2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

AiAReSeg: Catheter Detection and Segmentation in Interventional Ultrasound using Transformers

Alex Ranne[1], Yordanka Velikova[2], Nassir Navab[2], and Ferdinando Rodriguez y Baena[1]

***Abstract*— This work proposes a state-of-the-art transformer architecture to detect and segment catheters in axial interventional Ultrasound image sequences. The network architecture was inspired by the Attention in Attention mechanism, temporal tracking networks, and introduced a novel 3D segmentation head that performs 3D deconvolution across time. To train the network, we introduce a new data synthesis pipeline that uses physics-based catheter insertion simulations, along with a convolutional ray-casting ultrasound simulator to produce synthetic ultrasound images of endovascular interventions. The proposed method is validated on a hold-out validation dataset, thus demonstrated robustness to ultrasound noise and a wide range of scanning angles. It was also tested on data collected from silicon aorta phantoms, thus demonstrated its potential for translation from sim-to-real. This work represents a significant step towards safer and more efficient endovascular surgery using interventional ultrasound.**

I. INTRODUCTION

Cardiovascular disease is the most common cause of death in the world, accounting for 17.9 million deaths per annum [1]. Traditionally, open surgery is performed to expose the diseased vasculature, which poses significant trauma for the patient. As an alternative, computer-assisted minimally invasive endovascular surgery has been widely adopted due to its benefits of reducing patient recovery time, and lower risk of infection, thus saving costs for healthcare providers, and more importantly saving lives.

In endovascular surgery, catheters and guidewires are steered, under Fluoroscopic guidance, through tortuous vessel trees to reach their desired destination [2]. During navigation, staff and patient are exposed to prolonged periods of ionising radiation, which increases the risk of developing cancer. During an intervention, in order to visualise the vessels, the patient is also injected with a radiopaque dye (Digital Subtractive Angiography, DSA), which is harmful for the kidneys. On the other hand, this system still lacks the ability to obtain feedback on real-time instrument positions relative to the vasculature. This may introduce additional risks for the patient, as there may be frequent and unintentional contacts between these instruments and the vessel wall, with the consequent risks of perforation, dissection, thrombosis and embolization [3].

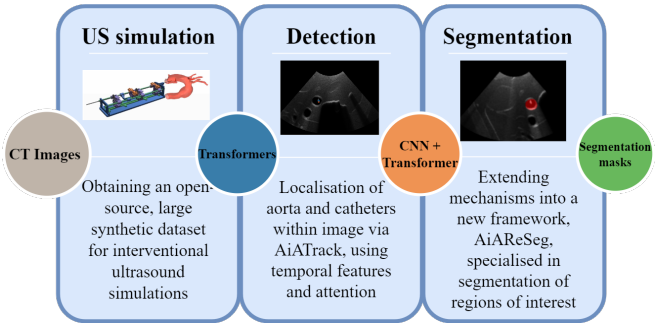

Fig. 1. Main workflow pipeline of proposed system. Stage 1: Data synthesis via physics engine and ray casting. Stage 2: Detection of critical anatomy locations. Stage 3: Semantic segmentation

Alternatively, intraoperative Ultrasound imaging (iUS) offers a non-ionising solution for visualisation. In comparison to Fluoroscopy, US imaging has been a popular tool in diagnosis and aneurysm screening [4], due to the high tissue contrast, temporal resolution, and efficacy [5]. In surgery, clinicians have applied it in conjunction with or completely replacing Fluoroscopy, in endovascular aneurysm repair [6], [7], Balloon Angioplasty [8], and Electrophysiology [9].

In order to monitor the instruments, the surgeon must detect the tip position of the catheter in US images, which poses a significant challenge. To begin with, the spatial resolution of an US image is limited by the number of elements in the transducer, and determined by the signal frequency [10]. In order to examine deep into the target tissue, the clinician must lower the frequency as waves with higher wavelength experience less attenuation [11]. However, this is at the expense of spatial resolution. Secondly, the noisy nature of ultrasound makes interpretation of images difficult, since images contain clutter, shadowing, and reverberation artifacts. Consequently, labeling and interpreting the image requires expert knowledge in order to relate the physical anatomy with the image, which may vary in quality, resolution, intensity, and acquisition protocols, and are not standardized. The progress in deep learning for US instrument detection and segmentation in endovascular procedures offers an opportunity for the field to reform. With the development of object detection networks, researchers have identified their potential in finding objects of varying sizes, and streamlining their workflow. Designing a suitable architecture for this task, and acquiring sufficient data to do so are two challenges that need to be solved.

In this paper, we propose a novel three-step framework to overcome the lack of intraoperative ultrasound data of a

This work was supported by the UKRI CDT in AI for Healthcare under Grant EP/S023283/1, the ICL-TUM Joint Academy of Doctoral Studies (JADS) program, and the TUM Global Incentive Fund. Github Repository: https://github.com/alex-613/AiAReSeg

[1]Alex Ranne and Ferdinando Rodriguez y Baena are with the Hamlyn Centre for Robotic Surgery, Imperial College London, SW7 2AZ, UK (e-mail: {alex.ranne17, f.rodriguez}@imperial.ac.uk) (Corresponding author: Alex Ranne).

[2]Yordanka Velikova and Nassir Navab are with the Chair for Computer Aided Medical Procedures and Augmented Reality (CAMP), Technical University of Munich, 85748 Garching, Germany (e-mail: {dani.velikova, nassir.navab}@tum.de)

cathetarisation, required to train our network. In Sect. III, we propose to generate synthetic iUS data with instruments inside by fusing a physics engine into an existing CT-to-US simulator, thus generating mechanically realistic scans. Once generated, this data was then used to train a novel detection and segmentation architecture (AiAReSeg), which we propose in in Sect. III-C. Finally, in Sect. IV, we evaluate the trained model on a hold out validation set of simulated US data, but also with aortic phantom images, which resemble a true surgical environment to a closer degree.

II. Related works

A. State of the art deep learning architectures

1) Detection: In terms of architecture, most popular networks fall into two categories: convolution-based (CNN) or attention-based. In CNN-based systems, Faster-R-CNN [12] leveraged the power of its region proposal network (RPN) to select regions of interest, prior to passing such features into fully connected layers for bounding box prediction. The network achieved real-time performance since the need for hand-picked anchor points (found in Fast-R-CNN [13]) was removed. However, following the introduction of attention in the Transformer architecture [14], vision transformers became a strong contender for CNNs as they can learn global dependencies from across the image with a patch-based approach, then concatenating the attention maps together to form the final prediction [15]. Much more recently, researchers have continued to evolve the field by combining the benefits of both worlds, fusing a CNN feature extractor with the attention mechanism. From this idea emerged numerous variants of the transformer, such as the Detection Transformer (DETR) [16], which used a ResNet50 backbone for feature extraction, before feeding its output into a transformer that provided embeddings corresponding to various objects in the scene. Using the bipartite matching loss [17], the network minimised the difference between a prediction output and the ground truth in a class-specific manner.

2) Semantic Segmentation: In the context of semantic segmentation, the CNN-based UNet [18] architecture and its adaptations have also performed exceptionally well in several medical ultrasound segmentation tasks [19], [20]. Following the introduction of the nnUNet pipeline [21], which proposed an all-in-one pre-processor, parameter and model selection pipeline, the performance of UNet has been further refined. With that said, nnUNet does not operate in real-time, making it not suitable for high-speed US ($<$10fps). On the other hand, attention-based segmentation networks have also seen much success, such as with the DETR to perform panoptic segmentation tasks [16], or in the case of the Segmentation Transformer (SETR) [22], which removed the ResNet50 backbone, but used a Sigmoid activation function to generate segmentation masks.

3) AiATrack - Learning with temporal features: Thus far, aforementioend models only use spatial features learned in a single frame to make the prediction. This may be sufficient for good-quality images, but may fail when there is occlusion due to shadows or artifacts. To solve this problem, we drew inspiration from clinicians, who rely on prior knowledge from the previous position of the aorta to reposition the probe and relocate the lost targets. This concept was previously captured in the AiATrack framework [23], where a ResNet50-Transformer framework was used together with a corner-predictor based bounding box head. However, the final box prediction still only draws information from the transformer decoder outputs, instead of across the entire sequence of data. We believe we can further improve this invention to operate on even more challenging tasks, such as locating a small catheter's cross-section in sequences of highly variable US images.

B. State of the art in US image simulation

Despite the impressive results achieved by deep learning, the majority of network architectures are supervised, learning from an extensive number of labeled ground truths, which for Ultrasound is not readily available due to the difficulties faced in acquisition. However, there are publicly available large sets of pixel-level labeled CT volumes, which can be translated into simulated US image/label pairs [24] where further data augmentation can be added via applying rotation, brightness jitter, random shadowing, and artificial tissue deformations, etc. In this way, a large training dataset can be generated for the pretraining of deep learning architectures, allowing the networks to learn domain-specific feature extraction, before retraining on a significantly smaller, real dataset. Evidence of successful transfer shown previously in Velikova et al.'s work [25], [26] motivates this idea.

III. Methodology

A. CT Data selection

Labeled CT volumes of 8 men and women were acquired from the publicly available dataset on Synapse [27]. The labels of bone, fat, skin and lungs were added in the label map to allow for the simulation to function. The detailed process of generating the interventional US is detailed below.

B. Physics-based catheterisation simulation

Since obtaining a large dataset for the initial training of a deep learning architecture is time consuming, we are proposing a new US data simulation pipeline for generating interventional data, which is otherwise only attainable from the operating room environment or via a phantom.

In literature, there are two ways to generate US data: finite difference solutions of the wave equation [28], [29], or ray-casting through an image volume, semantically labelled with its acoustic attenuation properties [26]. Since the former method requires solving a large system for each frame, it typically requires significantly longer computation time. Thus, the second method, albeit not as accurate, was selected. The simulator selected uses a hybrid ray-tracing convolutional method to define an anatomical representation that mimics the texture of real US images, define anisotropic properties, generate artifacts, and provide tissue contrast that allows regions of interest to be easily discernable.

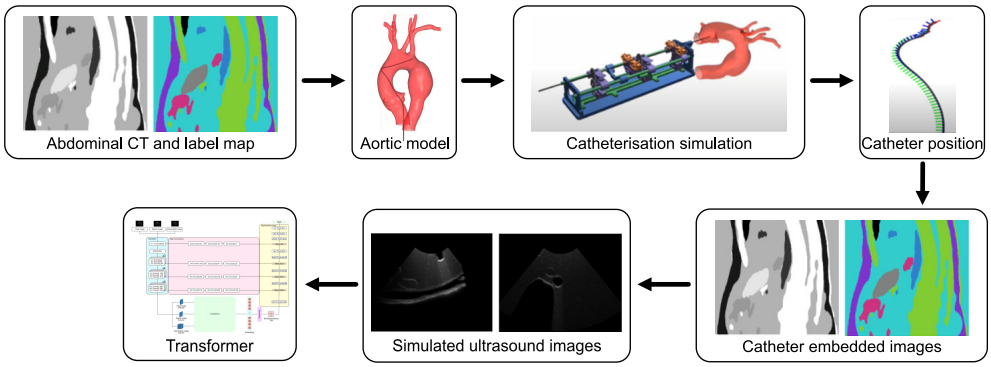

Fig. 2. Pipeline for synthetic ultrasound images generation from CT volumes. The aorta model is extracted, imported into the catheterisation engine, with the catheter positions exported as time series, then redrawn on the CT, and passed into a ray-tracing based simulator to generate the finished image

In order to generate a dataset consisting of catheters, we repurposed an open source catheterisation simulator developed by Jianu et al. [30], which is able to recreate high fidelity catheter-aorta mechanical interaction simulations. CathSim is built in the MuJoCo physics engine (DeepMind, London, UK) [31], [30], which is a powerful package that can perform real time multi-joint dynamics computations with contacts using a C based API.

The final preprocessing pipeline is detailed in Fig. 2. CathSim renders the mesh models of the aorta and the catheter separately, while the aorta mesh was divided into 1024 convex hulls, decomposed using the V-HACD algorithm. The decomposed hulls were transformed into the same coordinate frames as the simulated environment via Blender v3.2.1 (Blender Foundation, Amsterdam, Netherlands) and imported into CathSim. The insertion simulation was performed with a linear translation speed of 0.1m/s, and inserted for 1,000 time increments, where each increment represents 1/60th of a second, and positions of the catheter were sampled at regular intervals along its body, and exported into a time series csv, which was transformed back into the CT's coordinate system. Finally, the simulator was initialised with multiple splines on the surface of the patient's torso, and the splines were tilted at angles of 0, $\pm$ 30 and $\pm$ 60 degrees, and a sweep of 1,000 images were generated for each angle.

C. AiAReSeg Architecture

Attention in Attention + ResNet for Segmentation (AiAReSeg) is a novel segmentation architecture that is adapted from AiATrack [23], shown in Fig.4. The main architecture consists of three components, the attention-in-attention module, the transformer architecture, and the outer convolution-deconvolution layers. Attention-in-attention (AiA) was first proposed by Gao et al.'s work [23], where the authors observed that each query-key pair generated an independent attention map. The original attention mechanism used the following dot-product equation:

$$Attn(\mathbf{Q},\mathbf{K},\mathbf{V}) = (\text{Softmax}\left(\frac{\bar{\mathbf{Q}}\bar{\mathbf{K}}^T}{\sqrt{C}}\right)\bar{\mathbf{V}})\mathbf{W_o} \quad (1)$$

Where $\bar{Q} = QW_q$, $\bar{K} = KW_k$, $\bar{V} = VW_v$, where Q, K, V are the queries, keys and values, respectively, while W_q, W_k, W_v denotes the learnable weight arrays for the query, keys and values, $\mathbf{W_o}$ are the output weights, and C is the channel size. In the case of a noisy dataset with distracting backgrounds, the model may become confused from clutter in the scene, leading to poor predictions. However, it was noted that the attention weights near regions of interest were significantly higher, and pixels in such regions were of more interest than pixels with high attention weights further away. Thus, the designed AiA module applied attention again on the attention map M to filter out distant weights, which can be represented as:

$$InAttn(\mathbf{M}) = (\text{Softmax}\left(\frac{\bar{\mathbf{Q}}'\bar{\mathbf{K}}'^T}{\sqrt{D}}\right)\bar{\mathbf{V}}')(\mathbf{1}+\mathbf{W}'_\mathbf{o}) \quad (2)$$

Where $\bar{Q}', \bar{K}', \bar{V}'$ are intermediate weighted queries, keys and values, which were feature vectors taken from columns in $\mathbf{M}$, while D represents the intermediate channel size, defined in this case as the height of the attention map.When combined, the AiA module computes the following:

$$AiA(\mathbf{Q},\mathbf{K},\mathbf{V}) = (\text{Softmax}\left(\mathbf{M}+\text{In}(\mathbf{M})\right))\bar{\mathbf{V}})\mathbf{W_o} \quad (3)$$

AiATrack consists of three input branches, the initial frame, the current search frame, and selected intermediate

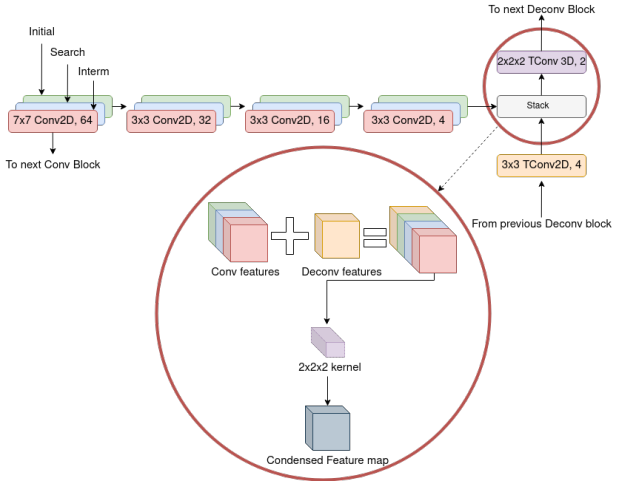

Fig. 3. Closeup of the 3D deconvolution pipeline. The three input feature maps are color coded in green, blue, and green for the initial, search and intermediate frames, respectively. The three frames are stacked with the output from the previous deconvolutonal block (amber), then deconvolution is performed with a 3D 2x2x2 kernel.

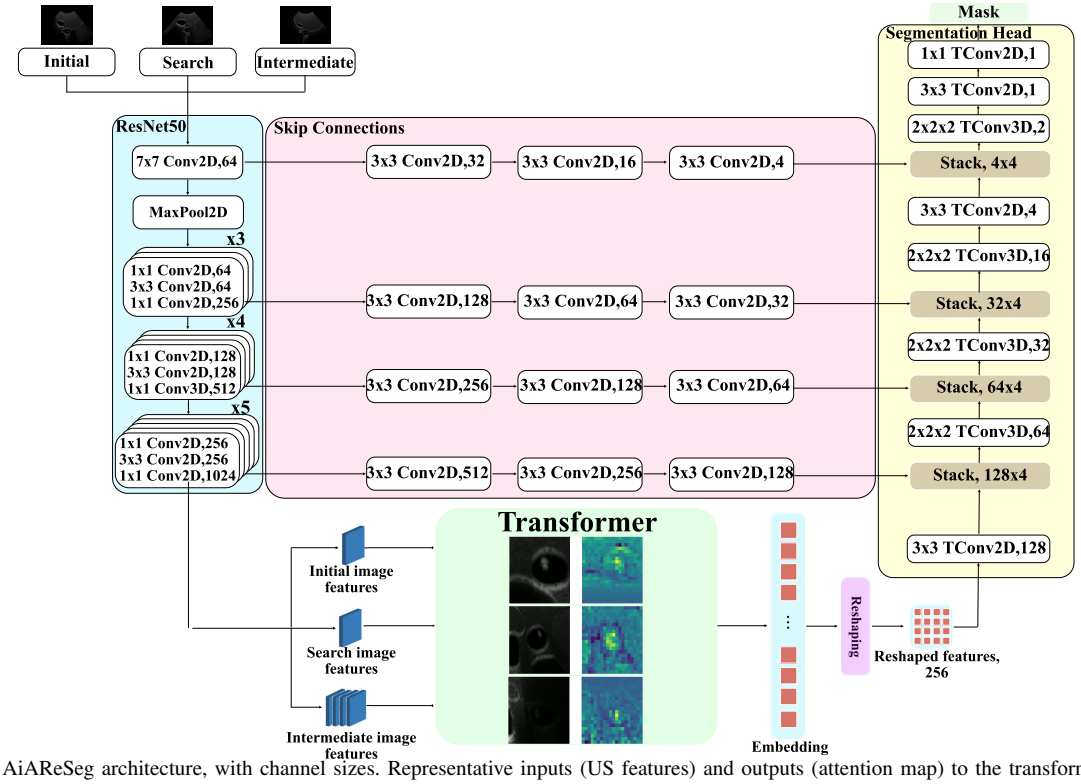

Fig. 4. The AiAReSeg architecture, with channel sizes. Representative inputs (US features) and outputs (attention map) to the transformer are shown.

frames. All frames pass through the ResNet50 feature extractor before performing self-attention in the transformer encoder, which searches for correlation within the feature array (Q=K=V). Thereafter, the feature maps from each branch were combined during the long-term (LT), and short-term (ST) cross-attention modules, thus enable learning across time (Q≠K≠V). During inferencing, the network incorporated a checking algorithm that stores high-performing prior examples (classified by the Dice metric > 0.7) in its memory, and then call upon them when predicting the current frame by concatenating it to the value. Note that the ResNet50 and the transformer encoders on each branch share their weights.

Having been inspired by AiATrack, AiAReSeg (Fig.4) aims to reconstruct segmentation masks from attention feature maps instead of bounding boxes. However, due to the nature of convolutional layers, spatial information inside of an image has been lost, and needs to be reintroduced. We retained the 3 branched ResNet/CNN (UNet) encoder to learn global features at different scales in different frames, then made two major innovations. First, we implemented a skip connection module to propagate the features to various locations in the segmentation head. During propagation, feature maps from ResNet were processed with intermediate convolutional layers to gradually reduce its channel size, such that it matches those from the deconvolution layer's intermediate outputs. Then, as shown in Fig. 3, the feature maps are stacked along a new dimension, generating (H,W,T) sized images, where T represents the time, before it was passed into a 3D convolutional layer to reduce T to 1, prior to further stacking from the next skip connection branch. We created this new dimension to help the network preserve disimiliar features, and learn 3D convolutional filters that best selects the important features across the sequence, before condensing it into the new upsampled 2D feature map.

In order to adapt AiAReSeg to both aorta and catheter, we had combined Dice coefficient (DSC), Binary Cross Entropy (BCE), and Mean Squared Error (MSE) loss functions, and weighted their importance with factors of 5, 2 and 2, respectively. The Dice loss measures the similarity between predicted and ground truth masks, and encourages models to produce more accurate segmentation results, and handles class imbalance. We also used the BCE loss to assign higher probabilities to the correct class and lower probabilities to the incorrect class, and the MSE to minimise the pixel-to-pixel distance between the ground truth and the prediction.

IV. Experimental Evaluation

We divided the evaluation of this pipeline into two phases: evaluation on hold out simulated image set, and on unseen phantom data. This test examines the capability of the network in generalising to unseen datasets, with the latter being a closer representation of the real patient anatomy.

To evaluate the performance of our system, we selected a handful of common and top-performing detection and segmentation models in literature. Most notably, this includes the Faster-R-CNN and DETR for detection, compared against the performance of AiATrack, while for segmentation we selected the standard UNet, and a clustering-based approach, which is explained in Sect. IV-C.

The models were trained and evaluated on both detection and segmentation of the aorta and catheter, evaluated separately. The reason for this choice is that analysis of the aorta is easier to perform, as it is unique and large in the input image. Catheters, on the other hand, are significantly

more challenging to detect due to the noisy nature of the background, since their shape and intensity range can easily be confused with artifacts or other features, thus affecting performance. In addition, their small size also created a significant class imbalance between the feature and the background, making the dice loss highly volatile.

We evaluated the tracking models using the average precision (AP) metric, defined as the area under the precision-recall curves, evaluated at different intersections over union (IOU) thresholds between 50 and 95%, including their average to form the mean AP score. On the segmentation side, we used the Dice metric (DSC), which indicates the degree of overlap, and the mean absolute error (MAE), which represents the distance from each pixel to its ground truth.

A. Training details

Experiments were conducted on a workstation with NVIDIA GeForce RTX3080, 32GB RAM, Intel core i7 (10700K). The physics-based simulations were performed on MuJoCo 2.10, where mesh models were decomposed into convex hulls using V-HACD [32]. The US simulations were generated on the ImFusion Suite (ImFusion GmbH, Munich, Germany), where the ray-casting algorithm was implemented [26]. 8 torso CT images of men and women were selected from the Synapse dataset, and used for catheterisation. Catheterisation simulation was performed for a total duration of 60 seconds, where a data recording of the catheter positions was performed at 4mm intervals along its body to provide a reasonable spatial resolution for reconstruction. During US simulation, the transducer was programmed to follow a predefined spline, and ray-casting simulation were performed at 0.1mm increments along the line. To increase data variability, this spline was also rotated by $\pm 30^o$, and re-projected onto the volume surface, creating different viewing angles of the anatomy.

B. Phantom data collection details

A small set of 2D testing images were collected manually in a free-hand manner from a ZONAE Z.One Ultra Ultrasound machine, using a C8-3 (3D) transducer at a scanning depth of 10cm. Axial view scanning was performed on an Elastrat silicon-based aortic arch phantom, immersed in lukewarm saline solution. To mimic a catheterisation procedure, we inserted a Merit Medical 5F vertical catheter at the distal end of the phantom, then followed the tip of the catheter with the US probe. We collected 5 US sequences with varying lengths, ranging between 400 - 700 frames.

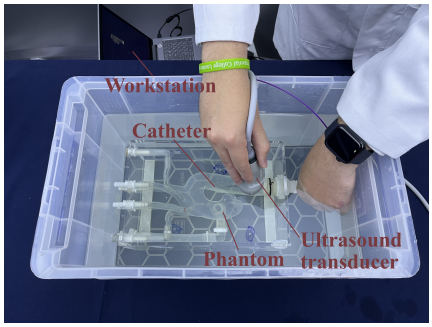

Fig. 5. Experimental setup with Ultrasound and aortic arch phantom

TABLE I

EVALUATION ON SYNTHETIC DATA: DETECTION MODELS

	Aorta			Catheter			FPS
Model Name	**AP**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**	
DETR	72.40	98.80	98.80	22.56	**77.10**	4.03	**38.5**
Faster-R-CNN	89.05	98.93	98.82	12.70	46.60	2.01	15.0
AiATrack	**94.77**	**100**	**100**	**22.86**	70.99	**6.33**	35.4

TABLE II

EVALUATION ON PHANTOM DATA: DETECTION MODELS

	Aorta			Catheter		
Model Name	**AP**	**AP50**	**AP75**	**AP**	**AP50**	**AP75**
DETR	1.4	5.3	0.3	NA	NA	NA
Faster-R-CNN	12.1	23.7	12.6	0.9	5.9	0.1
AiATrack	**45.7**	**100**	**82.62**	**14.3**	**63.93**	**3.79**

C. Model specific details

AiATrack: A patch of size 5^2 was cropped from the frame, and resized into a common dimension of 320 x 320 pixels. A ResNet-50 backbone was used [33] for feature extraction, downsampling the input until a size of 20x20 was achieved. Each feature map was flattened and passed into the transformer. A 4-head attention module was used, with the inner AiA module reducing the channel dimension of queries and keys to 64. The final prediction head used 3 Conv-BN-ReLU layers, a PrPooling layer and 2 fully connected layers. The model was pretrained for 300 epochs on the LaSOT dataset [34], then for an additional 200 epochs on the synthetic US dataset, both at a learning rate of 10^{-4}.

DETR: A standard DETR, with weights pretrained for 500 epochs on the COCO2017 dataset was used in this application [35]. The COCO dataset consists of more than 200,000 images consisting of over 80 categories of objects, thus equipping the model with the necessary feature extraction filters. The model is retrained on ultrasound data for 100 epochs. The learning rate in both cases is 10^{-5}.

Faster-R-CNN: An implementation of Faster-R-CNN from the Detectron2 library was used [36]. A ResNet50 backbone was used, together with a feature pyramid network. A COCO2017 pre-trained model was retrained on our own dataset for 100 epochs, with a learning rate of 10^{-5}.

AiAReSeg: Our proposed AiAReSeg framework offers training in an end-to-end manner. However, in order to accelerate the process of training, model weights prior to the final segmentation head were initialized with weights from an AiATrack model, pretraiend with 300 epochs on LaSOT at a learning rate of 10^{-5}.

UNet: A standard UNet from Ronneberger et al.'s work [18] was implemented with MONAI [37] and trained for 100 epochs with a learning rate of 10^{-3}.

Clustering Based evaluation: We also designed a partially unsupervised workflow to extract the catheter given a valid aorta segmentation. In this case, an US image was first filtered with the proposed aorta segmentation mask from AiAReSeg, extracting the aorta and its embedded catheter. This patch was then thresholded at a 70% intensity level, before a K-means clustering was performed (K=2). The final cluster selection was done based on the magnitude of the variance of each cluster ($VAR_{rms} = \sqrt{(var_x)^2 + (var_y)^2}$), where the cluster with the smallest variance was selected.

TABLE III

EVALUATION ON SYNTHETIC DATA: SEGMENTATION MODELS

	Aorta		Catheter		FPS
Model Name	**DSC**	**MAE**	**DSC**	**MAE**	
UNet	88.95	0.00258	80.06	**0.00010**	**150.1**
AiAReSeg	**91.92**	**0.00213**	**83.10**	0.00014	22.8
Clustering	-	-	46.17	0.24	10.2

TABLE IV

EVALUATION ON PHANTOM DATA: SEGMENTATION MODELS

	Aorta		Catheter	
Model Name	**DSC**	**MAE**	**DSC**	**MAE**
UNet	32.30	0.037	20.39	0.00068
AiAReSeg	**34.11**	**0.032**	**62.51**	**0.00018**
Clustering	-	-	25.80	0.26

V. RESULTS

Results from tracking experiments are shown in Tab. I, which presents the findings for simulations, whereas results from benchtop phantom trials are presented in Tab.II. In all cases, AiATrack demonstrated the highest level of mean AP, with a score of 94.77 for aorta and 22.86 for catheter tracking. While AiATrack surpassed all other models across all AP IOU thresholds for aorta tracking, it fell short of the DETR's AP50 of 77.10 (vs 70.99). Nevertheless, it still outperformed the DETR on average. When applied to phantom trials, the DETR and Faster-R-CNN struggled to generalise its performance to these images, with DETR yielding an especially poor mAP performance of 1.4. The same observation was made with catheter detection, where the DETR did not yield any metric for AP, while the Faster-R-CNN's performance were also poor. The AiATrack model's performance far exceeded both cases, at 45.7 and 14.3 for aorta and catheter detection respectively.

Similarly for segmentation, Tab. III presents testing of the model on simulation, and Tab. IV is for phantoms trials. We found that AiAReSeg's performance surpassed UNet in both aorta and catheter segmentation, in simulated (aorta: 91.92 vs 88.95, catheter: 83.10 vs 80.06) and phantom trials (aorta: 34.11 vs 32.30, catheter: 62.51 vs 20.39), indicating that the model was able to generalize to some degree from simulation to reality, without needing to retrain.

VI. DISCUSSIONS

From these results, we first observe that AiA-based systems yielded the highest level of performance across nearly all detection and segmentation tasks. With simulations, where the texture of generated images were similar to the training data, the detection model performed better on average and at the 50% and 75% thresholds. This indicated that within the same image domain, the model surpassed a selection of existing frameworks. This finding is within our expectations, as the model draws upon temporal information from across the sequences, effectively supplying the knowledge about how this feature is changing over time.

Furthermore, in neighbouring but different image domains (such as the phantom image domain), although the performance was severely impacted due to lack of retraining, AiATrack still surpassed its competitors, especially in the case of aorta detection, yielding an AP of 100 at 50% and 82.62 at 75% thresholds. For the more challenging catheter detection task, AiATrack's performance was still higher, in the case that the DETR and the Faster-R-CNN completely failed to generalise. These results indicated the robustness of the AiA framework in adapting to new domains. In the case that the model is provided with a small subset of images from this new domain, it is reasonable to assume that AiATrack will start training with more adapted weights (transferred from previous training examples) to this domain, and require less data to achieve similar levels of performance as Tab. I.

With segmentation, AiAReSeg used temporal features at the attention and reconstruction level as prior knowledge at different spatial scales to aid it in mask generation. As a result, the AiAReSeg architecture surpassed its UNet competitor in both aorta and catheter segmentation tasks in simulation, and in phantom trials. We recognise that due to the challenging nature of catheter semantic segmentation, where the mask label for each frame typically only consists of 20-100 pixels, the Dice metric is rather harsh in penalising the model, even where the absolute error between the model output and the ground truth is very low. Thus, when we examine the MAE metric, it was also noted that AiAReSeg was significantly better at minimising its distance with the ground truth in the phantom case (0.00018 vs 0.00068). Considering that catheter localisation in a clinical environment demands high accuracy, we believe that these results demonstrate the potential for our system to perform well when it is sufficiently retrained.

Finally, the poor performance for phantom aorta segmentation from both models was also investigated, and the main reason found was the significant difference in appearance of the tubular structure in simulation vs in phantom. While our chosen phantom mimics the mechanical properties of an aorta, and its aesthetic appearance, the acoustic behaviour of silicone is very different from reality. As a result, a phantom axial image has high intensity on the top surface of the aorta (indicating high reflectivity), while the lower surface is shadowed, creating a discontinuous tubular shape. This shape was not observed by the model during training using simulated data, hence confusing the networks.

VII. CONCLUSIONS

In this paper, we presented a solution to the data shortage problem in the field of interventional ultrasound, by presenting a bespoke data synthesis pipeline. Through experimentation, we have demonstrated that the dataset step towards bridging the gap between simulation and reality. Deep learning models trained with this dataset were able to exhibit satisfactory preliminary results on silicon phantoms without needing to retrain. These results pave the way for future works which verifies such models on real patient anatomy. We also present our innovation, the AiAReSeg architecture, which combines temporal information both when attention is applied, and during reconstruction in the 3D deconvolution layers. Injection of temporal information enhanced the model to become a competitive option for catheter segmentation tasks among its rivals.

REFERENCES

[1] JA Kaplan, JGT Augoustides, GR Manecke, T Maus, and DL Reich. Kaplans cardiac anesthesia: For cardiac and noncardiac surgery, 2017.

[2] Mohamed EMK Abdelaziz, Libaihe Tian, Mohamad Hamady, Guang-Zhong Yang, and Burak Temelkuran. X-ray to mr: the progress of flexible instruments for endovascular navigation. *Progress in Biomedical Engineering*, 3(3):032004, 2021.

[3] T Gregory Walker, Sanjeeva P Kalva, Kalpana Yeddula, Stephan Wicky, Sanjoy Kundu, Peter Drescher, B Janne d'Othee, Steven C Rose, and John F Cardella. Clinical practice guidelines for endovascular abdominal aortic aneurysm repair: written by the standards of practice committee for the society of interventional radiology and endorsed by the cardiovascular and interventional radiological society of europe and the canadian interventional radiology association. *Journal of Vascular and Interventional Radiology*, 21(11):1632–1655, 2010.

[4] Brant W Ullery, Richard L Hallett, and Dominik Fleischmann. Epidemiology and contemporary management of abdominal aortic aneurysms. *Abdominal Radiology*, 43(5):1032–1043, 2018.

[5] Thomas L Szabo. *Diagnostic ultrasound imaging: inside out*. Academic press, 2004.

[6] LK Von Segesser, B Marty, P Ruchat, M Bogen, and A Gallino. Routine use of intravascular ultrasound for endovascular aneurysm repair: angiography is not necessary. *European journal of vascular and endovascular surgery*, 23(6):537–542, 2002.

[7] Reinhard Kopp, Werner Zürn, Rolf Weidenhagen, Georgios Meimarakis, and Dirk A Clevert. First experience using intraoperative contrast-enhanced ultrasound during endovascular aneurysm repair for infrarenal aortic aneurysms. *Journal of vascular surgery*, 51(5):1103–1110, 2010.

[8] Masanori Wakabayashi, Sayaka Hanada, Hiroyuki Nakano, and Tsunemichi Wakabayashi. Ultrasound-guided endovascular treatment for vascular access malfunction: results in 4896 cases. *The journal of vascular access*, 14(3):225–230, 2013.

[9] James Michael Mangrum, James Paul Mounsey, Lai Chow Kok, John P DiMarco, and David E Haines. Intracardiac echocardiography-guided, anatomically based radiofrequency ablation of focal atrial fibrillation originating from pulmonary veins. *Journal of the American College of Cardiology*, 39(12):1964–1972, 2002.

[10] TC Hartshorne, CN McCollum, JJ Earnshaw, J Morris, and A Nasim. Ultrasound measurement of aortic diameter in a national screening programme. *European Journal of Vascular and Endovascular Surgery*, 42(2):195–199, 2011.

[11] Vivien Gibbs, David Cole, and Antonio Sassano. *Ultrasound physics and technology: how, why and when*. Elsevier Health Sciences, 2011.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[17] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[19] Yuanyuan Nie, Zhe Luo, Junfeng Cai, and Lixu Gu. A novel aortic valve segmentation from ultrasound image using continuous max-flow approach. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3311–3314. IEEE, 2013.

[20] Ruud JG Van Sloun, Regev Cohen, and Yonina C Eldar. Deep learning in ultrasound imaging. *Proceedings of the IEEE*, 108(1):11–29, 2019.

[21] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[22] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[23] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*, pages 146–164. Springer, 2022.

[24] P Rubi, E Fernandez Vera, J Larrabide, M Calvo, JP D'Amato, and I Larrabide. Comparison of real-time ultrasound simulation models using abdominal ct images. In *12th international symposium on medical information processing and analysis*, volume 10160, pages 55–63. SPIE, 2017.

[25] Yordanka Velikova, Walter Simson, Mehrdad Salehi, Mohammad Farid Azampour, Philipp Paprottka, and Nassir Navab. Cactuss: Common anatomical ct-us space for us examinations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 492–501. Springer, 2022.

[26] Mehrdad Salehi, Seyed-Ahmad Ahmadi, Raphael Prevost, Nassir Navab, and Wolfgang Wein. Patient-specific 3d ultrasound simulation based on convolutional ray-tracing and appearance optimization. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part II 18*, pages 510–518. Springer, 2015.

[27] Multi-atlas labeling beyond the cranial vault - workshop and challenge. https://www.synapse.org///!Synapse:syn3193805/wiki/89480, 2015.

[28] Jørgen Arendt Jensen. A new approach to calculating spatial impulse responses. In *1997 IEEE Ultrasonics Symposium Proceedings. An International Symposium (Cat. No. 97CH36118)*, volume 2, pages 1755–1759. IEEE, 1997.

[29] Bradley E Treeby, Elliott S Wise, Filip Kuklis, Jiri Jaros, and BT Cox. Nonlinear ultrasound simulation in an axisymmetric coordinate system using a k-space pseudospectral method. *The Journal of the Acoustical Society of America*, 148(4):2288–2300, 2020.

[30] Tudor Jianu, Baoru Huang, Mohamed E. M. K. Abdelaziz, Minh Nhat Vu, Sebastiano Fichera, Chun-Yi Lee, Pierre Berthet-Rayne, Ferdinando Rodriguez y Baena, and Anh Nguyen. Cathsim: An open-source simulator for endovascular intervention, 2023.

[31] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[32] Khalid Mammou. v-hacd: V-hacd - volume hierarchical approximate convex decomposition. `https://github.com/kmammou/v-hacd`, 2014.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[34] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin 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*, pages 5374–5383, 2019.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[36] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[37] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.