling, the aggregated feature map is scanned inversely to separate $\hat{z}$, $\hat{x}_i$, and $\hat{m}$. $\hat{z}$ and $\hat{m}$ are removed. The stage output $\hat{x}$ is then obtained by *folding* all $\hat{x}_i$. The tracking prediction is then obtained from the final $\hat{x}$ by a convolution prediction head.

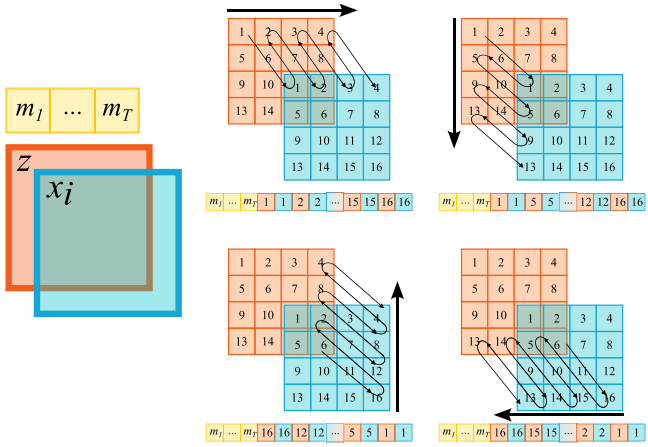

Fig. 2. A demonstration of the cross-map interleaved scan (CIS). CIS is performed in four directions, from which four sequences are regrouped separately. Assuming z and x_i are all in 4×4 with elements numbered $1 \sim 16$, the induced four sequences with m concatenated are shown.

B. Cross-Map Interleaved Scan (CIS)

There exist many scanning paradigms for Mamba [27]. However, there has yet to be a scanning method designed for tracking tasks to enhance the interaction between the template z and search x. The existing cross-correlation-based methods estimate the cross similarity with convolution, which possesses pixel-level interaction and positional inductive bias in nature. To perform SSM-based cross-correlation, although the model enjoys lossless long-range modeling of semantic information, the local pixel-level relationships between z and x can be unexpectedly neglected if they are simply concatenated together. In this work, before SSMX-Corr, the CIS is adopted to enhance pixel-wise interaction between z and x to better adapt to SSM-based cross-correlation. As shown in Fig. 2, CIS receives a template z, a submap x_i, and a motion descriptor m. Since the SSM modeling is unidirectional, four-directional scanning is adopted. For scanning in each direction, z and x_i are scanned pixel-by-pixel alternatively as demonstrated in Fig. 2. Four sequences are then induced, each with the motion descriptor m concatenated at the beginning. By performing CIS, the local pixels from template and search maps are regrouped to be adjacent, allowing the SSMX-Corr to be performed without losing local interaction and positional inductive bias.

C. SSMX-Corr: SSM is Convolution With Global Search

The regrouped sequences from CIS are then processed by SSM Cross-Correlation (SSMX-Corr). The X-Corr in existing trackers [10], [29] calculates the similarity map between template and search maps with convolution. This convolutional X-Corr operation f_{conv} is given by

$$f_{conv}(z, x) = z * x, \tag{1}$$

where $*$ denotes the convolution. Since convolutional X-Corr is a local operation, it only models local pixel-wise interaction yet fails to consider interrelationship between distant features. When local features from the small needle tip area are degraded by severe noise and artifacts, X-Corr can lose tracking easily without seeking hints from distant semantic features. Searching globally for potential semantic visual cues then turns out to be a promising paradigm inspired by segmentation-based needle trackers [17], [18], which obtain the needle's axial direction by segmentation.

To introduce long-range modeling capability into X-Corr, SSMX-Corr is proposed. SSMs are based on a continuous linear time-invariant (LTI) system that can be formulated as a linear ordinary differential equation (ODE) as follows

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}u(t),$$
$$y(t) = \mathbf{C}h(t), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ (state matrix), $\mathbf{B} \in \mathbb{R}^{N\times 1}$ (input matrix), and $\mathbf{C} \in \mathbb{R}^{1\times N}$ (output matrix). These two equations together map the input sequence $u(t) \in \mathbb{R}$ to output $y(t) \in \mathbb{R}$ through linear transformation of the latent states $h(t) \in \mathbb{R}^{N\times 1}$. To integrate the continuous system described in (2) into a digital system for deep learning integration, it needs discretization first. Δ is introduced as the step size at timescale. Equation (2) is then discretized as

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}u_t,$$
$$y_t = \mathbf{C}h_t, \tag{3}$$

where the discritized matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are given by $\bar{\mathbf{A}} = exp(\Delta \cdot \mathbf{A})$ and $\bar{\mathbf{B}} = (\Delta \cdot \mathbf{A})^{-1}(exp(\Delta \cdot \mathbf{A}) - I) \cdot \Delta\mathbf{B}$. During the discretization, the selective scan mechanism [24] is adopted to learn input-dependent parameters $\mathbf{B}, \mathbf{C}, \Delta$ from input $x \in \mathbb{R}^{L\times C}$, where L is the sequence length, C is the dimension. Parameters $\mathbf{B}, \mathbf{C}, \Delta$ are given by $\mathbf{B} = \text{Linear}(x) \in \mathbb{R}^{L\times N}$, $\mathbf{C} = \text{Linear}(x) \in \mathbb{R}^{L\times N}$, $\Delta = \text{Softplus}(\widetilde{\Delta} + \text{Linear}(x)) \in \mathbb{R}^{L\times C}$, where $\widetilde{\Delta}$ is a trainable parameter.

By reformulating (3) [32], the closed form of the output y_k at every time step k can be derived as

$$y_k = \mathbf{C}(\bar{\mathbf{A}})^k\bar{\mathbf{B}}u_0 + \mathbf{C}(\bar{\mathbf{A}})^{k-1}\bar{\mathbf{B}}u_1 + \cdots + \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}u_{k-1} + \bar{\mathbf{B}}u_k. \tag{4}$$

Then y can be formulated as the result of a convolution by extracting the coefficients into the SSM kernel $\bar{\mathcal{K}}$

$$y = \bar{\mathcal{K}} * u,$$
$$\bar{\mathcal{K}} = \left(\mathbf{C}\bar{\mathbf{A}}^j\bar{\mathbf{B}}\right)_{j\in[L]} \in \mathbb{R}^L = \left(\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}\right). \tag{5}$$

In this letter, input u is the scanned aggregation of m, z, x derived from CIS, which is defined by $u = Cat_{i=[1,n]}(CIS(z, x_i, m))$. The proposed SSMX-Corr f_{SSM} can then be written as

$$f_{SSM}(z, x, m) = \bar{\mathcal{K}} * u = \bar{\mathcal{K}} * Cat_{i=[1,n]}\left(CIS(z, x_i, m)\right), \tag{6}$$

which is a convolution between the concatenated map from scanned aggregation $CIS(z, x_i, m)$ and the kernel $\bar{\mathcal{K}}$. Note that $\bar{\mathcal{K}}$ comes from the parameterization of input (z, x, m). Thus, SSMX-Corr can be regarded as a *self-convolution* of z, x, m, yet it enjoys SSM's long-range modeling capabilities. With SSMX-Corr, tip tracking can benefit from *global search* between x and z. Distant semantic features from the needle shaft can be learned to guide tracking. The sequence modeling diagram in SSMX-Corr also facilitates better motion-vision information fusion than existing convolutional X-Corr.

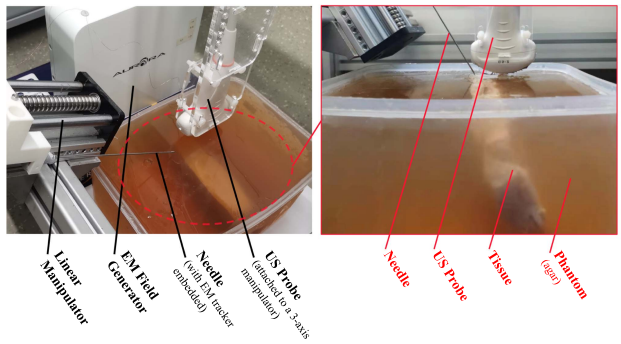

Fig. 3. Experimental setup. The in-plane case (imaging plane aligned parallel with needle trajectory) is shown as an example.

D. Implicit Low-Level Motion Descriptor

There exist some US needle trackers [6] that improve tracking performance by incorporating motion information. However, explicit motion integration by including a motion predictor makes it a non-end-to-end model, posing potential harm to its generalizability and training stability. An implicit low-level motion descriptor is introduced in this letter for robust tracking when the needle tip is intermittently invisible. Given a set of historical bounding boxes $B = \{\beta_1, \beta_2, \ldots, \beta_t\}$ (t is the time index), where $\beta_t = (w_t, h_t, cx_t, cy_t)$ is the bounding box defined by its width w_t, height h_t, and coordinate of the top-left corner (cx_t, cy_t), the low-level motion descriptor m_t is constructed using the local displacement, given by $m_t = (cx_t - cx_{t-1}, cy_t - cy_{t-1}) = (\Delta cx_t, \Delta cy_t)$. The motion descriptor m_t is in pixels units. This low-level local displacement smooths out the distraction from absolute information at the picture level, benefiting the model inferencing and generalizability. To store the historical motion during tracking, a motion queue at a length T is built with a FIFO (first-in-first-out) strategy. Note that for queue storage, the motion sequence is not sampled by time to keep the original motion step. The sequence of m is then formulated as $\{m_1, m_2, \ldots, m_T\}$ in the size of $T \times 2$, which is concatenated with z and x in SSMX-Corr for implicit information fusion.

III. Experiments and Results

A. Experimental Setup and Dataset Collection

The experimental setup is shown in Fig. 3. Ultrasound imaging was performed with a Verasonics Vantage 32 LE US machine with a Mindray C5-2 US probe. This probe has a transmit frequency of 3.5 MHz with 96 transducer elements. During imaging, a plane wave sequence with an imaging depth of 150 mm was utilized, and the sampling frequency was set to 25 MHz. An electromagnetic (EM) localization system (Aurora, NDI Inc.) was adopted to collect the needle tip's position as the ground truth, using an 18-gauge needle with a 5-DoF EM tracker embedded on its tip. The needle was attached to a linear manipulator for motorized insertion. The US probe was installed on a 3-axis manipulator to move the probe for needle visualization when needed. The phantom was made of agar of

3% mass concentration and the tissue was a slice of fresh pork. Silica powder was added to agar to simulate speckles in real anatomical tissue.

During the experiment, the data of the phantom-only experiment and tissue experiment were collected separately regarding three cases, i.e., in-plane (static), in-plane (moving), and out-of-plane [22]. For cases of in-plane (moving) and out-of-plane, the probe moved at the same speed as the horizontal velocity of needle insertion. For in-plane (static), only the needle moved. The needle insertion was performed with three angles (0°, 30°, and 60°) and three velocities (0.4, 1, and 2 mm/s) for each case. For each pair of angle and velocity, at least 6 procedures were conducted. Each needle insertion length is more than 120 mm to ensure a thorough procedure. 108 procedures have been done in total, including 51 in tissue and 57 in phantom, which results in 108 videos (492×856, 30 FPS) with 87330 frames. The pixel resolution of each frame (492×856) corresponds to a field-of-view of 150 mm $\times$ 260 mm. The ground truth of the needle tip position was acquired by the EM tracking system, which has been verified to have a satisfactory RMSE of 0.76 mm in our environment.

To train the model, the videos were first sampled by 10 to get 8733 images to remove redundancy. The dataset was then split into training (6113 samples), validation (873 samples), and testing (1747 samples) sets in the ratio 7:1:2. A video/setup-level data splitting was conducted to ensure images from the same video were not divided into different sets, and videos with different experiment setups were evenly divided into different sets. The program was implemented with PyTorch. All model training and inferencing was performed on a server running Ubuntu 22.04 with two Intel Xeon Platinum 8375 C CPUs and four NVIDIA RTX 4090 GPUs installed. Inferencing only involves one GPU. All models were trained with the same training strategy (350 epochs, batch size 32, AdamW optimizer), and they used the same input resolution of 384×384 for search images and 192×192 for template images. The learning rate was set to 3e-4 with a backbone learning rate of 3e-5, where both of them were dropped by a factor of 10 after 200 epochs. The backbone is trained together with the network but adopts a smaller learning rate. Scaling, blur, and position shifting were adopted as image augmentation. Gaussian noise was added to the historical motion sequence. The augmentations were performed dynamically during the training process.

B. Results

The evaluation was conducted based on three metrics that are commonly adopted in tracking evaluation, namely area under curve (AUC) [35], precision (P) [36], and normalized precision (P_{norm}) [36]. AUC reports the needle tracking success rate by evaluating the area under the Success Plot curve as defined in [35], which offers a comprehensive evaluation of tracking robustness. Higher AUC suggests that the tracker keeps the needle tip tracked for a larger portion of the frames. Beyond success rate by AUC, tracking error is evaluated by P and P_{norm} [36]. P evaluates the accuracy of predicted position measured by Center Location Error (CLE), which is the Euclidean distance between the predicted and ground truth positions. P is generated by calculating the percentage of frames where the CLE is less than a specific threshold, which is set to a conventional value of 20pixels. P_{norm} adjusts P to account for the scale of the object by normalizing CLE with the size of the ground-truth bounding box, providing a more consistent measure across varying object sizes. The ground-truth bounding box is defined based on the needle tip region that is visible in the ultrasound imaging. Higher P and P_{norm} can be interpreted as needle tip tracking with higher accuracy. In addition to these three metrics, the average tracking errors and standard deviations in millimeters were also reported to provide a better comparison in terms of physical metrics. Several state-of-the-art natural object trackers were retrained on the proposed dataset and used for comparison, including state-of-the-art transformer-based tracker MixFormerV2 [11], tracker with motion prompts (SwinTrack [34]), tracker with aggregated historical templates (STMTrack [33]), classical Siamese tracker SiamRPN++ [10] and its improved variants [19], [20], [21]. A US-based needle tracker by Yan et al. [6] was also used as a comparison.

As the results show in Tables I and II, the proposed MambaXCTrack achieves SOTA performance in almost all metrics of all evaluations while maintaining satisfactory inference speed. In the phantom environment with less background noise and more stable needle appearance, while most methods offer satisfactory tracking performance in the out-of-plane cases, a few trackers, including MambaXCTrack, distinguish themselves with superior performance in the in-plane cases. SwinTrack and MixFormerV2 were ranked second in the in-plane (static) and in-plane (moving) cases, respectively, while MambaXCTrack outperforms both to rank first in both cases. This achievement can be attributed to the higher modeling efficiency of the Mamba-based structure compared with the transformer architecture adopted in both SwinTrack and MixFormerV2. MambaXCTrack also achieves a 0.22 ± 0.14 mm tracking error and standard deviation (SD). This means that the estimated tip position can be maintained within the 18-gauge needle tip (with a cross-sectional radius of 0.625 mm) in almost all frames.

In the tissue environment that is much more challenging than the phantom, MambaXCTrack consistently achieves the best and second-best performance in nearly all evaluations. MambaXCTrack significantly outperforms Yan et.al. [6] by 125.5%, 104.8%, and 68.1% in AUC for all three cases. While both MambaXCTrack and the work in [6] integrate motion information to address the intermittent invisibility issues, the superior performance by MambaXCTrack is mainly attributed to longer modeling distance of SSMX-Corr than the transformer in [6] and better generalizability of an implicit low-level motion descriptor than an explicit motion predictor in [6]. As shown in Tab. II, MambaXCTrack achieves a tracking error of 0.49 ± 1.01 mm in tissue experiments, outperforming the second-best MixFormerV2 by 64.2% and 33.1%. The offset of the predicted position relative to the tip ground truth is kept within a distance of the needle cross-sectional radius, better than MixFormerV2, which has an average error of 2.2 times the needle radius.

In the average evaluation across both phantom and tissue, our tracker demonstrates superior performance across all metrics.

TABLE I
EVALUATION RESULTS ON MOTORIZED INSERTION IN PHANTOM AND TISSUE

Method	Phantom									Tissue									Mean		
	In-plane (static)			In-plane (moving)			Out-of-plane			In-plane (static)			In-plane (moving)			Out-of-plane					
	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P
SiamRPN++ [10]	50.7	68.5	72.9	43.1	41.4	68.8	64.5	79.0	84.1	29.9	30.4	50.1	44.3	32.8	81.0	66.0	85.9	92.0	48.8	55.7	73.6
SiamCAR [20]	60.0	79.6	85.5	51.7	53.0	85.4	62.1	80.5	83.9	42.8	51.2	75.1	46.4	34.1	80.2	71.3	90.3	98.1	55.5	65.0	84.7
SiamBAN [19]	70.5	86.6	93.5	66.0	79.3	100.0	74.4	96.0	100.0	53.2	62.8	79.4	56.4	63.9	81.0	66.6	83.1	87.9	64.6	78.6	90.6
SiamAttn [21]	62.3	87.1	90.4	72.5	95.0	99.9	74.7	97.8	100.0	40.4	53.5	64.6	51.6	64.1	90.4	69.2	93.1	98.3	62.1	82.0	90.2
Yan et.al. [6]	46.1	81.0	93.6	41.7	66.7	88.9	54.8	65.1	100.0	24.3	36.3	58.2	29.0	39.6	64.1	45.4	83.1	99.9	40.1	62.9	84.3
STMTrack [33]	38.1	90.2	98.8	42.4	68.5	98.1	42.3	97.2	100.0	42.8	61.9	79.5	37.8	50.2	72.8	51.7	92.3	97.6	42.4	76.8	91.7
SwinTrack [34]	73.8	92.6	99.8	65.5	76.6	100.0	73.2	96.0	100.0	52.2	61.2	79.7	56.0	64.5	81.2	75.4	93.0	98.5	66.1	80.6	93.6
MixFormerV2 [11]	71.6	93.4	97.6	73.6	96.0	99.9	78.4	97.9	100.0	44.2	56.3	66.5	60.9	75.2	92.6	74.5	94.0	99.5	67.0	85.6	92.5
MambaXCTrack	79.2	96.2	99.9	76.0	96.8	100.0	76.3	98.1	100.0	54.8	66.1	83.0	59.4	73.6	92.6	76.4	94.4	99.6	70.8	87.9	95.9

The methods with the best and the second best performance are noted in red and cyan color. All three metrics are reported in percentage (%).

TABLE II
AVERAGE ERRORS AND STANDARD DEVIATIONS (MM) OF THE TRACKING, WHICH ARE COMPUTED OVER THE TESTING SET WITH 1747 SAMPLES

Method	Phantom	Tissue	Mean	FPS
SiamRPN++ [10]	4.11±3.05	3.21±2.58	3.70±2.83	27.9
SiamCAR [20]	2.26±1.55	2.23±1.37	2.24±1.47	48.3
SiamBAN [19]	0.70±0.54	2.09±1.55	1.33±1.01	30.7
SiamAttn [21]	0.63±0.67	1.62±1.33	1.08±0.98	33.2
Yan et.al. [6]	0.98±0.94	2.83±2.21	1.84±1.53	24.0
STMTrack [33]	0.41±0.27	2.32±1.52	1.28±0.86	22.1
SwinTrack [34]	0.37±0.17	1.54±1.53	0.90±0.81	59.5
MixFormerV2 [11]	0.26±0.25	1.37±1.51	0.76±0.84	81.0
MambaXCTrack	0.22±0.14	0.49±1.01	0.34±0.55	34.9

Inference speed is reported in FPS.

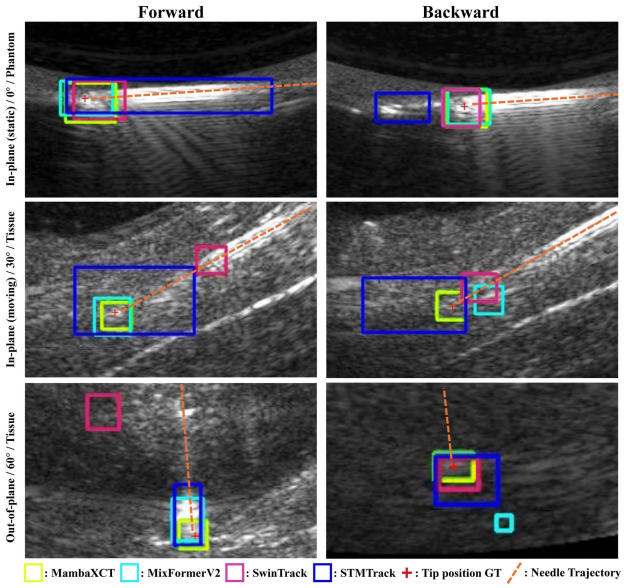

Fig. 4. Tracking demonstration of MambaXCTrack against three SOTA trackers in three scenarios. The view is zoomed in. The left column is the forward procedure, while the right is backward. The ground-truth tip position and needle trajectory are marked. See more examples in the supplementary video.

Specifically, it achieves an average error of 0.34 ± 0.55 mm, indicating that, in most frames, the tracker maintains accuracy within approximately one radius (0.635 mm), with error variations staying below 0.55 mm. This represents a substantial improvement over the second-best method, MixFormerV2, with reductions of 55.3% in average error and 34.5% in standard deviation.

A tracking demonstration is shown in Fig. 4 including three examples[1] that demonstrate different challenges in US needle tracking. The first example of In-plane (static) / 0° / Phantom in the first row shows a scenario with interference from needle traces. MambaXCTrack overcomes the interference from needle traces, which arise from the hollow regions created by needle retraction and have a very similar appearance to the needle tip. This interference has led to STMTrack's failures, whereas our tracker maintains successful tracking. For procedures done within the tissue, the distraction from noise and artifacts becomes more severe, and the needle visibility is degraded. The second example of In-plane (moving) / 30° / Tissue demonstrates this distraction and degradation, where the background is filled with noise from high-intensity speckles caused by variations in acoustic impedance. The backward procedure of this example also shows the intermittent invisible needle tip, since the imaging plane and needle tend to be misaligned in this dynamic case, where both probe and needle are moving. MambaXCTrack is the only one among the four that maintains successful tracking even when the tip is intermittently invisible. A similar case of tracking the intermittently disappearing needle is also demonstrated in the backward of out-of-plane / 60° / Tissue, where MixFormerV2 lost tracking but MambaXCTrack was able to keep the target tracked.

Our experiments show that MambaXCTrack runs at 34.9 FPS, which meets the real-time requirement for ultrasound needle tracking despite being slower than the MixFormerV2 and SwinTrack. However, our method consistently outperforms these methods in terms of needle tracking accuracy by leveraging pixel-level global modeling with SSM, which efficiently extracts rich semantic cues from neighboring regions. The linear complexity of SSM enables this pixel-level global operation, which can hardly be achieved by a transformer-based method without a significant increase in computational complexity. Therefore, MambaXCTrack achieves an appropriate balance between inference speed and accuracy, as clinical applications demand reliability more over speed gains when real-time thresholds can be met.

[1] Additional tracking demonstrations are provided in the supplementary video.

TABLE III
ABLATION STUDIES OF THE BASELINE MODEL (MAMBAXCTRACK) AGAINST SEVEN VARIATIONS

Method	Phantom			Tissue			Mean		
	AUC	P_{norm}	P	AUC	P_{norm}	P	AUC	P_{norm}	P
Baseline: MambaXCTrack	77.3	96.9	99.9	63.3	77.7	91.2	70.8	87.9	95.9
v_1: w/ CIS $[z, m, x_i]$	77.5 (+0.2)	96.6 (–0.3)	99.1 (–0.8)	62.8 (–0.5)	78.7 (+1.0)	89.1 (–2.1)	70.7 (–0.1)	88.2 (+0.3)	94.4 (–1.5)
v_2: w/o CIS, simply cat $[m, z, x]$	76.7 (–0.6)	97.1 (+0.2)	99.1 (–0.8)	61.6 (–1.7)	75.9 (–1.8)	87.9 (–3.3)	69.7 (–1.1)	86.8 (–1.1)	93.9 (–2.0)
v_3: w/o SSMX-Corr, w/ ConvX-Corr	73.4 (–3.9)	94.4 (–2.5)	97.2 (–2.7)	57.6 (–5.7)	73.8 (–3.9)	85.4 (–5.8)	66.1 (–4.7)	84.8 (–3.1)	91.7 (–4.2)
v_4: w/o m	71.9 (–5.4)	92.7 (–4.2)	95.0 (–4.9)	58.6 (–4.7)	76.5 (–1.2)	87.4 (–3.8)	65.7 (–5.1)	85.1 (–2.8)	91.4 (–4.5)
v_5: w/ m, but raw motion	75.5 (–1.8)	94.8 (–2.1)	97.2 (–2.7)	61.7 (–1.6)	78.1 (+0.4)	87.9 (–3.3)	69.0 (–1.8)	87.0 (–0.9)	92.9 (–3.0)
v_6: w/ m, $T = 120$	77.2 (–0.1)	96.9 (–0.0)	99.9 (–0.0)	61.9 (–1.4)	77.1 (–0.6)	89.7 (–1.5)	70.0 (–0.8)	87.7 (–0.2)	95.1 (–0.8)
v_7: w/ m, $T = 30$	74.5 (–2.8)	94.6 (–2.3)	97.4 (–2.5)	62.7 (–0.6)	79.8 (+2.1)	91.6 (+0.4)	69.0 (–1.8)	87.7 (–0.2)	94.7 (–1.2)

The results of the phantom and tissue experiments are reported as mean values of the three insertion techniques.

C. Ablation Study

Extensive ablation studies were conducted regarding the baseline model MambaXCTrack against seven variations v_1 to v_7, with their configurations and results shown in Table III. The first three were adopted to test the proposed SSMX-Corr and CIS. CIS is kept in v_1 but m is concatenated between z and x_i, different from the baseline ($[m, z, x_i]$), to evaluate different orders of information fusion. CIS is removed in v_2 and three feature maps are simply concatenated together into a sequence $[m, z, x]$. CIS is also removed in v_3 and SSMX-Corr is replaced by a convolutional X-Corr [10]. The results of these variations show the effectiveness of CIS and SSMX-Corr. Performance degradation in v_1 demonstrates that the tracking benefits more from motion information by concatenating m at the beginning of z and x_i. With CIS removed (v_2), performance degradation is observed in almost all evaluations, showing that tracking can be improved by building pixel-wise interaction between z and x. Performance drop is also observed in v_3, especially for the tissue experiments, where noise distraction is much more severe. It demonstrates that tracking under challenging environments is greatly improved by learning potential distant visual cues utilizing SSMX-Corr's long-range modeling capability.

Ablation regarding the proposed low-level motion descriptor was also carried out, namely v_4 to v_7. The motion descriptor is removed in v_4. v_5 keeps the motion descriptor but uses raw motion without preprocessing it like the baseline. Motion descriptors that cover different time durations are tested in v_6 (longer time, $T = 120$) and v_7 (shorter time, $T = 30$). A considerable degradation is reported in v_4, demonstrating that tracking is greatly enhanced by learning from historical low-level motion descriptors. Accuracy drops are also reported in almost all metrics of v_5, showing the improvement brought by extracting low-level descriptors from raw motion. The results of v_6 and v_7 demonstrate the length of baseline ($T = 60$), which covers 2 seconds, is a more suitable time span for motion descriptors in US needle tracking. This is likely because moderate-duration motion provides sufficient cues while avoiding interference from long-term noise.

D. Discussion

MambaXCTrack outperforms a series of SOTA trackers including CNN-based trackers and transformer-based trackers. Its superior performance can be attributed to two main factors. Firstly, the SSMX-Corr with CIS effectively addresses the unique domain characteristics of US images, i.e., *the local features from the target can become unpredictably unreliable, but searching for semantic cues over long distances significantly enhances tracking accuracy.* Studies using needle shaft segmentation for guidance [17], [18] suggest that while the needle tip may occasionally become obscured, the larger needle shaft remains partially visible, providing directional information. Instead of direct segmentation, SSM is employed to endow the model with long-range modeling capabilities, allowing it to learn from distant semantic features including the needle shaft. The existing ConvX-Corr adopted in SiamRPN++ [10] is a local pixel-wise operation that convolves x with kernel z. While it is considered effective for open-world tasks, its limited search range proves inadequate for US needle tracking, where images are often degraded by noise and artifacts, rendering local fine features unreliable. With SSMX-Corr, MambaXCTrack overcomes these challenges by identifying long-range semantic cues. By further integrating CIS, SSMX-Corr retains the advantages of ConvX-Corr, gaining positional inductive bias from this convolution-like operation.

The second aspect is the implicit low-level motion descriptor. *When the needle temporarily disappears, historical motion becomes more critical than visual information.* In such cases, relying on motion information rather than vision is a more effective strategy. Ablation studies demonstrate that integrating historical low-level motion descriptors with an appropriate time duration enhances tracking. SwinTrack [34] also incorporates motion information yet leads to worse performance than ours. This is partly because SwinTrack directly integrates raw motion sequences without calculating relative displacement like MambaXCTrack, which hinders the generalization. Additionally, its feature fusion is less efficient than our SSM, which is inherently suited for long-sequence modeling. STMTrack [33] adopts a template bank to store the historical frames, focusing on historical visual information rather than motion information. However, it performs even worse than SwinTrack [34]. This might be because of error accumulation due to aggregating historical US images, which contain severe noise and artifacts. Yet historical motion suffers less from error accumulation thanks to its simpler distribution and characteristic that does not contain visual features. In challenging environments, leveraging

simpler motion information is more effective than integrating visual features. However, while the tracker proposed by Yan et al. [6] also utilizes motion for ultrasound needle tracking, its performance is unsatisfactory. This suboptimal performance is likely attributed to its explicit motion prediction structure, which prevents the model from being fully end-to-end, resulting in error propagation and system instability. Directly predicting future motion can also cause degraded generalizability when encountering unseen videos, which contain motion that can be hard to be adapted by a simple motion predictor. On the contrary, MambaXCTrack leverages implicit motion integration to ensure an end-to-end structure with better generalization.

IV. Conclusion

In this letter, a US needle tracker MambaXCTrack with SSMX-Corr, CIS, and a low-level motion descriptor has been introduced for US needle tracking under challenging US imaging environments with noise, artifacts, and intermittent tip visibility. To the best of our knowledge, it is the first Mamba-based tracker for US needle tracking. Experiments and ablation studies show its effectiveness. However, MambaXCTrack requires a high-end GPU to achieve real-time inference, hindering deployment on low-end hardware. In future work, improvements will be made for better efficiency, including network pruning, model distillation, parallel optimization, etc. More experiments on manual needle insertion will also be conducted. The code implementation of the proposed method will be made publicly available in future work once ongoing developments are completed.

Acknowledgment

The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

[1] S. Liu et al., "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.
[2] M. D. Richardson, R. Ray, and R. R. Richardson, "Imaging modalities: Advantages and disadvantages," *Atlas Acquired Cardiovasc. Dis. Imag. Child.*, vol. 1, pp. 1–4, 2017.
[3] K. Mathiassen, D. Dall'Alba, R. Muradore, P. Fiorini, and O. J. Elle, "Robust real-time needle tracking in 2-D ultrasound images using statistical filtering," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 3, pp. 966–978, May 2017.
[4] M. Kaya, E. Senel, A. Ahmad, O. Orhan, and O. Bebek, "Real-time needle tip localization in 2D ultrasound images for robotic biopsies," in *Proc. Int. Conf. Adv. Robot.*, 2015, pp. 47–52.
[5] C. Shen et al., "Discriminative correlation filter network for robust landmark tracking in ultrasound guided intervention," in *Proc. Med. Image Comput. Comput. Assist. Intervention–MICCAI 2019*, 2019, pp. 646–654.
[6] W. Yan et al., "Learning-based needle tip tracking in 2D ultrasound by fusing visual tracking and motion prediction," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102847.
[7] H. Che et al., "Improving needle tip tracking and detection in ultrasound-based navigation system using deep learning-enabled approach," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 5, pp. 2930–2942, May 2024.
[8] A. Vaswani, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[9] Y. Zhang, P. Zheng, W. Yan, C. Fang, and S. S. Cheng, "A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 11125–11136.
[10] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4282–4291.
[11] Y. Cui et al., "MixFormerV2: Efficient fully transformer tracking," *in Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 58736–58751.
[12] Y. Zhang et al., "Motion-guided dual-camera tracker for endoscope tracking and motion analysis in a mechanical gastric simulator," 2024, *arXiv:2403.05146*.
[13] W. Yan, Q. Ding, J. Chen, K. Yan, R. S. -Y. Tang, and S. S. Cheng, "Visual tracking of needle tip in 2D ultrasound based on global features in a siamese architecture," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 4782–4788.
[14] W. Yan, R. Shing-Yan Tang, and S. S. Cheng, "Task-oriented network design for visual tracking and motion filtering of needle tip under 2D ultrasound," *IEEE Trans. Med. Imag.*, vol. 44, no. 4, pp. 1735–1749, Apr. 2025.
[15] C. Mwikirize et al., "Convolution neural networks for real-time needle detection and localization in 2D ultrasound," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, pp. 647–657, 2018.
[16] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Learning needle tip localization from digital subtraction in 2D ultrasound," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, pp. 1017–1026, 2019.
[17] A. M. Wijata, B. Pyciński, and J. Nalepa, "A needle in a (medical) haystack: Detecting a biopsy needle in ultrasound images using vision transformers," in *Proc. IEEE Int. Conf. Image Process.*, 2024, pp. 3017–3023.
[18] X. Hui et al., "Ultrasound-guided needle tracking with deep learning: A novel approach with photoacoustic ground truth," *Photoacoustics*, vol. 34, 2023, Art. no. 100575.
[19] Z. Chen et al., "SiamBAN: Target-aware tracking with siamese box adaptive network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5158–5173, Apr. 2023.
[20] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6269–6277.
[21] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, "Deformable siamese attention networks for visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6728–6737.
[22] A. Kimbowa et al., "Advancements in needle visualization enhancement and localization methods in ultrasound: A literature review," *Artif. Intell. Surg.*, vol. 4, no. 3, pp. 149–169, 2024.
[23] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Single shot needle tip localization in 2D ultrasound," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 637–645.
[24] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
[25] Q. Wang et al., "TrackingMamba: Visual state space model for object tracking," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 16744–16754, 2024.
[26] S. Lai et al., "MambaVT: Spatio-temporal contextual modeling for robust RGB-T tracking," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.
[27] H. Zhang et al., "A survey on visual mamba," *Appl. Sci.*, vol. 14, no. 13, 2024, Art. no. 5683.
[28] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.
[29] L. Bertinetto et al., "Fully-convolutional siamese networks for object tracking," in *Proc. Computer Vis.–ECCV 2016 Workshops*, Amsterdam, The Netherlands, 2016, pp. 850–865.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[31] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, 2018.
[32] A. Gu et al., "Combining recurrent, convolutional, and continuous-time models with linear state space layers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 572–585.
[33] Z. Fu, Q. Liu, Z. Fu, and Y. Wang, "STMTrack: Template-free visual tracking with space-time memory networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13774–13783.
[34] L. Lin et al., "Swintrack: A simple and strong baseline for transformer tracking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 16743–16754.
[35] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2411–2418.
[36] M. Muller et al., "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 300–317.