2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

# Visual Tracking of Needle Tip in 2D Ultrasound based on Global Features in a Siamese Architecture

Wanquan Yan, Qingpeng Ding, Jianghua Chen, Kim Yan, Raymond Shing-Yan Tang, and Shing Shin Cheng*

*Abstract*— **Ultrasound (US) is widely used in image-guided needle procedures. Correctly tracking the needle tip position in US images during the procedure plays an important role in improving the needle targeting accuracy and patient safety. This paper presents a leaning-based visual tracking network with a Siamese architecture, which makes full use of the attention mechanism to explore the potential of global features and takes advantage of an online target model prediction module to robustly track the needle tip in US images. Several self- and cross-attention modules are applied to learn global features from the whole US image. A discriminative target model is also learned as a complementary part to improve the discriminability of the proposed tracker. The template used during the tracking is updated frequently according to the tracking results to ensure that the tracker can always capture the latest characteristics of the appearance of the needle tip. Experimental results in both phantom and tissue showed that the proposed tracking network was more robust than other state-of-the-art visual trackers. The mean success rates of the proposed tracker are 7.1% and 9.2% higher than the second best performing visual tacker when the needle was inserted by motors and human hands in the tissue experiments.**

## I. INTRODUCTION

Needle-based percutaneous operations are often performed in minimally invasive surgeries (MIS) to deliver drugs, ablate tumors and perform biopsy. Knowing the accurate position of the needle tip in real time during surgical intervention is highly desired by surgeons, as it can help precise targeting and improve patient safety. To achieve this goal, the tip of the inserted needle should be tracked and imaged in real time. The widely used imaging modality includes: magnetic resonance imaging (MRI), computed tomography (CT), fluoroscopy and ultrasound (US). Taking the cost efficiency, exposure level to ionizing radiation, and real-time capability into consideration, ultrasound (US) is the best choice to provide intraoperative images. However, US imaging offers relatively low contrast and poor imaging quality, which will adversely impact US image-based target tracking.

Conventional visual tracking methods track the needle tip by using line detection algorithms. Morphological operations, including erosion and dilation, random sample consensus (RANSAC) and Gabor filter, are usually used to enhance the appearance of the needle axis at first. Then, the needle axis is segmented from the US images [1], [2] before other tip detection methods, including the circular Hough transform and blob detection, are applied to locate the needle tip [3], [4]. These methods are not robust when the needle tip cannot be clearly distinguished from the various background distractors. In practice, this situation often occurs, especially when the needle tip is inserted into biological tissues which features much background noise and speckles under US imaging. Kaya et al. [5] located the needle tip position by measuring the similarity between the current US frame and a needle template manually selected at the beginning of the tracking. This method directly uses the raw pixel values to calculate the similarity value without considering any additional delicate feature extraction. Our previous work [6] uses a tailored compressive tracking algorithm to track the needle tip. It relies on Harr-like features to determine the most similar region in the search images. The template matching and compressive tracking algorithms showed comparable performance in US image-based needle tip tracking.

Recently, deep features, especially convolutional features, have also been used in US image-based trackers. In [7], [8], they are used to find the region where the needle axis is located at before enhancement methods are employed to segment the needle axis. In other works [9]–[11], UNets are used to directly segment the needle axis without further post-processing. However, the needle axis is not always visible in US images. Even when the needle axis can be found, the appearance of the needle axis is often disconnected [7]. The invisibility and discontinuity of the needle axis hugely undermine the needle axis segmentation-based tip tracking algorithms. A few recent works attempt to enhance the needle tip appearance before applying the convolutional neural network (CNN) to locate the target position [12], [13] . However, their networks only contain a single branch, and the tracking results solely depend on the needle tip target model trained offline. As the real-time appearance of the needle tip in US images changes dramatically, these models are often unable to achieve satisfactory results because sampling a host

Research reported in this work was supported in part by Innovation and Technology Commission of Hong Kong (ITS/136/20, ITS/135/20, ITS/233/21, and ITS/234/21), in part by Research Grants Council (RGC) of Hong Kong (T45-401/22-N, CUHK 14217822), and in part by The Chinese University of Hong Kong (CUHK) Direct Grant. The content is solely the responsibility of the authors and does not reflect the views of the sponsors.

Wanquan Yan, Qingpeng Ding, Jianghua Chen, and Kim Yan are 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. *email: sscheng@cuhk.edu.hk

of different needle tip appearances in the offline training dataset is highly challenging in practice. In [14], [15], the long short-term memory (LSTM) networks are used to track the target by supplying the networks with multiple recent frames to learn additional cross-frame information. However, they rely on pre-operations to suppress the background noise and highlight the target. The networks used in these works are of basic architecture with weak feature extraction and background distractor suppression capability.

Based on the previous works and discussions, the challenges in US image-based needle tip tracking can be summarized as follows: 1) The US images contain a lot of distractors, including background noise, strong artifacts, and bright speckles, especially when imaging biological tissues. These characteristics severely limits clear visualization of the tiny needle tip target. 2) The appearance of the needle tip changes dramatically under US imaging, making it difficult to learn a robust appearance model to correctly represent the target's current appearance. The correct appearance model throughout the tracking process is a crucial cue for visual trackers to identify the needle tip's position through similarity comparison. 3) The needle tip may temporarily disappear from the US images due to occlusions caused by anatomic structures, and misalignment between US probe and needle axis. It should be noted that this third challenge cannot be addressed by a sole visual tracker.

In this work, we focus on developing a robust visual tracker with features that specifically address the first two challenges. The proposed tracker, taking advantages of attention modules and an online target model prediction module, is able to learn the global context from the entire image to track the needle tip under US imaging robustly. The network structure and loss function design also allow effective training process. Our contributions in this work include: 1) Proposing a Siamese neural network-inspired visual tracker for needle tip tracking. Unlike the single-branched networks in the existing works, the twin network architecture contains a template and a search branch. It learns object descriptors, instead of the specific target model, during training and locates the target through similarity measurement according to the cue provided in the template. The network thus can be trained on non-target specific datasets and then transferred to the US image-based needle tip tracking. This is important in medical image-based applications, where well-labelled datasets are scarce. 2) Using multiple self- and cross-attention modules to learn global contexts from the whole input US image. The noise in the whole US image, including regions near and far from the needle tip target, usually have similar statistic characteristics. The attention modules can learn global relationships between these features from the entire image to improve the tracker's ability to distinguish the target from the noisy background. 3) Introducing an online target model prediction module to predict a discriminative target model to increase the discriminability of the tracking network, and updating the template frequently to integrate the target's latest appearance into feature extraction for the target model. The rest of this work is organized as follows. In Section II, the details of the network structure are introduced and the related calculations in each module are defined. In Section III, the neural network training, experiment setup and coordinate registrations are provided. The experimental results and discussions are described in Section IV before a few conclusions are made in Section V.

## II. Network architecture

### A. Feature extraction block

An overview of the proposed tracking network is shown in Fig. 1. It contains two separate branches, a template branch (indicated by the blue arrows) and a search branch (indicated by the yellow arrows) to track the target. The inputs to these two branches are two independent image patches of different sizes, $\boldsymbol{z} \in \mathbb{R}^{3\times h\times w}$ and $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$. These two branches share the parameters of a feature extraction block which comprises a backbone network and a channel reduction layer. The first four stages of the ResNet50 [16] are used as the backbone network, with the convolution stride of the down-sampling layer in the fourth stage changing from 2 to 1 to obtain a high-resolution feature map. Furthermore, the $3 \times 3$ convolutions in the fourth stage are also modified to dilated convolutions with a stride of 2 to increase the receptive field. The extracted feature maps of these two branches are of sizes of $\hat{\boldsymbol{z}} \in \mathbb{R}^{C\times \frac{h}{s}\times \frac{w}{s}}$ and $\hat{\boldsymbol{x}} \in \mathbb{R}^{C\times \frac{H}{s}\times \frac{W}{s}}$ respectively, where $s$ is the scaling factor and $C$ is the channel number. Referring to the original ResNet50, after modification, the total output stride of the backbone network is reduced to 8 ($s = 8$), and $C = 1024$. The extracted feature maps then pass through a $1\times1$ convolutional layer to shrink the number of channels from $C = 1024$ to $d = 256$ to reduce the computational complexity.

### B. Attention mechanism

The attention module has the ability to find the inner relationships among features from every position of the feature map. It is thus suitable to learn the global contexts from the entire image. Inspired by the work in [17], the designed global feature aggregation block contains multiple self- and cross-attention modules. Specifically, two self-attention modules and a cross-attention module in each branch are used with an additional cross-attention module to fuse the features from the two branches at the end, as shown in Fig. 1. According to [18], multi-head attention can help the model to access the information from different representation subspaces, facilitating to learn a rich feature embedding. Therefore, multi-head attention was used in our module to learn features from various aspects. Given three matrices, namely query $Q$, key $K$ and value $V$, the attention is defined as the weighted sum of $V$:

$$\text{Atten}\left[Q, K, V\right] = \text{Softmax}\left(QK^T/\sqrt{d_k}\right)V \tag{1}$$

where $d_k$ is the dimension of the $K$ matrix. The Softmax operation is performed on each row such that the elements in each row act as weights of the value matrix.

**Self-attention:** The self-attention module is designed to extract global features from the feature map to increase the ability to discriminate the needle tip target from complex

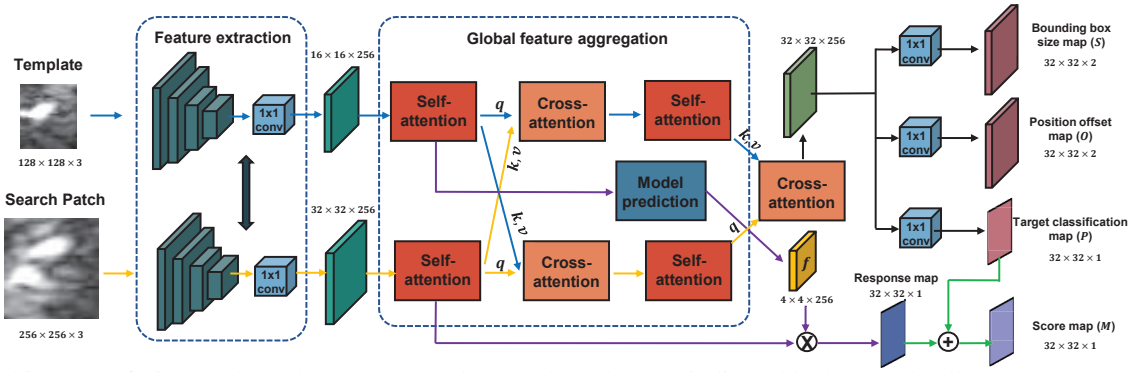


Fig. 1: Architecture of SiamGlobal. The template and search branches are indicated by blue and yellow arrows separately. The purple arrows show the pipeline of the model prediction, and the black arrows denote the classification and regression tasks.

background noise and strong distractors in the US images. The self-attention of the $i$-th head is defined as

$$S_{att_i} = \text{Atten}\left[\left(X + P_q\right) W_i^Q, \left(X + P_k\right) W_i^K, XW_i^V\right] \quad (2)$$

where $X \in \mathbb{R}^{N_x \times d}$ is the input matrix, $N_x$ is the multiplication of the width and height of the input feature map. $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$ and $W_i^V \in \mathbb{R}^{d \times d_v}$ are learnable weight matrices, and $d_v$ is the dimension of the $V$ matrix. $P_q$ and $P_k$ are spatial positional encodings of the $Q$ and $K$ matrices, and are generated according to [19]. The corresponding multi-head self-attention is defined as

$$S_{multihead} = \text{Concat}\left(S_{att_1}, S_{att_2}, \cdots, S_{att_h}\right) W^O \quad (3)$$

where $h$ is the number of heads and is set to 8, $W^O \in \mathbb{R}^{hd_v \times d}$. The attention of each head is concatenated along the channel dimension. As shown in Fig. 2, the multi-head self attention is then added with the input $X$ to form a residual connection. After that, a layer normalization is performed.

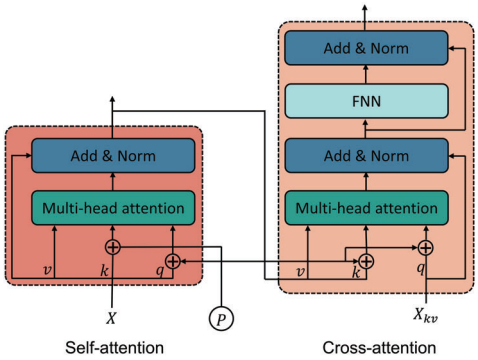


Fig. 2: Architecture of a self- and a cross-attention module. The cross-attention module illustrated in the figure located at the different branch with the self-attention module.

**Cross-attention:**This module is designed to aggregate the target features from both the template and the search branches. With the input from the same branch $X_{kv}$ and the other branch $X_q$ known, the cross-attention of the $i$-head is defined as

$$C_{att_i} = Atten\left[\left(X_q + P_q\right) W_i^Q, \left(X_{kv} + P_k\right) W_i^K, X_{kv} W_i^V\right] \quad (4)$$

Following the same way shown in Eq. 3, the multi-head cross-attention, $C_{multihead}$, can also be obtained. As shown in Fig. 2, different from the self-attention module, the cross-attention module has one more feed-forward network (FFN) followed by another residual connection and layer normalization. The FFN consists of two $1 \times 1$ convolutional layers with a ReLU activation functions in between. Here, $d_k = d_v = d/h$.

### C. Learning a discriminative target model

To enhance the noise suppression ability of the tracking network, an online target model prediction module is introduced. This module tries to predict a discriminative target model $\boldsymbol{f}$ by minimizing a discriminative learning loss:

$$L\left(\boldsymbol{f}\right) = \|r\left(\tilde{z} * \boldsymbol{f}, c\right)\|^2 + \lambda \|\boldsymbol{f}\|^2 \quad (5)$$

where $\tilde{z}$ is the extracted target features (the output of the first self-attention module in the template branch in our tracking network), $c$ is the ground-truth position of the target in the template image, $*$ denotes the convolutional operation, and $\lambda$ is a regularization factor. $r$ is a residual function that calculates the residual in every spatial location according to the convolutional result and the target's ground-truth position. The method proposed in [20] is adopted to minimize the loss function to obtain the target model $f$ online. The target model is represented as a kernel $\boldsymbol{f} \in \mathbb{R}^{K \times K \times d}$ of a convolutional layer, where $K$ is the size of the convolutional kernel ($K$ is set as 4 in our tracking network) and $d$ is the channel number of $\tilde{z}$. With the target model obtained, it is then convolved with the search feature, $\tilde{x}$ (output of the first self-attention module in the search patch) to obtain a response map, $\boldsymbol{R} = \tilde{x} * \boldsymbol{f}$, where $R \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 1}$. During the inference stage, the template is updated every time the peak value of the score map (see Subsection II-D) is larger than a predefined threshold $\delta$ to ensure high-quality target features are always stored in the template branch in the case of unclear or completely disappeared targets. During this template update, the target model is re-predicted to keep up-to-date with the needle tip's current appearance in the US images.

### D. Target classification and bounding box estimation

Instead of estimating the offset of the target position and the size of the bounding box in one regression head [21], here, the two tasks are estimated independently by two different heads (as shown in Fig. 1) such that each task can be evaluated by separate loss functions, allowing them to be trained effectively. A classification head is used to find the target position in

the output feature map. Each of these three heads contains a $1 \times 1$ convolutional layer followed by a sigmoid activation layer. The bounding box and position offset regression heads output two feature maps $S \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 2}$ and $O \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 2}$ to indicate the size of the bounding box and the offset of the target position. The output of the classification head is post-processed by adding a cosine window to suppress large displacement. The masked classification map is denoted as $P \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 1}$. The classification and response maps obtained in Subsection II-C are added together as the final score map:

$$M = (1 - w) P + w R \tag{6}$$

where $w$ is a weight to adjust the contributions of the target prediction module (here, $w$ is set as 0.4), and $M \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 1}$. The target position in the search path is determined by:

$$(x_c, y_c) = s\,(O\,(\arg\max(M)) + \arg\max(M)) \tag{7}$$

where $\arg\max(M)$ outputs the 2D position of the maximum value in the score map $M$. The size of the target's bounding box in the current search path is

$$(w_{bb}, h_{bb}) = (W, H) \cdot S\,(\arg\max(M)) \tag{8}$$

where $(\cdot)$ denotes element-wise multiplication.

### E. Loss function

The loss function proposed in [22] is referred to develop the training loss of the proposed tracking network. The training loss contains three different components: positioning error, $L_p$, offset error, $L_o$, and size prediction error of the bounding box $L_s$. Design of the positioning error follows the principle of focal loss [23] to solve the unbalanced number of positive and negative samples and focus on the hard-to-classify samples. Assuming that the ground-truth position of the target in the search patch is $\boldsymbol{p} \in \mathbb{R}^2$, a low resolution equivalent is calculated as: $\hat{\boldsymbol{p}} = \lfloor \frac{\boldsymbol{p}}{s} \rfloor$, where $\lfloor \cdot \rfloor$ denotes floor operation. Then, a Gaussian function centered at $\hat{\boldsymbol{p}}$,

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

is used to compare with the values in the score map, where $\sigma_p$ is a size-adaptive standard deviation [24]. The final expression of the positioning error is

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

where $M_{xy}$ is the score at position $(x, y)$ in the score map $M$. $\alpha$ and $\beta$ are hyper-parameters set as 2 and 4, respectively [24].

The position offset and the size of the bounding box are both trained by using a L1 loss function. The corresponding errors are defined as

$$L_o = \left|O_{\hat{\boldsymbol{p}}} - (\boldsymbol{p}/s - \hat{p})\right|, \quad L_s = \left|S_{\hat{\boldsymbol{p}}} - \hat{s}\right| \tag{11}$$

where $\hat{s} = \left(\frac{w_{gt}}{W}, \frac{h_{gt}}{H}\right)$, and $w_{gt}$ and $h_{gt}$ refer to the ground-truth size of the target's bounding box. Only errors in the ground-truth position are counted.

The total training loss is defined as:

$$L_{train} = L_p + \lambda_o L_o + \lambda_s L_s \tag{12}$$

where $\lambda_o$ and $\lambda_s$ are weights and both are set to 1.

## III. Network tracking and Experiment preparation

### A. Visual network training

The ResNet50 used in the feature extraction block was pretrained on the ImageNet [25]. The entire proposed visual tracking network was trained by using training splits of LaSOT [26], COCO [27], GOT-10K [28], TrackingNet [29] and ImageNet VID [25]. There were a total of 80 epochs during training, and each epoch contained 125 iterations, with the batch size being 40. The initial learning rate was set as 0.01 and decayed by 0.2 every 15 epochs. After training on these common datasets, the network was fine-tuned on a self-collected US image dataset, which contains more than 3,000 images collected from different experimental scenarios. During fine-tuning, there were a total of 40 epochs with only 75 iterations in each epoch. The initial learning rate was set as 0.0001 and the decay factor was 0.35. The sizes of the template and search patch are set to $128 \times 128$ and $256 \times 256$, which are smaller than the size of the original US images ($856 \times 492$), to ensure high computational efficiency. Our proposed tracking network was operated at around 20 frames per second (FPS) on a PC with Intel (R) Xeon(R) W-2102 CPU and RTX 5000 GPU to track the needle tip in all the experiments.

### B. Experimental Setup

To test the robustness of the proposed tracking network, the needle tip was tracked when the needle was inserted by a motorized linear stage and by human hands. An overall setup of the motorized needle insertion experiment is shown in Fig. 3. It mainly consists of a US machine (Vantage 32 LE, Verasonics, Inc., USA.) with a US probe (C5-2, Mindary, Inc., China), a 3-axis manipulator (LiTai Inc., China), a linear stage (LiTai Inc., China), an electromagnetic tracking system (NDI Inc., USA), an 18 gauge needle tool (Aurora needle, NDI Inc., USA), and a soft tissue gelatin phantom. The phantom was made according to the recipe in [30] with silica powder added to simulate the bright speckles caused by the anatomic structures in the US images. Chicken and pork were also used as the biological tissues in the experiments. No obvious differences in these two biological tissues were observed in terms of imaging quality, and all images acquired from them were used in the experiments.

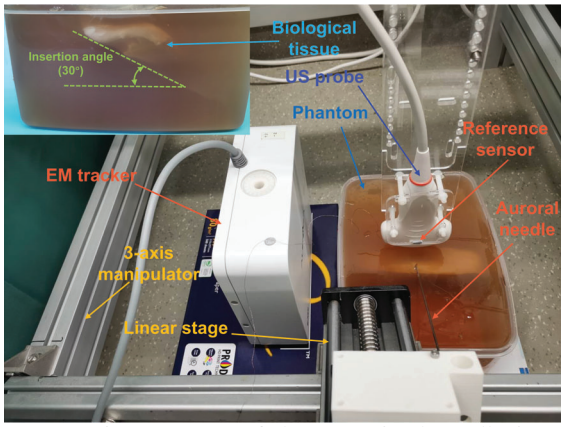


Fig. 3: Experiment setup of the motorized needle insertion experiments. (Inset: tissue experiment)

### C. Coordinate registration

The EM tracking system was used to provide the ground-truth position of the needle tip. To achieve this goal, a 6 DOF EM sensor was attached on the needle tip and an additional EM sensor was fixed on the US probe as a reference sensor. The needle tip's coordinates in the EM tracking frame $\boldsymbol{p}^{EM}$ can then be expressed as

$$\boldsymbol{p}^{EM} = T_{ref}^{EM} T_{US}^{ref} \boldsymbol{p}^{US} \tag{13}$$

where $\boldsymbol{p}^{US}$ is the tracked result (needle tip's coordinates) of our visual tracker. $\boldsymbol{p}^{EM}$ and the US probe reference coordinates $T_{ref}^{EM}$ can be directly read from the EM tracking system. $T_{US}^{ref}$ is the transformation matrix that connects the US frame to the EM tracking frame. Before the experiment, it was pre-calibrated according to the method proposed in [31]. With $T_{US}^{ref}$ known, $\boldsymbol{p}^{US}$ can be transformed to the EM tracking frame. The transformed values are compared with the ground truth $\boldsymbol{p}^{EM}$ to calculate the tracking error.

## IV. Experimental results and discussions

### A. Experimental procedures

During motorized needle insertion experiments, the US probe was held and moved by the 3-axis manipulator and the needle was inserted by a linear stage. The experiments were conducted in three different scenarios, namely In-plane-static, In-plane-moving and Out-of-plane. In the In-plane scenarios, the imaging plane of the US probe was parallel to the needle axis, while it was perpendicular to the needle axis in the Out-of-plane scenario. The needle stayed static in the In-plane-static scenario, and moved in the other two scenarios by following the movement of the US probe. In each scenario, the needle was inserted in three different angles (as shown in Fig. 3), namely 0°, 30° and 60°. A scenario with an insertion angle was called a case, and there are totally nine cases. During the experiments, each case was repeated at least nine times and the needle tip was tracked in both insertion and withdrawal processes. The needle tip was both tracked in the gelatin phantom and biological tissues. It should be noted that in the tissue experiments, a biological tissue was embedded inside the gelatin phantom, and the needle was inserted to puncture through the tissue. There are a total of 114 and 120 video sequences in the phantom and tissue experiments, respectively. During the manual insertion experiment, the US probe was held by human hands, and the needle was only inserted in the In-plane-static scenario with the biological tissue embedded in the gelatin phantom. As the insertion angle could not be precisely controlled by human hands in the manual insertion experiment, the insertion was roughly performed with three different insertion angles (i.e. small, medium, and large). The manual insertion experiments were performed by three different users and repeated about 270 times with different insertion angles. The experimental data, including US images and the ground-truth positions of the needle tip, were saved. Some state-of-the-art trackers, including ICTKF [6], template matching (TM) [5], Siammask [32], Tomp-101 [33], Stark [34], TrTr [19], TransT [17], were then run on this dataset to compare with our proposed visual tracking network. ICTKF and TM are state-of-the-art trackers used in US image-based target tracking. The rest are state-of-the-art trackers developed for object tracking in natural images, and have been fine-tuned in the same way described in Subsection III-A before being used for needle tracking in this work. These natural object trackers are used for comparison because the implementation codes of most deep learning trackers developed for US image-based target tracking and their training datasets are not publicly available, preventing accurate and fair comparison. It is worth noting that the ICTKF is a hybrid tracker proposed in our previous work [6], consisting of not only a visual tracker (Improved compressive tracking, ICT) but a Kalman filter (KF)-based motion predictor. All the other trackers, including our proposed tracker (SiamGlobal), are all visual tracks with no inter-frame motion information added.

### B. Experimental results

Some representative tracking results for three different cases are shown in Fig. 4. The subfigures in the first row show that there are complex background noises around the target and that the needle tip was nearly indistinguishable when it was inserted into the biological tissue. The needle tip was completely camouflaged in the noisy surroundings. The small patches in the top-right corner show the appearance of the needle tip in the current frame. It can also be found that the appearance of the needle tip experienced dramatic changes as the insertion continued. Although the noisy background posed an extremely difficult challenges to the needle tip tracking, our proposed tracker still robustly tracked the needle tip and correctly located it throughout the insertion process. The subfigures in the second row were collected from a phantom experiment, of which the background was relatively clean. However, since the appearance of the needle axis was discontinuous and the small segments of the needle axis had a highly similar appearance to the needle tip, the tracker was very likely to be trapped in these segments and stops moving with the needle tip. Subfigures in the third row show the tracking process in the Out-of-plane scenario during the tissue experiments. The contrast between the needle tip and the background was extremely low, and the needle tip could not be observed in some frames. This really challenges the discriminability of the trackers, which must have strong feature extraction abilities such that it can get sufficient context information and learn a robust target model. Our proposed tracker was shown to successfully tackle these challenges and correctly track the needle tip even in some of the most challenging conditions.

To quantitatively evaluate our proposed visual tracker, the tracking success rate of each case was defined. The needle tip in a video sequence is regarded as successfully tracked only if the tracking error of more than 95% of the frames in this video sequence is less than $\delta_e$. The percentage of successfully tracked videos to the total number of videos in the case is defined as the success rate of the tracker in this case. $\delta_e$ was set to 3 mm, which is in the range of generally acceptable targeting errors in clinical procedures [35]. The tracking success rates of the proposed tracker (SiamGlobal) and other state-of-the-

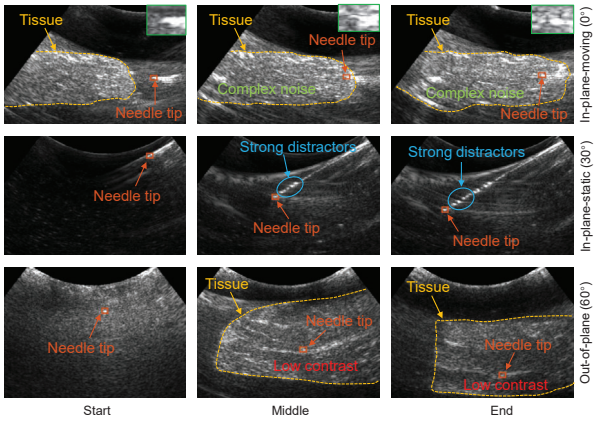


Fig. 4: Needle tip tracking by the proposed visual tracker in three different cases.

art trackers were compared for the motorized phantom and tissue experiments, as well as manual experiments, as shown in Tables I, II, and III respectively. The highest success rate in each case is highlighted in red, and the second and third best performing trackers are highlighted in green and blue.

TABLE I: Needle tip tracking success rates of different trackers for motorized needle operation in the phantom experiments

| Scenario | In-plane-static | | | | In-plane-moving | | | | Out-of-plane | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0° | 30° | 60° | *Mean* | 0° | 30° | 60° | *Mean* | 0° | 30° | 60° | *Mean* | |
| ICTKF | 80.1% | 81.5% | 81.7% | 81.1% | 86.3% | 90.1% | 93.1% | 90.2% | 100.0% | 100.0% | 100.0% | 100.0% | 90.4% |
| TM | 41.7% | 38.5% | 54.5% | 44.9% | 66.7% | 80.0% | 79.9% | 71.5% | 83.3% | 100.0% | 91.7% | 91.7% | 69.4% |
| Siammask | 50.0% | 38.5% | 25.0% | 37.8% | 75.0% | 86.7% | 100.0% | 87.2% | 100.0% | 100.0% | 91.7% | 97.2% | 74.1% |
| Tomp101 | 16.7% | 46.2% | 58.3% | 40.4% | 75.0% | 86.7% | 100.0% | 87.2% | 91.7% | 100.0% | 100.0% | 97.2% | 74.9% |
| TrTr | 33.3% | 46.2% | 66.7% | 48.7% | 66.7% | 93.3% | 100.0% | 86.7% | 100.0% | 91.7% | 100.0% | 97.2% | 77.5% |
| TransT | 66.7% | 76.9% | 58.3% | 67.3% | 66.7% | 73.3% | 100.0% | 80.0% | 100.0% | 100.0% | 91.7% | 97.2% | 81.5% |
| SiamGlobal | 58.3% | 69.2% | 83.3% | 70.3% | 58.3% | 93.3% | 100.0% | 83.9% | 100.0% | 100.0% | 100.0% | 100.0% | 84.7% |

TABLE II: Needle tip tracking success rates of different trackers for motorized needle operation in the tissue experiments

| Scenario | In-plane-static | | | | In-plane-moving | | | | Out-of-plane | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0° | 30° | 60° | *Mean* | 0° | 30° | 60° | *Mean* | 0° | 30° | 60° | *Mean* | |
| ICTKF | 41.4% | 64.1% | 73.3% | 59.6% | 63.5% | 54.4% | 70.2% | 62.7% | 86.2% | 82.9% | 94.5% | 87.9% | 70.1% |
| TM | 21.4% | 0.0% | 28.6% | 16.7% | 26.7% | 35.7% | 35.7% | 32.7% | 63.6% | 38.5% | 83.3% | 61.8% | 37.1% |
| Siammask | 57.1% | 54.5% | 68.9% | 60.2% | 50.0% | 71.4% | 35.7% | 52.4% | 100.0% | 70.0% | 91.7% | 87.2% | 66.6% |
| Tomp101 | 42.4% | 36.4% | 43.8% | 40.9% | 41.7% | 42.9% | 28.6% | 37.7% | 45.5% | 30.0% | 83.3% | 52.9% | 43.8% |
| TrTr | 57.1% | 45.5% | 72.7% | 58.5% | 58.3% | 85.7% | 42.9% | 62.3% | 90.9% | 70.0% | 91.7% | 84.2% | 68.3% |
| TransT | 35.7% | 36.4% | 37.5% | 36.5% | 41.7% | 85.7% | 35.7% | 54.4% | 36.7% | 50.0% | 75.0% | 53.9% | 48.3% |
| SiamGlobal | 57.1% | 63.6% | 90.9% | 70.5% | 71.1% | 85.7% | 57.1% | 71.3% | 90.9% | 70.0% | 91.7% | 84.2% | 75.4% |

TABLE III: Needle tip tracking success rates of different trackers under the "In-plane-static" scenario in the manual needle operation at small, medium, and large needle insertion angles in the tissue by three users

| Case | User 1 | | | | User 2 | | | | User 3 | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Small | Medium | Large | *Mean* | Small | Medium | Large | *Mean* | Small | Medium | Large | *Mean* | |
| ICTKF | 20.0% | 61.9% | 51.9% | 44.6% | 46.9% | 67.9% | 60.5% | 58.4% | 52.9% | 78.1% | 73.1% | 68.0% | 57.0% |
| TM | 42.3% | 37.5% | 44.7% | 41.5% | 51.5% | 46.2% | 61.3% | 53.0% | 46.2% | 54.1% | 76.6% | 59.0% | 51.2% |
| Siammask | 41.4% | 59.5% | 51.9% | 50.9% | 81.3% | 67.9% | 68.4% | 72.5% | 62.5% | 62.5% | 73.1% | 66.0% | 63.2% |
| Tomp101 | 37.2% | 38.1% | 25.9% | 33.8% | 43.8% | 60.7% | 34.2% | 46.2% | 37.5% | 31.3% | 53.8% | 40.9% | 40.3% |
| TrTr | 48.3% | 69.0% | 44.4% | 53.9% | 62.5% | 71.4% | 73.7% | 69.2% | 50.0% | 75.0% | 76.9% | 67.3% | 63.5% |
| TransT | 48.3% | 52.4% | 37.0% | 45.9% | 84.4% | 67.9% | 44.7% | 65.7% | 43.8% | 81.3% | 57.7% | 60.9% | 57.5% |
| SiamGlobal | 55.2% | 78.6% | 59.3% | 64.3% | 84.4% | 75.0% | 73.7% | 77.7% | 62.5% | 81.3% | 84.6% | 76.1% | 72.7% |

Referring to the tables above, the proposed tracker (SiamGlobal) obtains the highest success rates in most of the cases in the motorized-tissue experiments and all cases in manual-tissue experiments among all the visual trackers (i.e. ICTKF is not considered). The phantom experiment has relative good image quality, and the needle tip and its boundary can be clearly found in US images. Therefore, other state-of-the-art visual trackers performed equally well in some cases, especially in the Out-of-plane scenario where most trackers obtain 100% tracking success rate. The superiority of SiamGlobal over other state-of-the-art visual trackers is obvious in motorized and manual insertion in tissue experiments. In motorized-tissue experiments, the mean success rates of SiamGlobal was 10.3% and 9.0% higher than those of the second best performing visual tracker in the In-plane-static and In-plane-moving scenarios. They are also 10.4%, 5.2% and 9.2% higher than those of the second best performing visual tracker in the manual insertion experiments for Users 1, 2, and 3, respectively. To sum it up, the mean success rate of SiamGlobal was 3.2%, 7.1% and 9.2% higher than the second best performing visual tracker in all cases.

The Tomp101, TrTr and TransT were also based on the Transformer neural network, but our proposed tracker outperformed them to a large extent. Tomp101 ranks highest among these trackers in natural object tracking datasets, but it performed poorly in needle tip tracking experiments. This could be due to that the target state and test frame encoding used in the Tomp101 were not suitable for US images. Compared with TrTr, the proposed SiamGlobal applied multiple self- and cross-attention modules and thus have stronger global feature learning ability. TransT lacks a template update pipeline, and may not be able to adapt to the changing appearances of the needle tip. TM performed worst, although its template was updated. This is mainly because the raw pixel values used in this method have limited feature expression ability. Siammask obtains relatively good success rate, likely because that it also uses the global features during mask prediction.

The ICTKF showed competitive tracking ability in all three experiments, especially the phantom experiment. The KF-based motion prediction capability in ICTKF enhances its tracking performance over other visual trackers during temporary disappearance of the needle tip. Even without the motion prediction capability, our proposed SiamGlobal tracker with its highly superior visual tracking features showed competitive performance compared with ICTKF, especially in the biological tissue experiments as shown in Tables II and III.

## V. Conclusion and Future Work

In this paper, we present an attention-based target tracking neural network with a Siamese architecture to track the needle tip in US images in different scenarios with different insertion angles when the needle was inserted both by motors and human hands. The proposed tracking network uses the attention mechanism to fully extract global features in US images, and the discriminative target model acted as a complementary part to learn a robust target model. The ever-updated template ensures the model can always capture the latest target features during tracking. The results of both motorized and manual insertion experiments show that the proposed tracker performed best in most cases, especially in the tissue experiments where the US image contains complex strong distractors. In our future work, we will integrate the proposed visual tracking network with a motion prediction module to improve the tracking accuracy of surgical tools and anatomical landmarks under US images.

## References


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

[2] M. Kaya and O. Bebek, "Gabor filter based localization of needles in ultrasound guided robotic interventions," in *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*, Oct. 2014, pp. 112–117, iSSN: 1558-2809.

[3] G. J. Vrooijink, A. Denasi, J. G. Grandjean, and S. Misra, "Model predictive control of a robotically actuated delivery sheath for beating heart compensation," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 193–209, 2017.

[4] B. Konh, B. Padasdao, Z. Batsaikhan, and S. Y. Ko, "Integrating robot-assisted ultrasound tracking and 3d needle shape prediction for real-time tracking of the needle tip in needle steering procedures," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 17, no. 4, p. e2272, 2021.

[5] 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.

[6] 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.

[7] 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.

[8] S. Mukhopadhyay, P. Mathur, A. Bharadwaj, Y. Son, J.-S. Park, S. R. Kudavelly, S. Song, and H. Kang, "Deep learning based needle tracking in prostate fusion biopsy," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 11598. SPIE, 2021, pp. 605–613.

[9] 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.

[10] J. Gao, P. Liu, G.-D. Liu, and L. Zhang, "Robust needle localization and enhancement algorithm for ultrasound by deep learning and beam steering methods," *Journal of Computer Science and Technology*, vol. 36, no. 2, pp. 334–346, 2021.

[11] D. J. Gillies, J. R. Rodgers, I. Gyacskov, P. Roy, N. Kakani, D. W. Cool, and A. Fenster, "Deep learning segmentation of general interventional tools in two-dimensional ultrasound images," *Medical Physics*, vol. 47, no. 10, pp. 4956–4970, 2020.

[12] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Single shot needle tip localization in 2d ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 637–645.

[13] ——, "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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] M. Zhao, K. Okada, and M. Inaba, "Trtr: Visual tracking with transformer," *arXiv preprint arXiv:2105.03817*, 2021.

[20] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6182–6191.

[21] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4282–4291.

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

[23] 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.

[24] D. J. C. LAW H, "Detecting objects as paired keypoints," *Lecture Notes in Computer Science*, pp. 765–781, 2018.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] G. J. Vrooijink, M. Abayazid, S. Patil, R. Alterovitz, and S. Misra, "Needle path planning and steering in a three-dimensional non-static environment using two-dimensional ultrasound images," *The International journal of robotics research*, vol. 33, no. 10, pp. 1361–1374, 2014.

[31] 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.

[32] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.

[33] C. Mayer, M. Danelljan, G. Bhat, M. Paul, D. P. Paudel, F. Yu, and L. Van Gool, "Transforming model prediction for tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8731–8740.

[34] 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.

[35] 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.