Task-oriented network design for visual tracking and motion filtering of needle tip under 2D ultrasound

Wanquan Yan, Raymond Shing-Yan Tang, and Shing Shin Cheng

***Abstract*— Needle tip tracking under ultrasound (US) imaging is critical for accurate lesion targeting in US-guided percutaneous procedures. While most state-of-the-art trackers have relied on complex network architecture for enhanced performance, the compromised computational efficiency prevents their real-time implementation. Pure visual trackers are also limited in addressing the drift errors caused by temporary needle tip disappearance. In this paper, a compact, task-oriented visual tracker, consisting of an appearance adaptation module and a distractor suppression module, is first designed before it is integrated with a motion filter, namely TransKalman, that leverages the Transformer network for Kalman filter gain estimation. The ablation study shows that the mean tracking success rate (i.e. error <3mm in 95% video frames) of the visual tracker increases by 25% compared with its baseline model. The complete tracking system, integrating the visual tracker and TransKalman, outperforms other existing trackers by at least 5.1% in success rate and 47% in tracking speed during manual needle manipulation experiments in ex-vivo tissue. The proposed real-time tracking system will potentially be integrated in both manual and robotic procedures to reduce operator dependence and improve targeting accuracy during needle-based diagnostic and therapeutic procedures.**

***Index Terms*— Ultrasound imaging, needle tracking, motion filtering**

I. INTRODUCTION

NEEDLE-based procedures are commonly performed in minimally invasive surgeries (MIS), such as biopsy, drug delivery and local target therapy, to reduce postoperative pain and shorten recovery time [1]. In these procedures, it is important to accurately track the needle tip's position to guide it to the desired location while avoiding vital structures such as blood vessels. The trackers are usually developed based on intraoperative images, including magnetic resonance imaging (MRI), computed tomography (CT), fluoroscopy, and ultrasound (US). In these imaging modalities, US is the most commonly used due to its ionizing radiation-free, real-time, and portable characteristics. However, US images inherently have poor image quality and low contrast. The needle tip often has low visibility in US images, and sometimes may even disappear due to occlusions and misalignment between the US probe and the needle axis. These factors make the US image-based needle tip tracking a highly challenging task.

The challenges in US image-based needle tip tracking can be roughly summarized into three categories: 1) Dramatic appearance changes of the needle tip throughout the course of the needle motion. Most state-of-the-art visual trackers achieve object tracking by comparing the similarity between the target template and the candidate regions in the tracked frame. However, the ever changing appearance of the needle tip makes its template outdated [2], which undermines the tracking accuracy and leads to the loss of target. 2) Strong noise, bright speckles and artifacts in the background. The low contrast of the US images causes low visibility of the needle tip and unclear boundaries between the target and backgrounds [3]. The strong distractors have an appearance similar to that of the needle tip around the target. 3) Temporary disappearance of the needle tip. The needle tip may completely disappear from the current image due to occlusions caused by anatomical structures, misalignment between the US imaging plane and the needle axis [4], and poor transducer-skin contact. Pure visual trackers would fail to obtain any useful information about the target during such disappearance of the needle tip, and thus sudden drift is highly likely to occur.

In this work, we introduce a new US image-based needle tip tracking system, combining a visual tracker and a motion filter, as shown in Fig. 1. The contributions of our work can be summarized as follows:

- Designing a task-oriented visual tracking network that contains a feature extraction module with attention neural network, an appearance adaptation module, and a distractor suppression module. The feature extraction module and attention neural network extract rich and global semantic features from the input images. The appearance adaptation module efficiently builds and updates a template bank to save diverse appearance of the needle tip to adapt to the target appearance changes. The distractor suppression module suppresses distractors and highlights the real target features.

Research reported in this work was supported in part by Research Grants Council (RGC) of Hong Kong (T45-401/22-N, CUHK 14217822, and CUHK 14207823) and in part by Innovation and Technology Commission of Hong Kong (ITS/234/21, ITS/233/21, ITS/235/22, and Multi-scale Medical Robotics Center). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

Wanquan Yan is currently with the State Grid XinJiang Electric Power Co., Ltd. Supervoltage Branch. He was previously with the Department of Mechanical and Automation Engineering and T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong. 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. 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, and Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Hong Kong. Corresponding author's email:sscheng@cuhk.edu.hk

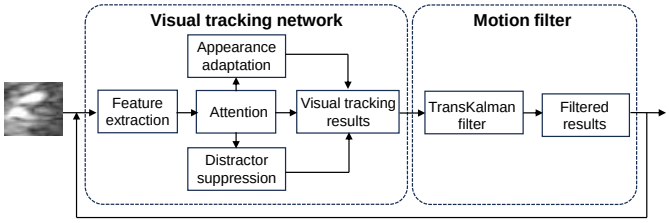

Fig. 1: Overview of the proposed task-oriented US image-based needle tip tracking system, whereby a motion filtering module is cascaded to a visual tracking network to postprocess the visual tracking results.

- Introducing TransKalman, a Transformer neural network-based motion filter that follows the design principles of the Kalman filter. The Transformer network is used to estimate the gain matrix used in the Kalman filter. The filter then smooths the position data output by the visual tracker when the needle tip can be clearly found, suppresses the unwanted sudden drifts, and provides a reasonable estimation of the target's state when the needle tip disappears.

Compared with existing works, most of which achieve improved needle tracking performance by stacking sophisticated architectures to build powerful feature extraction networks, this is the first work that designs compact, independent modules in a visual tracking network to address effectively and efficiently the specific challenges of the needle tip appearance changes and strong distractors during US-based needle tracking, thus achieving both high tracking robustness and computational efficiency. Compared with the motion filter in our previous work [14], which uses the Transformer network to directly estimate the final tracking results and is thus prone to drift error under noisy input data, the Transformer network and Kalman filter are combined in this work to form a robust motion filter that explores the inter-frame motion law in both linear and non-linear motions from a long input sequence, thus more robustly suppressing disturbances from unreliable visual tracking results.

The rest of the work is divided into several sections: In Section II, related work in pure US image-based and motion information-integrated target tracking methods are introduced. The proposed visual tracker and motion filter are described in Section III and IV. Experimental results and discussions are provided in Section V before a concise conclusion, and some future work is offered in Section VI.

II. Related work

A. Visual tracking

In the early work, the needle tip in US images is mainly tracked by using different image filters. For example, the Gaussian and median filters are usually first used to denoise US images; morphology operations, including dilation and erosion, are then applied to enhance the needle axis [5]; next, thresholding filters are used to binarize US images. Finally, line detection methods, such as random sample consensus (RANSAC) [6], or circle detectors, such as the Hough circle transform have also been used to detect the needle axis [7] or the needle tip, respectively. In other works, the Gabor filter is used to highlight pixels along the needle insertion direction [8] to improve the accuracy of the needle axis identification. Many of these methods perform segmentation of the needle axis before locating its tip. However, due to the poor quality of US images, the needle axis is usually not continuous and only partially visible [4], leading to failure of needle axis segmentation. Furthermore, the needle tip sometimes looks like an isolated part from the needle axis and is highly likely to be regarded as noise [9]. Therefore, the performance of these methods is generally poor.

Machine learning-based trackers have become the state-of-the-art trackers for US image-based target tracking in recent years. Raw pixel values are used in the template matching method proposed in [2] to directly find the most similar regions in the input frame to locate the needle tip. Some handcrafted features are also designed to locate the target. The harr-like featurese [10] and Log-Gabor features [8] are use to take the advantage of online classifier and Adaboost respectively, to track the target. Although some improvements have been obtained, the design of handcrafted features requires specific domain knowledge, and trackers which are only trained on these manually designed features lack the adaptability to track the needle tip in complex noisy environments.

Deep features, which can be automatically learned by deep neural networks, can describe the target from different aspects. They have shown great potential to significantly improve the tracking performance in clinically realistic scenarios. Some works based on UNet, convolutional neural network (CNN) and fully convolutional network (FCN) have been proposed to accurately segment the needle axis before other conventional methods are deployed to locate the needle tip [4], [11]. A spatial and channel "Squeeze and Excitation" was proposed in [12] to learn spatial information of the target and locate the needle tip. A bootstrap resampling method was proposed in [11] to facilitate the network training process to obtain better tracking performance. However, these needle axis segmentation-based methods still cannot correctly identify the entire needle axis when it is only partially visible in US images, leading to the failure of tracking. There are also some works that proposed different neural networks to directly locate the needle tip or landmarks in US images [9], [13]. Trackers in these works are still developed on the basis of basic and general neural networks, which are not elaborately designed according to the special characteristics of US images.

In our previous work [14], a visual tracking module was designed based on the encoder and decoder of the Transformer neural network to reinforce and propagate the appearance features of the target. Two different sets of Gaussian-shaped masks were designed to highlight the target features and suppress background distractors, respectively. A discriminative correlation filter (DCF) and a computationally expensive bounding box regression head were applied to obtain the final tracking box. While the tracking system achieves highly satisfactory performance, the complexity of the network architecture increases the computational cost, leading to the inability to offer real-time tracking performance. Besides, the

template update strategy highly relies on the assumption that the next target appearance is similar to its current appearance, leading to the templates being not sufficiently diverse to account for drastic appearance change.

The visual tracking network proposed in this work fully takes advantage of the advanced Siamese neural network and attention modules to design a compact feature extraction architecture. An online appearance adaptation module and a distractor suppression module are specifically designed to efficiently handle the problems of dramatic appearance changes of the needle tip and strong background distractors. Instead of segmenting the needle axis before detecting the needle tip, our proposed visual tracker directly locates the needle tip from the input US images, and the discontinuity of the needle axis does not affect the performance of our visual tracker.

B. Motion filtering

In previous studies of US image-based target tracking, the Kalman filter has often been combined with image filters to remove noise in the visual tracking results. It improves the tracking performance by integrating the motion information of the target into the tracking process according to the predefined motion models. In [15], the Kalman filter was used to filter the noise and cope with outliers to improve the tracking accuracy. In [16], the Kalman filter was applied to obtain the trajectory of the target to track the target when occlusion occurs. In [17], visual tracking results obtained from two different algorithms were fused by the Kalman filter to filter noise. The work in [10] used the Sage-Husa adaptive Kalman filter to adapt to the unstable measurement noises in the visual tracking results. In these works, the insertion velocity of the needle tip was assumed to be constant, and thus a linear model was used to describe the needle's movement. However, this does not describe needle motion under manual handling. Furthermore, the predefined values of the state and measurement noise matrices are crucial in the successful application of Kalman filter. In US image-based needle tip tracking, the statistic characteristics of the measurement noise are not stable under different imaging quality and thus difficult to be accurately estimated before tracking [10].

In learning-based trackers, limited works have attempted to explore the target's motion information between successive frames to improve the tracking performance. Most of them learn the temporal-spatio features of the target from several input images [18], [19]. The works take several templates, in which the target can be clearly found, as input and concatenate them to extract the spatial and temporal features independently. There are no fixed temporal patterns in the selection of templates in the time domain, as they are sampled mainly considering the image quality. Therefore, the sequential information of across the frames is undermined, and the network cannot learn rich motion information of the target. Trackers proposed in [3], [20] appended a long short-term memory (LSTM) recurrent neural network to a CNN to extract temporal information. The input to them was a sequence of consecutive plurality of frames. Due to the huge computational cost of LSTM, the number of frames contained in the input sequence cannot be large (less than five in most trackers). As a result, the neural network may not be able to find the implicit motion laws of the target from the noisy visual tracking results.

In our previous work [14], a motion filter based on a neural network was proposed to filter the position of the needle tip output by a visual tracking module. It assumes the position of the needle tip in the next time step can be estimated through a Gaussian distribution based on its current position. The neural network is then used to estimate the parameters of the conditional distribution. As the motion filtering only relies on the input historical position data, it can be subject to deteriorated performance when the input data are noisy.

There are other works that track the needle tip by analysing the movement of the needle caused by the needle insertion. The spectral analysis is used in [21], [22] to detect the needle tip in US images by analyzing the small movement of the needle tip caused by the natural hand tremors. This method has no requirement on the visibility of the target in US images. However, the minute tremor induced on the needle tip will also be transferred to the tissues around it, and the oscillated biological tissues will interfere with the phase analysis Furthermore, this method can only be used in situations when the US probe is stably fixed [23]. If the US probe is manually held, the slight movement of the probe will adversely impact image-based phase analysis, thus limiting the clinical feasibility of this method. The optical flow has also been used to track the needle tip [24] by analysing the dynamic intensity changes of pixels in US images caused by the movement of the needle. This method only performs well when the intensity change of pixels associated with the needle motion transitions smoothly. Furthermore, this method assumes that the neighboring pixels in an image always belong to the same features and move together [9]. These limitation and assumption restrict the application of optical flow-based method in the field of US image-based needle tip tracking, especially during manual needle manipulation.

Unlike the abovementioned methods, our motion filter follows the design principle of the Kalman filter. It calculates the gain matrix used in Kalman filer by using a Transformer network which does not require knowing the distribution of the state and measurement noises in advance. While the LSTM and Transformer are both suitable for processing time sequential data, the parallel processing ability of the Transformer allows input of large sequential data, facilitating the learning of inner relationships between the parallel input data. The long input sequence also allows a new movement model of the needle to be robustly learned, enabling our motion filter to handle the problem of model mismatch to a significant extent.

III. Visual Tracking

An overview of our visual tracking neural network is shown in Fig. 2. It consists of a feature extraction module, an attention neural network, an appearance adaptation module, and a distractor suppression module.

1) Feature extraction module: The feature extraction module consists of a modified version of the ResNet-50 network [25], which has shown great feature extraction ability, and a

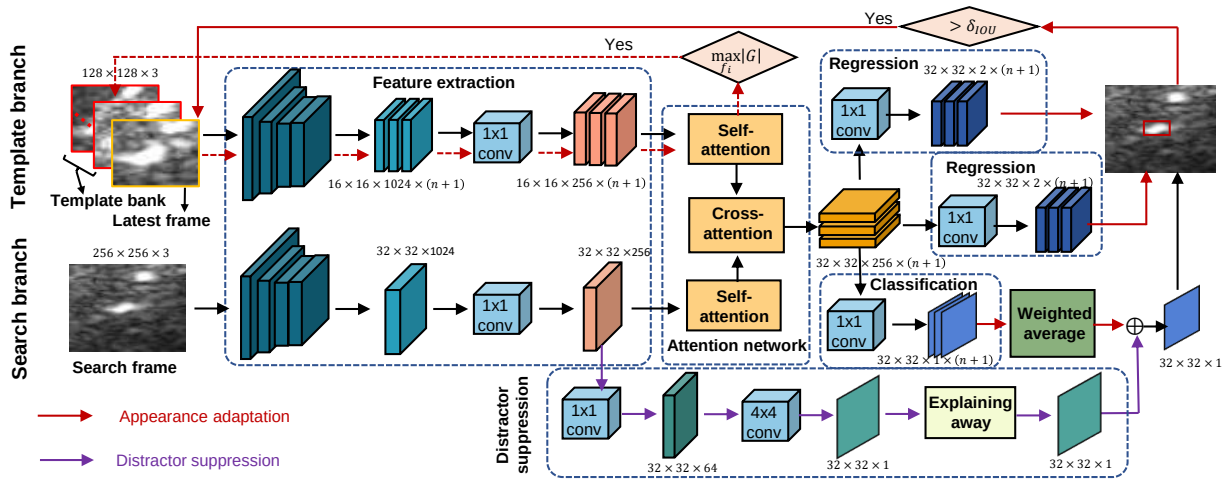

Fig. 2: Architecture of the visual tracking network. The red and purple arrows indicate appearance adaptation and distractor suppression modules, respectively. The dashed red arrows represents that the update of the template bank was conducted when at the end of tracking, and tracked results are used in the updating process.

1×1 convolutional block. Specifically, the last stage of the ResNet-50 is removed and the output of the fourth stage is used in the following computation. The convolutional stride of the downsampling layer in the fourth stage is changed from 2 to 1 to obtain high-resolution features. The 3×3 convolutional layers in the fourth stage are also changed to dilation convolution with a stride of 2 to increase the receptive field. The additional 1×1 convolutional layer is used to reduce the number of channels of extracted features to reduce computational loads in the following attention network. The network parameters in the feature extraction module are shared by the template and search branches. The outputs of the feature extraction module of the template and search branches are then $\boldsymbol{T} \in \mathbb{R}^{(n+1)\times C\times \frac{w}{s}\times \frac{h}{s}}$ and $\boldsymbol{S} \in \mathbb{R}^{C\times \frac{W}{s}\times \frac{H}{s}}$, where w, h, W and H are the spatial size of the templates and the search image ($w = h = 128$, $W = H = 256$); C is the channel number ($C = 256$); s=8 is the stride of the feature extraction module after modification; n is the number of templates in the template bank, and the 1 in $n+1$ denote the template obtained from the last frame (introduced in the III-.3).

2) Attention neural network: The attention neural network contains two self-attention blocks which are separated into the template and search branches independently, and a cross-attention block which takes the outputs of the two self-attention blocks as input to measure the similarity between the templates and the search image. The inputs to the attention block are assigned as $\boldsymbol{X}_q \in \mathbb{R}^{N_q\times d}$ and $\boldsymbol{X}_{kv} \in \mathbb{R}^{N_{kv}\times d}$. Similar to [26], the corresponding quarry ($\boldsymbol{Q}$), key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$) are defined and used to calculate the attention:

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}] = [(\boldsymbol{X}_q + \boldsymbol{P}_q)\boldsymbol{W}_q, (\boldsymbol{X}_{kv} + \boldsymbol{P}_{kv})\boldsymbol{W}_k, \boldsymbol{X}_{kv}\boldsymbol{W}_v] \quad (1)$$

$$\text{Atten}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_k}\right)\boldsymbol{V} \quad (2)$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v \in \mathbb{R}^{d\times d'}$ are learnable parameter matrices, and Atten $\in \mathbb{R}^{N_q\times d'}$. Similar to [27], $\boldsymbol{P}_q$, $\boldsymbol{P}_{kv}$ are position embeddings, and are only added to the quarry and key matrices. For multi-head attention:

$$\text{Multi-Head} = \text{Concat}(\text{Atten}_1, \cdots, \text{Atten}_h)\boldsymbol{W}_o \quad (3)$$

where h is head number ($h = 8$); $\boldsymbol{W}_o \in \mathbb{R}^{hd'\times d}$; $d' = \frac{d}{h}$.

For the self-attention block in the template branch, the input matrices $\boldsymbol{X}_q = \boldsymbol{X}_{kv} = \boldsymbol{T}'$, where $\boldsymbol{T}' \in \mathbb{R}^{N_T\times C}$ is the reshaped version of the template feature $\boldsymbol{T}$, and $N_T = (n+1)\times \frac{w}{s}\times \frac{h}{s}$. Thus, in the attention block, $N_q = N_{kv} = N_T$, and $d = C$. Similarly, for the self-attention block in the search branch, $\boldsymbol{X}_q = \boldsymbol{X}_{kv} = \boldsymbol{S}'$, where $\boldsymbol{S}' \in \mathbb{R}^{N_S\times C}$ and $N_S = \frac{W}{s} \times \frac{H}{s}$. For the cross-attention module, $\boldsymbol{X}_q$ and $\boldsymbol{X}_{kv}$ are the outputs of the multi-head self-attention blocks in the template and search branches, respectively. The residual connection and feedforward layers after the attention blocks are the same as those in the Transformer architecture [26].

3) Appearance adaptation module: In our visual tracker, an appearance adaptation module, which consists of a template bank, is proposed to save different templates of the needle tip as diverse as possible. In this way, it is more likely to find a highly similar template that has a similar appearance to the needle tip in the current frame. Assign the output feature vector of the multihead attention in the template branch as $f \in \mathbb{R}^{\frac{w}{s}\frac{h}{s}C\times 1}$. Mathematically, to enrich the target information in the template bank is to maximize the volume $\Gamma(f_1, \cdots, f_n)$ of the parallelotope formed by the encoded feature vectors in the template bank [28]. The similarity between templates T_i and T_j can be measured by calculating the convolution between the corresponding encoded features: $f_i * f_j$. Doing this for all templates in the template bank, we can get a Gram matrix:

$$G(f_1, \cdots, f_n) = \begin{bmatrix} f_1 * f_1 & f_1 * f_2 & \cdots & f_1 * f_n \\ \vdots & \vdots & \ddots & \vdots \\ f_n * f_1 & f_n * f_2 & \cdots & f_n * f_n \end{bmatrix} \quad (4)$$

It has been shown that the determinant of the Gram matrix is the square of the n-dimentional volume Γ of the constructed parallelotope. Therefore, we can get the following relationship:

$$\max_{f_i} \Gamma(f_1, \cdots, f_n) \propto \max_{f_i} |G(f_1, \cdots, f_n)| \quad (5)$$

Increasing the number of templates in the template bank can increase the determinant of the Gram matrix, as long as

$n < \frac{w}{s}\frac{h}{s}C$. However, to ensure high computational efficiency, the number should be kept within a certain limit and fixed. A new template is only added to the template bank when it can increase the determinant of the Gram matrix by replacing an existing one. However, if only the determinant is considered, the template bank will finally end up with irreverent image patches, as the background and noisy patches are much more dissimilar to the needle tip than the target patch, and the tracker will finally drift away. To fix this problem, the template manually selected at the beginning of the tracking T_1, which is the only fully trusted template, is always saved, and then, only when the new patch is similar to the initial template to some extent is added to the template bank:

$$f_i * f_1 \;>\; \delta_{det}\; f_1 * f_1 \tag{6}$$

where δ_{det} is a hyperparameter.

The appearance of the needle tip in the last frame is always the most similar to that in the current frame, and a correct appearance model can significantly improve the tracking accuracy. Therefore, the target template cropped from the last frame should be combined with the templates in the template bank to track the needle tip. However, when drift occurs, the latest template may contain more background or noise features, which will on the contrary undermine the visual tracker. To address this problem, the intersection over union (IOU) of the tracked bounding boxes of the current frame and the template which has the highest classification score in the template bank was calculated. If the IOU value is higher than a predefined threshold δ_{IOU}, the template cropped from the current frame is saved and concatenated with templates in the template bank to track the target in the next step. The appearance adaptation module is indicated by the red arrows in Fig 2. The red dashed arrows indicate that the update of the template bank was carried out when the tracking of the current frame ended, and only the template cropped from the current frame is passed to the template branch to perform update.

4) Distractor suppression module: During tracking, not only is the target appearance model important, but the distractors models are also useful. With the features of strong distractors saved, they compete to explain every pixel in the input image. In this case, the distractors are more likely to be correctly identified by the saved distractor models, and the probability of identifying them as the target will be suppressed. This can be achieved by the explaining away method proposed in [29], and was used in our visual tracking network to suppress distractors.

The distractor suppression module contains an online target classification network followed by the explaining away method (indicated by the purple arrows in Fig. 2). Inspired by the work in [30], the online target classification network contains a 1×1 and a 4×4 convolutional layers. The parameters of these two convolutional layers are learned online through the meta-learning proposed in [30]. The network takes the features extracted by the feature extraction module and outputs a score map of the target in the current search frame. The explaining away method was then applied to suppress distractors and highlight the potential target area. The detailed implementation of the explaining away method can be found in [29].

5) Target localization and loss function design: There are one classification and two regression heads. The output of the attention network is passed through these heads to generate a classification map and two regression maps to indicate positions ($\boldsymbol{P}$), displacement offsets ($\boldsymbol{O}$), and the size of the bounding boxes ($\boldsymbol{B}$) of the target, where $\boldsymbol{P} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 1\times n}$, $\boldsymbol{O} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 2\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 2\times n}$.

The score maps output by the classification head are then post-processed by adding a cosine window to them to suppress large displacements. Next, to fully make use of all the templates, the score maps are weighted by the maximum value in the corresponding maps before they are summed together to form a weighted map $\boldsymbol{P}_w \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}}$. The distractor suppression module also outputs a score map $\boldsymbol{P}_d \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}}$, which is then integrated with $\boldsymbol{P}_w$ to locate the target: $\boldsymbol{P} = \gamma\boldsymbol{P}_w + (1-\gamma)\,\boldsymbol{P}_d$, where γ, used to adjust the contributions of the distractor suppression module to the target localization task, was empirically set to 0.65. Although each template has independent offset and bounding box regression maps, only regression maps of the latest template are used to locate the target, as the appearance of the target in this template is the most similar to its real appearance at the current time step. The target position and the size of the bounding box are finally determined by

$$(x_c,\; y_c) = s\left(\arg\max\left(\boldsymbol{P}\right) + \boldsymbol{O}'\left(\arg\max\left(\boldsymbol{P}\right)\right)\right) \tag{7}$$

$$(w_{bb},\; h_{bb}) = (W, H)\cdot\boldsymbol{B}'\left(\arg\max\left(\boldsymbol{P}\right)\right) \tag{8}$$

where $\arg\max\left(\boldsymbol{P}\right)$ outputs the 2D position of the maximum value in the score map $\boldsymbol{P}$, and $\boldsymbol{O}'$ and $\boldsymbol{B}'$ represent the offset and bounding box regression maps of the latest template.

The design of the loss function contains three components [31]: the positioning error L_p, the offset estimation error L_o and the size prediction error of the bounding box L_b. Assume that the ground-truth position of the needle tip in the current search patch is $\boldsymbol{p} \in \mathbb{R}^2$, and a corresponding low-resolution equivalent $\hat{\boldsymbol{p}} = \lfloor \frac{\boldsymbol{p}}{s} \rfloor$ can be used to represent its position on the score map. Then, a Gauss function

$$\hat{\boldsymbol{P}} = \exp\left(-\left(\left(p_x - \hat{p}_x\right)^2 + \left(p_y - \hat{p}_y\right)^2 / \left(2\sigma_p^2\right)\right)\right) \tag{9}$$

centered at $\hat{\boldsymbol{p}}$ is used to calculate the focal loss [32]:

$$L_p = -\sum_{x,y}\begin{cases}\left(1-P_{xy}\right)^{\alpha}\log\left(P_{xy}\right), & \hat{P}_{xy}=1\\ \left(1-\hat{P}_{xy}\right)^{\beta}\left(P_{xy}\right)^{\alpha}\log\left(1-P_{xy}\right), & \text{Otherwise}\end{cases} \tag{10}$$

where σ_p is a size-adaptive standard deviation [31], $\boldsymbol{P}_{xy}$ is the value of the score map $\boldsymbol{P}$ at position $(x,\; y)$. α and β are two hyperparameters set to 2 and 4, respectively.

The position offset and the size of the bounding box are both trained using L1 loss functions and are defined as

$$L_o = \left|\boldsymbol{O}'_{\hat{\boldsymbol{p}}} - \left(\boldsymbol{p}/s - \hat{p}\right)\right|, \quad L_b = \left|\boldsymbol{B}'_{\hat{\boldsymbol{p}}} - \hat{b}\right| \tag{11}$$

where $\hat{b} = \left(\frac{w_{gt}}{W}, \frac{h_{gt}}{H}\right)$, and w_{gt} and h_{gt} are ground-truth size of the bounding box. The total loss function is defined as

$$L = L_p + L_o + L_b \tag{12}$$

IV. Motion Filtering

1) Kalman filter: When adopting the Kalman filter, the state and measurement vectors of the tracking system are assigned as $\boldsymbol{X}$ and $\boldsymbol{Z}$, where $\boldsymbol{Z} = [x, y]$ represents the position of the needle tip in the US image obtained from the visual tracker, and $\boldsymbol{X} = [x, v_x, y, v_y]$ represents the estimated position of the needle tip and the insertion velocities along the X- and Y-directions. The tracking system can thus be described

$$\begin{cases} \boldsymbol{X}(t) = \boldsymbol{F}\hat{\boldsymbol{X}}(t-1) + \boldsymbol{V} \\ \boldsymbol{Z}(t) = \boldsymbol{H}\hat{\boldsymbol{X}}(t) + \boldsymbol{W} \end{cases} \tag{13}$$

where $\boldsymbol{F}$ and $\boldsymbol{H}$ are the state and measurement transition matrices; $\boldsymbol{V}$ and $\boldsymbol{W}$ are white Gaussian noises. The corresponding covariance matrices of $\boldsymbol{V}$ and $\boldsymbol{W}$ are $\boldsymbol{Q}$ and $\boldsymbol{R}$. The filtering process of the Kalman filter can be divided into two different phases:
(1) State update:

$$\hat{\boldsymbol{X}}(t|t-1) = \boldsymbol{F}\hat{\boldsymbol{X}}(t-1|t-1) \tag{14}$$

$$\boldsymbol{P}(t|t-1) = \boldsymbol{F}\boldsymbol{P}(t-1|t-1)\boldsymbol{F}^T + \boldsymbol{Q} \tag{15}$$

where $\boldsymbol{P}$ is the error covariance matrix.
(2) Measurement update:

$$\boldsymbol{S}(t) = \boldsymbol{H}\boldsymbol{P}(t|t-1)\boldsymbol{H}^T + \boldsymbol{R} \tag{16}$$

$$\boldsymbol{K}(t) = \boldsymbol{P}(t|t-1)\boldsymbol{H}^T\boldsymbol{S}(t)^{-1} \tag{17}$$

$$\hat{\boldsymbol{X}}(t|t) = \hat{\boldsymbol{X}}(t|t-1) + \boldsymbol{K}(t)\left[\boldsymbol{Z}(t) - \boldsymbol{H}\hat{\boldsymbol{X}}(t|t-1)\right] \tag{18}$$

$$\boldsymbol{P}(t|t) = [\boldsymbol{I} - \boldsymbol{K}(t)\boldsymbol{H}]\boldsymbol{P}(t|t-1)[\boldsymbol{I} + \boldsymbol{K}(t)\boldsymbol{H}] - \boldsymbol{K}(t)\boldsymbol{R}\boldsymbol{K}'(t) \tag{19}$$

where $\boldsymbol{S}$ is the innovation covariance matrix, $\boldsymbol{K}$ is the gain matrix, and $\hat{\boldsymbol{X}}(t|t)$ is the estimated state value of the target.

2) TransKalman: TransKalman is designed based on the principle of the Kalman filter, with the gain matrix directly obtained by an attention-based neural network. Inspired by the work in [33], the input in TransKalman includes four components: the observation difference ΔZ_t, the innovation difference $\Delta\tilde{Z}$, the state difference $\Delta\hat{\boldsymbol{X}}_{t-1}$, and the state update difference $\Delta\tilde{\boldsymbol{X}}_{t-1}$. The detailed definitions of them are as follows:

$$\Delta\boldsymbol{Z}(t) = \boldsymbol{Z}(t) - \boldsymbol{Z}(t-1) \tag{20}$$

$$\Delta\tilde{\boldsymbol{Z}}(t) = \boldsymbol{Z}(t) - \boldsymbol{H}\hat{\boldsymbol{X}}(t|t-1) \tag{21}$$

$$\Delta\hat{\boldsymbol{X}}(t) = \hat{\boldsymbol{X}}(t|t) - \hat{\boldsymbol{X}}(t-1|t-1) \tag{22}$$

$$\Delta\tilde{\boldsymbol{X}}(t) = \hat{\boldsymbol{X}}(t|t) - \boldsymbol{F}\hat{\boldsymbol{X}}(t-1|t-1) \tag{23}$$

By concatenating these variables as a long input vector, $\boldsymbol{X}_{cat}^t = [\Delta\boldsymbol{Z}(t), \Delta\tilde{\boldsymbol{Z}}(t), \Delta\hat{\boldsymbol{X}}(t-1), \Delta\tilde{\boldsymbol{X}}(t-1)] \in \mathbb{R}^n$ where $n = 12$, $\boldsymbol{X}_{cat}^t$ is then combined with a set of historical input vectors before input into the TransKalman network to obtain the Kalman gain matrix $\boldsymbol{K}(t)$. The parallel data processing ability of the Transformer enables the gain matrix to be properly learned from a long input data sequence. It should be noticed that the differences in state and state update in $\boldsymbol{X}_{cat}^t$ at time t are $\Delta\hat{\boldsymbol{X}}(t-1)$ and $\Delta\tilde{\boldsymbol{X}}(t-1)$ as the value of $\hat{\boldsymbol{X}}(t|t)$ is not available at the beginning. With the gain matrix known, the target state value is estimated according to Eq. (18).

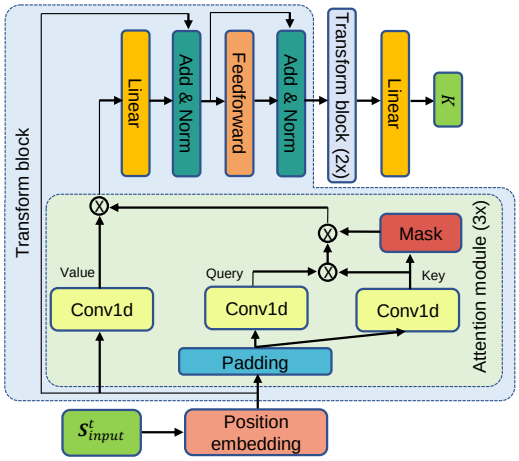

Fig. 3: Architecture of TransKalman. For simplicity, only one of the three parallel attention module is shown and only one of the detailed structure of the three stacked transform blocks is plotted.

The detailed architecture of TransKalman is shown in Fig. 3. A series of input vectors are arranged in the order of time to form an input sequence $\boldsymbol{S}_{input}^t = [\boldsymbol{X}_{cat}^{t-T+1}, \cdots, \boldsymbol{X}_{cat}^{t-1}, \boldsymbol{X}_{cat}^t]$, where T is the sequence length that represents how far the historical data is taken into consideration to influence the estimation of the current state value (T was set to 30 in our experiments). The arranged sequence first passes through a position embedding layer to integrate the sequential information into each input vector. The embedded input sequence is then processed by a multi-head attention module, and for simplicity, only one head is shown in Fig. 3. As $\boldsymbol{X}_{cat}^t$ is a one dimensional (1D) vector, the value ($\boldsymbol{V}_h \in \mathbb{R}^{T\times n}$), query ($\boldsymbol{Q}_h \in \mathbb{R}^{T\times n}$) and key ($\boldsymbol{K}_h \in \mathbb{R}^{T\times n}$) matrices in the attention module are calculated by using the 1D convolution, where the subscript denotes the h-th head. The attention of the h-th head is then calculated by:

$$\boldsymbol{O}_h = \text{Softmax}\left(\boldsymbol{Q}_h(\boldsymbol{K}_h)^T/\sqrt{n} \times M\right)\boldsymbol{V}_h \tag{24}$$

where $\boldsymbol{O}_h \in \mathbb{R}^{T\times n}$. Different from the attention equation shown in Eq. (2), here, a mask matrix M with elements in the upper triangle equal to negative infinity is used to ensure that future data do not have an impact on the attention calculation in the current time step. As the mask filters out future data, TransKalman can work properly from the first frame of tracking, and does not require special initialization or waiting time to collect a long input sequence, as in [33].

The attention of each head is calculated in parallel and concatenated together. There are a total of 3 heads, and the final attention is expressed as $\boldsymbol{O}_{set} = [\boldsymbol{O}_1, \boldsymbol{O}_2, \boldsymbol{O}_3] \in \mathbb{R}^{T\times 3n}$. Apart from the attention module, there are also linear, normalize and feedforward layers in the transform block. The concatenated attention is first reshaped by the linear layers to $\mathbb{R}^{T\times n}$ and then processed by the rest of the layers. To fully explore the temporal information, there are three transform

blocks that stack together. The output of the last transform block is processed by a linear layer to generate the gain matrix.

TransKalman does not rely on the process and measurement noise matrices, Q and R, to estimate the target's state value and thus does not require these noises to obey Gaussian distribution, which is usually not the case in US image-based target tracking [10]. Therefore, it is more robust when combined with the visual tracker to finish the tracking task. Furthermore, the use of the attention mechanism enables the estimator to fully explore the underlying motion law of the target from historical data. The state and measurement transition matrices are not used during the estimation of the Gain matrix, which also enables TransKalman to handle model misidentification to a certain extent.

V. Experiment Results and Discussions

A. Experimental setup and procedures

An overview of the experimental setup is shown in Fig. 4. In our experiments, the needle was manipulated both by motors and human hands. The US images were collected using the Vantage 32 LE system (Verasonics, Inc., USA) and the C5-2 US probe (Mindary, Inc., China). The video sequences collected from both motorized and manual needle manipulation experiments were eventually used to test our proposed tracking system. An 18-gauge sensorized needle tool (Aurora Needle, NDI, Inc., USA) was used for the needle insertion during the experiments. It has a 5 DOF electromagnetic (EM) sensor embedded in its tip, such that the ground-truth position of the needle tip can be obtained by using the corresponding EM tracking system (NDI, Inc., USA). The root mean square error (RMSE) of the positional accuracy of the needle tool in our experimental setup were tested to be 0.76 mm$\pm$0.13 mm, respectively, which is slightly larger than the position accuracy provided by the manufacturer (0.57 mm). During the experiments, the needle was inserted into both phantom and ex-vivo biological tissues. The phantom was made according to the recipe in [15] with silica powder added to simulate the speckles in the US images, and the biological tissues were pork and chicken.

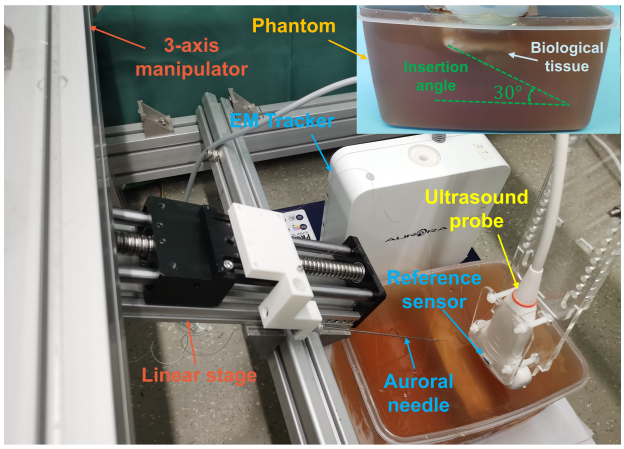

Fig. 4: Setup of the motorized needle insertion experiments.

During the motorized experiments, a 3-axis manipulator (LiTai Technology, Inc., China) was used to hold the US probe to follow the movement of the needle tip, and a linear stage (LiTai Technology, Inc., China) was used to translate the needle. The needle was inserted in three different scenarios, namely, In-plane-static, In-plane-moving, and Out-of-plane. In the In-plane scenarios, the needle axis stayed in the US imaging plane, while in the Out-of-plane scenario, the needle axis was perpendicular to the US imaging plane and only the needle tip could be found in US images. The US probe was moved with the needle tip in the In-plan-moving and Out-of-plane scenarios, and kept static in the In-plane-static scenario. In each scenario, the needle was inserted at three different angles ($0°$, $45°$, $60°$) with three different insertion velocities (0.4 mm/s, 1 mm/s and 2 mm/s). These angles are commonly used in biopsy surgeries, including breast and prostate [39]. A scenario with a specific insertion angle was called a case, and there are a total of 9 cases. In each case, the needle tip was tracked in both insertion and the withdrawal processes, and each case was repeated at least 9 times. There are a total of 114 and 105 video sequences collected in the phantom and tissue experiments, respectively.

During the manual experiments, three individuals were invited to insert the needle with one hand and hold the US probe static with the other hand. Due to the imprecise operation of the human hands, the insertion angles were roughly divided into three categories: small (S, $-10° \sim 10°$), medium (M, $30° \sim 45°$) and large (L, $60° \sim 75°$). It should be noted that the needle was only inserted in the In-plane-static scenario in these manual experiments, because it is highly challenging for humans to move the US probe to follow the needle tip in either in-plane-moving or out-of-plane scenario. Each person was required to repeat at least 15 times at each angle, and a total of 283 video sequences were collected in the tissue experiments.

B. Coordinate Registration

The ground-truth position of the needle tip could be obtained directly from the embedded EM sensor in the needle tip represented in the EM tracking coordinate system Φ_{em}. To map it to the US image's coordinate system Φ_{us}, a coordinate registration procedure was implemented. According to the method proposed in [40], an additional 6 DOF EM sensor was attached to the US probe. The RMSE of the position accuracy of the sensor reported by the manufacturer is 0.48 mm, but it was tested to be 0.63$\pm$0.11 mm with our own experimental setup. The transformation matrix T_{ref}^{em} that transforms the position of the needle tip $\boldsymbol{p}^{em}$ from Φ_{em} to the reference sensor's coordinate system Φ_{ref} can then be directly calculated using the pose data of the 6 DOF sensor in the EM tracking system. The position of the needle tip in Φ_{us} was denoted as $\boldsymbol{p}^{us}$, and the relationship between $\boldsymbol{p}^{em}$ and $\boldsymbol{p}^{us}$ can be expressed as:

$$\boldsymbol{p}^{em} = T_{ref}^{em}\, T_{us}^{ref}\, \boldsymbol{p}^{us} \tag{25}$$

where $\boldsymbol{p}^{em}$ could be read directly from the EM tracking system and $\boldsymbol{p}^{us}$ was the position located by our proposed visual tracking system. Only the transformation matrix T_{us}^{ref} was unknown. With the relative position of the reference sensor

to the US probe fixed, the value of T_{us}^{ref} was fixed, and can be calculated using the SVD method proposed in [40].

C. Network training

The visual tracking and the TransKalman networks were trained independently in this work because they are responsible for different tasks. Training splits sampled from publicly available object tracking datasets, including TrackingNet [34], COCO [35], LaSOT [36], GOT-10K [37] and ImageNet VID [38], were first used to train the visual tracking network. There were a total of 5,000 images in each split and the batch size was set at 40. The network was trained for 80 epochs with the initial learning rate set at 0.01, and decayed by 0.2 every 15 epochs. The network was then fine-tuned on a self-collected US image-based needle tip tracking dataset under motorized needle manipulation consisting of 5,000 images. The dataset was collected in the same experimental setup shown in Fig. 4. At the beginning of each fine-tuning epoch, 500 images were randomly selected from the dataset for the purpose of validation and the rest were used for training. The total training epoch, initial learning rate and the decay factor in the fine-tune process were set to 50, 1×10^{-4} and 0.35, respectively. The dataset used to train and validate TransKalman contains two different parts, including computer-simulated needle motion and actual needle motion under manual manipulation. 1) To simulate the movement of the needle tip with constant velocity (inserted by motors) in 2D plane, a simulated dataset was collected by using Matlab. The simulated needle velocity ranges from +2 mm/s to -2 mm/s, where the positive and negative signs represent the insertion and withdrawal of the needle. Two sets of Gaussian noises which obey $\mathcal{N}(0, 0.1^2)$ and $\mathcal{N}(0, 0.1^4)$ were then added to the simulated positions and velocities in each time step. The sequence length of each trajectory was set to 1,000 and a total of 5,000 trajectories were simulated. 2) During the manual insertion and withdrawal experiments, an 18 gauge Aurora Needle (NDI, Inc., USA) with a 5 DOF electromagnetic (EM) sensor embedded in the tip was used, and the movement of the needle tip was recorded. In each data sequence, the needle was inserted or withdrawn about 10 cm in around 2 minutes, and a total of 422 data sequences were collected. The TransKalman network was trained on this hybrid dataset with 1200 epochs and batch size set to 16. The Adam optimizer was used during the training with the initial learning rate and the decay factor set to 1×10^{-3} and 1×10^{-5}, respectively. This hybrid dataset was also randomly split into 90% for training and 10% for validation.

D. Ablation study and performance comparison of the visual tracking network

To show the effectiveness of each module (attention module, appearance adaptation module, and distractor suppression module) in the visual tracking network, an ablation study was conducted. A performance index named "success rate" was proposed to quantitatively evaluate the tracking performance. The success rate of a tracker in a particular case was defined as the ratio of the number of successfully tracked video sequences to the total number of video sequences in a particular case. The tracking of a video was recognized as successful if the Euclidean distance tracking error of 95% frames in the video was less than 3 mm, which is in the range of acceptable needle targeting error in percutaneous procedures [41].

Before performing the ablation study, the optimal number of templates, n, in the template set should be determined, considering its impact on the tracking success rate and tracking speed. Fig. 5 shows the change of the tracking success rate and the tracking speed by TO visual of the needle tip during the motorized insertion experiments in the tissue when n is set to different values. It can be observed that an increase in n leads to improvement in the success rate. This is mainly because the more diverse the template set is, the probability of finding a template which has most similar appearance to the needle tip's current appearance increases, thus enhancing the tracking accuracy. However, when $n > 14$, the success rate does not experience significant increase while the tracking speed continues to decrease. To achieve a balance in achieving high tracking success rate and high tracking speed, the value of n was set to 14.

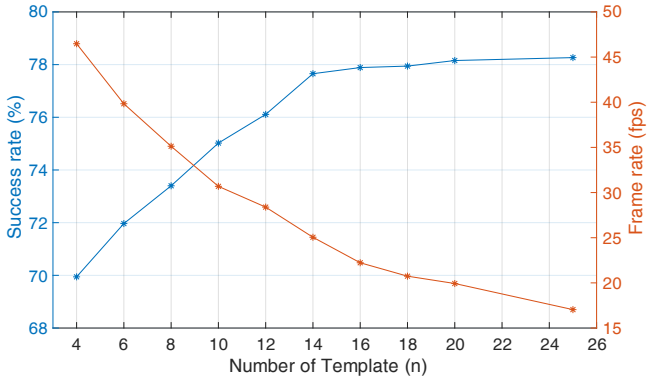

Fig. 5: Tracking success rate and tracking speed when n is set to different values in motorized needle experiments in the tissue.

During the ablation study of the visual tracking network, the tracking success rates of the attention module solely (Att), attention with appearance adaptation modules (Att + AppAda), attention with distractor suppression modules (Att+DisSpp) and the complete task-oriented visual tracking network (TO visual or Att+AppAda+DisSpp) in the motorized insertion in tissue experiments are shown in Table I. The highest success rate in each case is highlighted in red.

TABLE I: Comparison in the tracking success rate (%) during ablation study of the visual tracking network in the motorized needle insertion experiments in tissue

Scenario	In-plane-static (%)				In-plane-moving (%)				Out-of-plane (%)				Mean
	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	
Attention (Att)	28.6	50.9	81.8	53.7	66.7	71.4	50.0	62.7	81.80	70.0	91.7	81.2	65.9
Att+AppAda	57.1	63.6	90.9	70.5	66.7	85.7	71.4	74.6	100.0	70.0	91.7	87.2	77.5
Att+DisSpp	57.1	63.6	90.9	70.5	75.0	85.7	78.5	79.7	100.0	80.0	91.7	90.6	80.3
TO visual	71.4	76.3	90.9	79.5	75.0	85.7	78.6	79.7	100.0	70.0	91.7	87.2	82.2

The addition of either the appearance adaptation and distractor suppression modules improves the tracking success rate compared to using the attention module only. The success rate increases in each case except the Out-of-plane-60° case, where the attention module is already able to achieve a high success rate (91.7%). The appearance adaptation and distractor suppression modules improve the visual tracker from different aspects. In In-plane-static-30°, both the success rates of Att+AppAda and Att+DisSpp are 63.6%, but the combination of them results in a success rate of 76.3%, which is higher than either of them. Compared to Attention, the improvements in the success rate of Att+AppAda+DisSpp are significant in the In-plane-static and In-plane-moving scenarios at 48.0% and 27.1%, respectively. Without specific annotation, the differences in the success rate are denoted in the relative term in the following analysis. The overall success rate of TO visual in the motorized needle experiment in tissue is 82.2%, which increases by 24.7% compared to the network with the attention module only.

TO visual has also been compared to our previously proposed learning-based visual tracking module (LR visual) [14] in terms of the tracking success rate in the motorized needle insertion experiments in tissue in Table II. Significant performance improvement can be observed in nearly all cases. The tracking success rate increases by the highest (32.2%) in the In-plane-moving scenario. The improvement in the overall success rate is 23.2% (66.7% vs. 82.2%), verifying the superiority of TO visual when compared to LR visual.

TABLE II: Comparison in the success rates (%) by TO visual and our previous work (LR visual) [14] in the motorized needle insertion experiments in tissue

Scenario	In-plane-static (%)				In-plane-moving (%)				Out-of-plane (%)				Mean
	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	
LR visual	54.2	60.8	75.3	63.4	50.4	71.4	59.1	60.3	66.4	71.9	91.7	76.6	66.7
TO visual	71.4	76.3	90.9	79.5	75.0	85.7	78.6	79.7	100.0	70.0	91.7	87.2	82.2

E. Evaluation of TransKalman motion filter

1) Effect of TransKalman on the overall tracking system: To show the performance improvement boosted by the addition of TransKalman, the success rates of TO visual and TO system (TO visual+TransKalman) are compared in the motorized needle insertion in tissue experiments, as shown in Table III. With the addition of TransKalman filter, the mean success rate increases in every scenario, and the overall mean success rate increases by 6.3% (87.4% vs. 82.2%). The largest increase (28.6%) is found in the Out-of-plane scenario when the insertion angle is 30°.

TransKalman helps to robustly track the needle tip not only by filtering the visual tracking results but also by providing reasonable position estimation when the needle tip temporarily disappears from the US images. Fig. 6(a) shows the tracking results of TO visual and TO system when the needle tip disappeared from US images. Fig. 6(b) shows some snapshots of the tracking results. It can be seen that the needle tip completely disappeared from Frame 671 to 679. During this period, as the visual tracking network could not get any useful visual information of the target from the current US images, TO visual suddenly drifted away from its original insertion path to a noisy distractor, as seen by the large displacement shown in Fig. 6(a) and the green bound box shown in Fig. 6(b). In contrast, TransKalman in TO system correctly estimated the needle tip's real position and suppressed the drift. With no target-similar structures contained in the estimated bounding box, the template update process was stopped and the template bank was not contaminated by the distractors. Therefore, when the needle tip reappeared, the visual tracking network in TO system correctly found the target, as shown in Fig. 6(b) Frame 730. The same phenomenon occurred again when TransKalman suppressed the large drift at Frame 754 and the visual tracking network re-detected the needle tip at Frame 773.

TABLE III: Comparison in the tracking success rate between TO visual and TO system (TO visual+TransKalman) in the motorized needle insertion experiments in tissue

Scenario	In-plane-static (%)				In-plane-moving (%)				Out-of-plane (%)				Mean
	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	
TO visual	71.4	76.3	90.9	79.5	75.0	85.7	78.6	79.7	100.0	70.0	91.7	87.2	82.2
TO system	85.7	75.0	90.9	83.9	88.9	85.7	78.6	84.4	100.0	90.0	91.7	93.9	87.4

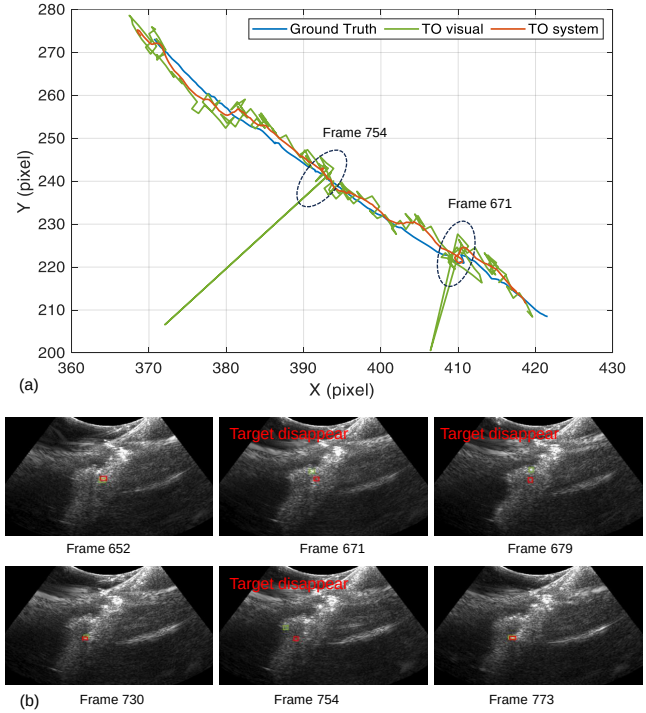

Fig. 6: (a) Ground truth trajectory of the needle tip and tracking results by TO visual and TO system (TO visual+TransKalman) (b) Snapshots of the tracking results by TO visual and TO system.

2) Comparing TransKalman and Kalman filter: A simulation study was performed to first compare the performance of TransKalman and the Kalman filter in terms of their convergence speed in nonlinear motions in a manual needle insertion

experiment. The input to the filters was the ground truth of needle tip's trajectory with noise added to simulate the noisy measurement data obtained from the visual tracking network. The covariance matrix of the added noise was set to $4\boldsymbol{I}_{2\times2}$ where $I_{2\times2}$ is a 2×2 identity matrix. The Monte Carlo experiment was performed by repeating the filtering process 100 times for each filter to obtain the corresponding RMSE values. It can be seen from Fig. 7 that the RMSE

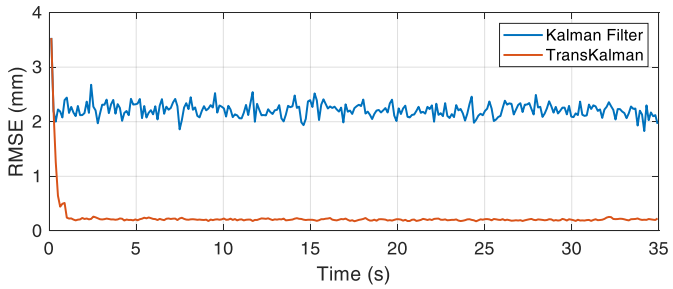

Fig. 7: RMSE of the Kalman filter and the proposed TransKalman in the case where the needle is manually inserted with nonlinear velocity.

of the TransKalman filter converges quickly until it reaches about 0.25 mm. The RMSE of the Kalman filter stays around 2.2 mm, which is slighter larger than the magnitude of the added noise. Its inability to converge is mainly due to a mismatch between the Kalman filter model built based on a constant velocity assumption and the actual motion of the needle tip.

To show the superior characteristics of TransKalman against the Kalman filter in estimating motions caused by nonlinear velocity, both TransKalman and Kalman filter were concatenated to TO visual before comparing their tracking performance in the manual needle manipulation experiments. It should be noted that the optimal initial values of Q and R matrices for TO visual+Kalman filter were determined by a grid search method. Fig. 8(a) shows the change of needle insertion velocity against time in one of the experiments. The needle was first inserted forward by the human hand and withdrawn by a short distance before it was inserted forward again. During this process, the needle insertion velocity was constantly adjusted by the human operator to make sure the needle tip was visible in the US images. Fig. 8(b) shows the ground truth, the tracking result of TO visual, TO system (TO visual+TransKalman) and TO Kalman (TO visual+Kalman filter) in the US image coordinate system in both horizontal (X) and vertical (Y) directions. It can be seen that TO Kalman was able to robustly track the needle tip for a longer period of time than TO visual. However, when the needle insertion velocity experienced a dramatic nonlinear change around 76 s, the Kalman filter failed to adapt to this change and provided inaccurate filtering results, leading to increase in the tracking error thereafter, as shown in Fig. 8(c). Under TO system, the tracking error was constantly maintained at a low level, as shown in Fig. 8(d). The strong adaptability of TransKalman in TO system to the nonlinear needle velocity can also be validated during the transition phases between the needle insertion and withdrawal processes. It was able to

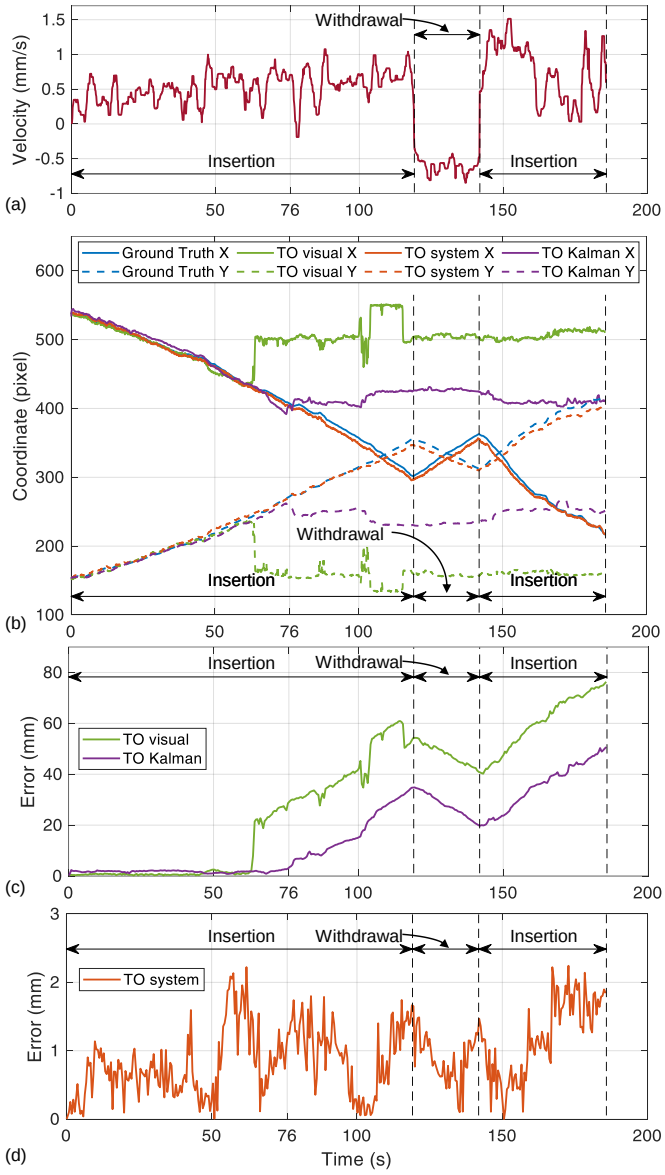

Fig. 8: (a) The needle insertion velocity against time. (b) The tracking results with respect to the ground truth of the TO visual, TO system (TO visual + TransKalman), and TO Kalman (TO visual + Kalman filter) in both horizontal (X) and vertical (Y) directions.(c) Tracking error of TO Visual and TO Kalman. (d) Tracking error of TO System.

quickly adapt to the motion direction changes and provided robust filtering results to suppress drift. As shown in Table IV, the overall mean success rate of TO system is higher than TO Kalman by 33.0%, validating the significantly greater tracking robustness of the proposed TransKalman filter over the Kalman filter during manual needle manipulation.

3) Comparing TransKalman filter and LR motion: To further demonstrate the state-of-the-art performance of the TransKalman filter, the motion filter (LR motion) proposed in our previous work [14] was concatenated to the TO visual to form a new tracking system called TO LR motion which was then used to compared with TO system. As seen in Table IV, TO

system (TO visual+TransKalman) outperforms TO LR motion (TO visual+LR motion) in nearly all cases and increases the overall mean success rate by 4.9%. LR motion only relies on the historical coordinate data of the target to estimate its current position, while TransKalman makes use of not only the historical coordinate data, but also other data differences (i.e., $\Delta \boldsymbol{Z}, \Delta \tilde{\boldsymbol{Z}}, \Delta \hat{\boldsymbol{X}}, \Delta \tilde{\boldsymbol{X}}$) to filter the target's current position. Furthermore, instead of directly estimating the current position of the needle tip, TransKalman estimates a Gain matrix first before filtering the target's position by following the data filtering principles in the Kalman filter, which helps to increase its robustness.

TABLE IV: Tracking success rates of TO Kalman, TO LR motion, and TO system to compare the effect of Kalman filter, LR motion, and TransKalman filter in robust needle tip tracking during the manual needle insertion in tissue experiments

Case	User 1 (%)				User 2 (%)				User 3 (%)				Mean
	Small	Medium	Large	*Mean*	Small	Medium	Large	*Mean*	Small	Medium	Large	*Mean*	
TO Kalman	48.3	64.3	59.3	57.3	59.4	71.4	63.2	64.7	50.0	65.6	80.8	65.5	62.5
TO LR motion	63.3	75.2	59.3	65.9	91.2	82.1	83.5	85.6	84.6	87.5	86.4	86.2	79.2
TO system	65.5	81.0	70.4	72.3	93.8	85.7	86.8	88.8	84.6	87.5	92.3	88.1	83.1

F. Evaluation of the overall tracking system performance

Figure 9 shows some snapshots of the tracking results of TO system (TO visual + TransKalman filter). The figures in the first row of Fig. 9 are sampled from the phantom experiment in the Out-of-plane scenario with the needle insertion angle equal to $0°$, and the small figures in the right-bottom corners show the appearance of the needle tip in the current frames. It can be observed that the appearance of the needle tip experienced dramatic changes throughout the tracking process. Since there was not much noise in the background in the phantom experiments, most trackers could handle this appearance change even with inaccurate target templates. However, when the background noise was strong with a lot of distractors surrounding the needle tip (as shown in the figures in the second and third rows of Fig. 9), the inaccurate target template would adversely affect the needle tip tracking results. Figures in the third row of Fig. 9 are sampled from the tissue experiments in the In-plane-moving scenario with the insertion angle being $60°$. It can be seen that the needle tip even disappeared from the US images in some frames. In such cases, a sudden huge drift occurred, and without other information, most trackers were likely to be trapped with the distractors. In our experiment, TO system successfully tracked the needle tip even in the most challenging scenarios in the last two rows. The proposed appearance adaptation and distractor suppression modules successfully updated the template bank and suppressed the strong noise in the background. TransKalman then provided a reasonable estimation of the needle tip's position when it temporarily disappeared.

To quantitatively evaluate the tracking robustness of TO system, its tracking success rates were compared against several state-of-the-art trackers, including ICTKF [10], Prdimp [42], SiamBan [43], Stark [18], AiATrack [44], KeepTrack [45] and LR tracking [14], in both motorized and manual needle insertion experiments with the results shown in Table V, VI and VII. The p-values calculated using the paired t-test were displayed to show the statistical significance of the difference in the success rates when comparing the different tracking methods against TO system. The first, second and third best performing trackers in each case are highlighted in red, green, and blue.

According to Table V, in the motorized phantom experiments, both our proposed TO system and the LR tracking share the highest mean success rate, which is 3.8% and 10.9% higher than the third and fifth trackers (ICTKF and AiATrack) respectively. Without their respective motion filters, TO visual and LR visual also performed well with the second and fourth highest success rates. The overall mean success rates of other trackers are all above 60%, and nearly 100% in the Out-of-plane scenario. As shown in Fig. 9, the US images in the phantom experiments had relatively high quality, and the needle tip in the Out-of-plane scenario was highly distinguishable from the background noise. Such idealized testing conditions do not demand strong requirements on the noise suppression and appearance adaptation abilities of the visual trackers. Therefore, most trackers were able to obtain highly promising tracking results. Our proposed TO visual and TO system only showed minor improvement in the success rate over other trackers, which does not truly demonstrate the advantages of our proposed trackers. Furthermore, the needle tip often disappeared for a significant period of time during the initial stage of the needle insertion. The TransKalman filter integrated in TO system might not have sufficient time to acquire enough accurate historical position data of the needle tip to estimate its motion law, limiting its positive impact on the tracking success rate.

The superiority in the tracking robustness of our proposed TO system can be more clearly observed in the tissue experiments. As seen in Table VI, the overall mean success rate of TO system increases by 5.7% and 11.5%, compared to the second and fourth best performing trackers, namely LR tracking and SiamBan. TO system outperforms LR tracking in a majority of cases, especially in the In-plane-static-$0°$ case, where the success rate increases by a significant 23.1% (85.7% vs. 69.6%). It should be noted that TO visual obtains the third best performance by only relying on our proposed task-oriented visual tracking network. Most of the other trackers achieve mean success rates below 52% in the tissue experiments, which are a significant performance degradation compared with those in the phantom experiments. This is mainly due to the inability of these trackers to overcome the more complex noise and stronger distractors in the tissue under US imaging compared to the phantom.

In the manual needle insertion experiments in the tissue, TO system achieves the highest overall mean success rate of 83.1%, which is 5.1% and 30.6% higher than those of the second and fourth best trackers, namely LR tracking and SiamBan, as shown in Table VII. It should again be noted that TO visual outperforms other pure visual information-based state-of-the-art trackers to a large extent with an increase of 4.4% over its closest competitor (i.e. SiamBan).

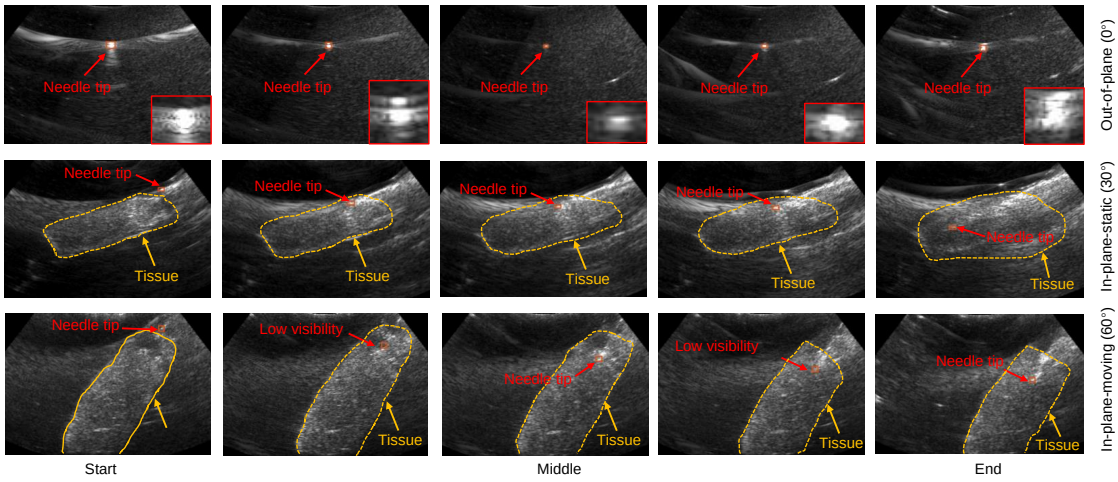

Fig. 9: Tracking results of TO system in three different scenarios. Snapshots in the first row are sampled from phantom experiments and snapshots in the second and third rows are sampled from tissue experiments.

TABLE V: Success rates and p-values of the proposed TO system and other trackers in the motorized needle insertion in phantom experiments

Scenario	In-plane-static (%)				In-plane-moving (%)				Out-of-plane (%)				Mean (%)	p-value
	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*		
LR visual	75.0	76.9	83.3	78.4	83.3	93.3	85.7	87.4	91.7	100.0	100.0	97.2	87.7	0.031
TO visual	75.0	76.9	83.3	78.4	91.7	93.3	92.9	92.6	100.0	100.0	100.0	100.0	90.3	0.045
PrDimp	41.7	38.5	66.7	48.9	41.7	80.0	92.9	71.5	100.0	100.0	91.7	97.2	72.6	<0.01
SiamBan	66.7	30.8	33.3	43.6	50.0	93.3	92.9	78.7	100.0	100.0	100.0	100.0	74.1	<0.01
Stark	33.3	30.8	33.3	32.5	33.3	66.7	92.9	64.3	83.3	75.0	100.0	86.1	61.0	<0.01
AiATrack	50.0	76.9	66.7	64.5	75.0	86.7	100.0	87.2	100.0	100.0	100.0	100.0	83.9	0.041
KeepTrack	58.3	30.8	58.3	49.1	66.7	93.3	92.9	84.3	100.0	91.7	100.0	97.2	76.9	<0.01
ICTKF	83.3	76.9	83.3	81.2	83.3	86.7	92.9	87.6	100.0	100.0	100.0	100.0	89.6	0.029
LR tracking	81.3	83.1	86.9	83.8	92.1	98.0	95.3	95.1	100.0	100.0	100.0	100.0	93.0	0.048
TO system	75.0	84.6	91.7	83.8	91.7	93.3	100.0	95.1	100.0	100.0	100.0	100.0	93.0	-

TABLE VI: Success rates and p values of the proposed TO system and other trackers in the motorized needle insertion in tissue experiments

Scenario	In-plane-static (%)				In-plane-moving (%)				Out-of-plane (%)				Mean (%)	p-value
	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*	0°	30°	60°	*Mean*		
LR visual	54.2	60.8	75.3	63.4	50.4	71.4	59.1	60.3	66.4	71.9	91.7	76.6	66.7	<0.01
TO visual	71.4	76.3	90.9	79.5	75.0	85.7	78.6	79.7	100.0	70.0	91.7	87.2	82.2	0.049
PrDimp	28.6	16.7	63.6	36.3	33.3	71.4	28.6	44.4	36.4	40.0	91.7	56.0	45.6	<0.01
SiamBan	78.6	66.7	72.7	72.7	77.8	92.9	64.3	78.3	90.9	70.0	91.7	84.2	78.4	<0.01
Stark	28.6	33.3	45.5	35.8	22.2	57.1	21.4	33.6	45.5	40.0	66.7	50.7	40.0	<0.01
AiATrack	28.6	25.0	36.4	30.0	22.2	85.7	21.4	43.1	45.5	40.0	66.7	50.7	41.3	<0.01
KeepTrack	28.6	33.3	54.5	38.8	77.8	85.7	14.3	59.3	27.3	50.0	91.7	56.3	51.5	<0.01
ICTKF	42.9	66.7	72.7	60.7	66.7	57.1	71.4	65.1	81.8	80.0	91.7	84.5	70.1	0.013
LR tracking	69.6	79.4	85.5	78.2	80.6	75.2	81.1	79.0	91.6	86.5	94.5	90.9	82.7	0.041
TO system	85.7	76.3	90.9	83.9	88.9	85.7	78.6	84.4	100.0	90.0	91.7	93.9	87.4	-

TABLE VII: Success rates and p-values of the proposed TO system and other trackers in the manual needle insertion in tissue experiments

Case	User 1 (%)				User 2 (%)				User 3 (%)				Mean (%)	p-value
	Small	Medium	Large	*Mean*	Small	Medium	Large	*Mean*	Small	Medium	Large	*Mean*		
LR visual	46.5	48.8	58.1	51.1	71.4	72.3	66.7	70.1	52.3	65.6	78.3	65.4	62.2	<0.01
TO visual	53.4	66.7	59.3	59.8	74.5	72.3	69.2	72.0	50.0	71.3	80.8	67.4	66.4	<0.01
PrDimp	44.8	50.0	40.7	45.2	71.9	64.3	68.4	68.2	50.0	65.6	76.9	64.2	59.2	<0.01
SiamBan	41.4	61.4	55.6	52.8	78.1	68.6	68.4	71.7	56.2	65.6	76.9	66.3	63.6	<0.01
Stark	10.3	33.3	18.5	20.7	34.4	46.4	23.7	34.8	18.8	37.5	34.6	30.3	28.6	<0.01
AiATrack	13.8	42.9	33.3	30.0	43.8	42.9	39.5	42.0	37.5	53.1	46.2	45.6	39.2	<0.01
KeepTrack	20.7	40.5	37.0	32.7	56.2	57.1	50.0	54.5	43.8	65.6	73.1	60.8	49.3	<0.01
ICTKF	20.7	59.5	48.1	42.8	46.9	71.4	55.3	57.9	50.0	75.0	73.1	66.0	55.6	<0.01
LR tracking	63.3	69.0	63.8	65.4	93.8	82.1	81.6	85.8	82.4	87.5	88.5	86.1	79.1	0.001
TO system	65.5	81.0	70.4	72.3	93.8	85.7	86.8	88.8	84.6	87.5	92.3	88.1	83.1	-

TABLE VIII: Tracking errors of different trackers in the motorized needle insertion in tissue experiments

Scenario	In-plane-static (mm)			In-plane-moving (mm)			Out-of-plane (mm)		
	Mean	Std	95%ile	Mean	Std	95%ile	Mean	Std	95%ile
PrDimp	1.76	0.58	2.37	1.75	0.81	2.65	1.34	0.60	1.90
SiamBan	1.75	0.46	2.60	1.71	0.99	2.85	0.95	0.46	1.63
Stark	2.30	1.60	2.86	2.15	0.71	2.88	2.18	1.27	2.47
AiATrack	2.26	1.03	2.72	2.25	1.42	2.99	1.90	1.19	2.74
KeepTrack	2.09	1.25	2.62	2.36	1.27	2.87	1.65	0.72	2.58
ICTKF	2.32	1.21	2.40	2.08	1.70	2.81	1.59	0.70	2.55
LR tracking	1.67	0.95	2.45	1.20	1.30	2.68	0.87	0.54	1.65
TO system	1.36	0.40	1.65	1.71	0.63	2.38	1.12	0.40	1.69

TABLE IX: Tracking errors of different trackers in the manual needle insertion in tissue experiments

Case	User 1 (mm)			User 2 (mm)			User 3 (mm)		
	Mean	Std	95%ile	Mean	Std	95%ile	Mean	Std	95%ile
PrDimp	1.61	0.63	2.64	1.41	0.48	2.45	1.39	0.64	2.65
SiamBan	1.56	0.74	2.68	1.41	0.45	2.52	1.43	0.64	2.49
Stark	1.94	0.55	2.65	1.49	0.51	2.45	1.66	0.73	2.72
AiATrack	1.87	0.72	2.68	1.54	0.60	2.61	1.47	0.52	2.63
KeepTrack	1.54	0.62	2.62	1.52	0.61	2.72	1.54	0.49	2.52
ICTKF	1.41	0.69	2.40	1.34	0.42	2.26	1.53	0.80	2.87
LR tracking	1.31	0.63	2.23	1.03	0.41	1.99	1.10	0.40	1.78
TO system	1.34	0.57	2.06	1.12	0.42	1.73	1.11	0.45	1.72

The ICTKF, that integrates the Kalman filter shows advanced tacking performance than many other trackers in the motorized phantom and tissue experiments, but only ranked seventh in the manual experiments. This is mainly because the needle insertion velocity in the manual experiments often varied greatly in both magnitude and direction, rendering the constant motion model in ICTKF ineffective in capturing the real-time movement of the needle tip. While TO system and LR tracking both include a nonlinear motion filter, the former outperforms the latter in nearly all cases. The motion filter in LR tracking solely relies on interpretation of the historical position data to directly estimate the tracking results, easily leading to drift error when the visual tracking results are inaccurate for a short period of time. In contrast, the use of the TransKalman to estimate the gain matrix and the updating of the filtered results by following the filtering principles in the Kalman filter enables our proposed TO system to have stronger noise identification and suppression abilities.

The tracking errors of these trackers in both motorized and manual needle insertion in tissue experiments are shown in Table VIII and IX, respectively. It should be noted that the tracking errors are computed only based on successfully tracked (i.e., error $<$ 3mm) videos of each tracker, and thus the mean tracking error for each tracker in each scenario is always less than 3 mm. The TO system and the LR tracking offer the smallest mean errors in all scenarios in these tissue experiments. Both also offer the smallest standard deviations, especially in the manual needle insertion experiments, mainly because that they have integrated an effective motion filter that smooths out the tracking trajectory. It should also be observed that the 95 percentile error for the TO system is the smallest in nearly all scenarios, validating its most consistently accurate tracking performance.

Furthermore, TO system achieves a tracking speed of 23 fps, which is 46.5% higher than that of LR tracking (15.7 fps). The significantly improved tracking speed is achieved largely due to the compact network architectures in the task-oriented modules and the use of three independent 1×1 convolutional layers to obtain the position of the needle tip and the size of the bounding box in the visual tracking network. The high tracking speed of TO system facilitates its integration into US image-guided procedures that requires real-time position feedback of the needle tip.

G. Comparison with other state-of-the-art trackers

As most existing tracking algorithms in the field of US-guided needle tip tracking are not open sourced, and the datasets used to evaluate their trackers are commonly self-collected and not publicly available, a rough comparison of our proposed task-oriented visual tracking and motion filtering system (i.e. TO system) has been made with these state-of-the-art trackers according to the results reported in their works. The detailed information can be found in Table X.

The method in [47] obtains similar tracking error as our proposed TO system, at slightly above 1 mm. However, its tracking success rate and tacking speed are less than 45% and less than 2 fps, respectively, which are significantly lower than TO system. Methods in [21], [22], [48] detect the needle tip by leveraging the needle tremor motion, induced by the operator's hand holding the needle base. Tested with 3600 evaluation images, these methods achieve similar tracking success rate and tracking error as TO system. However, their tracking speeds are extremely low at less than 1 fps. These methods sequentially analyze the phase change of pixels in US images from several successive frames to obtain the tremor information. This computationally expensive process significantly lowers the achievable tracking speed. Therefore, these trackers were designed for needle detection instead of real-time tracking throughout a clinical procedure. The works in [2], [11] achieve small tracking errors, but they were only tested on small datasets in the idealized gelatin/agar-based phantoms, which feature significantly higher image quality and much less noisy background than the ex-vivo biological tissue under US imaging. They also do not offer acceptable tracking speed for real-time tracking. The tracker in [49] uses the appearance information of the needle axis to track the needle tip and shows 100% tracking success rate and real-time tracking performance (28.6 fps). However, these results are obtained from just 38 images in an agar-based phantom experiment. Furthermore, as highlighted previously, the needle axis in US images is usually not continuous and the needle tip can sometimes appear as an independent part. This phenomenon is extremely serious in challenging real-world cases where the background noise in US images is strong and the visibility of the needle axis is poor. In these cases, the appearance information-based trackers may not be able to correctly identify the entire needle axis, potentially leading to inaccurate detection of the needle tip. In general, except the tremor-based trackers, all the other state-of-the-art trackers have a strict requirement on visibility of the needle tip at all

TABLE X: Comparison with state-of-the-art US-based trackers

Tracker	Tracking clue	Experimental environment	Insertion mode	Needle tip visibility	Size of evaluation dataset	Success rate	Tracking error (mm)	Tracking speed
[47]	Intensity change of needle tip	Chicken breast tissue	Manual insertion	Yes	200	44%	1.04 ± 0.36	1.8 fps
[21]	Tremor	Porcine tissue	Manual insertion	No	3,600	81%	1.69 ± 1.54	-
[22]	Tremor	Porcine tissue	Manual insertion	No	60	100%	1.21 ± 1.47	0.85 fps
[48]	Tremor	Porcine tissue	Manual insertion	No	3,600	84%	1.78 ± 2.20	0.11 fps
[2]	Appearance of needle tip	Gelatin-based phantom	Motorized insertion	Yes	245	-	0.68 ± 0.21	7.1 fps
[11]	Appearance of needle axis	Agar-based phantom	Motorized insertion	Yes	20	-	0.7	0.46 fps
[49]	Appearance of needle axis	Agar-based phantom	Motorized insertion	Yes	38	100%	1.22 ± 1.22	28.6 fps
TO system	Appearance of needle tip and its position history	Pork/Chicken tissue	Manual insertion	No*	$> 250,000$	83.1%	1.19 ± 0.48	23 fps

* Temporary invisibility of the needle tip in US images can be handled by the TO system.

times, and cannot obtain any useful visual information of the target when the needle tip disappears from the US images for a short period of time, leading to complete tracking failure or a large drift error that cannot be recovered. Taking into account various important evaluation metrics simultaneously, our proposed TO system has shown to be the only real-time tracking system that achieves both state-of-the-art tracking success rate and tracking error under the most challenging environment (i.e. ex-vivo biological tissue) with more than 250,000 evaluation images (>400 needle insertion video sequences). It is also uniquely capable of performing robust and smooth tracking when the needle tip is temporarily not visible due to its incorporation of the TransKalman motion filter. It leverages the understanding of the the motion information of the needle tip to provide reasonable estimation during the needle tip disappearance in both linear and nonlinear needle motions, significantly increasing the tracking robustness to address different challenging scenarios in real-world practice.

VI. Conclusion

In this paper, a needle tip tracking system under US imaging has been developed, combining a visual tracker with various concise, task-specific modules and TransKalman as a motion filter for the visual tracking results. A comprehensive set of ablation and evaluation studies validates the significant improvement brought by the proposed individual modules in the visual tracker and TransKalman, and the superior tracking robustness and accuracy of the overall tracking system against other state-of-the-art trackers in both motorized and manual needle manipulation experiments in ex-vivo tissues. All these were achieved by the tracking system at a tracking speed more than 20fps, leading to its real-time tracking capability under US imaging. In the future work, the tracking system will be integrated with advanced anatomical target segmentation algorithm and an effective control strategy to develop an image guided surgical navigation framework, potentially enabling a fully automated needle-based percutaneous procedure.

VII. Acknowledgments

We would like to acknowledge Yuelin Zhang from the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong for his contribution to the experimental evaluation in this work.

References

[1] K. Mohiuddin and S. J. Swanson, "Maximizing the benefit of minimally invasive surgery," *Journal of surgical oncology*, vol. 108, no. 5, pp. 315–319, 2013.

[2] M. Kaya, E. Senel, A. Ahmad, and O. Bebek, "Visual needle tip tracking in 2D us guided robotic interventions," *Mechatronics*, vol. 57, pp. 129–139, 2019.

[3] C. Mwikirize, A. B. Kimbowa, S. Imanirakiza, A. Katumba, J. L. Nosher, and I. Hacihaliloglu, "Time-aware deep neural networks for needle tip localization in 2d ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 5, pp. 819–827, 2021.

[4] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Convolution neural networks for real-time needle detection and localization in 2d ultrasound," *International journal of computer assisted radiology and surgery*, vol. 13, no. 5, pp. 647–657, 2018.

[5] M. Kaya and O. Bebek, "Needle localization using gabor filtering in 2d ultrasound images," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 4881–4886.

[6] P. Chatelain, A. Krupa, and M. Marchal, "Real-time needle detection and tracking using a visually servoed 3d ultrasound probe," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1676–1681.

[7] S. H. Okazawa, R. Ebrahimi, J. Chuang, R. N. Rohling, and S. E. Salcudean, "Methods for segmenting curved needles in ultrasound images," *Medical image analysis*, vol. 10, no. 3, pp. 330–342, 2006.

[8] C. R. Hatt, G. Ng, and V. Parthasarathy, "Enhanced needle localization in ultrasound using beam steering and learning-based segmentation," *Computerized Medical Imaging and Graphics*, vol. 41, pp. 46–54, 2015.

[9] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Learning needle tip localization from digital subtraction in 2d ultrasound," *International journal of computer assisted radiology and surgery*, vol. 14, no. 6, pp. 1017–1026, 2019.

[10] W. Yan, Q. Ding, J. Chen, Y. Liu, and S. S. Cheng, "Needle tip tracking in 2d ultrasound based on improved compressive tracking and adaptive kalman filter," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3224–3231, 2021.

[11] A. Pourtaherian, F. Ghazvinian Zanjani, S. Zinger, N. Mihajlovic, G. Ng, H. Korsten *et al.*, "Improving needle detection in 3d ultrasound using orthogonal-plane convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 610–618.

[12] J. Y. Lee, M. Islam, J. R. Woh, T. Washeem, L. Y. C. Ngoh, W. K. Wong, and H. Ren, "Ultrasound needle segmentation and trajectory prediction using excitation network," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 3, pp. 437–443, 2020.

[13] Y. Zhang, X. Dai, Z. Tian, Y. Lei, Y. Chen, P. Patel, J. D. Bradley, T. Liu, and X. Yang, "Liver motion tracking in ultrasound images using attention guided mask r-cnn with long-short-term-memory network," in *Medical Imaging 2022: Ultrasonic Imaging and Tomography*, vol. 12038. SPIE, 2022, pp. 156–161.

[14] W. Yan, Q. Ding, J. Chen, K. Yan, R. S.-Y. Tang, and S. S. Cheng, "Learning-based needle tip tracking in 2d ultrasound by fusing visual tracking and motion prediction," *Medical Image Analysis*, vol. 88, p. 102847, 2023.

[15] G. J. Vrooijink, M. Abayazid, S. Patil, R. Alterovitz, and S. Misra, "Needle path planning and steering in a three-dimensional non-static en-

vironment using two-dimensional ultrasound images," *The International journal of robotics research*, vol. 33, no. 10, pp. 1361–1374, 2014.

[16] L. Zhou and J. Zhang, "Combined kalman filter and multifeature fusion siamese network for real-time visual tracking," *Sensors*, vol. 19, no. 9, p. 2201, 2019.

[17] M. Kaya, E. Senel, A. Ahmad, and O. Bebek, "Visual tracking of multiple moving targets in 2d ultrasound guided robotic percutaneous interventions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1996–2002.

[18] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 448–10 457.

[19] H. Saribas, H. Cevikalp, O. Köpüklü, and B. Uzun, "Trat: Tracking by attention using spatio-temporal features," *Neurocomputing*, vol. 492, pp. 150–161, 2022.

[20] P. Huang, G. Yu, H. Lu, D. Liu, L. Xing, Y. Yin, N. Kovalchuk, L. Xing, and D. Li, "Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking," *Medical physics*, vol. 46, no. 5, pp. 2275–2285, 2019.

[21] P. Beigi, R. Rohling, T. Salcudean, V. A. Lessoway, and G. C. Ng, "Detection of an invisible needle in ultrasound using a probabilistic svm and time-domain features," *Ultrasonics*, vol. 78, pp. 18–22, 2017.

[22] P. Beigi, R. Rohling, S. E. Salcudean, and G. C. Ng, "Casper: computer-aided segmentation of imperceptible motion—a learning-based tracking of an invisible needle in ultrasound," *International journal of computer assisted radiology and surgery*, vol. 12, no. 11, pp. 1857–1866, 2017.

[23] P. Beigi, S. E. Salcudean, G. C. Ng, and R. Rohling, "Enhancement of needle visualization and localization in ultrasound," *International journal of computer assisted radiology and surgery*, vol. 16, pp. 169–178, 2021.

[24] E. Ayvali and J. P. Desai, "Optical flow-based tracking of needles and needle-tip localization using circular hough transform in ultrasound images," *Annals of biomedical engineering*, vol. 43, pp. 1828–1840, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8126–8135.

[28] A. Sauer, E. Aljalbout, and S. Haddadin, "Tracking holistic object representations," *arXiv preprint arXiv:1907.12920*, 2019.

[29] B. Gao and M. W. Spratling, "Explaining away results in more robust visual tracking," *The Visual Computer*, pp. 1–15, 2022.

[30] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4660–4669.

[31] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[33] G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. Van Sloun, and Y. C. Eldar, "Kalmannet: Neural network aided kalman filtering for partially known dynamics," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1532–1547, 2022.

[34] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 300–317.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[36] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5374–5383.

[37] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[39] B.-M. He, R. Chen, Z.-K. Shi, G.-A. Xiao, H.-S. Li, H.-Z. Lin, J. Ji, H.-X. Peng, Y. Wang, Y.-H. Sun *et al.*, "Trans-perineal template-guided mapping biopsy vs. freehand trans-perineal biopsy in chinese patients with psa¡ 20 ng/ml: similar cancer detection rate but different lesion detection rate," *Frontiers in Oncology*, vol. 9, p. 758, 2019.

[40] H. Zhang, F. Banovac, A. White, and K. Cleary, "Freehand 3d ultrasound calibration using an electromagnetically tracked needle," in *Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display*, vol. 6141. SPIE, 2006, pp. 775–783.

[41] G. Ploussard, J. I. Epstein, R. Montironi, P. R. Carroll, M. Wirth, M.-O. Grimm, A. S. Bjartell, F. Montorsi, S. J. Freedland, A. Erbersdobler *et al.*, "The contemporary concept of significant versus insignificant prostate cancer," *European urology*, vol. 60, no. 2, pp. 291–303, 2011.

[42] M. Danelljan, L. V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7183–7192.

[43] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6668–6677.

[44] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "Aiatrack: Attention in attention for transformer visual tracking," in *European Conference on Computer Vision*. Springer, 2022, pp. 146–164.

[45] C. Mayer, M. Danelljan, D. P. Paudel, and L. Van Gool, "Learning target candidate association to keep track of what not to track," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 444–13 454.

[46] Y. Zhao, C. Cachard, and H. Liebgott, "Automatic needle detection and tracking in 3d ultrasound using an roi-based ransac and kalman method," *Ultrasonic imaging*, vol. 35, no. 4, pp. 283–306, 2013.

[47] C. Mwikirize, I. Nosher, Hacihaliloglu, and J. L., "Convolution neural networks for real-time needle detection and localization in 2D ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 5, pp. 647–657, 2018.

[48] P. Beigi, S. Rohling, S. E. Robert, and G. C. Ng, "Spectral analysis of the tremor motion for needle detection in curvilinear ultrasound via spatiotemporal linear sampling," *International journal of computer assisted radiology and surgery*, vol. 11, pp. 1183–1192, 2016.

[49] M. Kaya, E. Senel, A. Ahmad, O. Orhan, and O. Bebek, "Real-time needle tip localization in 2d ultrasound images for robotic biopsies," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 47–52.