2025 IEEE World Haptics Conference (WHC)
Suwon Convention Center, Suwon, Korea
July 8 ~ 11, 2025

FS-Net: An Encoder-Decoder Architecture for Catheter Segmentation and Contact Forces Estimation in Intracardiac Catheters

Pedram Fekri[1], Mehrdad Zadeh[2], Javad Dargahi[1]

***Abstract*—Surgeons require accurate catheter visualization and force estimation during catheter-based surgeries. Segmentation is crucial for both catheter visualization and force estimation purposes. However, using separate models for segmentation and force estimation is computationally costly. Recently, sensor-free vision/deep learning-based models have shown reasonable performance in estimating the applied forces during such surgeries, aiming to reduce the risk of surgical error. These models, however, require pre-processing to cleanly segment the catheter from the background in input images, which increases computational complexity and reduces system throughput. In this work, an encoder-decoder architecture is presented to simultaneously segment the catheter and estimate the applied forces in 2D. The presented method is designed to be deployed on a monoplane fluoroscopy machine. This multi-output network takes a raw image of the catheter's deflection and outputs both the segmented shape of the catheter and the estimated forces in 2D. Similar to object detection models, the network solves a classification problem for segmentation and a regression problem for force estimation. This integrated approach provides the estimated forces and segmented catheter shape within a single end-to-end model. Validation results show that the model accurately maps raw RGB images to the 2D force space and precisely segments the catheter.**

***Index Terms*—Semantic Segmentation, Multitask segmentation, Catheter Force estimation, Catheter Segmentation**

I. Introduction

Cardiac catheterization is a Minimally Invasive Surgery (MIS) that can be employed by surgeons for diagnostic and therapeutic purposes. In this procedure, a surgeon inserts a long flexible tube called a catheter into the vascular system of a patient e.g., from groin, neck, or shoulder under X-ray fluoroscopy imaging system to acquire information about heart muscles, heart valves and blood vessels [1]. Utilizing this procedure, surgeons can investigate the possibility of cardiac diseases e.g., heart failure, valve diseases and vessel blockages. They may also employ the catheter as a treatment for eclectic cardiovascular diseases such as angioplasty, stent placement and ablation [2]–[4]. Both biplane and monoplane fluoroscopy can be used for the aforementioned procedures. Biplane fluoroscopy provides enhanced spatial information by capturing images from two different angles, making it ideal for complex interventions like neurointerventions and intricate cardiac ablations. However, monoplane fluoroscopy is more commonly used due to its cost-effectiveness and simplicity. It is generally less expensive to acquire and maintain, making it accessible for many medical facilities, especially those with budget constraints. For many standard ablation procedures, the detailed spatial information from biplane systems is not necessary. Monoplane fluoroscopy, when combined with advanced mapping technologies like 3D electroanatomical mapping, offers sufficient accuracy and effectiveness while significantly reducing radiation exposure. These advantages make monoplane fluoroscopy a practical and efficient choice for a wide range of medical interventions [5], [6]. This work specifically focuses on data generated by monoplane vision systems due to their prevalent use.

Although the interventional catheterization treatment has shown positive efficacy, this procedure has some imperfections and safety issues. The hazardous circumstances fall into two general categories: 1- intangibility, and 2- catheter localization. As for the first category, regular catheters do not provide surgeons with haptic feedback or the sense of touch when the tip of the catheter touches anatomical lumens (e.g., heart or vessel tissues) [7], [8]. Generally, solutions like skill transfer and surgical maneuver validation are employed to address these surgical safety concerns [9], [10]. However, despite these approaches, the challenges of intangibility and accurate catheter localization remain significant obstacles in ensuring the utmost safety during such procedures [11].

Furthermore, inserting and controlling the catheter may cause unexpected movement due to the high variance shape of the vascular trees and elastic deformation of both catheter and blood vessels [12]. Coming to a collision with the vessel's wall during the insertion, the catheter may puncture or scratch the tissue throughout the path which can cause fatal bleeding. It may also slice off a part of an existing calcification or clot in the vessel. The blood stream can take the clot to the brain vessel and develop blood vessel blockages. It reduces the blood flow in the brain which may lead to stroke. The above-mentioned complications demand for assisting surgeons throughout the procedure by detecting and distinguishing the catheter from other narrow organs such as blood vessels withing the X-Ray images.

A. Related Work

These two complications have been addressed in the literature from two separate perspectives.1- force sensing, 2-

[1]Pedram Fekri and Javad Dargahi are with the Mechanical, Industrial and Aerospace Engineering Department, Concordia University, Montréal, Quebec, Canada p_fekri@encs.concordia.ca, dargahi@encs.concordia.ca

[2]Mehrdad Zadeh is with the Electrical and Computer Engineering Department, Kettering University, Flint, Michigan, USA mzadeh@kettering.edu

catheter localization and visualization.

Force Estimation: the majority of proposed methods in the literature attempt to provide the surgeon with the applied force information at the tip of the catheter. In [13], a learning-based model was proposed to map the catheter's deflection features into their corresponding generated force along x and y direction. The features are attained by a separate image processing-based feature extractor. However, the features represent catheter's deformations are not robust to different image variations. As an update on this method with an aim to resolve its drawback, in [14], a Convolutional Neural Network (CNN) was proposed to directly solve the force estimation problem without deploying further feature extraction phase. In [15], [16], it strives to extract the model of deflections and their corresponding forces through the synthetic images of the catheter generated by a simulation. The aforementioned methods have been designed to operate on a monoplane fluoroscopy machine (e.g., an experimental setup that replicates monoplane fluoroscopy functionality), mapping a single image of the catheter into the force space. These models predict the force along the x and y directions because the catheter's deflection cannot be observed in the z direction through a 2D image from monoplane fluoroscopy. In [17], a novel deep learning architecture called the Y-Net was proposed to calculate the applied forces at the tip of the catheter along x, y and z directions. This end-to-end network, simultaneously receives two images of a catheter from two angles and outputs the forces in 3D. Obviously, this method is intended to be deployed on a biplane fluoroscopy system, which is not the focus of this study.

Catheter Segmentation: previously reviewed force estimators process images with cleaned shapes of the catheter, outputted by a segmentation model. In fact, the images on which they are trained are RGB images captured from an experimental setup. The segmentation method removes shadows as well as other objects, such as force sensors, from the scene. Similarly, in a real operation room, distinguishing the inserted catheter from lumens in an X-Ray not only aids surgical visibility and measurements but also prepares images for force estimators. Catheter segmentation has employed methods ranging from image processing algorithms and CNN-based to transformer-based networks [18]. For instance, studies like [13] have used thresholding to extract the catheter's distal shaft shape. Additionally, networks such as the Fully Convolutional Network, SegNet, U-Net, Hr-Net and Mask Region-based Convolutional Neural Network have been prominent in medical semantic segmentation and specifically in segmenting catheters in fluoroscopy images [19]–[28].

Multitask Force Estimation and Catheter Segmentation: As noted, the issues related to catheter force estimation and visualization have traditionally been addressed separately. Typically, a learning-based force estimator relies on segmented catheter shapes provided by image processing or deep learning methods. In surgical settings, this often necessitates using two distinct networks: one for segmenting the catheter in the images and another for generating inputs for force estimation. The lack of a unified end-to-end solution that can both segment the catheter and estimate forces simultaneously increases computational demands and may impair real-time performance due to the need to run two large networks concurrently.

Addressing the aforementioned drawback, H-Net was recently proposed as a multi-task architecture for biplane fluoroscopy systems to address the combined challenges of catheter segmentation and 3D force estimation. By processing two X-Ray images from different angles simultaneously through two parallel encoder-decoder sub-networks, H-Net enables accurate 3D segmentation and reconstruction. It integrates segmentation and force estimation in a unified network, predicting forces along the x, y, and z axes while reducing computational complexity and improving real-time performance [29].

As an extension built on H-Net, in this work, we presented a method called FS-Net (Force Segmentation Network) which simplifies the architecture for use with monoplane fluoroscopy systems, focusing on 2D force estimation and single-view catheter segmentation. FS-Net employs a streamlined design with a single encoder-decoder structure, featuring one segmentation head and one force estimation head. This adaptation eliminates the need for dual-image input and 3D reconstruction, making FS-Net computationally efficient and well-suited for standard clinical environments using monoplane setups. By targeting 2D force estimation directly and integrating segmentation, FS-Net retains the benefits of an end-to-end solution while addressing the practical constraints of monoplane systems. Consequently, this method is pioneering in its ability to simultaneously tackle two critical tasks in interventional catheterization using a single deep learning architecture, while maintaining state-of-the-art performance in both force estimation and catheter segmentation.

II. Multi-modal Deep Encoder-Decoder Network

As previously mentioned, FS-Net is designed to address two challenging and crucial problems in interventional catheterization procedures using a single architecture: 1) force estimation, and 2) catheter segmentation. Typically, a deep learning-based force estimator processes a segmented and clean shape of a catheter's distal shaft. It uses a convolutional-based feature extractor to identify features of deflections. Subsequently, these extracted features are mapped to the corresponding applied forces at the tip by solving a regression problem. Conversely, for catheter segmentation in raw images (such as X-Rays), it is common to use a semantic segmentation model. The outputs from these models can not only be fed into the aforementioned force estimator but also be used for visualization purposes, thereby enhancing surgeons' visibility during procedures.

The FS-Net is a single-input, multi-output network that takes raw images as input and outputs both the segmented plane of the catheter and the applied forces in 2D (this work [30] is an example of a multi-output network). Essentially, the FS-Net incorporates both a classification head for semantic segmentation of the catheter and a regression head for force estimation. As an encoder-decoder network, the encoder acts as a shared feature extractor for both heads. The bottleneck's embedding, derived from the encoder, is fed into both the

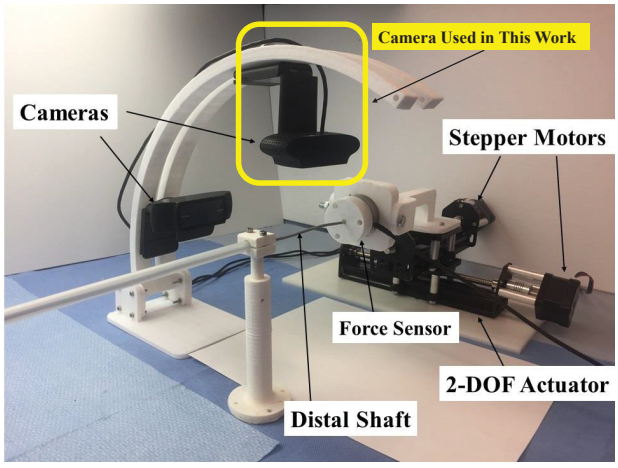

Fig. 1. This is the designed setup replicating a biplane fluoroscopy system. In this work, the images captured by the camera circled in yellow are used to train and test FS-Net.

decoder and the classification head. The decoder then upscales these features to generate the segmentation map of the catheter. Simultaneously, the force estimation head transforms the input features into the 2D force space. However, the bottleneck might contain features of objects other than the catheter's deflection, which, as previous studies suggest, can impact the accuracy of the prediction [13], [14], [17]. Similar to H-Net, the classification head receives inputs not only from the encoder but also from the decoder's feature maps. The next section provides a detailed review of the data preparation process [29].

A. Experimental Setup and Data Compilation

In our study, we utilized RGB images from an experimental setup as a feasibility study to investigate the simultaneous estimation of force and segmentation of the catheter [17]. This approach was necessitated by the impracticality of collecting real X-Ray images with corresponding force information. The use of RGB images allowed us to evaluate the potential of our proposed method under controlled conditions, providing initial insights and validating the concept before moving to more complex and clinically relevant imaging modalities. To this end, as shown in Fig. 1, the data was acquired using a setup that included two Logitech-C920 cameras, a bi-directional catheter (Boston Scientific Blazer II XP), and a force sensor (ATI Mini40). During the experiments, the catheter was pressed against the surface of the force sensor, while the cameras recorded the deflections and the sensor simultaneously measured the applied forces in the x, y and z directions. This setup aimed to mimic the real-world scenario of catheterization, in which a surgeon inserts a catheter through a patient's blood vessels under the guidance of a biplane imaging system.

Our proposed network, being a multi-output architecture, is designed to simultaneously address semantic segmentation and force estimation challenges. Consequently, each sample in the compiled dataset includes an unsegmented RGB image of the catheter's distal shaft (the last 10cm leading to the tip) from the top camera, as the input for the FS-Net (as shown in Fig. 1). The outputs or targets for each sample are the segmented images of the catheter and the measured forces along the x and y axes. In fact, the goal for the segmentation head is to remove any objects other than the catheter (e.g., shadows, 3D printed parts, and force sensors). A visualization for the revised dataset, information on the prepared data as well as the force statistics will be provided in Section III.

B. Methodology

Recent advances in AI have significantly transformed various aspects of vascular interventional surgery (VIS), including robotic instrument delivery, force perception through haptic feedback, surgical navigation with multimodal image fusion, and virtual surgical systems. These developments leverage deep learning for tasks such as optimizing catheter manipulation, enhancing tactile sensing, and enabling real-time surgical guidance through multimodal imaging analysis [31]. In line with this broader progress, recent deep learning models have significantly transformed learning-based force estimation solutions, aiming to advance the development of sensor-free catheters [13], [14], [17]. Unlike model-based approaches [32]–[36], these methods model the applied force at the catheter's tip w.r.t the shape of the catheter's deflections depicted in images. Consequently, a dataset obtained from realistic or simulated experiments is necessary to train these models. These models predominantly utilize a CNN architecture to translate deflection images into feature space. Once these embeddings are obtained, a regressor is then able to predict the forces in the output. In order to diminish the impact of irrelevant objects (e.g., noises or anatomical lumens) on the precision of the model, the input images to the aforementioned models need to be the segmented catheters. Semantic segmentation is a common solution through which the shape of the catheter is distinguished from the background [19], [20], [37]–[39]. The output of such models can be used both visualization purposes and as the input to learning-based force estimator. As a matter of fact, there are two separate networks that solve force estimation together.

The FS-Net (an extension of H-Net [29]) is a CNN-based architecture that solves the problems of force estimation and catheter shape segmentation using a single network. Fig 2 demonstrates the FS-Net architecture. Inspired by the H-Net, this network has an encoder-decoder architecture with a single input and two outputs designed for a monoplane vision system [29]. It has a classification and a regression head that segments the catheter and estimates the force based on the catheter's deflections respectively. The encoder is an image feature embedding extractor which is shared between both the classification head and the decoder. As shown in Fig 2, a raw image of size $I \in R^{(h \times w \times c)}$ is passed to the encoder as an input at a time. The encoder down-scales the input through 4 blocks (b). Each block contains two successive 2D convolution layers with the following equation:

$$A_n^l = Conv2D(f_n^l, I^{l-1}) = f_n^l * I^{l-1} = \\ (\sum_i \sum_j f_n^l[i,j] \times I^{l-1}[h-i, w-j]) + b_n^l \quad (1)$$

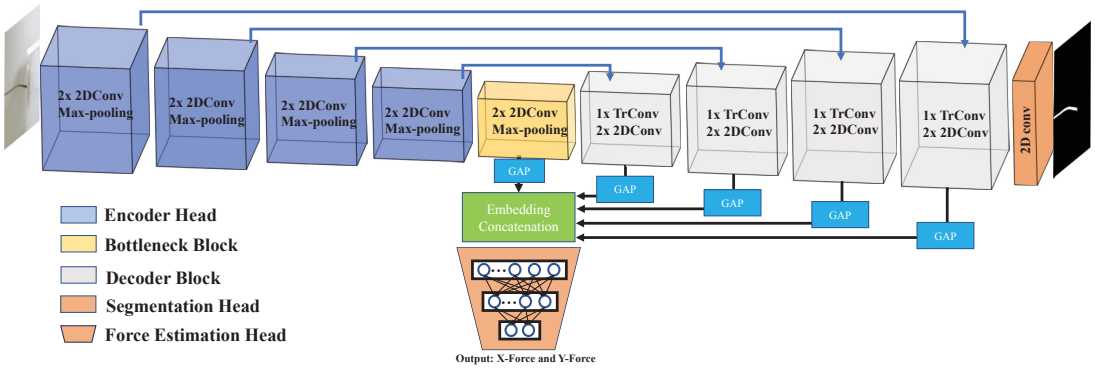

Fig. 2. The diagram illustrates the FS-Net's architecture. The network contains an encoder and a decoder in which the encoder is shared between the segmentation and force estimation head.

where A is the feature map generated by convolution ("$*$") between filter f and input I. Each convolution layer l has n number of filters $f_n^l \in R^{i,j}$ with the size of $i \times j$ as well as bias $b_n^l \in R^n$ added to the filters. Each convolution layer l outputs feature maps $A_n \in R^{h,w}$ of size $h \times w$ that goes into a ReLU activation function as below:

$$\hat{A}_n^l = ReLU(A_n^l) = max(0, A_n^l[h, w]) \tag{2}$$

However, the output $\hat{A}$ for all convolution layers in each block is still in the size of the inputs as the filter f convolves with stride $s = 1$. Furthermore, a copy of feature maps outputted by the second convolution layer of a block ($\hat{A}_{copy}^b \in R^{h_b \times w_b}$) is preserved for reinforcing the corresponding decoder layer's input. This process will be discussed in more detail in the decoder part. The final component of each block is a max-pooling layer with stride $s = 2$ that comes after the activation function of the second convolution layer in order to reduce the features' dimension and pass them to the next block as the input.

The output of the fourth block is fed to the bottleneck of the model in which it applies two 2D convolutions with ReLU activation functions without changing the size of the input. In fact, the bottleneck generates inputs for both the regression head (e.g., force estimation head) and the model's decoder. A copy of the the bottleneck feature maps go into the first layer of the decoder while another copy ($\hat{A}_n^{btn}$) is directly fed to the regression head. Considering that the regression head starts with a dense layer, it is requisite to feed it with a 1D feature vector. To this end, a global average pooling is applied to every feature map spit out by the bottleneck so as to generate a part of a 1D feature V named vector $V_{btn} \in R^{n \times 1}$ as follows:

$$v_{btn}[n] = \frac{1}{h \times w} \sum_h \sum_w \hat{A}_n^{btn}[h, w] \tag{3}$$

The equation above turns each feature map ($h \times w$) into a single scalar so that n feature maps constitute a vector of size n (V_{btn}) as a part of a input vector V to the regression head. The other part of V will be completed by combining the features provided by the decoder. This procedure will be explained in the decoder section as well.

As previously discussed, the encoder is fed by the raw and noisy images of the catheter's deflections, and it strives to extract high-level, rich features throughout the depth of the network and pass the maps to the bottleneck. At this point, the bottleneck generates a part of the regression head V_{btn} as well as the input for the decoder. The goal for the decoder is to up-scale the feature maps received from bottleneck with the aim to reconstruct an output matrix in the shape of the input image to the encoder. Having this objective in mind, $\hat{A}$ is the input to the decoder which is constituted by 4 deconvolution blocks [40]. Each block has a corresponding block in the encoder where the output shape of the encoder block equals the input shape of the corresponding decoder block. Starting from the bottleneck, the first layer of the first block is a deconvolution (convolution transpose) layer that up-scales the input to the shape of the corresponding encoder block input [40]. To be more precise, in a convolution operation, an output matrix element (or pixel) is calculated by convolving the filter with a region of the input. Conversely, in a convolution transpose operation, this process is reversed.

As explained earlier, the output of the Conv2D transpose of each decoder's block of the FS-Net is represented by $U^b \in R^{h_b \times w_b}$ in the same size of the corresponding block in the encoder before applying the max-pooling layer $\hat{A}_{copy}^b$. The input of the convolutional layer subsequent to the convolution transpose layer is calculated as follows:

$$\hat{U}^b = \hat{A}_{copy}^b \oplus U^b \tag{4}$$

where $\oplus$ denote the concatenation of A^b and U^b along the channels. $\hat{U}^b$ is an input of a 2D convolution layers with ReLu activation followed by an analogues layer. Moreover, a copy of the second convolution layer's output (the last layer of a decoder block) goes into a global average pooling layer (3) in order to convert feature maps to an embedding of size n in which each component represents a map (v_{dec}^b). In this case, each block b has two outputs: the first one ($\hat{U}^b$) is the input to the following decoder block $b + 1$ and the second one is an embedding vector v_{dec}^b that goes into the regression head. The last decoder block is deemed as the classification head. It produces n-channel feature maps in the size of the input image to the encoder. Subsequently, a 1×1 convolution layer converts the n channels to a single channel as follows:

$$\sigma(out_{cls}) = \frac{1}{1 + e^{-out_{cls}}} \tag{5}$$

Where $out_{cls} \in R^{h \times w \times 1}$ is the output of a 1×1 convolution layer and a Sigmoid activation function $\sigma(out_{cls})$ squashes the value of each component in the output matrix (e.g., each pixel) between 0 and 1. In other words, the head solves the problem as a binary classification that distinguishes between a catheter's body and other regions. The classification head optimizes the following binary cross entropy loss function:

$$L_{cls}(out_{cls}, t_{cls}) = \\ -\frac{1}{N}\sum_{i}^{N} t^i_{cls} log(\sigma(out_{cls})^i) + (1 - t^i_{cls}) log(1 - \sigma(out^i_{cls})) \tag{6}$$

Regression on the other hand, is fed by 2 component of the FSNet: the bottleneck (V_{btn}) and the decoder (v^b_{dec}). In accordance with the explanation provided, each v^b_{dec} is a vector obtained by a Global Average Pooling. The input of the regression head is calculated as follows:

$$\vec{V}_{reg} = \vec{V}_{btn} \oplus [\|^b_1 \vec{v}^b_{dec}] \tag{7}$$

where $\|^b_1$ denotes concatenating the feature embedding of the decoder's blocks. It results in a larger vectors that is constituted by all 4 embeddings of the decoder. Additionally, $\vec{V}_{btn}$ is concatenated with the aforementioned vector which leads to a single vector $\vec{V}_{reg}$ as the input to the regression head with the aim to estimate the contact forces. The regression head is composed of three successive dense layer with a relu activation function (except for the output head which has a linear activation) as below:

$$\hat{A}^l = ReLu(W^l x^{l-1} + b^l) \tag{8}$$

in the first dense layer, weights matrix $W^l \in R^{m \times n+m}$ and biases vector $b^l \in R^{m \times 1}$ where m is the number of units, n is the length of V_{btn} and m is the size of V_{dec} that make V_{reg} together which is a vector of size $n + m$. It is worth mentioning that for the first dense layer $x^{l-1} = V_{reg}$.

Since the output of the regression head is a vector of size 2 indicating the predicted force along x and y direction, the last dense layer e.g., the output layer encompasses 2 units so as to map the input of the regression head to the 2D force space. To this end, the head employs a Mean Squared Error (MSE) loss function to minimize the error of the regression output as follows:

$$L_{reg}(\hat{A}^l, t_{reg}) = \frac{\sum_{k=1}^{d} (\hat{A}^l_k - t^k_{reg})^2}{d} \tag{9}$$

in the loss function above $\hat{A}^l$ is the predicted forces vector along x and y from the output layer ($d = 2$) while t is the actual force vector. As discussed, the FS-Net has two separate heads with their own loss functions. In other words, the both heads contributes to the parameters update process in an end-to-end manner. The rate of contributions for both losses can be regulated by their corresponding weight.

$$L_{total} = \beta_1 L_{cls} + \beta_2 L_{reg} \tag{10}$$

However, in the FS-Net the weights for the classification head (segmentation) and the regression head (force estimation) are set equally so that $\beta_1 = \beta_2 = 0.5$. The loss weights were determined through empirical trial and error during preliminary experiments. We observed that using equal weights for the classification and regression losses led to more stable training and better convergence. Finally, the Root Mean Squared Propagation (RMSprop) optimization algorithm is considered to solve the problem and optimize L_{total} [41].

TABLE I
THE STATISTICS OF RECORDED FORCES IN THE COMPILED DATASETS.

Force	Samples	Mean	std	Min	Max
x	19500	0.160038	0.049717	-0.02143	0.26340
y	19500	-0.243149	0.075354	-0.33935	0.00006

III. EVALUATION AND DISCUSSION

This section focuses on detailing the data preparation, model configuration and performance evaluation for FS-Net. Section A will thoroughly review the process of datasets preparation for training, testing, and validating. As FS-Net is a multi-modal network producing outputs in two distinct formats, this subsection will address the procedures for preparing both semantic segmentation and force data for training and testing the model. In subsection A, the configuration of FS-Net will also be described for both the training and inference phases. Section B will evaluate the performance of FS-Net from two different perspectives: catheter segmentation and force estimation tasks. It will compare the force estimation component with four state-of-the-art methods in the field, while benchmarking the segmentation component against three commonly used semantic segmentation architectures in the medical imaging domain.

A. Dataset Preparation and Model Configuration

As previously highlighted in Section II-A, in contrast to the Y-Net and H-Net, this work inputs a single RGB image at a time into the FS-Net [17], [29]. Consequently, the images compiled, which depict the deflection of the catheter obtained from the experimental setup, were left unaltered. The annotations for each image were derived from the output of thresholding algorithms employed in the Y-Net [17]. In fact, the other methods available in the benchmark of this study (e.g., ResNet, ANN-based, and SVR-based) all use segmented images, whereas FS-Net utilizes these images as annotations for the segmentation head. Therefore, the dataset comprises 19,500 samples, each including an RGB image ($224 \times 224 \times 3$), an annotation ($224 \times 224 \times 1$), and the corresponding force vector of size 2 in the x and y directions.

Table I reports the statistics of forces along x and y. In the annotation planes, the catheter's region is marked as 1, while the rest, representing the background, are marked as 0. To ensure impartiality, the training, test, and validation sets in this study are identical to those used in the Y-Net. The dataset, consisting of 19,500 samples, was reshuffled and divided as follows: the training set contains 80% of the data, equivalent

to 15,600 samples, while the remaining 20% was equally split between the test and validation sets. Furthermore, the performance of FS-Net was evaluated using the test proposed in [17] in order to assess the model's generalization and robustness when exposed to camera displacement. To simulate potential displacements of the vision system, both rotation (randomly sampled between $\pm 20^{\circ}$) and scaling (randomly sampled between $\pm 20\%$) augmentations were applied to all samples during both training and testing phases (depending on the configuration, as explained in the following section). Each image was augmented with one instance of random rotation and one of random scaling, ensuring consistent variability across the dataset. This strategy was designed to enhance the model's robustness to geometric transformations commonly encountered in clinical scenarios.

FS-Net processes a single RGB image through its encoder, featuring four blocks each with two successive convolution layers using 32 filters of size 3×3 and a stride of 1 followed by a ReLU activation. Feature maps are downscaled by a max-pooling layer with (2×2) and stride 2 down-scales, while the bottleneck's stacked convolution layers increase the feature map count to 128 with the same kernel size and stride. The decoder is a mirror image of the encoder with four blocks. Each block consists of a 2D convolution transpose layer with 32 kernels (3×3) and a stride of 2 followed by two convolution layers encompassing 32 kernels of size 3×3 and a stride of 1. The activation function used in these layers is a ReLu. The final block, the classification head, employs a single 1×1 filter and a sigmoid activation to construct the segmentation plane. Each decoder block produces a 32-length embedding, which, when combined with $\vec{V}btn$, forms a 256-size vector ($\vec{V}reg$) for the regression head. This head consists of two dense layers with 64 and 32 units, respectively, leading to outputs passing through a ReLu activation and being mapped into a 2D force space by a two-unit output layer with linear activation. The following bullet points explain the process in more detail:

- Decoder Block 1: $(\mathrm{H} \times \mathrm{W} \times 32) \rightarrow$ GAP $\rightarrow$ 32-dimensional vector
- Decoder Block 2: $(\mathrm{H} \times \mathrm{W} \times 32) \rightarrow$ GAP $\rightarrow$ 32-dimensional vector
- Decoder Block 3: $(\mathrm{H} \times \mathrm{W} \times 32) \rightarrow$ GAP $\rightarrow$ 32-dimensional vector
- Decoder Block 4: $(\mathrm{H} \times \mathrm{W} \times 32) \rightarrow$ GAP $\rightarrow$ 32-dimensional vector
- Bottleneck output ($\mathbf{V}_{\text{btn}}$): $(\mathrm{H} \times \mathrm{W} \times 128) \rightarrow$ GAP $\rightarrow$ 128-dimensional vector
- Concatenated vector ($\mathbf{V}_{\text{reg}}$): [32 × 4 + 128] = 256 dimensions

FS-Net was trained on a compiled set with 32 samples per batch, a learning rate of 1×10^{-4}, and over 80 epochs. Model performance was monitored with validation in each epoch to implement early stopping for overfitting prevention.

B. Results and Discussions

As highlighted earlier, receiving an image of the catheter's distal shaft, the FS-Net accomplishes 2 major tasks in a single end-to-end architecture: 1- catheter segmentation 2- force estimation. To this end, it is warranted to analyze the model's performance from the two aforementioned perspectives separately. In light of assessing the model from both force estimation and segmentation head, the trained FS-Net is fed by the unseen test set in the inference mode. The performance of the force estimation head is examined by the following metrics in Table II: Mean Absolute Errors (**MAE**), Mean Squared Errors (**MSE**), Root Mean Squared Errors (**RMSE**), R^2 and the ratio of RMSE (R) and the average of the maximum (M) forces in the available directions ($\mathbf{R}/\mathbf{M}$). Using the results obtained from the metrics mentioned above for the FS-Net, a benchmark was established to compare FS-Net's quality in predicting contact force at the tip of the catheter along the x and y directions outputted by the force estimation head.

TABLE II
THE TABLE BENCHMARKS THE PERFORMANCE OF THE FS-NET'S FORCE ESTIMATION HEAD COMPARING WITH THE LITERATURE.

Method	FD	MSE	MAE	RMSE	R^2	R/M
MLP-based [13]	2	4.49e-05	0.0040	-	0.98	-
SVR-based [13]	2	6.27e-05	0.0046	-	0.98	-
ResNet [14], [42]	2	-	-	0.025	-	0.033
FS-Net	2	2.55e-05	0.0036	0.0047	0.99	0.036
FS-Net (Aug1)	2	3.54e-05	0.0044	0.0059	0.98	0.045
FS-Net (Aug2)	2	3.80e-05	0.0046	0.006	0.98	0.046
FS-Net (Aug3)	2	3.35e-04	0.014	0.0183	0.90	0.13

To ensure fairness, three 2D learning-based force estimation methods from the literature were selected, each of which had been previously trained and evaluated on the same dataset (same experimental setup) used in this study. The first two methods listed in the table (e.g., Multi Layer Perceptron (MLP) and Support Vector Regression-based (SVR) [13]) are the results of a study by which it maps the extracted features of a catheter's distal shaft within an image to their respective force vector along x and y. The extracted feature are attained by an image processing-based technique. Comparing these models with the FS-Net without augmentation (fifth method in the table), MSE and MAE exhibit markedly lower errors for the FS-Net.

In contrast to the first and second methods in the benchmark, the ResNet approach used by Fekri *et al.* [14] does not necessitate a separate feature extraction phase, owing to the inherent characteristics of CNN-based networks. However, the results reported for both RMSE and R/M indicate a significantly higher estimation error for the ResNet-based method. Similar to FS-Net, the predicted Force Dimension (FD) for the aforementioned methods equals two, as they are designed to be fed by a single image of the catheter from a monoplane vision system. Not only does ResNet, but also the other two methods reviewed in the literature, require a segmentation method to supply a clear shape of the catheter as input. In contrast, FS-Net can accurately estimate the force by

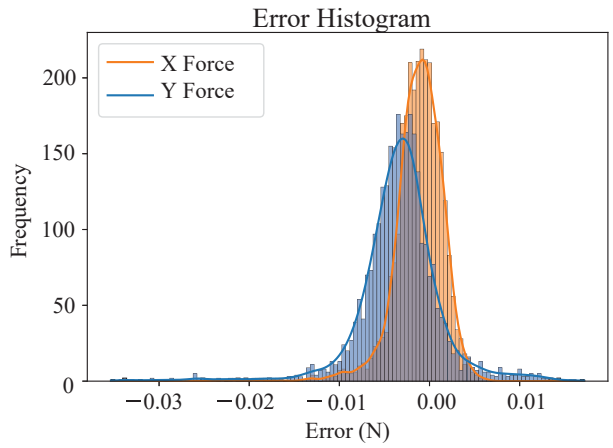

Fig. 3. This histogram illustrates the distribution of errors in the 2D force estimations made by FS-Net on the test set

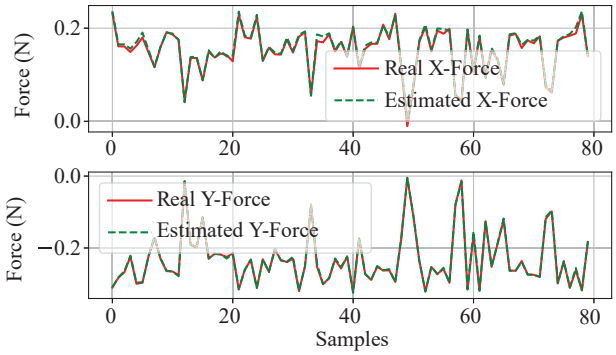

Fig. 4. The diagram demonstrates the predicted forces in x and y, given 80 RGB images from the test set.

processing the real (RGB) non-segmented image of a catheter. This capability reduces the computational complexity of the model by eliminating a separate segmentation phase, while the model still surpasses other state-of-the-art learning-based force estimation methods. Fig. 3 depicts the distribution of the error obtained from the deviation between the FS-Net predictions and the target forces along x and y. As suggested in [17], given that the desired mean value of the error distribution is above zero, a mechanism can be implemented to evaluate the model's error at various camera positions relative to the catheter. This aims to identify the optimal position that yields the minimum prediction error. Furthermore, this issue can also be addressed by compiling data from a camera rotating around the catheter. Fig. 4 contains two subplots which demonstrates the FS-Net's predictions on 80 samples in the test set along with their corresponding targets for the force along the two directions.

The last three rows of Table II examine the impact of augmentation on the throughput of FS-Net. FS-Net (Aug1) represents a configuration in which the model was trained on an augmented training set and also validated on an augmented test set. This configuration mimics a real situation in the operating room, featuring an uncontrolled workspace and turbulence in the vision system in both training phase and inference engine. The synthetic movement and vibration in the setup has slightly increased the error of prediction in comparison with the regular FS-Net. Aug-2 denotes the FS-Net variant that was fed an augmented training set during training but was evaluated on a non-augmented test set. Compared to the previous test, this represents an approximate 4.5% increase in the MAE. However, the model's generalization further decreased in the final configuration, Aug-3, where it showcases the results of FS-Net trained on a standard training set and tested on an augmented test set. The results reveal the capability of FS-Net in handling turbulence in the system when trained on the augmented training set.

On the other hand, the FS-Net's segmentation head's performance was evaluated using accuracy and the mean Intersection over Union (**mIoU**) metric. This was compared with three commonly used semantic segmentation architectures in medical applications: FCN, U-Net, and Hr-Net [19], [37], [38], [43]–[45]. The first two methods, namely FCN and U-Net, were re-implemented from scratch. In the FCN implementation, a VGG-16 backbone was integrated with an FCN-8s segmentation head. For both networks, the final layer was designed identically to the FS-Net segmentation head, focusing on a binary classification problem. These networks were trained on the compiled training set using a batch size of 32 and a learning rate of 1×10^{-8} across 150 epochs.

However, for the benchmark, an existing implementation of Hr-Net, already trained on the Cityscapes dataset, was utilized [46], [47] while FS-Net was trained from scratch on our dataset without significant prior knowledge. This pre-trained model was fine-tuned on our training set. Table III presents the benchmark results on our test sets. As can be seen, all models have shown reasonable performance, indicating that the data is not too complex for the models. However, the mIoU reflects that FS-Net has outperformed FCN and has a performance roughly similar to U-Net. Hr-Net [38] has demonstrated superior mIoU compared to FS-Net which could be attributed to the fact that Hr-Net was already trained on a large dataset before being fine-tuned on our dataset.

The second column of the table contrasts the models based on the number of their trainable parameters, expressed in millions. According to the results, FS-Net features a lightweight architecture with significantly fewer parameters. This aspect is crucial when deploying the model on edge devices with limited resources. To evaluate the efficiency of our model for real-time applications, we measured its inference latency and throughput on GPU platform using an input size of 224×224×3. The evaluation was performed over 100 iterations following 10 warm-up runs to ensure stable performance. On an NVIDIA GeForce RTX 2080 GPU, the model achieved an average inference latency of 14.95 ms, corresponding to a throughput of 66.87 frames per second (FPS). The model's computational complexity was estimated at approximately 5.71 GFLOPS, demonstrating a favorable trade-off between accuracy and efficiency for deployment across diverse hardware settings, including both high-performance and resource-limited environments. It is also worth mentioning that none

TABLE III
THE TABLE REPORTS THE PERFORMANCE OF THE FS-NET SEGMENTATION HEAD COMPARED WITH THE LITERATURE.

Models	FD	params (m)	Acc	mIoU
FCN [19]	0	134	99.2	94.0
U-Net [37]	0	34	99.8	95.7
Hr-Net [38]	0	9.636	98.8	96.1
FS-Net	2	0.436	99.8	95.5
FS-Net - Aug1	2	0.436	99.8	95.7
FS-Net - Aug2	2	0.436	99.7	94.5
FS-Net - Aug3	2	0.436	99.5	89.0

of the methods presented in Table III (except for FS-Net) have a multitask architecture capable of estimating force as a separate task in addition to segmentation (FD = 0). Fig. 5 displays the segmentation results of all networks evaluated in the benchmark across five input images. The first two rows present the original inputs and their corresponding annotations. As illustrated, the discrepancy in mIoU between FS-Net and Hr-Net is not visually significant.

Similar to Table II, the last three rows of Table III investigate the impact of augmentation on the segmentation results for FS-Net. FS-Net - Aug1 demonstrates improved performance and generalization compared to FS-Net with no augmentation. Additionally, the performance of FS-Net trained on augmented data and tested on non-augmented data (Aug-2) is remarkable. However, FS-Net - Aug 3 reaches the same conclusion as the force estimation benchmark: the model trained on regular data shows lower generalization capability when exposed to disturbances such as setup vibration and camera turbulence. Fig. 6 has plotted the loss and accuracy trend of FS-Net across the training in 80 epochs. The regression loss represents the prediction precision on both training and validation set while the accuracy of the segmentation head (cls) has shown in a separate axis. The diagram has exhibited an smooth training process with no sign of over-fitting as the evaluation metrics and losses for both training and validation set caught up with no fluctuation and deviation. Although the classification head has reached to a convergence point quickly, due to negligible scale of the forces, the regression head required more time to converge properly.

It is worth mentioning that the implementation of FS-Net, FCN and U-Net as well as training and all validations were conducted on an Ubuntu 20.04 machine with an NVIDIA 2080 GPU. Also, in this work, we did not apply any domain adaptation techniques during training or testing. However, domain adaptation and generalization methods can play a critical role in improving model robustness across varying clinical conditions. Integrating such approaches in future work may enhance the model's reliability and applicability in real-world medical settings [48]. Lastly, incorrect segmentation or force estimation may lead to clinical risks, such as excessive tissue

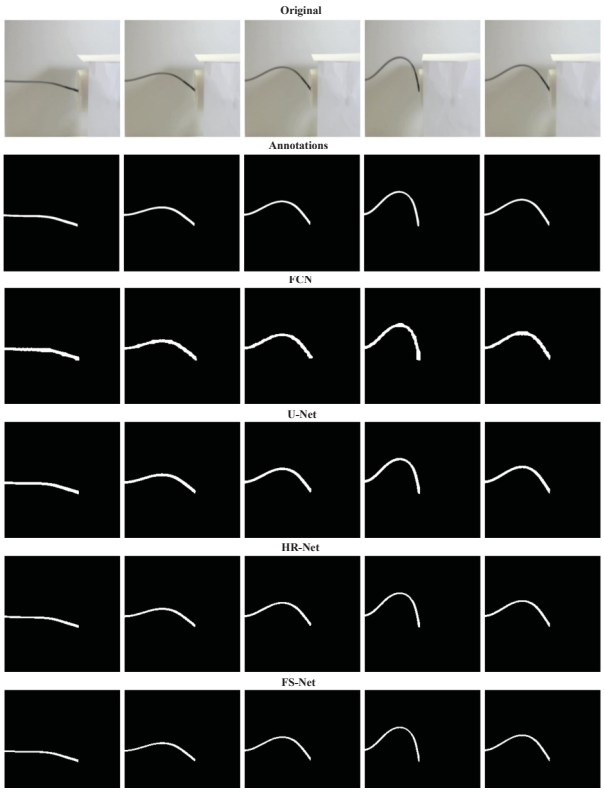

Fig. 5. The diagram demonstrates 5 samples from the test set and the prediction results of the methods in Table III.

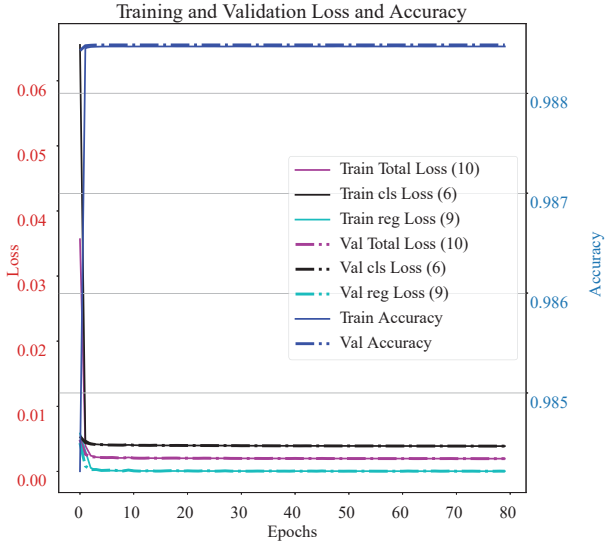

Fig. 6. The diagram plots the training and validation loss and accuracy.

contact or misinterpretation of catheter position. Although our current model does not include uncertainty estimation or failure detection, future work will incorporate techniques, safety thresholds, and failure case analysis to enhance reliability [10].

IV. CONCLUSIONS

In this work, we introduced an extension of H-Net called FS-Net, a convolutional encoder-decoder architecture with one

input and two outputs, designed to simultaneously handle catheter segmentation and force estimation at the tip. This state-of-the-art model efficiently estimates forces directly from uncleaned RGB images and segments the catheter's shape, offering a comprehensive solution for both tasks without added computational complexity. We evaluated FS-Net against three top force estimators and three leading medical image segmentation models, with FS-Net consistently showing superior accuracy in all benchmarks. The primary limitation is its inability to estimate forces along the z axis when processing dual images. Future work will focus on refining FS-Net's architecture to address this issue. Also, we plan to extend the proposed model to support multiclass segmentation by incorporating a synthetic X-ray image generator capable of adding anatomical structures to the scene. This will enable the segmentation head to not only identify the catheter but also classify additional anatomical parts, while the force estimation branch retains focused attention on the catheter to preserve estimation accuracy.

REFERENCES

[1] X. Hu, A. Chen, Y. Luo, C. Zhang, and E. Zhang, "Steerable catheters for minimally invasive surgery: a review and future directions," *Computer Assisted Surgery*, vol. 23, no. 1, pp. 21–41, 2018.

[2] Y. R. Manda and K. M. Baradhi, *Cardiac Catheterization Risks and Complications*. StatPearls Publishing, Treasure Island (FL), 2021. [Online]. Available: http://europepmc.org/books/NBK531461

[3] R. Mehta, K.-J. Lee, R. Chaturvedi, and L. Benson, "Complications of pediatric cardiac catheterization: A review in the current era," *Catheterization and Cardiovascular Interventions*, vol. 72, no. 2, pp. 278–285, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ccd.21580

[4] G. Ndrepepa and H. Estner, "Ablation of cardiac arrhythmias—energy sources and mechanisms of lesion formation," in *Catheter Ablation of Cardiac Arrhythmias*. Springer, 2006, pp. 35–53.

[5] C.-S. Hong, Z.-C. Chen, K.-T. Tang, and W.-T. Chang, "The effectiveness and safety between monoplane and biplane imaging during coronary angiographies," *Acta Cardiol Sin*, vol. 36, no. 2, pp. 105–110, Mar. 2020.

[6] D. M. Leistner, L. S. Schlender, J. Steiner, A. Erbay, J. Klotsche, P. Schauerte, A. Haghikia, U. Rauch-Kröhnert, D. Sinning, A. Lauten, H.-C. Mochmann, C. Skurk, U. Landmesser, and B. E. Stähli, "A randomised comparison of monoplane versus biplane fluoroscopy in patients undergoing percutaneous coronary intervention: the RAMBO trial," *EuroIntervention*, vol. 16, no. 8, pp. 672–679, Oct. 2020.

[7] M. Y. Lo, P. Sanders, P. Sommer, J. M. Kalman, U. R. Siddiqui, S. Sundaram, C. Piorkowski, N. Olson, S. M. Madej, and D. N. Gibson, "Safety and effectiveness of a next-generation contact force catheter: Results of the tactisense trial," *JACC: Clinical Electrophysiology*, 2021.

[8] C.-F. Chen, X.-F. Gao, M.-J. Liu, C.-L. Jin, and Y.-Z. Xu, "Safety and efficacy of the thermocool smarttouch surroundflow catheter for atrial fibrillation ablation: A meta-analysis," *Clinical cardiology*, vol. 43, no. 3, pp. 267–274, 2020.

[9] P. Fekri, J. Dargahi, and M. Zadeh, "Deep learning-based haptic guidance for surgical skills transfer," *Frontiers in Robotics and AI*, vol. 7, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.586707

[10] P. Fekri, P. Setoodeh, F. Khosravian, A. Safavi, and M. H. Zadeh, "Towards deep secure tele-surgery," in *Proceedings of the International Conference on Scientific Computing (CSC)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2018, pp. 81–86.

[11] D. C. Shah and M. Namdar, "Real-time contact force measurement: a key parameter for controlling lesion creation with radiofrequency energy," *Circulation: Arrhythmia and Electrophysiology*, vol. 8, no. 3, pp. 713–721, 2015.

[12] S. Famouri, P. Fekri, M. Roshanfar, and J. Dargahi, "Towards surgical skill modeling in cardiac ablation using deep learning," in *2023 27th International Conference on Methods and Models in Automation and Robotics (MMAR)*, 2023, pp. 216–221.

[13] H. Khodashenas, P. Fekri, M. Zadeh, and J. Dargahi, "A vision-based method for estimating contact forces in intracardiac catheters," in *IEEE 1st International Conference on Autonomous Systems, ICAS*. IEEE, 2021.

[14] P. Fekri, H. R. Nourani, M. Razban, J. Dargahi, M. Zadeh, and A. Arshi, "A deep learning force estimator system for intracardiac catheters," in *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2021, pp. 1–6.

[15] M. Roshanfar, P. Fekri, and J. Dargahi, "A deep learning model for tip force estimation on steerable catheters via learning-from-simulation," *EasyChair*, no. 10364, 6 2023.

[16] ——, "Toward autonomous cardiac catheterization through a parametric finite element simulation with experimental validation," *ICAS 2023*, p. 23, 3 2023.

[17] P. Fekri, H. Khodashenas, K. Lachapelle, R. Cecere, M. Zadeh, and J. Dargahi, "Y-net: A deep convolutional architecture for 3d estimation of contact forces in intracardiac catheters," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3592–3599, 2022.

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[21] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[22] A. Nguyen, D. Kundrat, G. Dagnino, W. Chi, M. E. M. K. Abdelaziz, Y. Guo, Y. Ma, T. M. Y. Kwok, C. Riga, and G.-Z. Yang, "End-to-end real-time catheter segmentation with optical flow-guided warping during endovascular intervention," 2020. [Online]. Available: https://arxiv.org/abs/2006.09117

[23] M. Gherardini, E. Mazomenos, A. Menciassi, and D. Stoyanov, "Catheter segmentation in x-ray fluoroscopy using synthetic data and transfer learning with light u-nets," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105420, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260719312301

[24] H. Yang, C. Shan, A. F. Kolen, and P. H. N. de With, "Weakly-supervised learning for catheter segmentation in 3d frustum ultrasound," *CoRR*, vol. abs/2010.09525, 2020. [Online]. Available: https://arxiv.org/abs/2010.09525

[25] C. Baur, S. Albarqouni, S. Demirci, N. Navab, and P. Fallavollita, "Cathnets: Detection and single-view depth prediction of catheter electrodes," in *Medical Imaging and Augmented Reality*, G. Zheng, H. Liao, P. Jannin, P. Cattin, and S.-L. Lee, Eds. Cham: Springer International Publishing, 2016, pp. 38–49.

[26] I. Ullah, P. Chikontwe, H. Choi, C. H. Yoon, and S. H. Park, "Synthesize and segment: Towards improved catheter segmentation via adversarial augmentation," *Applied Sciences*, vol. 11, no. 4, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/4/1638

[27] A. Nguyen, D. Kundrat, G. Dagnino, W. Chi, M. E. M. K. Abdelaziz, Y. Guo, Y. Ma, T. M. Y. Kwok, C. Riga, and G.-Z. Yang, "End-to-end real-time catheter segmentation with optical flow-guided warping during endovascular intervention," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9967–9973.

[28] I. Abdulhafiz and F. Janabi-Sharifi, "A hybrid approach to 3d shape estimation of catheters using ultrasound images," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 1912–1919, 2023.

[29] P. Fekri, M. Zadeh, and J. Dargahi, "H-net: A multitask architecture for simultaneous 3d force estimation and stereo semantic segmentation in intracardiac catheters," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 844–851, 2025.

[30] H. Chang, P. Xu, H. Yao, J. Li, X. Xin, and M. Guizani, "Nonprobe adaptive compensation for optical wireless communications based on orbital angular momentum," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.

[31] B. Ren, Y. Zhao, J. Zhang, H. Li, K. Li, and J. Zhang, "The critical technologies of vascular interventional robotic catheterization: A review," *IEEE Sensors Journal*, vol. 23, no. 24, pp. 30 051–30 069, 2023.

[32] M. Khoshnam, M. Azizian, and R. V. Patel, "Modeling of a steerable catheter based on beam theory," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 4681–4686.

[33] J. Till, V. Aloi, and C. Rucker, "Real-time dynamics of soft and continuum robots based on cosserat rod models," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 723–746, 2019.
[34] R. J. Webster, J. M. Romano, and N. J. Cowan, "Mechanics of precurved-tube continuum robots," *IEEE Transactions on Robotics*, vol. 25, no. 1, pp. 67–78, 2008.
[35] M. Kouh Soltani, S. Khanmohammadi, and F. Ghalichi, "A three-dimensional shape-based force and stiffness-sensing platform for tendon-driven catheters," *Sensors*, vol. 16, no. 7, p. 990, 2016.
[36] J. Back, L. Lindenroth, K. Rhode, and H. Liu, "Three dimensional force estimation for steerable catheters through bi-point tracking," *Sensors and Actuators A: Physical*, vol. 279, pp. 404–415, 2018.
[37] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on arXiv:1505.04597 [cs.CV]). [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a
[38] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *CoRR*, vol. abs/1908.07919, 2019. [Online]. Available: http://arxiv.org/abs/1908.07919
[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.
[40] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2528–2535.
[41] T. Tieleman, G. Hinton *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385
[43] H. Yang, C. Shan, A. F. Kolen, and P. H. N. de With, "Efficient catheter segmentation in 3d cardiac ultrasound using slice-based fcn with deep supervision and f-score loss," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 260–264.
[44] A. M. Aleong, A. Berlin, J. Borg, J. Helou, A. Beiki-Ardakani, A. Rink, S. Raman, P. Chung, and R. A. Weersink, "Rapid multi-catheter segmentation for magnetic resonance image-guided catheter-based interventions," *Medical Physics*, 2024. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.17117
[45] Y. Wang, H. K. Lam, Z.-G. Hou, R.-Q. Li, X.-L. Xie, and S.-Q. Liu, "High-resolution feature based central venous catheter tip detection network in x-ray images," *Med Image Anal*, vol. 88, p. 102876, Jun. 2023.
[46] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.
[47] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[48] M. Noori, M. Cheraghalikhani, A. Bahri, G. A. V. Hakim, D. Osowiechi, M. Yazdanpanah, I. B. Ayed, and C. Desrosiers, "Fds: Feedback-guided domain synthesis with multi-source conditional diffusion models for domain generalization," 2024. [Online]. Available: https://arxiv.org/abs/2407.03588