frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 17 January 2025
DOI 10.3389/fonc.2024.1519536

OPEN ACCESS

EDITED BY
Bilgin Kadri Aribas,
Bülent Ecevit University, Türkiye

REVIEWED BY
Çetin İmamoğlu,
Ankara Onkoloji Eğitim ve Araştırma Hastanesi, Türkiye
Yaşar Türk,
İstanbul Sağlık ve Teknoloji Üniversitesi Şişli Kolan International Hospital, Türkiye

*CORRESPONDENCE
Zhemin Zhuang
zmzhuang@stu.edu.cn
Huancheng Zeng
hczeng1989@126.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 30 October 2024
ACCEPTED 30 December 2024
PUBLISHED 17 January 2025

CITATION
Zhang Q, Chen J, Wang J, Wang H, He Y, Li B, Zhuang Z and Zeng H (2025) Needle tracking and segmentation in breast ultrasound imaging based on spatio-temporal memory network. *Front. Oncol.* 14:1519536. doi: 10.3389/fonc.2024.1519536

Needle tracking and segmentation in breast ultrasound imaging based on spatio-temporal memory network

Qiyun Zhang[1†], Jiawei Chen[1†], Jinhong Wang[2], Haolin Wang[1], Yi He[3,4], Bin Li[5], Zhemin Zhuang[1*] and Huancheng Zeng[6*]

[1]College of Engineering, Shantou University, Shantou, Guangdong, China, [2]Department of Ultrasound, Shantou Chaonan Minsheng Hospital, Shantou, Guangdong, China, [3]Shantou University Medical College, Shantou, Guangdong, China, [4]Department of Ultrasound, Shantou Central Hospital, Shantou, Guangdong, China, [5]Product Development Department, Shantou Institute of Ultrasonic Instruments, Shantou, Guangdong, China, [6]The Breast Center, Cancer Hospital of Shantou University Medical College, Shantou, Guangdong, China

Introduction: Ultrasound-guided needle biopsy is a commonly employed technique in modern medicine for obtaining tissue samples, such as those from breast tumors, for pathological analysis. However, it is limited by the low signal-to-noise ratio and the complex background of breast ultrasound imaging. In order to assist physicians in accurately performing needle biopsies on pathological tissues, minimize complications, and avoid damage to surrounding tissues, computer-aided needle segmentation and tracking has garnered increasing attention, with notable progress made in recent years. Nevertheless, challenges remain, including poor ultrasound image quality, high computational resource requirements, and various needle shape.

Methods: This study introduces a novel Spatio-Temporal Memory Network designed for ultrasound-guided breast tumor biopsy. The proposed network integrates a hybrid encoder that employs CNN-Transformer architectures, along with an optical flow estimation method. From the Ultrasound Imaging Department at the First Affiliated Hospital of Shantou University, we developed a real-time segmentation dataset specifically designed for ultrasound-guided needle puncture procedures in breast tumors, which includes ultrasound biopsy video data collected from 11 patients.

Results: Experimental results demonstrate that this model significantly outperforms existing methods in improving the positioning accuracy of needle and enhancing the tracking stability. Specifically, the performance metrics of the proposed model is as follows: IoU is 0.731, Dice is 0.817, Precision is 0.863, Recall is 0.803, and F1 score is 0.832. By advancing the precision of needle localization, this model contributes to enhanced reliability in ultrasound-guided breast tumor biopsy, ultimately supporting safer and more effective clinical outcomes.

Discussion: The model proposed in this paper demonstrates robust performance in the computer-aided tracking and segmentation of biopsy needles in ultrasound imaging, specifically for ultrasound-guided breast tumor biopsy, offering dependable technical support for clinical procedures.

KEYWORDS

computer-aided diagnosis, breast cancer, ultrasound, punch biopsy, needle segmentation

1 Introduction

In the realm of modern medicine, ultrasound-guided breast tumor biopsy is a cost-effective, convenient, and safe diagnostic method commonly employed for obtaining tissue samples for histological examination and pathological analysis (1). This technique aids physicians in confirming the origin and nature of breast lesions, as well as in monitoring disease progression (2). Ultrasound imaging is frequently employed to guide needle biopsy in real-time, enabling physicians to accurately navigate the needle to the target site while minimizing complications and preventing damage to surrounding tissues (2–5). However, the unique characteristics of ultrasound imaging, such as its low signal-to-noise ratio and dependence on the beam-to-needle orientation, introduce challenges in maintaining consistent visibility of the needle during procedures.

During the imaging process, those electronically controlled ultrasound beams which perpendicular to the needle produce strong specular reflections, thus enhancing needle visibility (6, 7). However, this optimal condition is often not maintained during the biopsy. When the needle direction is not perfectly perpendicular to the ultrasound beams or deviates from the ultrasound plane, it may result in unclear or completely invisible needle imaging (8). Furthermore, the needle's visibility diminishes with increasing insertion depth due to the attenuation of ultrasound beams (9). These challenges are exacerbated by the needle's small size, low contrast with surrounding tissues, and motion artifacts, which can mislead inexperienced physicians and increase the likelihood of inaccurate biopsy sampling and additional surgical risks (10).

To address these issues, during ultrasound-guided biopsy procedures where physicians manually manipulate the needle, real-time and precise computer-aided needle segmentation and tracking are essential. Such technologies can assist physicians in accurately locating the needle while adjusting the needle's position and angle in real-time, thereby significantly improving both the safety and accuracy of the procedure. Particularly for clinical novices and early-career physicians, AI-assisted tools play a pivotal role in reducing technical barriers, increasing confidence, and enhancing procedural outcomes. By enabling real-time feedback, these systems ensure optimal needle trajectory and alignment, even under suboptimal imaging conditions.

Previous research has explored various approaches to needle segmentation and tracking in ultrasound images. Device-based solutions, such as electromagnetic trackers (11), have been proposed, but image-based methods are generally more suitable for clinical settings due to their ease of integration. Traditional computer vision and machine learning techniques based on image have been extensively investigated. For example, Novotny et al. (12) proposed using Principal Component Analysis (PCA) to integrate prior knowledge with ultrasound data for enhanced representation. Ding et al. (13) developed a template-based method for preprocessing ultrasound images to increase contrast between the needle tip and surrounding tissues, and applied the Gaussian Flow Lines (GFL) algorithm to detect the needle edge. Zhou and Qiu et al. (14, 15) proposed a 3D Hough transform method using distance metrics to optimize fitting results, while Kaya and Senel et al. (16, 17) introduced a two-stage Gabor filter method for needle tip localization and optimized estimation of insertion angles. Although these methods offer valuable insights, they are limited by their dependency on image quality, high computational demands, and sensitivity to variations in needle shape and size.

With rapid advancements in deep learning technology, its application in computer vision and medical image analysis has become increasingly widespread, leading to the emergence of deep learning methods for needle segmentation in ultrasound images. Geraldes and Rocha et al. (18, 19) used Multi-Layer Perceptrons (MLPs) to segment needle in 2D ultrasound images, demonstrating the feasibility of deep learning for detecting needle in challenging ultrasound images. Pourtaherian et al. (20) proposed an Orthogonal Plane Convolutional Neural Network (OPCNN) to detect needle positions in 3D ultrasound images. Yang et al. (21) introduced the VOI-CNNs model, which utilizes a three-plane convolutional neural network for needle segmentation. Lee et al. (22) developed a segmentation-based tracking model that integrates spatial and channel "Squeeze and Excitation" (scSE) modules, and Yang et al. (23) proposed a Direction-Fused Fully Convolutional Network (DF-FCN) to train models both along and perpendicular to the needle axis. While these methods reduce the need for manual intervention required by traditional approaches, they typically rely on large amounts of annotated training data and computational resources. To improve detection efficiency, Mwikirize et al. (24) proposed a region-based Fast R-CNN for detecting needle in 2D ultrasound images. Ronneberger et al. (25) introduced U-Net, a fully convolutional network that leverages multi-scale semantic information from ultrasound images

to enhance detection accuracy. Additionally, many similar works have focused on detailed optimizations: Zhang et al. (26) proposed an attention-based U-Net for multi-needle segmentation and localization; Yang et al. (27) introduced a 3D patch-wise method to segment needle in 3D ultrasound images by dividing the 3D image into small blocks; and Bi et al. (28) developed a method to explicitly separate anatomical and domain features by calculating mutual information in the latent space, improving the generalization ability of segmentation models.

However, despite these advancements, existing deep learning methods for needle localization in ultrasound-guided procedures struggle to effectively extract features from ultrasound images with low signal-to-noise ratios and complex backgrounds. Moreover, they fail to leverage the inter-frame relationships present in needle video sequences, overlooking subtle displacements and morphological changes of the needle in complex tissue backgrounds. Consequently, these limitations result in insufficient localization accuracy and affect the safety and precision of clinical operations.

To address these limitations, we propose a novel Spatio-Temporal Needle Attention Network (STNAN), which achieves accurate segmentation, tracking, and prediction of the needle's dynamic trajectory. Unlike conventional methods, STNAN leverages the inter-frame relational information inherent in ultrasound video sequences to enhance its tracking capabilities. Subsequently, we propose a feature encoder that integrates CNN and Transformer architectures, enabling the model to extract multi-scale features and local information from ultrasound images while capturing long-range dependencies across image sequences. Additionally, we design an innovative feature processing method that incorporates optical flow estimation to extract subtle displacement and morphological features of the needle in complex tissue environments. Furthermore, the proposed approach offers substantial potential for clinical implementation, particularly in assisting novice physicians, by providing additional support and enhancing operator confidence.

The main contributions of this paper are as follows:

1. To leverage the correlation information between frames in video sequences, we introduce a memory bank structure to store features extracted from ultrasound needle video sequences, thereby effectively processing and utilizing temporal information.
2. To address the challenges posed by ultrasound images with low signal-to-noise ratios and complex backgrounds, we propose a semantic feature encoder that integrates CNN and Transformer structures. This encoder not only extracts multi-scale features and local information from the images but also captures long-range dependencies within them.
3. To accurately capture and track the needle's motion trajectory, we propose a feature processing method that integrates optical flow estimation, capable of extracting subtle displacements and morphological changes of the needle.

2 Materials and methods

The natural scene model Space-Time Correspondence Networks (STCN) provides an effective framework for capturing spatio-temporal correspondences (29). However, it is confronted with two challenges: due to substantial domain differences between medical and natural scene images, the basic feature extraction approach may be insufficient for capturing critical information in medical imaging, resulting in mismatched features. Additionally, common global matching patterns can lead the model to mistakenly segment distant objects with similar appearances. To address these challenges, we proposed Spatio-Temporal Needle Attention Network (STNAN) based on STCN, which employs the negative squared Euclidean distance to effectively model spatio-temporal correspondence, avoiding the need to re-encode mask features for each frame. The model utilizes a memory bank to store feature information from the previous frames. Furthermore, we introduce a novel hybrid architecture feature encoder and integrate optical flow information from adjacent frames as supplementary features to perceive motion information. The network structure is depicted in Figure 1.

STNAN takes a video sequence and the annotation of the first frame as input, processes the video frame by frame, and establishes a Feature Memory Bank. The network includes two encoders: C&T Query Encoder and Memory Encoder. The C&T Query Encoder takes the current frame of the ultrasound image as the query frame input, extracts the key feature of the query frame, denoted as k^Q. Meanwhile the Memory Encoder takes the previous images and corresponding masks as input, extracts the associated value feature, denoted as v^M.

During the sequence query process, features are extracted only once per frame. After completing a query, the key feature k^Q of the current query frame is used as the key feature k^M of the memory frame for the next query, reducing computational overhead. The key feature k^Q of the query frame is compared with the key feature k^M of the memory frame, and the affinity value *aff* between them is calculated using the negative squared Euclidean norm. The corresponding value feature v^M is retrieved from the Feature Memory Bank. The key features k^Q and corresponding k^M are processed through our designed Optical Flow Regression module to obtain the optical flow vector *flow*. This vector, along with the affinity value *aff*, is utilized to calculate the value feature of the query frame, denoted as v^Q. Finally, the mask of the current frame's target object can be generated through the decoder.

2.1 Feature extraction with hybrid architecture of C&T query encoder

For ultrasound needle images characterized by low signal-to-noise ratios and complex backgrounds, fully extracting the needle's features is essential for improving network accuracy. Convolutional Neural Networks (CNNs) are highly effective at extracting deep, discriminative features from image data, particularly excelling at

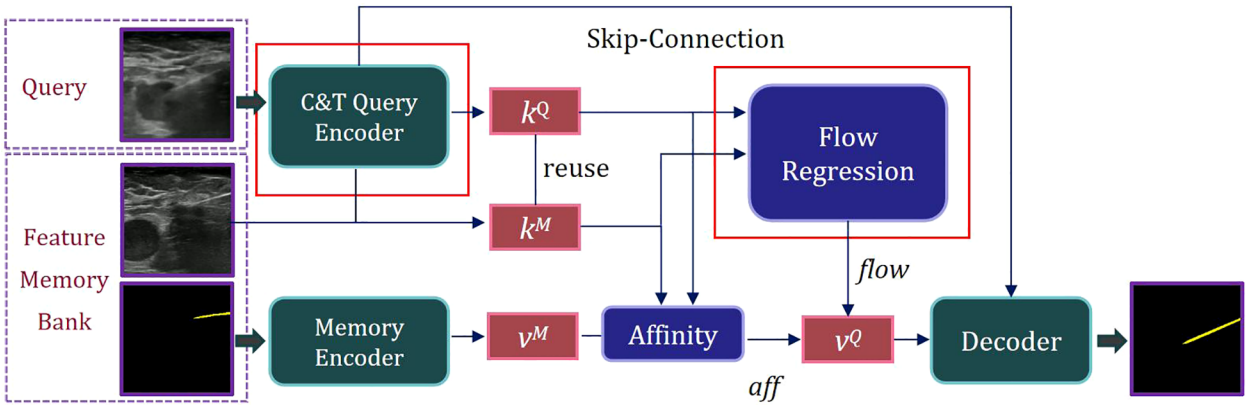

FIGURE 1
Overview of the proposed STNAN for needle segmentation.

capturing local information and multi-scale features. However, CNNs have limitations in handling long-range dependencies, and experience a significantly increase in computational complexity as network depth grows. On the other hand, Transformers effectively capture long-range dependencies in images through self-attention mechanisms but perform less sensitively in processing local structural details. Additionally, Transformers entail significant computational resources when dealing with high-resolution images and exhibit relatively weaker capabilities in extracting local feature.

To capitalize on the strengths of both CNNs and Transformers, we propose a hybrid architecture called C&T Query Encoder, which integrates CNN and Transformer components. This architecture utilizes CNNs' capabilities in local feature extraction and multi-scale information processing, alongside Transformers' ability to capture long-range dependencies, thereby achieving more comprehensive and efficient feature extraction in image analysis tasks. Its structure is illustrated in Figure 2.

The design of the C&T Query Encoder follows these steps: The input image is first processed through three convolutional layers for initial processing, downsampling, and local feature extraction while preserving spatial information. The output of these initial convolutional layers is then fed into four combined modules, each consisting of a Patch Embedding layer and i Blocks. These modules are alternated to apply Self-Attention at various scales, capturing connections between different regions in the image. The Patch Embedding layer performs two-fold downsampling through a convolutional layer to extract low-resolution and multi-scale features. Each Block contains residual structures with both Depth-Wise Separable Convolution (DW Conv) and a Lightweight Multi-Head Self Attention module (LMHSA). These modules are interspersed with Layer Norm (LN), ensuring stable feature extraction.

DW Conv significantly reduces computational complexity and the number of parameters by separating the convolution operation into depthwise and pointwise convolutions, thereby enhancing computational efficiency. In each Block, we employ residual structures, initially implement a DW Conv with a 3 × 3 kernel size, followed by the integration of the LMHSA module, and subsequently reapply another DW Conv. This configuration effectively extracts local features from the feature map by combining the advantages of depthwise and pointwise convolutions. Additionally, the introduction of residual connections facilitates efficient gradient propagation throughout the network, thus enhancing training stability. The overall goal is to improve translation invariance in computer vision tasks while maintaining computational efficiency.

Following the first DW Conv module, Layer Normalization is applied to the feature map, enhancing training stability and convergence speed. The normalized feature map is then fed into the LMHSA module, as illustrated in Figure 3.

We employ a size-controllable DW Conv to reduce the resolution of the input feature, thereby decreasing the size of the K and V subspace feature maps and lowering the computational load of self-attention mechanism. The input feature is then mapped to three subspaces—V, K, and Q— using three distinct linear spatial mapping matrices W^V, W^K, and W^Q, and the attention value is calculated as follows:

$$attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V$$

where d represents the dimension of the key vector

After the LMHSA module, we further apply Layer Normalization to the feature map. This is followed by another DW Conv to process the featuremap, extracting local information.

Through the above design, the model can effectively extract semantic features, thus providing robust feature support for subsequent segmentation tasks.

2.2 Flow regression estimation based on search window

Deep stereo matching methods can leverage multi-view images to compute depth information, thereby improving the model's ability to perceive the needle's spatial position. Since stereo

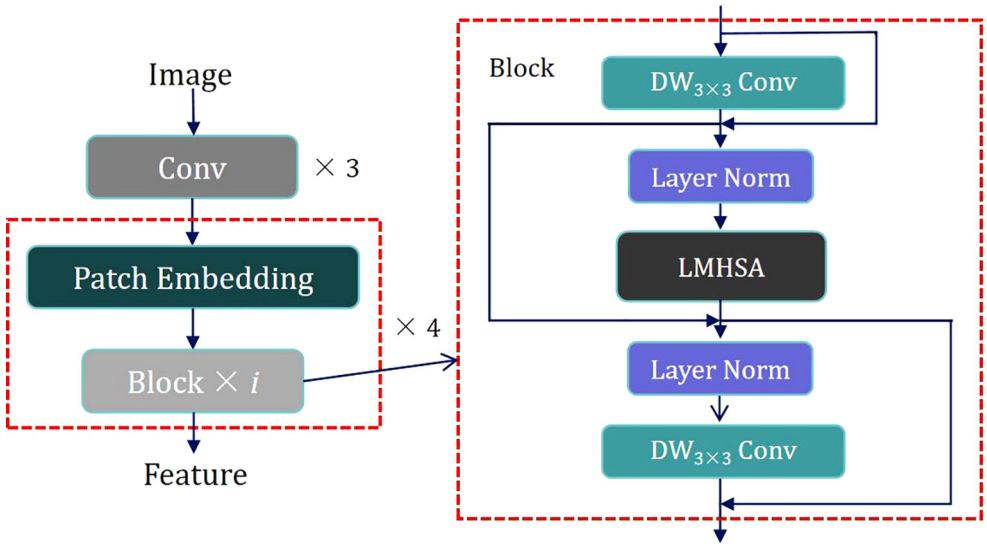

FIGURE 2
Scheme of the proposed Hybrid Architecture of C&T Query Encoder.

matching can be considered a specific case of optical flow, the matching cost learning in optical flow estimation is equally applicable to needle motion estimation. Optical flow estimation captures subtle motion information between adjacent frames, which is essential for detecting the needle's minute displacements. By estimating optical flow, the model gains a deeper understanding of the needle's motion patterns, enabling more precise trajectory tracking during segmentation.

In our deep stereo matching approach, we introduce a matching cost learning mechanism to enhance the accuracy of optical flow estimation by learning the matching cost between adjacent frames.

As shown in Figure 4, the Semantic Feature F_t of the current frame is superimposed with the features F_{t-1} under each hypothesis to obtain U × V Semantic Fusion Feature Maps $F_u(p)$:

$$F_u(p) = F_t(p)||F_{t-1}(p+u)$$

where $p = (x, y)$ represents the original position coordinates, || denotes superposition in the channel dimension, and u represents the displacement hypothesis.

In the feature map, for each pixel, we limit the scan matching range to a local window with determined scale containing U × V pixels. By hypothesizing various displacement directions of each pixel, we obtain corresponding $F_u(p)$, as shown in Figure 5.

As seen in the Figure 5, each pixel has U × V displacement hypotheses.

On this basis, the semantic fusion feature map $F_u(p)$ is fed into the Displacement Aware Cost Learning (DACL) module. The DACL module comprises a series of convolutional and deconvolutional layers to extract features from $F_u(p)$. Using a CNN-based 2D matching network, the module calculates local correlations and aggregates matching costs under U × V different displacement hypotheses, producing a comprehensive matching

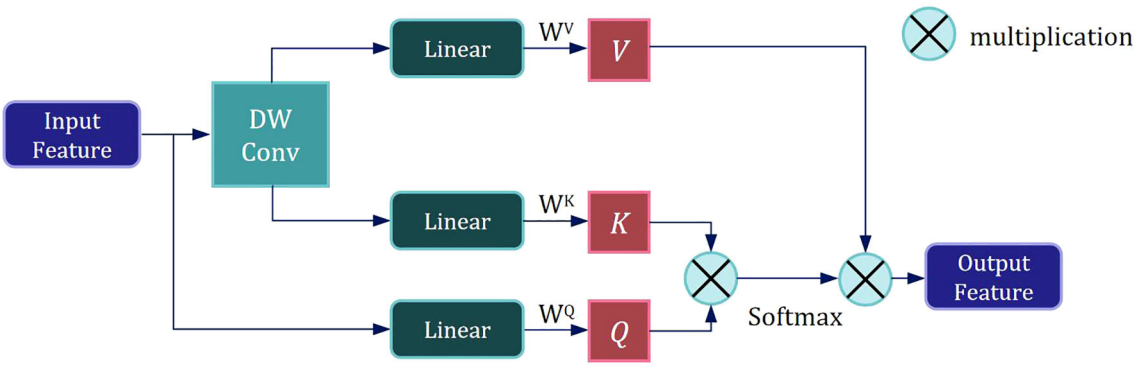

FIGURE 3
Scheme of the proposed Lightweight Multi-Head Self Attention module.

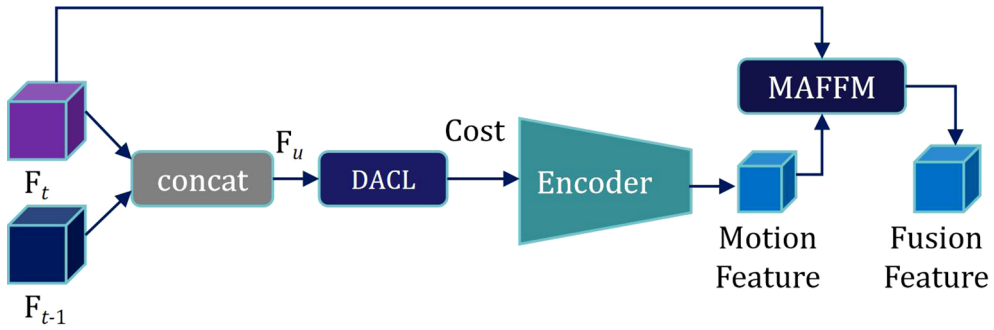

FIGURE 4
Scheme of the proposed feature processing method incorporating optical flow estimation.

cost, denoted as *Cost*. Furthermore, the DACL module reweights the matching cost through a 1 × 1 convolution, calculated as follows:

$$Cost(u,p) = Conv_{1\times1}(\sum_{u=(U,V)} M(F_u(p)))$$

where M represents the 2D matching network.

Subsequently, motion features from the previous and current frames are extracted by encoding. Motion features are then fused with the semantic feature F_t of the current frame through the Motion-attention Based Feature Fusion Module (MAFFM), as illustrated in Figure 6.

We begin by applying a 1 × 1 convolution to reduce the dimensionality of the Motion Feature, followed by smoothing them using a Sigmoid activation function. A broadcast mechanism is then employed to reweight the Semantic Feature F_t of the current frame. Finally, we superimpose F_t onto itself to obtain a Motion Attention Feature, which is utilized for subsequent mask propagation. This design enables the model to effectively exploit inter-frame relationships within the needle sequence, allowing it to accurately capture the needle's motion trajectory.

2.3 Real-time ultrasound needle dataset

In the field of ultrasound-guided biopsy, particularly for breast tumor puncture, there is a significant shortage of publicly available datasets for needle segmentation. As a result, most existing studies rely on non-public datasets derived from non-human samples. These datasets often do not accurately reflect real surgical conditions and typically focus on single-frame images, overlooking the continuity of needle movement during the biopsy process. To address this gap, in collaboration with physicians from the Ultrasound Imaging Department at the First Affiliated Hospital of Shantou University, we collected ultrasound biopsy video data from 11 patients of varying ages, creating a real-time segmentation dataset specifically tailored for ultrasound-guided needle puncture in breast tumors. Each video was recorded at a frame rate of 25 frames per second. The institutional review board approved this study, and the requirement to obtain informed consent was waived (approval number: B-2022-182).

To ensure temporal continuity and reduce redundant features, we extracted frames at random intervals of 5 to 10 frames, selecting approximately 250 consecutive frames from each video. Each frame contains a single needle, annotated by experienced ultrasound imaging physicians to obtain mask images as dataset labels. We used nine video frame sequences (approximately 2,250 images) for the training set and the remaining two sequences (approximately 500 images) for the test set. All images were uniformly resized to 448 × 448 pixels.

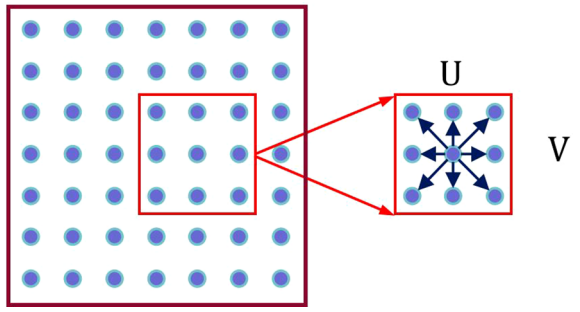

FIGURE 5
Illustration of the displacement search box.

2.4 Loss function and evaluation metrics

Biopsy needle in ultrasound images appear as thin, elongated straight lines and are significantly outnumbered by background pixels. To address this issue of imbalanced samples, we adopt Binary Cross Entropy (BCE) as the loss function. The calculation of BCE is as follows:

$$L_{BCE} = -(y * \log(p) + (1-y) * \log(1-p))$$

where y represents the true label of the sample (1 for biopsy needle, 0 for background), and p represents the probability that the model predicts the sample as a biopsy needle.

To comprehensively assess the segmentation performance, we use several evaluation metrics, including Intersection over Union, Dice coefficient, Precision, Recall, and F1-score.

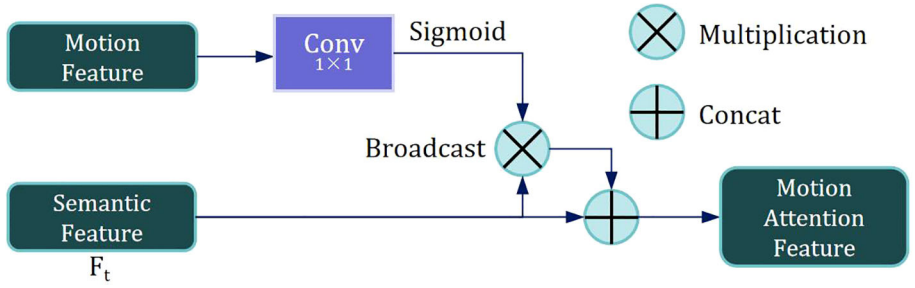

FIGURE 6
The structure of the Motion-attention based feature fusion module (MAFFM).

The calculation formulas for *IoU*, *Dice*, *Precision*, *Recall*, and *F1-score* are as follows:

$$IoU = \frac{TP}{TP + FP + FN}$$

$$Dice = \frac{2TP}{2TP + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Considering the real-time requirements of ultrasound-guided biopsy procedures, we measured the inference speed of the model on the test set, recording both the number of frames segmented per second and the average time required for segmenting each frame.

2.5 Training parameters and training method settings

Our model was trained on an NVIDIA GeForce RTX 3090 GPU using the Adaptive Moment Estimation (Adam) optimizer. The training was conducted with a batch size of 4, an initial learning rate of 10^{-5}, and a learning rate decay factor of 0.2.

To enhance the continuity of the model's segmentation and tracking performance, we employed the following training strategy: During each iteration, four video sequences were randomly selected from the training set and we sequentially select five images as training samples, with the first image providing the ground truth mask label. Memory features were extracted from the first frame and its corresponding mask label to predict the mask of the second frame, and the features from both frames were stored in the memory repository. These fused memory features were then used to predict the masks for subsequent frames in sequence. For all frames except the first, the predicted masks were compared with the ground truth labels for loss calculation, followed by backpropagation and parameter updates in the network. To improve the model's robustness, we randomly reverse the temporal order of the five continuous frames during training, enabling the model to learn both forward and backward predictions.

To prevent redundant memory features and maintain efficient inference speed, we adopted the default memory storage interval of STCN, storing semantic and mask features every 5 frames.

3 Results

3.1 Parameter optimization experiments

We introduces the C&T Query Encoder, which allows for the stacking of a variable number of Blocks, thereby facilitating a balance between model accuracy and inference speed. We systematically explored the impact of the number of Blocks *i* in the C&T Query Encoder on model performance and validated the performance of the STCN+C&T model under various configurations through experimentation. Table 1 presents the performance metrics of the STCN+C&T model when varying numbers of Blocks are stacked in the C&T Query Encoder. In terms of the performance metrics, multiple standard metrics including IoU, Dice, Precision, Recall, F1 scores and FPS are used for evaluating image segmentation tasks.

In Table 1, the first column represents the number of Blocks in the four stacked structures of the C&T Query Encoder. As the number of Blocks increases, the IoU and Dice scores of the STNAN model improve, while inference speed decreases correspondingly.

TABLE 1 Performance comparison of STCN+C&T model with different numbers of blocks.

The number of blocks	IoU	Dice	Precision	Recall	F1	FPS
(2, 2, 6, 2)	0.718	0.806	0.840	**0.803**	**0.821**	**60.17**
(2, 2, 10, 2)	0.712	0.803	0.838	0.799	0.818	55.41
(3, 3, 12, 3)	0.709	0.802	0.845	0.790	0.816	52.53
(3, 3, 16, 3)	**0.720**	**0.809**	**0.854**	0.790	0.820	46.52

The bold values provided in the table below indicate the best-performing metrics in comparison.

The experiments reveal that when the number of Blocks is set to 3, 3, 16, and 3 respectively, the STCN+C&T model achieves optimal performance on the test set, with an IoU of 0.720, Dice of 0.809, F1 score of 0.820, and FPS of 46.52, significantly surpassing the 25 FPS ultrasound biopsy video data provided by the hospital, thereby meeting real-time operational requirements.

Furthermore, we investigated the effect of varying displacement search window sizes in optical flow estimation on model performance. On a semantic feature map with a size of 12 × 12, we set various displacement search window sizes and conducted experiments under the optimal Block configuration, with the results shown in Table 2.

Since the displacement search window is centered on each pixel and has an odd-numbered side length, we implemented four window size configurations: 5 × 5, 7 × 7, 9 × 9 and 11 × 11. The results indicate that the performance of the model improves when the displacement search window size increases from 5 × 5 to 9 × 9; When the size continues to increase to 9 × 9, the model effect decreases. We speculate that this decline occurs because the needle displacement between some frames is large, and a smaller search window cannot adequately capture these variations, resulting in some needle pixels in the past frames not matching the correct points in the current frame. As the search window size continues to increase, although more information can be captured, it may also lead to matching noise points or irrelevant points in the background, thereby reducing the accuracy of the matching. Ultimately, when the displacement search window size is 9 × 9, the model achieves its optimal performance, which can also satisfy real-time operational requirements.

3.2 Ablation study

This paper introduces several improvements to the original STCN model, proposing the STNAN model. The effectiveness of each improvement module to improve the model's performance was validated through a series of experiments. Specifically, we compared the performance of the original STCN model with that of the model incorporating each improvement module, with the results shown in Table 3.

The performance metrics of the original STCN model is as follows: IoU is 0.696, Dice is 0.793, Precision is 0.861, Recall is 0.761, F1 score is 0.808, and FPS is 60.79, with a single-frame segmentation time of 0.016 seconds.

TABLE 2 Performance of STNAN with different displacement search window sizes.

Size of window (U, V)	IoU	Dice	Precision	Recall	F1	FPS
U=5, V=5	0.712	0.801	0.855	0.775	0.813	**38.59**
U=7, V=7	0.722	0.810	0.851	0.798	0.824	37.74
U=9, V=9	**0.731**	**0.817**	**0.863**	**0.803**	**0.832**	35.57
U=11, V=11	0.720	0.809	0.852	0.796	0.823	30.26

The bold values provided in the table below indicate the best-performing metrics in comparison.

TABLE 3 Ablation study results of the STNAN model.

Model	IoU	Dice	Precision	Recall	F1	FPS
STCN	0.696	0.793	0.861	0.761	0.808	**60.79**
STCN+C&T	0.720	0.809	0.854	0.790	0.820	46.52
STCN+C&T +FR(STNAN)	**0.731**	**0.817**	**0.863**	**0.803**	**0.832**	35.57

The bold values provided in the table below indicate the best-performing metrics in comparison.

After adding C&T Encoder and FR modules, significant improvements are demonstrated across recall (0.761 vs 0.803), precision (0.861 vs 0.863), IOU (0.696 vs 0.731) and Dice(0.793 vs 0.863).In conclusion, the proposed improvement strategies in this paper have significantly enhanced the performance of the model in ultrasound needle tracking and segmentation tasks.

3.3 Comparative experiments

Table 4 presents the results of the proposed STNAN against all the compared methods over the datasets. Comparisons between our STNAN and recently favored frameworks for medical image segmentation were conducted, benchmarking against convolutional baseline models such as U-Net, VGG16 and FCN32. We further perform comparison against the STCN segmentation networks.

The results demonstrate that our STNAN model surpasses the performance of all other methodologies. This highlights the significant advantages of STNAN in terms of segmentation accuracy and overall quality, indicating its strong suitability for meeting the clinical requirements of biopsy procedures.

To visually illustrate the performance differences between various models in handling ultrasound needle tracking and segmentation tasks, we selected several ultrasound needle images from the test set in sequential order, emphasizing cases where the grayscale visibility of the biopsy needle is relatively weak. Furthermore, we conducted a visual analysis of each model's segmentation and tracking results, as depicted in Figure 7.

TABLE 4 Comparison with various models with respect to efficiency and segmentation metrics.

Model	IOU	Dice	Precision	Recall	F1	FPS
PSPNet	0.233	0.348	0.579	0.291	0.387	29.15
FCN32	0.563	0.689	0.762	0.676	0.716	19.93
UNet	0.511	0.623	0.716	0.660	0.687	23.63
SegNet	0.399	0.548	0.621	0.527	0.570	15.50
VGG16+UNet	0.508	0.649	0.789	0.667	0.723	22.90
ResNet50 +FCN32	0.566	0.697	0.737	0.684	0.710	9.96
STCN	0.696	0.793	0.861	0.761	0.808	**60.79**
STNAN(Ours)	**0.731**	**0.817**	**0.863**	**0.803**	**0.832**	35.57

The bold values provided in the table below indicate the best-performing metrics in comparison.

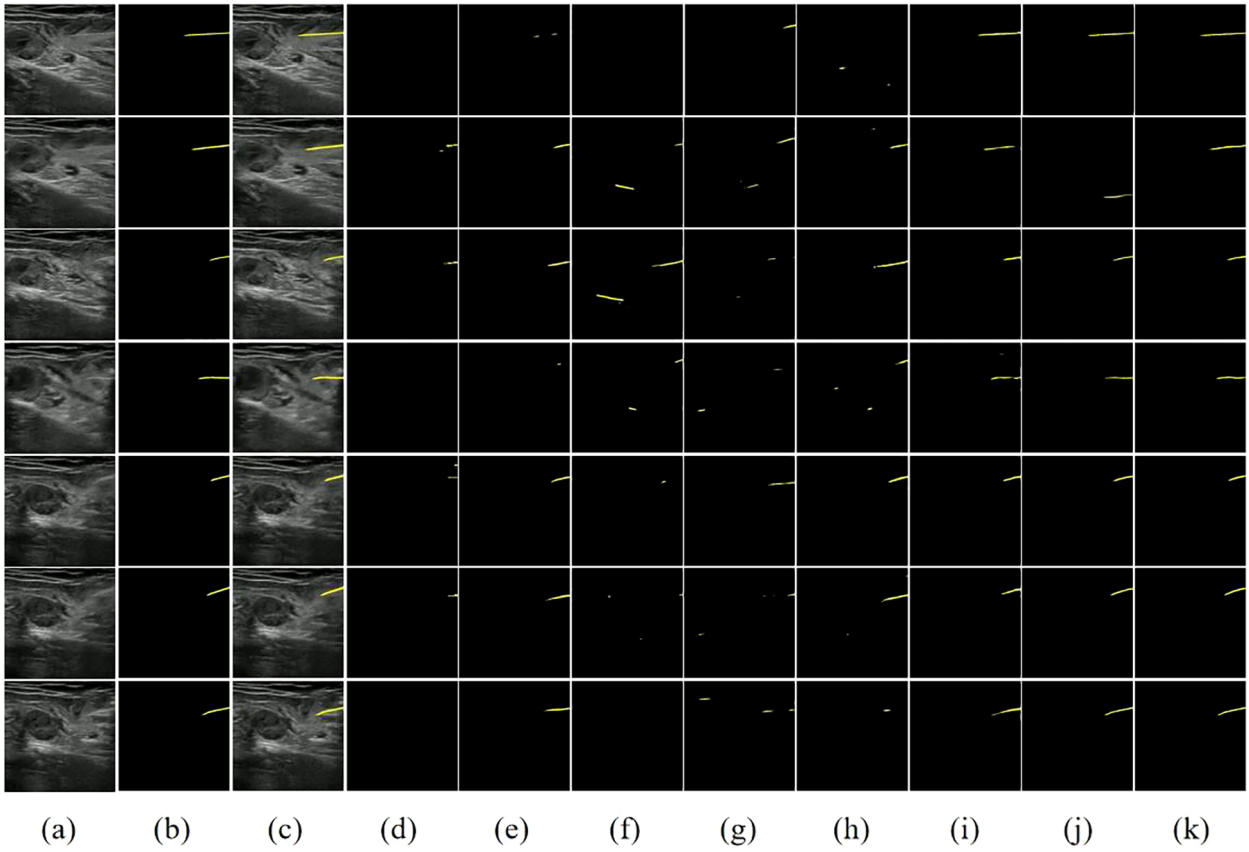

FIGURE 7
Visual comparison of different methods on our dataset. **(A)** Image. **(B)** Ground Truth. **(C)** Image+GT. **(D)** PSPNet. **(E)** FCN32. **(F)** UNet. **(G)** SegNet. **(H)** VGG16+UNet. **(I)** ResNet50+FCN32. **(J)** STCN. **(K)** STNAN. Yellow indicates the prediction of needle.

The first two columns of Figure 7 display the original ultrasound images from the test set alongside their corresponding ground truth labels. The third column displays the superimposed ultrasound images along with ground truth data, while the remaining columns show the segmentation results of each model. Classic semantic segmentation models often struggle with challenging ultrasound needle tracking and segmentation tasks, frequently failing to accurately segment the needle and tenting to misidentify needle-like tissues in the image as the needle, resulting in an increased incidence of false positives in the segmentation results. The STCN model considers the spatio-temporal correlation between frames, improving the segmentation results to some extent. However, it still suffers from false positives due to its reliance on global matching techniques, which may undermine its effectiveness during clinical procedures.

In contrast, our STNAN model demonstrates superior performance in accurately identifying and tracking needle while maintaining a lower rate of false positives. It is attributed to the proposed semantic feature encoder for low signal-to-noise ratio ultrasound images and the optical flow regression estimation method. These advancements make STNAN highly promising for real-world clinical applications, as it effectively minimizes false positives while achieving precise segmentation of the puncture needle.

4 Discussion

This paper introduces a novel Spatio-Temporal Needle Attention Network (STNAN) to address key challenges in ultrasound-guided needle biopsy. By integrating Convolutional Neural Networks (CNNs) and Transformer architectures, STNAN effectively extracts multi-scale features and captures long-range dependencies. Additionally, the model incorporates an optical flow estimation mechanism to detect subtle displacements and morphological changes of the needle in complex tissue environments. Experimental results demonstrate that STNAN significantly outperforms existing methods in enhancing the accuracy and continuity of needle localization, thereby improving both the safety and precision of clinical procedures.

The proposed model is designed as an AI-assisted tool to support less experienced physicians during ultrasound-guided biopsy procedures. In practice, needle visibility is often compromised due to factors such as suboptimal angles, deviations from the ultrasound plane, and signal attenuation at greater depths, which pose significant challenges for novice operators. STNAN provides real-time needle tracking and segmentation, offering additional guidance and improving operator confidence.

The dataset used to train STNAN consisted of 11 patients, yielding approximately 2750 image frames. During model training,

STNAN achieved convergence without signs of overfitting, demonstrating that the data was sufficient for learning key features of needle localization. However, we acknowledge that expanding the dataset could further enhance the model's generalizability, especially when applied to more diverse clinical scenarios.

STNAN shows excellent performance in ultrasound-guided needle localization tasks, providing robust technical support for clinical operations. In future work, we will focus on further optimizing the model's real-time performance and exploring its broader applications in ultrasound-guided interventions. Expanding training datasets with diverse clinical cases is expected to improve the model's generalizability, and advanced techniques such as data augmentation and transfer learning will be employed to optimize model performance. Additionally, incorporating trajectory prediction features could further assist operators in planning needle movements, paving the way for safer and more precise procedures in clinical practice.

Data availability statement

The dataset fully protects patient privacy, has been approved by the ethics committee, and is intended for academic research only. Requests to access the datasets should be directed to Zhemin Zhuang, zmzhuang@stu.edu.cn.

Ethics statement

The studies involving humans were approved by The First Affiliated Hospital of Shantou University School of Medicine Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

QZ: Conceptualization, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JC: Conceptualization, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JW: Data curation, Writing – original draft, Writing – review & editing. HW: Conceptualization, Writing – original draft, Writing – review & editing. YH: Data curation, Resources, Writing – original draft, Writing – review & editing. BL: Data curation, Resources, Writing – original draft, Writing – review & editing. ZZ: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. HZ: Data curation, Resources, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China under Grant (No. 82071992), the Youth Fund of Cancer Hospital of Shantou University Medical 356 College (N0.2023A006), the Basic and Applied Basic Research Foundation of Guangdong Province (No. 2020B1515120061), the Medical Scientific Research Foundation of Guangdong Province of China (No. A2024154), and the Talent Support Program of the First Affiliated Hospital of Shantou University Medical College(No. YCTJ-2023-09).

Acknowledgments

We would like to express our gratitude to Shantou Ultrasound Instrument Research Institute Co., Ltd. for providing the equipment and data, and to Cancer Hospital of Shantou University Medical College and Shantou Chaonan Minsheng Hospital for providing the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Ciliberti V, Maffei E, D'Ardia A, Sabbatino F, Serio B, D'Antonio A, et al. Combined fine needle aspiration cytology and core needle biopsy in the same setting: A two-years' experience. *Cytopathology*. (2024) 35:78–91. doi: 10.1111/cyt.13318

2. Lisotti A, Cominardi A, Conti Bellocchi MC, Crinò SF, Larghi A, Facciorusso A, et al. Repeated endoscopic ultrasound-guided fine-needle biopsy of solid pancreatic lesions after previous nondiagnostic or inconclusive sampling. *Digestive Endoscopy*. (2024) 36:615–24. doi: 10.1111/den.14686

3. Vilas-Boas F, Ribeiro T, Macedo G, Dhar J, Samanta J, Sina S, et al. Endoscopic ultrasound-guided through-the-needle biopsy: A narrative review of the technique and its emerging role in pancreatic cyst diagnosis. *Diagnostics*. (2024) 14:1587. doi: 10.3390/diagnostics14151587

4. Gopakumar H, Puli SR. Value of endoscopic ultrasound-guided through-the-needle biopsy in pancreatic cystic lesions. *A Systematic Rev Meta-Analysis J Gastrointestinal Cancer*. (2024) 55:15–25. doi: 10.1007/s12029-023-00949-w

5. Vrooijink GJ, Abayazid M, Misra S. (2013). Real-time three-dimensional flexible needle tracking using two-dimensional ultrasound, in: *2013 IEEE International conference on robotics and automation*, , Vol. p. pp. 1688–93. Karlsruhe, Germany: IEEE. doi: 10.1109/ICRA.2013.6630797

6. Barr RG. Improved needle visualization with electronic beam steering: proof of concept. *Ultrasound Quarterly*. (2012) 28:59–64. doi: 10.1097/RUQ.0b013e3182585fea

7. Che H, Qin J, Chen Y, Ji Z, Yan Y, Yang J, et al. Improving needle tip tracking and detection in ultrasound-based navigation system using deep learning-enabled approach. *IEEE J Biomed Health Inf*. (2024) 28(5):2930–42. doi: 10.1109/JBHI.2024.3353343

8. Grube S, Latus S, Behrendt F, Riabova O, Neidhardt M, Schlaefer A. Needle tracking in low-resolution ultrasound volumes using deep learning. *Int J Comput Assisted Radiol Surg*. (2024) 19:1975–81. doi: 10.1007/s11548-024-03234-8

9. Kimbowa A, Pieters A, Tadayon P, Arora I, Gulam S, Pinos A, et al. Advancements in needle visualization enhancement and localization methods in ultrasound: a literature. *Art Int Surg*. (2024) 4:149–69. doi: 10.20517/ais.2024.20

10. Stone MB, Moon C, Sutijono D, Blaivas M. Needle tip visualization during ultrasound-guided vascular access: short-axis vs long-axis approach. *Am J Emergency Med*. (2010) 28:343–7. doi: 10.1016/j.ajem.2008.11.022

11. Zhao Z, Tse ZTH. An electromagnetic tracking needle clip: an enabling design for low-cost image-guided therapy. *Minimally Invasive Ther Allied Technologies*. (2019) 28:165–71. doi: 10.1080/13645706.2018.1496939

12. Novotny PM, Cannon JW, Howe RD. (2003). Tool localization in 3D ultrasound images, in: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003: 6th International Conference*, Montréal, Canada, November 15-18, 2003. pp. 969–70. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-39903-2_127

13. Ding M, Cardinal HN, Guan W, Fenster A. Automatic needle segmentation in 3D ultrasound images. In: *Medical imaging 2002: visualization, image-guided procedures, and display*. SPIE (2002). p. 65–76. doi: 10.1117/12.466907

14. Zhou H, Qiu W, Ding M, Zhang S. Automatic needle segmentation in 3D ultrasound images using 3D improved Hough transform. In: *Medical imaging 2008: visualization, image-guided procedures, and modeling*. SPIE (2008). p. 688–96. doi: 10.1117/12.770077

15. Qiu W, Yuchi M, Ding M, Tessier D, Fenster A. Needle segmentation using 3D Hough transform in 3D TRUS guided prostate transperineal therapy. *Med physics*. (2013) 40:042902. doi: 10.1118/1.4795337

16. Kaya M, Bebek O. (2014). Gabor filter based localization of needles in ultrasound guided robotic interventions, in: *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*. pp. 112–7. Santorini, Greece: IEEE. doi: 10.1109/IST.2014.6958456

17. Kaya M, Bebek O. (2014). Needle localization using gabor filtering in 2D ultrasound images, in: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4881–6. Hong Kong, China: IEEE. doi: 10.1109/ICRA.2014.6907574

18. Geraldes AA, Rocha TS. (2014). A neural network approach for flexible needle tracking in ultrasound images using kalman filter, in: *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, . pp. 70–5. Sao Paulo, Brazil: IEEE. doi: 10.1109/BIOROB.2014.6913754

19. Rocha TS, Geraldes AA. (2014). Flexible needles detection in ultrasound images using a multi-layer perceptron network, in: *5th ISSNIP-IEEE Biosignals and Biorobotics Conference (2014): Biosignals and Robotics for Better and Safer Living (BRC)*, pp. 1–5. Salvador, Brazil: IEEE. doi: 10.1109/BRC.2014.6880999

20. Pourtaherian A, Ghazvinian Zanjani F, Zinger S, Mihajlovic N, Ng G, Korsten H, et al. (2017). Improving needle detection in 3D ultrasound using orthogonal-plane convolutional networks, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, . pp. 610–8. Springer. doi: 10.1007/978-3-319-66185-8_69

21. Yang H, Shan C, Kolen AF, de With PH. (2018). Catheter detection in 3d ultrasound using triplanar-based convolutional neural networks, in: *2018 25th IEEE International Conference on Image Processing (ICIP)*, . pp. 371–5. Athens, Greece: IEEE. doi: 10.1109/ICIP.2018.8451586

22. Lee JY, Islam M, Woh JR, Washeem TM, Ngoh LYC, Wong WK, et al. Ultrasound needle segmentation and trajectory prediction using excitation network. *Int J Comput Assisted Radiol Surgery*. (2020) 15:437–43. doi: 10.1007/s11548-019-02113-x

23. Yang H, Shan C, Kolen AF, de With PH. (2019). Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn, in: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, . pp. 1122–6. Venice, Italy: IEEE. doi: 10.1109/ISBI.2019.8759420

24. Mwikirize C, Nosher JL, Hacihaliloglu I. Convolution neural networks for real-time needle detection and localization in 2D ultrasound. *Int J Comput assisted Radiol surgery*. (2018) 13:647–57. doi: 10.1007/s11548-018-1721-y

25. Ronneberger O, Fischer P, Brox T. (2015). U-net: Convolutional networks for biomedical image segmentation, in: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference*, , October 5-9, 2015. pp. 234–41. Munich, Germany: Springer, proceedings, part III 18. doi: 10.1007/978-3-319-24574-4_28

26. Zhang Y, Lei Y, Qiu RL, Wang T, Wang H, Jani AB, et al. Multi-needle localization with attention U-net in US-guided HDR prostate brachytherapy. *Med physics*. (2020) 47:2735–45. doi: 10.1002/mp.14128

27. Yang H, Shan C, Kolen AF, de With PH. (2019). Automated catheter localization in volumetric ultrasound using 3D patch-wise U-Net with focal loss, in: *2019 IEEE International Conference on Image Processing (ICIP)*, . pp. 1346–50. Taipei, Taiwan: IEEE. doi: 10.1109/ICIP.2019.8803045

28. Bi Y, Jiang Z, Clarenbach R, Ghotbi R, Karlas A, Navab N. (2023). MI-SegNet: Mutual information-based US segmentation for unseen domain generalization, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, . pp. 130–40. Cham: Springer. doi: 10.1007/978-3-031-43901-8_13

29. Cheng HK, Tai YW, Tang CK. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Adv Neural Inf Process Systems*. (2021) 34:11781–94.