Online Instance Segmentation and Reconstruction of Ultrasound Vascular Videos

Jiuan Chen[1,2], Mingcong Chen[3,5], Sili Zou[4], Jianjin Wu[4], Gaofeng Meng[1,2,5], Hongbin Liu[1,2,5,6]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems(MAIS),
Institute of Automation, Chinese Academy of Sciences,Beijing,China
[3] Department of Biomedical Engineering, City University of Hong Kong, Hong Kong, China
[4] Vascular & Endovascular Surgery Department, Second Affiliated Hospital of Naval Medical University, Shanghai, China
[5] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation,
Chinese Academy of Sciences, Hong Kong,China
[6] School of Biomedical Engineering and Imaging Sciences, King's College London, London SE1 7EU, UK

***Abstract*—The application of ultrasound in interventional surgery faces many challenges due to its lack of clarity. While existing algorithms can process single-frame ultrasound images efficiently, they still suffer from poor accuracy and discontinuous detection. In this paper, we propose an online instance segmentation network for ultrasound-guided interventional surgery Videos. When detecting the current frame, the fusion of previous frames enhances the accuracy and continuity of the segmentation. Meanwhile, real-time 3D reconstruction of vessels and interventional instruments is also achieved through the collaboration of robotic arms. Furthermore, we construct a novel dataset for vascular interventions. It accurately labels vessels and interventional instruments in 112 ultrasound videos, making it suitable for tasks related to the detection and segmentation of vascular ultrasound images. Experiments demonstrate that the proposed network improves detection accuracy by 11.0% mAP75 compared to the state-of-the-art method.**

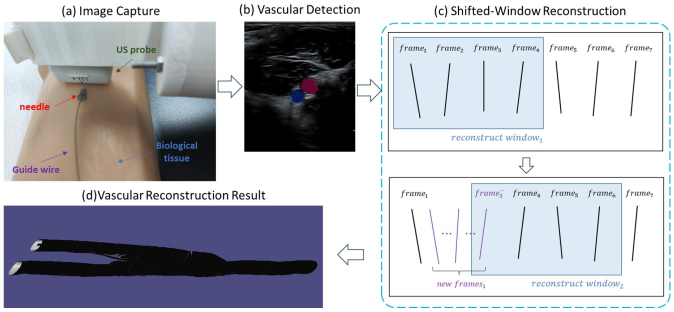
***Index Terms*—Ultrasound, Vascular interventional surgery, Video instance segmentation, Reconstruction**

I. Introduction

Vascular interventional surgery has increasingly become an integral part of modern medicine, offering treatment for patients with arterial blockages, aneurysms, and other vascular diseases [1], [2]. This method heavily relies on precise imaging techniques, particularly X-rays and ultrasonography, to ensure accurate placement of catheters, guidewires, or other interventional instruments.

Although ultrasonography offers real-time soft tissue visualization, X-rays are still favored in interventional surgeries due to their enhanced clarity with contrast agents and superior resolution for metallic instruments. However, the radiative nature of X-rays entails potential risks from prolonged exposure. ultrasonography presents a feasible alternative to mitigate these concerns. However, for ultrasonography to replace X-rays in vascular intervention, a series of challenges must be addressed as follows:

- Ultrasound image is constrained by the probe length, limiting its field of view and making it less comprehensive than single X-rays image.



(a) Image Capture
US probe
needle
Guide wire
Biological tissue
(b) Vascular Detection
(c) Shifted-Window Reconstruction
$frame_1$ $frame_2$ $frame_3$ $frame_4$ $frame_5$ $frame_6$ $frame_7$
$reconstruct\ window_1$
$frame_1$ $frame_3^-$ $frame_4$ $frame_5$ $frame_6$ $frame_7$
...
$new\ frames_1$
$reconstruct\ window_2$
(d)Vascular Reconstruction Result

Fig. 1. Overview of the whole online segmentation and reconstruction system. (a) ultrasound scanning with the collaboration of a robot arm (b) online instance segmentation of the vascular video (c) real-time 3D reconstruction using shifted window (d) reconstruction result of the vessels.

- Ultrasound performs well in imaging soft tissues, but is limited in imaging dense objects such as bone or metal interventional instruments.
- Ultrasound imaging frequently yields indistinct tissue borders, with targets intermittently or persistently vanishing in individual frames.

This work is aim to enhance the diagnostic capabilities of ultrasonography through the application of video instance segmentation. Accurate segmentation and detection results enhance real-time three-dimensional vascular reconstructions, thereby improving diagnostic and interventional precision. To achieve this, we develop a dynamic segmentation and reconstruction system, as illustrated in Fig. 1, in collaboration with a robot arm. Our contributions in this work include:

- Constructing a novel video instance segmentation dataset [3] of vascular interventions. This dataset includes categories of vascular structures and interventional instruments, such as guidewires, needles.
- Proposing an end-to-end online instance segmentation network for ultrasound video analysis, which can enhance segmentation accuracy and tracking continuity.
- Proposing a real-time 3D reconstruction algorithm for visualizing the interior of vessels during interventional

procedures. The reconstruction is facilitated by a robotic arm equipped with an ultrasound probe, setting the foundation for future advancements in automated ultrasonography with robotic assistance.

II. Related Works

A. Segmentation Using Deep Learning

Segmentation networks typically adopt an encoder-decoder architecture. The Fully Convolutional Network (FCN) [4] adheres to this design pattern, and many researchers apply it to medical image segmentation. Christ et al. [5] proposed a cascaded FCN to enhance the accuracy of liver segmentation. This design's strength is in using different filters at each stage, significantly enhancing segmentation quality. Similarly, Wu et al. [6] explored the potential of cascaded FCN for fetal boundary detection in ultrasound images.

U-Net [7] is an enhancement and extension of the FCN. With minimal training data, U-Net demonstrates exceptional performance in medical image segmentation tasks. It was employed for ultrasound image segmentation in many works [8]- [10]. However, these studies still focused on single-frame images processing, neglecting temporal information. Cicek et al. developed a 3D U-Net [11] to enrich the U-Net architecture with more spatial information. They posited that adjacent 2D image slices convey nearly identical information, leading to the idea of volumetric segmentation. Nevertheless, processing video data as 3D volumes is only feasible for offline videos. To better harness the temporal information in ultrasound images, Seo et al. [12] utilized Long Short-Term Memory neural networks (LSTM) to track vessel wall motion in ultrasound imaging. However, the computational complexity of LSTM might introduce latency, which is not ideal for real-time video processing.

B. Ultrasound Image Reconstruction

The volume rendering technique is one of the most efficient solutions for 3D medical image reconstruction. Dong et al. [13] proposed a novel method for freehand 3D ultrasound reconstruction. They used a 3D Approximate Nearest Neighbor (ANN) algorithm to improve the computational efficiency. Chen et al. [14] presented a 3D ultrasound volume reconstruction approach based on a kernel regression model. This method maps each pixel from the sampled images to the corresponding voxel in the reconstructed volume data, followed by kernel regression to perform optimization. Moon et al. [15] used piecewise-smooth Markov Random Field (MRF) model in their work. Compared to traditional geometric interpolation-based methods, this approach excels in noise reduction and boundary preservation. Beyond freehand techniques, there are also solutions that collaborate with robotic arms. Jiang et al. developed an automated ultrasound scanning system by analyzing ultrasound image quality using a neural network [16], [17]. Chen et al. [18] improved the scanning process by incorporating force control to ensure smoother scans and built an automated ultrasound scanning system. Suligoj et al. [19] achieved 3D reconstruction of ultrasound images with the assistance of a robotic arm guided by a depth camera.

III. Methods

A. Network Architecture

Given the limitations in accuracy and continuity observed when conducting segmentation on individual images, our work introduces an online instance segmentation approach for videos. Inspired by IDOL [20], an attention mechanism is utilized to extract salient image features. Subsequently, contrastive learning is applied to generate more discriminative instance embeddings as shown in Fig. 2.

Transformer module: The feature maps obtained from backbone are feed into the Deformable DETR [21] module, where the attention mechanism focuses more precisely on specific regions rather than the whole image. The input object queries are transformed into $D_o \in \mathbb{R}^{N \times d}$, where N represents the number of instances to be decoded, and d represents the dimensions of instance embeddings.

Output heads: The class head maps embeddings to category probabilities, while the box head, through a 3-layer FFN and linear projection, outputs four-dimensional bounding box coordinates. Inspired by CondInst [22], a mask branch emerges from Transformer encoder to provide multi-scale feature maps F_{mask} for the mask head. The center position of the bounding box for position-sensitive feature extraction is also fused with F_{mask} to provide positional guidance during mask generation.

Instance association: Inspired by SimOTA [23], the module computes the Intersection over Union (IoU) between predictions and the ground truth. Then it sums the highest 10 IoUs to obtain a value referred to as D_k as:

$$D_k = \max\left(1, \sum\nolimits_{P_i \in \text{Top10}_{\text{IoU}}} IoU(GT, P_i)\right) \tag{1}$$

where P denotes the prediction of the network. Then the module employ the cost function to identify the smallest top D_k samples as positive examples, while considering all others as negative examples. The cost function is as follows:

$$cost = \mathcal{L}_{OTA}^{cls} + \alpha \mathcal{L}_{OTA}^{reg} \tag{2}$$

where L_{OTA}^{cls} denotes the classification loss, and L_{OTA}^{reg} is utilized to quantify the divergence between the predicted bounding boxes and ground truths.

Contrastive head: To ensure that embeddings of the same instances in different images become as similar as possible, a contrastive head, consisting of a lightwesight FFN, is employed. Before being sent to the contrastive head, the embeddings of the reference frame will be dynamically divided into positive and negative samples by the instance association module. Then the contrastive loss $\mathcal{L}_{contrast}$ can be calculated as:

$$\mathcal{L}_{contrast} = \log\left[1 + \sum_{e_r^+} \sum_{e_r^-} \exp\left(e_k \cdot e_r^- - e_k \cdot e_r^+\right)\right] \tag{3}$$

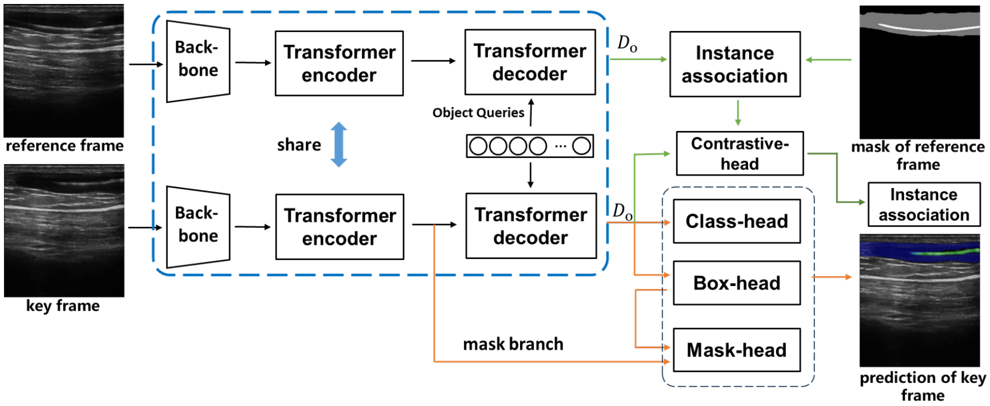


Fig. 2. Architecture of training network we propose (US-VIS). The blue box indicates feature extraction through dual-branch backbone and transformer modules. The orange arrows denote tasks performed by different output heads, and the green arrows show the instance association using contrastive learning.

where e_r^+ and e_r^- are positive and negative embeddings in the reference frame. The e_k is the embeddings of key frame. All of them are projected into a new vector space by the contrastive head, where embeddings of the same object instance are closer.

B. Loss Function

Assuming that N instances are predicted in a single image, the optimal match between predictions and ground-truth is computed using the Hungarian algorithm [24]. For the prediction of the bounding box $\hat{bi}$ corresponding to the i-th instance and the ground truth bounding box b_i, the $\mathcal{L}_{box}$ can be calculated with a composite of $\mathcal{L}_1$ loss and the Generalized Intersection over Union (GIoU) [25] loss $\mathcal{L}_{giou}$ as:

$$\mathcal{L}_{box} = \sum_{i}^{N} \left[\lambda_{b1} \| \, b_i - \hat{b}_i \, \|_1 + \lambda_{b2} . \mathcal{L}_{giou}\left(b_i, \hat{b}_i\right)\right] \tag{4}$$

whereλ_{b1} andλ_{b2} denote weights that can be set manually. And the $\mathcal{L}_{mask}$ can be calculated as:

$$\mathcal{L}_{mask} = \sum_{i}^{N} \left[\mathcal{L}_{Dice}\left(m_i, \hat{m}_i\right) \; + \; \mathcal{L}_{Focal}\left(m_i, \hat{m}_i\right)\right] \tag{5}$$

where m_i and $\hat{m}_i$ are the ground-truth and predictions of the mask for the i-th instance, respectively.

Finally, the total loss $\mathcal{L}$ can be computed as:

$$L = \mathcal{L}_{cls} + \mathcal{L}_{box} + \lambda_1 \mathcal{L}_{\text{mask}} + \; \lambda_2 \mathcal{L}_{contrast} \tag{6}$$

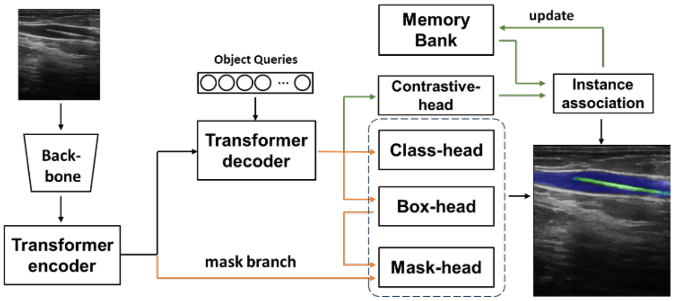


Fig. 3. Architecture of US-VIS model used to predict on the current frame

where $\mathcal{L}_{cls}$ is computed using the cross-entropy function. The default values for the loss weights λ_1 and λ_2 are 1.0 in our experiments, respectively.

C. Network Inference

The trained contrastive head is capable of extracting more discriminative instance embeddings from current frame. Furthermore, a memory bank is established to preserve these highly discriminative embeddings as depicted in Fig. 3. Given M instances' embeddings already stored and N instances predicted in the current frame, the similarity between a predicted instance embedding e_i and a memory bank's stored instance embedding $\bar{e}_j$ is computed as follows:

$$f\left(i, j\right) = \left[\frac{\exp\left(\bar{e}_j \cdot e_i\right)}{\sum_{m=1}^{M} \exp\left(\bar{e}_m \cdot e_i\right)} + \frac{\exp\left(\bar{e}_j \cdot e_i\right)}{\sum_{n=1}^{N} \exp\left(\bar{e}_j \cdot e_n\right)}\right] /2 \tag{7}$$

If $f\left(i, j\right)$ exceeds a certain threshold (set to 0.5 in our case), we consider the currently predicted e_i to be a match with the historically predicted e_j in memory bank. If no match is found but e_i has a high prediction score obtained by the class head, then a new identity is assigned for it. After the match of all the instances predicted in the current frame, the strategy for updating the embeddings of the j-th instance $\bar{e}_j$ is determined based on two factors: the frame interval t between the image F_t in the memory bank and the current frame, and the predicted score of $\mathbf{e}_j^t$ on the F_t. The formula is as follows:

$$\bar{e}_j = \frac{\sum_{t=1}^{T} e_j^t \times \left(score_j^t \times \frac{T}{t}\right)}{\sum_{t=1}^{T} score_j^t \times \frac{T}{t}} \tag{8}$$

where T denotes the size of memory bank, which we set to 5 in our experiments. The $score_j^t$ denotes the prediction score of the j-th instance on the F_t.

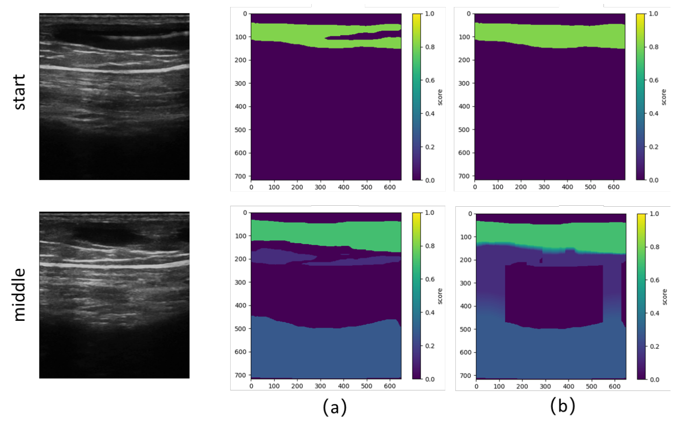


Fig. 4. (a) Predictions of two frames in the same video (b) The interpolated Vascular Score map

D. Segmentation Enhancement

In ultrasound imagery, instruments represent small targets with indistinct features, leading to potential misidentification by the network. This method applies nonlinear interpolation to the vascular prediction score map V, utilizing it to weight the prediction scores of instruments D. Consequently, the proximity of an instrument to regions with higher vascular prediction scores directly enhances the confidence level of its predicted presence.The new prediction of instruments can be calculated as:

$$score_D = \frac{\sum(V \times D)}{\sum D} \tag{9}$$

E. 3D Reconstruction

As shown in the Fig. 1, The robotic arm drives the ultrasound probe, capturing images via a stream capture device. In order to accomplish the 3D reconstruction, the 2D imaging coordinate needs to be converted to 3D robot coordinates. The probe's lateral axis is designated as $\vec{p_t}$ and its normal direction as $\vec{p_n}$. Given the 2D image coordinates of the pixel as $\begin{bmatrix} u & v \end{bmatrix}$, the three-dimensional coordinates of the pixel in the end-effector system $Pixel_{base}$ can be computed as:

$$Pixel = u\vec{p_t} + v\vec{p_n} \tag{10}$$

$$\begin{bmatrix} Pixel_{base} & 1 \end{bmatrix} = {T_r}^{-1} \begin{bmatrix} Pixel & 1 \end{bmatrix} \tag{11}$$

where T_r is calculated based on the robot's end position. Then real-time 3D reconstruction can be performed by shifted-window Bezier interpolation [26]. The Bezier interpolation is used to interpolate smooth curves or surfaces between a given set of control points. Bezier curve can be defined using n control points to shape its contour as:

$$B(t) = \sum_{i=0}^{n-1} C_{n-1,i}(1-t)^{n-i-1}t^i P_i \tag{12}$$

where P_i represents the coordinates of the control point in the base coordinate system,$B\left(t\right)$ denotes the interpolated point, and t varies within $[0, 1]$, indicating its position on the Bezier curve.

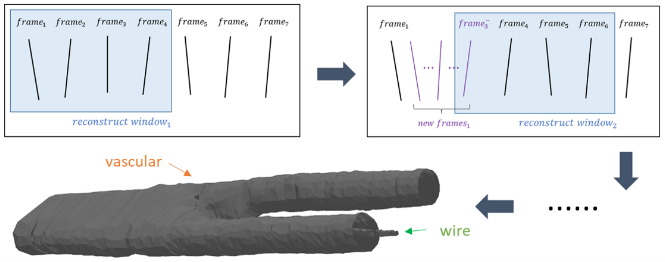


Fig. 5. Dynamic reconstruction of blood vessels and guidewires

Upon acquiring four image frames with a robotic arm, 3D reconstruction is conducted on the instance segmentation results, dynamically rebuilding each instance's mask in three dimensions. As depicted in Fig. 5, shifted-window Bezier interpolation is utilized for each quartet of frames, specifically $[\text{frame}_1\ \text{frame}_2\ \text{frame}_3\ \text{frame}_4]$. After processing the first reconstruction window with four frames, and the third frame frame_3 is replaced by the interpolated frame_3^- of first window. Interpolation is then applied to the next window, $\left[\text{frame}_3^-\ \text{frame}_4\ \text{frame}_5\ \text{frame}_6\right]$. When the four frames in one window are reconstructed, the result is displayed, and subsequently, the next window undergoes new reconstruction. This process achieves the visual effect of dynamic 3D reconstruction.

IV. Experimental Result and Discussions

A. Dataset construction

This novel dataset we construct contains common imaging categories observed in vascular interventional surgeries, such as vessels, guidewires, and needles. Additionally, it includes various imaging scenarios like in-plane and out-plane. Besides,

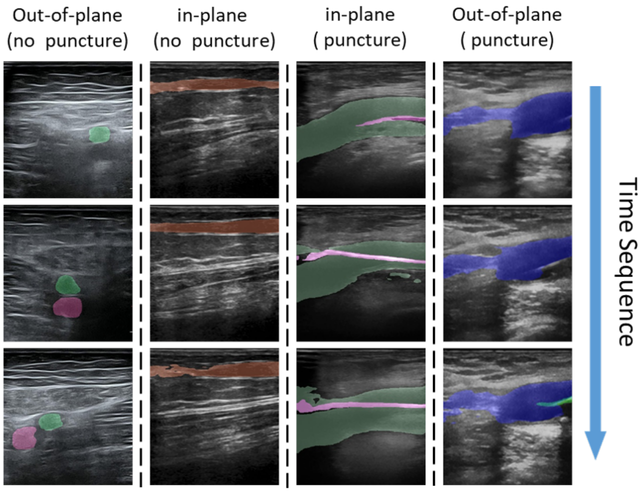


Fig. 6. Segmentation results by proposed online video instance segmentation network in different cases.

TABLE I
ACCURACY OF IMAGES WITH INTERVENTIONAL DEVICES

methods	**Out-of-plane**				**In-plane**				**Mean**			
	mAP50	**mAP75**	**mAR**	**mDice**	**mAP50**	**mAP75**	**mAR**	**mDice**	**mAP50**	**mAP75**	**mAR**	**mDice**
MA-Net	72.7	39.5	53.4	83.9	65.3	44.7	74.2	77.7	69	42.2	63.8	80.8
U-Net	77.3	53.2	65.8	82.3	60.2	44.2	73	79.8	68.7	48.7	69.4	81.1
SegFormer	77.1	55.6	63.1	84.3	63.3	43.4	79.5	86.3	70.2	49.5	71.3	85.3
TransUNet	80.8	66.8	82.3	84.1	71.8	55.1	77.1	**86.6**	76.3	60.9	79.8	**85.3**
US-VIS	**85.1**	**80.7**	**88.5**	**85.2**	**77.1**	**62.7**	**81.6**	83.2	**83.5**	**71.7**	**85.1**	84.2

TABLE II
ACCURACY OF IMAGES WITHOUT INTERVENTIONAL DEVICES

methods	**FPS**	**mAP50**	**mAP75**	**mAR**	**mdice**
MA-Net	17.8	55.7	23.5	61.5	66.6
U-Net	**28.4**	61.1	27.8	63.1	68.7
SegFormer	21.2	49.2	26.4	65.1	69.8
TransUNet	3.54	64.8	38.1	67.3	69.1
US-VIS	10.3	**68.1**	**47.1**	**70.9**	**70.3**

TABLE III
EFFICIENCY OF RECONSTRUCTION ALGORITHMS

Methods	FPS	Output Size
Point cloud with mesh generation	**26.8**	3.2M
Ours	21.7	**1.1M**

the images are captured from different anatomical regions, including the carotid and femoral artery.

We collected vascular ultrasound videos during routine ultrasound examinations and interventional procedures using the linear probe with an ultrasound system (Angell technology Ltd, China). To date, we have gathered 40 cases, totaling 112 ultrasound interventional videos. Additionally, we collected and processed another non-surgery 2,000 vascular ultrasound images. Inspired by the Youtube-VIS2021 [3] data format, we established the US-Vascular-VIS dataset for tasks related to the detection and segmentation of interventional ultrasound.

B. *Experimental settings and procedures*

The network US-VIS we proposed was trained based on the US-Vascular-VIS dataset on a workstation (DELL, Nvidia RTX3090Ti). Several Out-performing networks that have been widely used in the medical image field, including SegFormer [27], MA-Net [28], UNet [10], and TransUNet [29], were ran on the same dataset to compare them with our proposed network.

For non-puncture conditions, we used 10 in-plane and 10 out-of-plane ultrasound videos captured from four patients for testing, and the result is shown in the Table I. In the puncture condition, we used 10 in-plane ultrasounds captured from three patients for testing. The result is shown in the Table II.

In assessing the reconstruction algorithm's effectiveness, we initially undertook offline experiments, comparing it against ultrasound point cloud reconstruction [19] and facet generation using nearest neighbors algorithm [13].

C. *Experimental results*

We present representative detection and segmentation results depicted in Fig 6. Our model demonstrated impressive performance in real-time target tracking and re-detection of lost targets in ultrasound videos.

In terms of detection accuracy, our proposed model has better performance. Table I quantitatively shows the online detection results of ultrasound videos without interventional instruments. The average accuracy of vascular detection by our proposed network structure achieved a mAP75 score of 71.7%, which is 10.8% higher than the second best one. Especially in the out-of-plane situation, our network achieved a mAP75 score of 80.7%. Table II displays the online detection results of videos with interventional instruments. The appearance of small targets, such as metallic interventional instruments, in ultrasound images makes detection more challenging. The network we proposed achieved is 11.0% higher than the second best one.

In terms of segmentation accuracy, the TransUNet outperformed ours by 3.4% in dice score on large vessels (in-plane images). The employment of the global self-attention mechanism from Transformers within TransUnet contributes significantly to its superior performance in segmentation tasks. The network we proposed, which employs Deformable Detr [21], achieves a 0.5% higher accuracy in detecting small instruments. The utilization of Deformable DETR enhances the network performance for small objects, such as interventional instruments, due to its specialized attention mechanism.

While accuracy improves, the attention mechanism reduces the inference speed significantly. The TransUNet achieves 3.54 FPS on an NVIDIA 3060 graphics card, while our proposed network reaches 10.3 FPS. The U-Net achieves an inference speed of 28.4 FPS because of its simpler network structure.

The accuracy of the reconstruction algorithm relies on the precision of vessel segmentation, thus the evaluation of reconstruction efficiency is the main focus of our experiments. As shown in Table III, we achieved a processing speed of 21.7 FPS, with a final output file size of 1.1 MB for single-vessel reconstruction. Our algorithm is a bit slower in comparison, but the output file is smaller, allowing the reconstruction results to be displayed more easily and dynamically. Considering that we used a mature and encapsulated codebase when implementing the point cloud reconstruction algorithm, so the results are within acceptable limits.Of course, we will explore more alternative evaluation metrics for reconstruction in the future.

V. Conclusion

In this paper, we proposed an end-to-end online instance segmentation network to process interventional vascular ultrasound videos. Instead of processing single-frame images, we explored the use of historical frames to improve the accuracy and continuity of segmentation. Based on the accurate segmentation, we realized the real-time 3D reconstruction of vessels with the collaboration of the robot arm, enabling surgeons or robotic systems to acquire an intuitive insight into the intravascular state. Meanwhile, we built a novel dataset that can be used in multiple deep learning tasks for vascular interventional ultrasound. In the future, we will expand our segmentation and reconstruction efforts to include more interventional instruments, such as balloon catheters , and perform further experiments with the assistance of robots.

Acknowledgment

The authors gratefully acknowledge the support by the Institute of Automation, Chinese Academy of Sciences, Second Affiliated Hospital of Naval Medical University and the InnoHK programme.

References

[1] S. Toggweiler, R. Gurvitch, J. Leipsic, D. Wood, A. Willson, R. Binder, A. Cheung, J. Ye, and J. Webb, “Percutaneous aortic valve replacement: vascular outcomes with a fully percutaneous procedure,” in Journal of the American College of Cardiology, vol. 59, no. 2, 2012, pp. 113-8.

[2] C. Grimaldi, F. di Francesco, F. Chiusolo, R. Angelico, L. Monti, P. Muiesan, and J. de Ville de Goyet, “Aggressive prevention and preemptive management of vascular complications after pediatric liver transplantation: A major impact on graft survival and long-term outcome,” in Pediatric Transplantation, vol. 22, 2018.

[3] L. Yang, Y. Fan, and N. Xu, “Video instance segmentation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5188-5197.

[4] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, 2015, pp. 640-651.

[5] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, et al., “Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, 2016.

[6] L. Wu, X. Yang, S. Li, et al., “Cascaded Fully Convolutional Networks for automatic prenatal ultrasound image segmentation,” in 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017. IEEE.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[8] C. Chu, J. Zheng, and Y. Zhou, “Ultrasonic thyroid nodule detection method based on U-Net network,” in Computer methods and programs in biomedicine, vol. 199, 2020, pp. 105906.

[9] M. Amiri, R. Brooks, and H. Rivaz, “Fine-Tuning U-Net for Ultrasound Image Segmentation: Different Layers, Different Outcomes,” in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, 2020, pp. 2510-2518.

[10] C. G. Morales, J. Yao, T. Rane, R. Edman, H. Choset, and A. Dubrawski, “Reslicing Ultrasound Images for Data Augmentation and Vessel Reconstruction,” in 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 2710-2716.

[11] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016.

[12] J. Seo, L. S. Nguon, and S. Park, “Vascular wall motion detection models based on long short-term memory in plane-wave-based ultrasound imaging,” in Physics in medicine and biology, 2023.

[13] J. Dong, W. Cong, D. Ai, Y. Chu, Y. Huang, H. Song, Y. Jiang, Y. Wang, M. Li, and J. Yang, “Multiresolution Cube Propagation for 3-D Ultrasound Image Reconstruction,” in IEEE Transactions on Computational Imaging, vol. 5, 2019, pp. 251-261.

[14] X. Chen, T. Wen, and X. Li, et al., “Reconstruction of freehand 3D ultrasound based on kernel regression,” in BioMed Eng Online, vol. 13, 2014, p. 124.

[15] H. Moon, G. Ju, S. Park, and H. Shin, “3D freehand ultrasound reconstruction using a piecewise smooth Markov random field,” in Computer Vision and Image Understanding, vol. 151, 2016, pp. 101-113.

[16] Z. Jiang, Z. Li, M. Grimm, M. Zhou, M. Esposito, W. Wein, W. Stechele, T. Wendler, and N. Navab, “Autonomous robotic screening of tubular structures based only on real-time ultrasound imaging feedback,” IEEE Transactions on Industrial Electronics, vol. 69, no. 7, pp. 7064–7075, 2021.

[17] Z. Jiang, Y. Gao, L. Xie, and N. Navab, “Towards autonomous atlas-based ultrasound acquisitions in presence of articulated motion,” IEEE Robotics and Automation Letters, 2022.

[18] M. Chen, Y. Huang, J. Chen, T. Zhou, J. Chen, and H. Liu, “Fully Robotized 3D Ultrasound Image Acquisition for Artery,” in 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 2690-2696, 2023.

[19] F. Suligoj, C. M. Heunis, J. Sikorski, and S. Misra, “RobUSt–an autonomous robotic ultrasound system for medical imaging,” IEEE Access, vol. 9, pp. 67456-67465, 2021.

[20] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. L. Yuille, and X. Bai, “In Defense of Online Models for Video Instance Segmentation,” in European Conference on Computer Vision, 2022.

[21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable Transformers for End-to-End Object Detection,” arXiv preprint arXiv:2010.04159, 2020.

[22] Z. Tian, C. Shen, and H. Chen, “Conditional convolutions for instance segmentation,” in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pp. 282-298, Springer, 2020.

[23] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, “Ota: Optimal transport assignment for object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 303-312, 2021.

[24] H. W. Kuhn, “The Hungarian method for the assignment problem,” Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83-97, 1955.

[25] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 658-666, 2019.

[26] Q. Huang, Y. Huang, W. Hu, and X. Li, “Bezier interpolation for 3-D freehand ultrasound,” IEEE Transactions on Human-Machine Systems, vol. 45, no. 3, pp. 385-392, 2014.

[27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in Advances in Neural Information Processing Systems, vol. 34, pp. 12077-12090, 2021.

[28] T. Fan, G. Wang, Y. Li, H. Wang, “MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation,” IEEE Access, vol. 8, pp. 179656-179665, 2020.

[29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “TransUNet: Transformers make strong encoders for medical image segmentation,” arXiv preprint arXiv:2102.04306, 2021.