

Multimodal Learning-Based Frameworks for Sensor-Free Force-Aware Perception and Autonomous Navigation in Cardiac Catheterization

Pedram Fekri

A Thesis

in

The Department

of

Mechanical, Industrial, and Aerospace Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Mechanical Engineering) at

Concordia University

Montréal, Québec, Canada

January 2026

© Pedram Fekri, 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Pedram Fekri**

Entitled: **Multimodal Learning-Based Frameworks for Sensor-Free Force-Aware
Perception and Autonomous Navigation in Cardiac Catheterization**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Mechanical Engineering)

complies with the regulations of this University and meets the accepted standards with respect to
originality and quality. Signed by the Final Examining Committee:

Dr. Ferhat Khendek Chair

Dr. Mahdi Tavakoli External Examiner

Dr. Muthukumaran Packirisamy Examiner

Dr. Arash Mohammadi Examiner

Dr. Mojtaba Kheiri Examiner

Dr. Javad Dargahi Supervisor

Dr. Mehrdad Zadeh Co-supervisor

Approved by

Muthukumaran Packirisamy, Chair
Department of Mechanical, Industrial, and Aerospace Engineering

March 2026

Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Multimodal Learning-Based Frameworks for Sensor-Free Force-Aware Perception and Autonomous Navigation in Cardiac Catheterization

Pedram Fekri, Ph.D.

Concordia University, 2026

Cardiac catheterization is a procedure in minimally invasive cardiovascular interventions, where safe and effective performance relies on accurate force awareness, catheter visualization, and navigation within complex vascular geometries. This doctoral research investigates learning-based, sensor-free frameworks using standard clinical imaging modalities, with the objective of advancing toward autonomous catheter-based systems. The work is conducted as a multidisciplinary study at the intersection of artificial intelligence, deep learning-based computer vision, medical robotics, mechatronics, and autonomous systems, based on the premise that catheter deflections observed from multiple imaging viewpoints encode sufficient information to infer three-dimensional contact forces and support goal-directed navigation. A lightweight fusion-based convolutional neural network is first proposed to estimate three-dimensional contact forces directly from stereo catheter images without reliance on physical sensors or explicit mechanical modeling. This is extended to a multitask encoder-decoder architecture that simultaneously performs catheter segmentation and force estimation from biplane fluoroscopic images within a single end-to-end framework, supported by a synthetic X-ray image generator designed to resemble clinical fluoroscopy. The framework is further extended through a Vision Transformer-based architecture with cross-attention for joint stereo segmentation and force estimation. Finally, autonomous catheter navigation is explored through a goal-conditioned, multimodal vision-to-action model trained using imitation learning for perception-driven catheter steering in a physical robotic setup. Overall, this dissertation demonstrates that force estimation, catheter perception, and navigation can be learned within sensor-free frameworks, supporting future autonomous robotic catheterization systems.

Acknowledgments

I am deeply grateful for the opportunity to pursue my doctoral studies at Concordia University, an experience that I owe to Montréal and to Canada, a country that has shaped both my academic and personal life over the past years. Conducting this research within such a diverse, supportive, and intellectually stimulating environment has been a privilege for which I am sincerely thankful.

Above all, I wish to express my deepest gratitude to my wife, Nika Abedi. This journey would not have been possible without her unwavering support, sacrifices, and boundless patience. Throughout the long and demanding years of this PhD, she carried far more than her share of life's responsibilities, often placing her own needs aside to support mine. Her dedication, compassion, strength, and belief in me sustained me through moments of exhaustion, uncertainty, and challenge. This thesis is as much a reflection of her commitment and love as it is of my own work, and I share this achievement with her wholeheartedly.

I dedicate this thesis to the memory of my beloved late mother, Noushin Hashemi, whom I lost only three months after beginning my PhD journey. She devoted her entire life to me as her only child, offering unconditional love, encouragement, and sacrifice until her final days. It remains my greatest unfulfilled wish that she could be here to witness this moment. Although she is no longer physically present, her spirit, guidance, and devotion have accompanied me throughout this journey. This work stands as a tribute to her life and her endless love.

I would like to express my sincere appreciation to my supervisor, Dr. Javad Dargahi, for his patience, guidance, and continuous support throughout my doctoral studies. His deep expertise, high standards, and commitment to excellence played a fundamental role in shaping my research direction and professional growth. I am grateful for the trust he placed in me and for the intellectual

freedom he provided, which allowed this work to mature and evolve.

I am also very thankful to my co-supervisor, Dr. Mehrdad Zadeh, for his invaluable support, particularly during the early stages of my PhD. His guidance, encouragement, and willingness to help in all aspects of my academic journey were instrumental in establishing a strong foundation for my research. I would like to thank the researchers at ReachLab, Department of Computer Science, Kettering University, as well as collaborators at Navistar in Michigan, USA, for providing me with the opportunity to work on perception systems for self-driving vehicles. This experience significantly broadened my research perspective and technical expertise.

I extend my sincere appreciation to the members of my PhD examination committee, Dr. Muthukumar Packirisamy, Dr. Arash Mohammadi, Dr. Mojtaba Kheiri, and Dr. Mehdi Tavakoli, for generously dedicating their time, insight, and expertise to reviewing my work and providing valuable feedback.

I am grateful to the Department of Mechanical Engineering at Concordia University, and in particular to Dr. Brandon Gordon, for his trust and for offering me the valuable opportunity to serve as a part-time faculty member. Teaching under his guidance was a rewarding experience that greatly enriched my academic development.

I would also like to acknowledge the generous financial support I received through the Fondation Arbour PhD Scholarship, awarded to me on two occasions. I am especially grateful to Dr. Marine Hadengue for her support and encouragement, and to Ms. Diane de Champlain for her exceptional assistance and kindness throughout this process.

I am thankful for the opportunity to expand my research through industrial collaborations. I sincerely appreciate PlumFind and my colleagues there, particularly Dr. Oz Gurtuna, for their support and collaboration. I also extend my gratitude to Dr. Fiona Zhao and Dr. Mutahar Safdar from McGill University for their role in establishing a meaningful and productive collaboration.

I am grateful for my time at Zebra Technologies and the Zebra Technologies Research Chair, where I worked as an AI researcher alongside outstanding scientists and engineers whose insights greatly influenced my research. I would like to particularly thank Mr. Arnaud Lina, Director of Engineering, Dr. Gregory Strubel, my research colleagues at LIVIA, École de technologie supérieure (ÉTS), and Dr. Christian Desrosiers for their support and collaboration.

I would also like to thank Dr. Majid Roshanfar, Dr. Thomas Looi, and Dr. Dale Podolsky from The Wilfred and Joyce Posluns Centre for Image Guided Innovation and Therapeutic Intervention (PCIGITI) at the Hospital for Sick Children (SickKids), University of Toronto, for their efforts in establishing and supporting a collaborative research relationship.

Finally, I acknowledge the continued support of Concordia University, the School of Graduate Studies, and the Gina Cody School of Engineering. I am thankful to my colleagues at the Surgical Robotics Lab (SRL) for their camaraderie and shared experiences, especially Mr. Farzad Famouri. I also extend my sincere appreciation to the staff of the Engineering Department, particularly Ms. Leslie Hossein, for her invaluable administrative support and kindness throughout my doctoral studies.

Contents

List of Figures	xi
List of Tables	xvi
1 Introduction	1
1.1 Background	1
1.1.1 Minimally invasive surgery (MIS)	1
1.1.2 MIS applications	2
1.2 Interventional cardiac catheterization	3
1.2.1 Manual Catheterization	4
1.2.2 Robotic-assisted catheterization	5
1.3 Catheterization Bottlenecks	7
1.3.1 Lack of Haptic Feedback and Force Control	7
1.3.2 Mechanical Navigation, Vascular Tortuosity and Vessel Injury	8
1.3.3 Imaging-Related Challenges and Radiation Exposure	8
1.3.4 Procedural Time and Operator Ergonomics	9
1.4 State-of-the-arts: Force Sensing, visual perception and navigation	9
1.4.1 Force Sensing	10
1.4.2 Autonomous Navigation	13
1.5 Research Motivations, Objectives and Scopes	18
1.5.1 Force Sensing Limitations	18
1.5.2 Catheter Navigation Limitations	19

1.5.3	Research Objectives	21
1.5.4	Research Scopes	21
1.6	Thesis Contributions	23
1.7	Publications	24
1.8	Thesis Layout	27
2	Y-Net: A CNN for 3D Estimation of Contact Forces in Intracardiac Catheters	29
2.1	Introduction	29
2.1.1	Related Work	30
2.1.2	Contributions	32
2.2	Deep Convolutional 3D Force estimator	33
2.2.1	Experimental Setup	34
2.2.2	Dataset Compilation	37
2.2.3	Methodology	37
2.3	Evaluation and Discussion	41
2.3.1	Dataset Preparation and Model Configuration	41
2.3.2	Results and Discussions	44
2.4	Conclusion	48
3	H-Net: A Multitask Architecture for Simultaneous 3D Force Estimation and Stereo Semantic Segmentation in Intracardiac Catheters	51
3.1	Introduction	51
3.1.1	Related Work	53
3.1.2	Contributions	55
3.2	Multi-Task Encoder-Decoder CNN	56
3.2.1	Synthetic Data Generation and Data Preparation	57
3.2.2	Methodology	59
3.3	Evaluation and Discussion	64
3.3.1	Dataset Preparation and Model Configuration	64
3.3.2	Results and Discussions	67

3.4	Conclusion	73
4	TransForSeg: A Multitask Stereo ViT for Joint Stereo Segmentation and 3D Force Estimation in Catheterization	75
4.1	Introduction	75
4.1.1	Related Work	76
4.1.2	Contributions	78
4.2	Multi-Input Multi-Output Encoder–Decoder Transformer	80
4.2.1	Dataset	80
4.2.2	Methodology	81
4.3	Results and Discussions	86
4.3.1	Implementation Details	86
4.3.2	Quantitative Results	87
4.3.3	Ablation Study	90
4.4	Conclusion	97
5	DINO-CVA: A Multimodal Goal-Conditioned Vision-to-Action Model Toward Autonomous Catheter Navigation	99
5.1	Introduction	99
5.1.1	Related Work	102
5.1.2	Contributions	104
5.2	Experimental Setup and Data Compilation	105
5.2.1	Robotic Platform for Catheter Navigation	106
5.2.2	Robot Interface Software	107
5.3	Dino-CVA: A Vision-Action model for Catheter Navigation	108
5.3.1	Data Compilation	109
5.3.2	Methodology	109
5.4	Results and Discussions	114
5.4.1	Data Preparation and Model Configuration	115
5.4.2	Results	120

5.4.3 Ablation Study	121
5.4.4 Limitations	123
5.5 Conclusion	125
6 Conclusion and Future Work	126
6.1 Conclusion	126
6.2 Future Work	128
Bibliography	130

List of Figures

Figure 1.1	Schematic of a cardiac ablation catheter insertion path via femoral access [1].	4
Figure 1.2	CorPath GRX Neurovascular System.	6
Figure 1.3	TactiCath Contact Force Ablation Catheter.	11
Figure 1.4	THERMOCOOL [®] catheter (Biosense Webster).	12
Figure 2.1	The designed experimental setup is shown in 4 subplots. This setup is equipped with a regular catheter, two cameras, two linear actuators and a force sensor to bend the catheter's distal shaft in different shapes.	35
Figure 2.2	Joint distribution of the recorded contact force components along the x , y , and z axes. The diagonal subplots show the marginal histograms of each force component, while the off-diagonal scatter plots depict the pairwise relationships between the force components.	36
Figure 2.3	Architecture of the proposed Y-Net. The filter sizes and stride values are indicated on the corresponding layer blocks. The network consists of two parallel components that process image streams acquired from the top and side views, respectively. Each component has its own independent set of learnable parameters at every layer.	38
Figure 2.4	A sample of dataset records consisting of paired images acquired from the top and side views of an ablation catheter. The second row shows the corresponding images after the segmentation step performed using a thresholding algorithm. . . .	42
Figure 2.5	The diagram illustrates the training and validation loss curves for the different Y-Net configurations.	47

Figure 2.6	This diagram illustrates the histogram of estimation errors for the three-dimensional forces predicted by Y-Net on the test set.	48
Figure 2.7	The diagram presents the predicted contact forces along the x , y , and z directions for 80 samples from the test set. Each sample consists of a paired set of images capturing the catheter deflection from two views.	49
Figure 3.1	Overview of the synthetic X-ray image generation process using RGB images and their corresponding thresholded catheter masks.	58
Figure 3.2	Overview of the H-Net architecture. The network consists of two parallel sub-networks with shared encoder-decoder parameters, two segmentation heads, and a central force estimation head. Given two input X-ray images, H-Net simultaneously produces two catheter segmentation maps and estimates the applied forces in three dimensions.	60
Figure 3.3	The diagram demonstrates two samples from RGB, synthetic XRay-1 and XRay-2 dataset. The outputs of H-Net for each sample are shown as well. For each dataset, H-Net was already trained on its train set.	65
Figure 3.4	The diagram demonstrates the predicted forces in x , y and z , given 80 input samples of XRay-2's test set.	70
Figure 3.5	Histograms of force estimation errors generated by H-Net are shown for each dataset in subplots (a)–(c). Subplots (d)–(f) depict the corresponding training and validation loss and accuracy curves.	71

Figure 4.1	Overview of the TransForSeg architecture. The model consists of a Vision Transformer (ViT)-based encoder and decoder with shared weights. Two X-ray images are provided as input, one processed by the encoder and the other by the decoder. Each image is divided into non-overlapping patches, a [CLS] token is appended, and positional embeddings are added to form the input token sequences. Both the encoder and decoder comprise T stacked Transformer blocks, with the final block producing the corresponding output embeddings. The decoder output embeddings are fused with the encoder embeddings through an embedding fusion (cross-attention) layer. The [CLS] token from the decoder is passed to an MLP head for 3D contact force estimation, while the remaining token embeddings from both the encoder and decoder are reshaped and upsampled by a shared CNN-based decoder to generate the catheter segmentation maps.	82
Figure 4.2	Error histograms of the force estimation head for TransForSeg (denoted as w , indicating models with segmentation heads) and TransForcer (denoted as w/o , indicating models without segmentation heads) across all evaluated datasets. In each subplot, t and s denote the Tiny and Small model configurations, respectively.	92
Figure 4.3	Training and validation loss curves. Subplots (a)–(d) show the segmentation loss for the Tiny and Small configurations of TransForSeg, while subplots (e)–(h) present the force estimation loss for TransForSeg (denoted as w , models with segmentation heads) and TransForcer (denoted as w/o , models without segmentation heads).	93
Figure 4.4	Qualitative results across 3 out of 6 noise conditions for 3 input samples. The first row displays the original (clean) samples, followed by the corresponding segmentation outputs produced by the model in the second row.	96

Figure 5.1	Overall experimental setup. The system includes a transparent vessel phantom positioned beneath a top-mounted camera, a robotic platform with an actuation unit mounted on a linear rail, and a joystick-based controller for teleoperation. A buckling-prevention guide ensures stable catheter insertion. The graphical user interface enables parameter configuration and provides real-time visualization of the catheter within the phantom.	107
Figure 5.2	Close-up of the robotic platform. Three stepper motors, each driven by a microstepping driver, enable three degrees of freedom (DoFs): catheter translation, distal tip steering, and base rotation. The system is mounted on an optical breadboard, with 3D-printed fixtures securing the catheter during operation.	108
Figure 5.3	Architecture of the DINO-CVA model. The framework receives video frames of the catheterization task together with motor states (joystick inputs). The frames are encoded using a frozen DINOv2 backbone and organized into a temporal sequence. The motor states are linearly projected to match the visual embedding space. After positional embeddings are applied, cross-attention fuses the two modalities, while spatial and temporal dependencies are modeled by a causal transformer. The resulting fused representation is then conditioned on the target frame through a gated fusion network, and the action head predicts the next controller state s_{t+1} . . .	110
Figure 5.4	The figure shows nine predefined starting and ending points, indicated by red markers on the artificial vasculature. The images also illustrate representative visual inputs used to train and evaluate the DINO-CVA model. The starting location was kept identical across all experiments, while the target endpoints varied among the nine defined scenarios.	114
Figure 5.5	Violin plots illustrating the empirical distribution of motor states across both data-splitting protocols (episode-based and scenario-based) for the training, validation, and test sets. The plots highlight variability in catheter control inputs and reflect differences in control behavior across scenarios and execution episodes. . .	116

Figure 5.6 True-predicted plots and error histograms for knob, rotation, and translation control using DINO-CVA. Episode-based results are shown in (a–f): (a) knob true-predicted, (b) rotation true-predicted, (c) translation true-predicted, (d) knob error distribution, (e) rotation error distribution, and (f) translation error distribution. Scenario-based results are shown in (g–l): (g) knob true-predicted, (h) rotation true-predicted, (i) translation true-predicted, (j) knob error distribution, (k) rotation error distribution, and (l) translation error distribution. The true-predicted plots illustrate alignment between predicted and ground-truth motor states, while the histograms depict the distribution of prediction errors across tasks. 119

List of Tables

Table 2.1	The statistics of recorded forces (N) in the compiled datasets.	43
Table 2.2	The table reports the performance of the Y-Net comparing with the learning-based methods in the literature.	45
Table 3.1	The table reports the performance of the H-Net comparing with the learning-based methods in the literature.	68
Table 3.2	The table reports the performance of the H-Net’s segmentation heads comparing with the literature.	72
Table 4.1	Performance of Force Estimation Across Datasets	88
Table 4.2	Performance of Catheter Segmentation Across Datasets	89
Table 4.3	Ablation: Model with and without segmentation	91
Table 4.4	Performance of Models Across Different Noises	95
Table 5.1	Data split statistics for scenario-based and episode-based datasets.	117
Table 5.2	Performance comparison of the proposed model and LSTM baseline on episode-based and scenario-based datasets.	118
Table 5.3	Ablation study on the scenario-based dataset. The table reports the performance of the model and its variants when goal or modality components are removed.	122

Nomenclature

2D	Two-Dimensional
3D	Three-Dimensional
AF	Atrial Fibrillation
ANN	Artificial Neural Network
ASDs	Atrial Septal Defects
BC	Behavioral Cloning
BCE	Binary Cross Entropy
CNN	Convolutional Neural Network
CRT	Cardiac Resynchronization Therapy
DCNN	Deep Convolutional Neural Network
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-Network
FFN	Feed-Forward Network
GAIL	Generative Adversarial Imitation Learning
GGF	Gated Goal Fusion
IAGN	Guided Autonomous Guidewire Navigation

IL Imitation Learning

IoU Intersection over Union

IRL Inverse Reinforcement Learning

LfD Learning from Demonstration

LSTM Long Short-Term Memory

MAE Mean Absolute Error

mDice Mean Dice Coefficient

MHCA Multi-Head Cross-Attention

MHSA Multi-Head Self-Attention

MICS Minimally Invasive Cardiac Surgery

MIS Minimally Invasive Surgery

MLLMs Multimodal Large Language Models

MLP Multilayer Perceptron

MSE Mean Squared Error

NLP Natural Language Processing

PCI Percutaneous Coronary Intervention

PPO Proximal Policy Optimization

RAMIS Robot-Assisted Minimally Invasive Surgery

RCI Robot-Assisted Cardiovascular Intervention

RL Reinforcement Learning

RMSE Root Mean Squared Error

RMSprop Root Mean Squared Propagation

SAC Soft Actor-Critic

SOTA State of the Art

SVR Support Vector Regression

VATS Video-Assisted Thoracoscopic Surgery

ViT Vision Transformer

VLA Vision-Language-Action

VR Virtual Reality

WM World Model

Chapter 1

Introduction

1.1 Background

1.1.1 Minimally invasive surgery (MIS)

Minimally Invasive Surgery (MIS) refers to surgical techniques performed through very small incisions or natural body orifices, instead of the large incisions used in traditional open surgery. By minimizing the cutting of skin and underlying tissues, MIS methods aim to reduce the physical trauma of surgery. In practice, surgeons use specialized instruments and technologies such as endoscopes and catheters to access internal organs via these small “keyhole” openings. This approach results in far less disruption to healthy tissue compared to a conventional open procedure [2]. The clinical benefits of MIS for patients have been well documented. In general, MIS is associated with significantly reduced postoperative pain and a lower risk of complications during and after surgery. Patients who undergo MIS techniques tend to experience less blood loss and minimal scarring due to smaller incisions. A key advantage is faster recovery, hospital stays are usually shorter and patients return to normal activities sooner than they would after open surgery. For example, one review notes that MIS leads to shorter hospitalizations, quicker recovery times, and even reduced overall stress on the immune system. Additional benefits include a lower risk of wound infections and often a decreased need for postoperative pain medication. These advantages, summarized in the literature, make MIS an attractive option where it can be safely used, as it can achieve outcomes comparable

to open surgery with significantly less surgical trauma to the patient [2].

1.1.2 MIS applications

Building on these advantages, MIS has been successfully adapted across many surgical specialties. In general surgery, procedures such as laparoscopic cholecystectomy, appendectomy, and colorectal resections have become standard of care, replacing traditional open methods in a majority of cases due to their superior recovery profiles [3]. In orthopedics, arthroscopy provides minimally invasive access to joints, allowing ligament reconstructions or cartilage repairs through small portals rather than open arthrotomy [4]. Thoracic surgery has likewise been revolutionized by Video-Assisted Thoracoscopic Surgery (VATS), now widely used for lung resections and mediastinal procedures, which avoids rib spreading and reduces postoperative morbidity [5]. In neurosurgery, endoscopic approaches to the pituitary and ventricular system provide access to critical intracranial structures without the morbidity of large craniotomies [6]. The unifying principle across all these domains is that MIS preserves procedural efficacy while minimizing collateral tissue damage.

Within this broad landscape, the application of minimally invasive approaches to cardiac surgery and cardiology represents a particularly impactful advancement. Unlike traditional open-heart operations that require full sternotomy, Minimally Invasive Cardiac Surgery (MICS) employs smaller thoracic incisions—such as mini-thoracotomies or partial sternotomies—to access the heart while sparing patients the trauma of splitting the breastbone [7]. Cardiac catheterization, a cornerstone of modern minimally invasive cardiology, involves advancing a long, flexible tube—a catheter—via a peripheral blood vessel into the heart chambers or coronary vasculature under imaging guidance. This procedure serves both diagnostic and therapeutic purposes. On the diagnostic side, it enables direct measurement of intracardiac pressures, assessment of cardiac output, visualization of coronary anatomy, evaluation of valvular function, and detection of shunts or intracardiac defects [8]. For instance, left and right heart catheterizations are employed to quantify pressures and assess hemodynamics in various disease states, including heart failure and valvular disease, while coronary angiography—a specific catheter-based technique—provides detailed imaging of the coronary arteries to identify blockages or lesions requiring intervention [9].

Therapeutically, cardiac catheterization has transformed how cardiovascular pathologies are managed. Procedures such as Percutaneous Coronary Intervention (PCI), including balloon angioplasty and deployment of stents, are routinely performed via catheter access to restore blood flow in stenosed coronary arteries [10]. Catheter-based techniques also facilitate valve interventions such as balloon valvuloplasty for stenotic valves or Transcatheter Aortic Valve Replacement (TAVR) for a failing aortic valve [11]. Additionally, catheterization can facilitate device closure of congenital defects such as Atrial Septal Defects (ASDs or PFOs), sampling of myocardial tissue via biopsy, and precise hemodynamic assessment in conditions such as pulmonary hypertension and cardiogenic shock [12]. This minimally invasive approach achieves critical diagnostic and therapeutic aims without the need for open-chest surgery, reducing patient morbidity and shortening recovery. In the United States alone, over a million cardiac catheterization procedures are performed annually, reflecting its central role in contemporary cardiac care. With advances in catheter technology and operator expertise, complication rates have decreased, making catheter-based interventions both safe and versatile in managing a wide range of cardiac diseases from ischemic heart disease to structural and hemodynamic disorders [13].

While minimally invasive techniques span a wide range of surgical and interventional domains, the remainder of this dissertation focuses specifically on their application within cardiac interventions. Owing to its clinical impact, technical complexity, and strong dependence on imaging, device navigation, and force interaction with delicate cardiac structures, interventional catheterization constitutes the primary focus of this work. The following section therefore concentrates on interventional cardiac catheterization, outlining its procedural workflow, technical challenges, and the specific problem setting addressed in this thesis.

1.2 Interventional cardiac catheterization

Cardiac catheterization harnesses a slender, flexible tube—the catheter, to navigate the vasculature and access the heart, enabling both diagnostic and therapeutic interventions without open surgery. The standard catheter design typically features a polymer shaft with reinforcing elements to balance flexibility and torque, facilitating smooth navigation through tortuous vessels. Modern

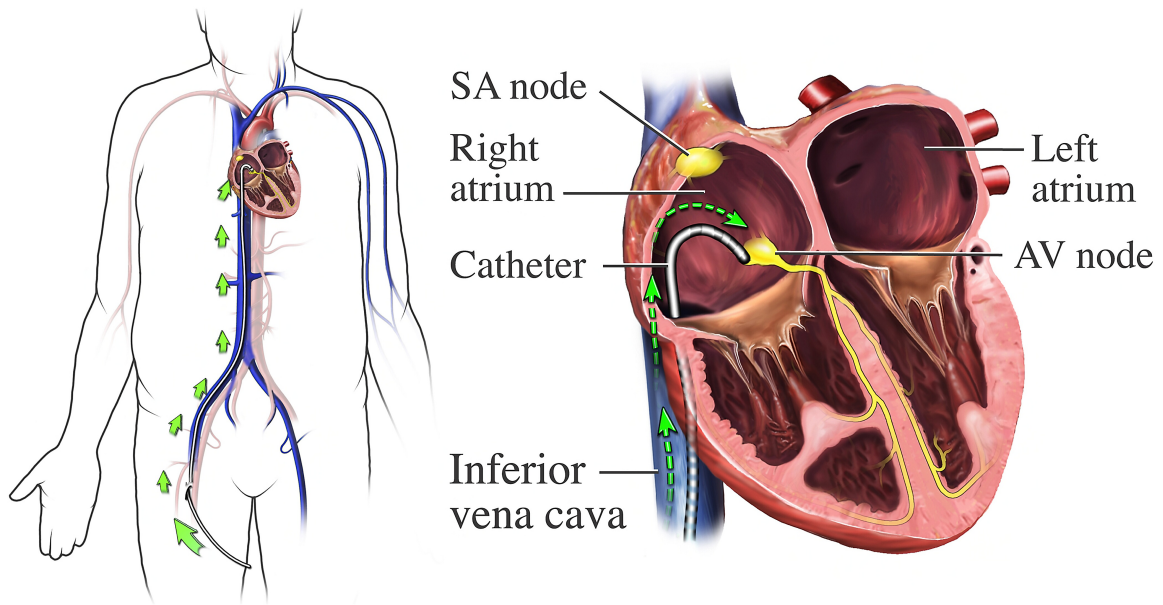


Figure 1.1: Schematic of a cardiac ablation catheter insertion path via femoral access [1].

variants include steerable distal segments controlled via pull-wires or tendon-based actuation, allowing operators to direct the tip toward precise targets such as arrhythmogenic foci for stimulation or ablation. Atrial Fibrillation (AF), one of the most common indications for catheter-based therapy, exemplifies the scale of conditions treated by these techniques. AF is the most frequent sustained arrhythmia, affecting an estimated 10.5 million adults in the United States (approximately 5% of the population). Projections suggest this number will grow to 12.1 million by 2030 in the U.S.. Globally, AF affects more than 33 million people, with prevalence continuing to rise due to aging demographics and increasing cardiovascular risk factors [14, 15, 16, 17, 18]. These statistics illustrate both the widespread application of cardiac catheterization and the significant patient population that benefits from such minimally invasive interventions.

1.2.1 Manual Catheterization

In manual catheterization, clinicians insert the catheter through entry points like the femoral or radial artery and manipulate its position under continuous fluoroscopic imaging (see Fig 1.1). Diagnostic procedures such as coronary angiography, pressure measurements, or intracardiac mapping rely on the operator’s tactile skill and visual feedback to safely navigate through cardiac anatomy.

Therapeutic interventions—including balloon angioplasty, stenting, septal defect closures, transcatheter valve repairs, and catheter ablation—build on these principles, applying device-specific mechanisms to treat underlying conditions. However, manual approaches pose challenges including cumulative radiation exposure to operators, physical fatigue, and limitations in precision of movement.

1.2.2 Robotic-assisted catheterization

To address these, robotic-assisted catheter systems have been introduced. Among the clinically deployed platforms is the Sensei Robotic System (Hansen Medical), featuring a remote control console and an articulating sheath that improves catheter stability and reach; it has demonstrated safety and feasibility in complex atrial fibrillation ablations. The Magellan Robotic System (Auris Health), derived from Sensei, provides enhanced steerability within peripheral vessels through multidirectional articulation at both catheter and sheath levels [19, 20, 21]. In addition, CorPath GRX (Corindus, now Siemens) enables robot-assisted percutaneous coronary intervention (see Fig. 1.2). It grants the operator precise control of guidewires, balloons, and stents via a remote console, significantly reducing radiation exposure while maintaining high procedural success rates. Clinical experience, including first-in-human uses in Japan, supports its safety and efficacy. Robot-assisted PCI has also been performed remotely via tele-operated sessions, underscoring its potential for expanded access to expert care [22, 23].

Magnetic navigation systems represent a distinct robotic approach. The Niobe (Stereotaxis) Magnetic Navigation System leverages external magnetic fields to direct catheter movements with high precision. In ablation procedures, Niobe has enabled refined catheter stability—yielding improved mapping resolution with fewer manual adjustments [24]. Other notable innovations include the Amigo Remote Catheter System (Catheter Precision), a simpler robotic accessory allowing physicians to manipulate standard electrophysiology catheters from outside the radiation field. The Amigo system supports basic movements—rotation, advancement, and deflection—while maintaining procedural efficacy and safety [25].

More recent research prototypes aim to enhance tactile feedback and autonomy. A magnetorheological fluid-based haptic catheter system offers operator force awareness via responsive feedback



Figure 1.2: CorPath GRX Neurovascular System.

for safer tissue contact. Another emerging platform, MAGiC RMN, demonstrates high success rates in treating arrhythmias with improved contact force control and safety [26, 27]. Reaching beyond physical hardware, simulator-based platforms like CathSim enable realistic training and algorithm development for autonomous navigation—capturing catheter–vessel interaction data including contact forces [28].

The effectiveness of any catheter-based system depends heavily on advanced imaging. Fluoroscopy remains the backbone of live guidance but its limitations—2D projection and radiation exposure—are addressed through electroanatomic mapping systems (e.g., CARTO, EnSite, Rhythmia), which construct 3D electrophysiological maps of heart chambers without continuous radiation. Anticipated machine learning–enhanced imaging promises even greater situational awareness and procedural precision.

1.3 Catheterization Bottlenecks

Registry data show that diagnostic (manual) coronary catheterization has a low vascular complication rate (approximately 0.7%), whereas Percutaneous Coronary Intervention (PCI) has a significantly higher rate—around 2.6–2.7% for elective PCI, escalating to as high as 6% in complex procedures [29, 30, 31]. In contrast, early evidence from robotic-assisted systems indicates comparable safety profiles. For instance, CorPath GRX has demonstrated similar procedural success and complication rates to manual PCI while significantly reducing operator radiation exposure. Robotic systems for AF ablation show promising reductions in fluoroscopy time, though the impact on perforation or stroke risk is not yet conclusively lower than manual methods [32].

When considered cumulatively, complication sources such as force misjudgment, vascular injury, embolism, radiation, and fatigue highlight inherent risks in both manual and robotic catheterization. Many of these risks stem from unintentional human error, underscoring the importance of examining the underlying causes. The following subsections review these root sources of complications in detail, providing a structured discussion of the specific human and procedural factors that compromise safety.

1.3.1 Lack of Haptic Feedback and Force Control

Despite the efficacy of interventional catheter procedures in treating complex cardiac conditions, a critical limitation persists: the operator lacks direct tactile feedback regarding the forces applied at the catheter tip. Surgeons cannot feel the mechanical interaction between the catheter and the delicate tissues of the heart or vascular walls, increasing the potential for harm [33, 34]. Clinical research indicates that the safe and effective contact force range during catheter ablation lies approximately between 0.1–0.4 N. Forces below this threshold may lead to incomplete lesion formation and ineffective arrhythmia treatment, whereas excessive force can cause tissue perforation, late-thrombus formation, or cardiac tamponade [35, 36]. Thus, even subtle misjudgments in applied force, due to the absence of touch, can dramatically impact procedural outcomes. Robotic-assisted interventions offer some improvement but do not fully restore the tactile perception lost in manual operation. Although these systems provide fine mechanical control and radiation protection,

they still deprive the operator of direct haptic cues. Surgeons must rely solely on visual indicators, such as catheter tip behavior on imaging, to infer contact force, introducing risk in delicate manipulations.

1.3.2 Mechanical Navigation, Vascular Tortuosity and Vessel Injury

Navigating through the complex, highly variable geometry of the vascular tree requires catheters to negotiate sharp curves, bifurcations, and spiral pathways. The elastic deformation of both catheter shafts and vessel walls can result in unpredictable catheter motion, complicating precise control. When the catheter inadvertently collides with the vessel wall, it can cause localized but potentially serious trauma—ranging from superficial endothelial denudation to full-thickness perforation, leading to internal bleeding or pseudoaneurysm formation [37]. Moreover, during navigation, the catheter edge or guidewire may dislodge calcific plaques or thrombi. Embolic materials can travel through the bloodstream and occlude critical distal vessels, most seriously, in the cerebral circulation, provoking ischemic strokes. Even micro-emboli, though subclinical, may contribute to cumulative cognitive decline in vulnerable patients [38].

1.3.3 Imaging-Related Challenges and Radiation Exposure

Fluoroscopy remains the standard imaging modality guiding catheterization, providing real-time X-ray visualization of the catheter relative to anatomical landmarks. While indispensable, fluoroscopy carries two primary limitations: Firstly, reliance on X-ray radiation exposes both patients and Operation Room (OR) staff to ionizing beam over a procedure’s duration. Cumulative radiation exposure increases the risk for stochastic effects including malignancy in patients, and cataracts or skin injuries in operators. Prolonged and repeated exposures particularly in complex or lengthy ablation cases, magnify these concerns [39, 40].

Secondly, fluoroscopy provides no tactile or capable sensory information regarding contact forces. The operator must interpret catheter movement and deformation visually without feeling resistance or compliance. This mismatch dramatically raises the risk of vessel injury: for instance, excessive force may lead to dissection or perforation at critical junctures such as arterial bifurcations or narrow passages.

1.3.4 Procedural Time and Operator Ergonomics

Endovascular navigation often occupies a substantial portion of interventional procedures, with prolonged cases demanding sustained concentration, precise hand–eye coordination, and static posturing under heavy lead protection. This prolonged focus contributes to operator fatigue especially in manual catheterization, which demands sustained concentration and precise hand-eye coordination. Moreover, numerous surveys and observational studies have shown that interventional cardiologists are at particularly high risk of musculoskeletal injury due to the physical demands of prolonged catheterization procedures. Operators often remain in static, awkward postures for extended periods while wearing heavy lead aprons, which markedly increases spinal load and joint stress. Large cross-sectional studies report that up to 84% of interventionalists experience at least one orthopedic injury during their career, with the lumbar and cervical spine being the most frequently affected regions, followed by the shoulders and hips.

These occupational hazards are compounded by the repetitive nature of catheter manipulation and the need for sustained hand–eye coordination under fluoroscopic guidance. In addition to orthopedic strain, the reliance on traditional lead aprons contributes to chronic back pain and fatigue, prompting calls for improved ergonomic practices and the development of lighter or lead-free protective solutions. Robotic-assisted platforms offer ergonomic relief and radiation shielding by allowing the operator to control instruments from a console. Nevertheless, current systems lack comprehensive haptic feedback and remain highly attention-demanding. The cognitive burden, monitoring imaging, controlling devices, and ensuring safety, under sustained conditions continues to result in clinician fatigue and increased mental workload [41, 42, 43].

1.4 State-of-the-arts: Force Sensing, visual perception and navigation

The intrinsic complexities of catheterization, coupled with the requirement for surgeons to perform gestures with sub-millimeter precision, have positioned this domain as a highly active area of research across multiple disciplines, including robotics and automation, sensing and materials science, and computer science. In response, a wide range of approaches has been proposed to address the inherent challenges of the procedure. Broadly, these efforts can be categorized into two major

research directions: (1) force sensing, which aims to provide accurate measurement and feedback of catheter–tissue interactions, and (2) semi-autonomous control, which seeks to enhance navigation precision and reduce operator burden through varying levels of robotic or algorithmic assistance. The following sections provide a detailed review of these approaches and their contributions to advancing catheter-based interventions.

1.4.1 Force Sensing

The ability to accurately sense forces applied along the x, y, and z axes, particularly at the catheter tip, is fundamental to improving the safety and efficacy of catheterization procedures. Real-time knowledge of catheter–tissue interactions not only enhances procedural precision but also helps to prevent complications such as perforation, dissection, or incomplete ablation lesions [44, 45]. Consequently, reliable detection and estimation of contact forces are critical to both guiding surgeons during delicate manipulations and providing timely warnings of potentially hazardous collisions. Broadly, two principal strategies have emerged for force measurement in catheter-based interventions. The first involves embedded sensors integrated directly into the catheter tip, enabling direct quantification of contact forces at the point of interaction. The second approach leverages distal shaft shape analysis, where force estimation is inferred indirectly from the catheter’s geometric deformation as it navigates vascular pathways. Both strategies present distinct advantages and technical challenges, and together they form the foundation of ongoing research into safer and more intelligent catheterization systems. In the following sections, these two paradigms are reviewed in detail, with particular emphasis on their underlying principles, representative implementations, and inherent limitations.

Embedded Sensors

Embedded force sensors are primarily integrated at the catheter tip and/or along the distal shaft to provide operators with enhanced tactile perception during intervention [46, 47, 48]. These sensors are commonly developed based on either electrical or optical transduction principles. For instance, an optical-based miniature force sensor was introduced in [49] with the objective of embedding it at the tip of an ablation catheter to enable three-dimensional force measurement. Similarly, Yokoyama

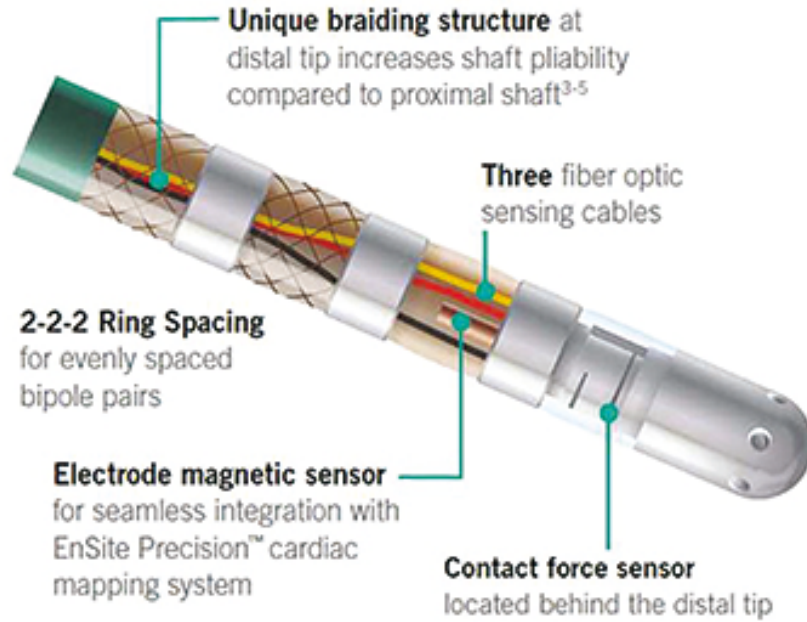


Figure 1.3: TactiCath Contact Force Ablation Catheter.

et al. [50] proposed a Fiber Bragg Grating (FBG)-based sensor, which demonstrated feasibility for integration at the catheter tip. This innovation has since been translated into a clinically approved device, the TactiCath™ Contact Force Ablation Catheter (Abbott, Abbott Park, Illinois, United States) (see Fig 1.3). Another commercially available example is the THERMOCOOL® Catheter (Biosense Webster), which incorporates contact force sensing capabilities (see Fig 1.4).

Despite their clinical benefits, sensor-embedded catheters remain significantly more expensive than conventional sensor-less designs. Furthermore, their implementation necessitates substantial modifications to the standard catheter architecture, introducing additional challenges in terms of manufacturing complexity, miniaturization, and regulatory approval [51]. These limitations continue to motivate research into alternative, sensor-less approaches for force estimation.



Figure 1.4: THERMOCOOL[®] catheter (Biosense Webster).

Sensor-free Methods

Sensor-free strategies have recently gained significant attention as they address many of the limitations inherent to sensor-embedded catheters [47]. These approaches rely on analyzing the deflections of the distal shaft in conventional, sensor-less catheters and estimating contact forces from image-derived features. Generally, sensor-free methods for force estimation can be classified into two categories: model-based approaches and learning-based approaches.

In **model-based methods**, the catheter’s mechanical properties are inferred from image observations using mathematical formulations such as Euler–Bernoulli beam theory, Cosserat rod theory, and multi-body dynamic models [52, 53, 54, 55, 56, 57, 58, 59, 60]. While these methods provide physically interpretable insights, their accuracy is often limited in practice due to uncertainties in parameters (e.g., catheter stiffness variations, vessel compliance, or boundary conditions). Such discrepancies between model assumptions and real-world catheter behavior remain a persistent challenge [61].

By contrast, **learning-based** approaches aim to directly map distal shaft deflections to force vectors using data-driven techniques. These methods require datasets of catheter images paired

with ground-truth force measurements to train models capable of predicting forces. For example, Khoshnam *et al.* proposed an algorithm that monitored variations in distal shaft curvature to estimate tip forces from images [62]. In another study, it introduced a machine vision-based feature extraction framework that transformed pixel-level deflection images into feature vectors, which were subsequently mapped to forces in the x, y plane using a Multi Layer Perceptron network (MLP) and Support Vector Regression (SVR) models [63]. However, this pipeline suffered from limited generalizability in distinguishing diverse deflection patterns, and its reliance on pixel-wise feature extraction resulted in high computational overhead.

To overcome these limitations, subsequent work explored deep learning architectures that bypass handcrafted feature extraction. A ResNet-based Convolutional Neural Network (CNN) was proposed that directly regressed catheter deflection images to force outputs [1, 64]. Such end-to-end methods offer improved robustness and scalability compared to traditional feature engineering. Nonetheless, most existing learning-based approaches remain constrained to two-dimensional image inputs, which restricts their ability to infer forces along the z -axis. This limitation arises because 2D projections cannot capture out-of-plane deflections. Furthermore, image-based pipelines inherently require reliable catheter detection or segmentation as a preprocessing step to isolate the shaft geometry, which introduces additional complexity and potential sources of error.

1.4.2 Autonomous Navigation

As discussed earlier, catheterization procedures may be performed manually or with the aid of guidewire manipulators. In both cases, successful navigation requires surgeons to attain a high level of proficiency, typically developed through years of training and gradual refinement of technical skills. This learning curve reflects the inherent complexity of maneuvering catheters through tortuous and patient-specific vascular anatomies, often under time-sensitive and high-stakes conditions.

To alleviate the cognitive load on operators and reduce reliance on manual trial-and-error maneuvers, recent research has focused on developing semi-autonomous and autonomous catheter navigation systems. These approaches aim to enhance efficiency by enabling the catheter to traverse complex vascular pathways with minimal manual intervention. In principle, such systems consist of two tightly coupled components: (1) a perception module, which provides real-time situational

awareness by interpreting the operative environment (e.g., vascular geometry, catheter position, tissue interactions), and (2) a control module, which translates this information into precise actuation commands to achieve safe and efficient navigation. The integration of these components enables more reliable catheter steering, with the potential to shorten procedure times, minimize errors, and improve overall procedural outcomes [65].

To address the limitations of manual control, advancements in autonomy for robotic catheterization have increasingly been explored through three state-of-the-art learning paradigms: Reinforcement Learning (RL), Inverse Reinforcement Learning (IRL), and Imitation Learning (IL). These methods seek to endow robotic systems with advanced decision-making capabilities, enabling them to perform complex navigation tasks with minimal human intervention. RL approaches allow robotic agents to learn optimal catheter navigation strategies through trial-and-error interactions with the environment, while IRL infers underlying reward structures from expert demonstrations to replicate human-level decision making. IL, on the other hand, leverages supervised learning of expert trajectories, enabling systems to mimic skilled operators directly. The following sections review representative works within each of these learning paradigms, with particular emphasis on their application to robotic catheter navigation.

Reinforcement Learning: (RL) has been extensively investigated for autonomous navigation in vascular and catheter-based interventions, offering adaptive decision-making capabilities derived from trial-and-error interactions within simulated or physical environments. Recent works have demonstrated its promise across a range of system designs and training paradigms. One study proposed a Deep Q-Network (DQN)-based RL framework for autonomous guidewire navigation in coronary interventions, using sequences of grayscale images around the guidewire tip as input. To accelerate convergence and improve policy quality, human demonstrations were incorporated during pre-training, thereby enhancing the agent’s performance in traversing complex vascular geometries. The method achieved high success rates in both 2D and 3D coronary artery phantoms, underscoring the feasibility of RL-driven navigation for coronary procedures [66].

Building on this foundation, another work introduced an autonomous guidewire feeding method for vascular interventions using the Soft Actor-Critic (SAC) algorithm in combination with YOLOv5s for real-time guidewire tip detection [67]. By processing tip position and image inputs, the system

generated push–pull and rotation actions to steer the guidewire within a vessel phantom. This framework effectively addressed challenges of modeling accuracy and training efficiency, achieving precise and stable guidewire control [68].

Beyond navigation alone, RL has also been integrated into training systems for skill transfer. A notable example is a Virtual Reality (VR)-based training platform employing the Deep Deterministic Policy Gradient (DDPG) algorithm. This system incorporated haptic feedback to transfer expert catheterization skills to novices in real time. By modeling expert behaviors and providing tactile guidance, the framework improved novice performance in catheter manipulation, thereby demonstrating the broader applicability of RL in surgical training [69].

Further advancements have focused on image-guided RL navigation. One approach, termed Image-Guided Autonomous Guidewire Navigation (IAGN), utilized the SAC algorithm augmented with the BDA-Star path planning algorithm and a pre-trained convolutional neural network (CNN) for feature extraction. This integration reduced reliance on manual demonstrations and mitigated slow convergence issues, enabling reliable and precise autonomous guidewire navigation during endovascular interventions [70].

Finally, to facilitate standardized research and reproducibility, CathSim, an open-source RL-compatible simulator for endovascular interventions, has been introduced. Built on the MuJoCo physics engine, CathSim provides a real-time training environment that supports advanced algorithms such as Proximal Policy Optimization (PPO) and SAC. By closely replicating catheter–vessel interactions and supporting both image- and data-driven inputs, CathSim enables robust training and validation of RL-based navigation strategies under controlled and reproducible conditions [28].

Imitation Learning: (IL) has gained increasing attention in robotic catheterization and related minimally invasive interventions, as it enables systems to learn expert-level behaviors directly from demonstrations rather than relying solely on handcrafted models or trial-and-error exploration. By leveraging expert trajectories, IL frameworks allow robotic agents to replicate complex surgical maneuvers with high precision and adaptability, thereby reducing the reliance on prolonged operator control and minimizing the learning curve for autonomous systems.

One notable study introduced an imitation learning-based approach for 3D/2D model-to-image registration in cardiac procedures, such as Cardiac Resynchronization Therapy (CRT). This method

aligned preoperative anatomical models derived from MRI or CT with intraoperative X-ray images by training a neural network to mimic optimal alignment actions demonstrated by experts. Trained in a supervised manner using expertly aligned image pairs, the model predicted corrective adjustments that improved both the accuracy and robustness of image-guided interventions [71].

Another framework, Collaborative Robot-Assisted Endovascular Catheterization, applied Generative Adversarial Imitation Learning (GAIL) to automate catheter manipulation. Expert demonstrations were segmented into state–action pairs and used to train a GAIL model, where the generator learned to replicate expert strategies and the discriminator distinguished between expert and generated actions. The learned policies were subsequently refined with Proximal Policy Optimization (PPO) and deployed on a robotic catheter system. This approach significantly improved navigation precision, reduced operator involvement, and demonstrated adaptability across varying anatomical models. The model processed inputs such as the robotic end-effector’s position, the target site (e.g., lesion or tumor), and expert action sequences, producing control actions for safe and efficient catheter guidance [72].

Beyond catheterization, IL has also been applied in Robot-Assisted Minimally Invasive Surgery (RAMIS). For instance, an automation framework for soft-tissue retraction combined GAIL with PPO to learn retraction strategies from a limited set of expert demonstrations. Initially trained in simulation, the model was later transferred to a physical robotic system, where it enabled autonomous retraction with human-like precision, exposing surgical targets such as tumors while adapting to dynamic tissue responses [73].

Further advancements have integrated IL with reinforcement learning for enhanced robustness. One such study introduced a robust path planning framework for robotic catheterization by combining GAIL with PPO-based RL. This hybrid approach allowed the model to learn from expert demonstrations while optimizing for task-specific objectives, such as minimizing collisions with vessel walls. Inputs included catheter position, target location, and real-time environmental data, while outputs corresponded to adaptive control commands that guided the catheter along safe and efficient trajectories. The inclusion of a Curriculum Learning module further enabled progressive training across increasingly complex environments, improving generalization to realistic vascular anatomies [74].

Inverse Reinforcement Learning: (IRL) extends beyond traditional imitation learning by seeking to infer the underlying reward function that governs expert behavior, rather than directly mimicking expert actions. This distinction enables IRL to capture the latent objectives of expert demonstrations, thereby producing more generalizable and adaptive control policies for complex surgical environments.

One study introduced an IRL-based algorithm for pre-operative path planning in intravascular and intracardiac catheter navigation during minimally invasive surgeries. The framework integrated Behavioral Cloning (BC) and Generative Adversarial Imitation Learning (GAIL) to implicitly approximate the underlying reward structure through adversarial learning. By inferring these objectives, the model generated navigation strategies capable of adapting to dynamic and patient-specific anatomies, surpassing the limitations of standard imitation learning. The system took as input the catheter’s position, the designated target location, and real-time environmental feedback, and produced as output a sequence of control actions for precise and safe catheter advancement. This approach improved both the safety and precision of catheter-based interventions by aligning decision-making policies with clinically relevant goals [75].

A related work applied IRL to autonomous navigation in mechanical thrombectomy, where the reward function was derived from expert demonstrations consisting of recorded state–action pairs. Under the assumption that experts act optimally with respect to an unknown reward function, IRL formulated an optimization problem to recover this function by identifying what rewards best explained observed behaviors. The environment was represented using features such as proximity to obstacles and distance to target sites, and the inferred reward function was subsequently used to train a reinforcement learning policy—implemented with the Soft Actor-Critic (SAC) algorithm—for autonomous catheter or guidewire navigation. This strategy demonstrated high precision and robustness, showing the potential of IRL to replicate expert-level decision-making while maintaining adaptability to diverse and evolving surgical conditions [76].

1.5 Research Motivations, Objectives and Scopes

Despite the remarkable progress of catheter-based interventions, significant challenges remain that limit both safety and efficacy. As outlined in the preceding discussion, current systems, whether manual or robotic, still suffer from drawbacks related to force perception, navigation through tortuous vasculature, and heavy reliance on fluoroscopic imaging. These limitations, which will be highlighted in the following sections, form the foundation for the research motivations of this dissertation, guiding the pursuit of solutions that address the unmet needs of precision, safety, and operator support in interventional cardiology. Accordingly, the following sections examine these challenges in detail, beginning with the limitations of existing force sensing approaches and subsequently addressing the difficulties associated with catheter navigation.

1.5.1 Force Sensing Limitations

Efforts to address the lack of haptic feedback in catheterization have led to three main categories of force sensing strategies: sensor-embedded catheters, model-based estimation, and learning-based image-driven methods. While each direction has shown promise, they also present important limitations that restrict their widespread clinical adoption.

Sensor-embedded catheters: Catheters equipped with miniaturized force sensors (e.g., fiber Bragg gratings or microelectromechanical sensors) can directly measure tip–tissue interactions and thereby restore tactile awareness to the operator. However, these devices are significantly more expensive compared to standard, sensor-less catheters, limiting their accessibility in routine practice. In addition, embedding sensors necessitates substantial changes to the design and structure of conventional catheters, raising concerns about manufacturability, regulatory approval, and long-term reliability under repeated sterilization and clinical use.

Model-based approaches: Physics-driven models that estimate forces from catheter shape and material properties represent a lower-cost alternative, as they avoid hardware modifications. Nevertheless, these methods are highly sensitive to uncertain parameters such as patient-specific anatomy, boundary conditions, and material nonlinearities. Even small deviations between modeled assumptions and the actual catheter–tissue interaction in vivo can significantly reduce estimation accuracy.

This persistent gap between theoretical models and authentic catheter behavior remains a fundamental challenge.

Learning-based methods: more recently, data-driven learning approaches have been introduced to overcome both hardware costs and modeling uncertainties. By training on real experimental datasets, these methods learn to map catheter deflection in 2D images to estimated tip forces. Although effective in many cases, they suffer from two key drawbacks. First, because the input is typically a single-plane 2D image, these methods cannot reliably capture variations along the depth (z) axis, making estimation of the full 3D force vector incomplete. Second, they require explicit detection or segmentation of the catheter within each frame to extract its deflection shape. This necessity adds a preprocessing stage that increases computational burden and reduces efficiency. Moreover, segmentation cannot be eliminated altogether since the surgeon simultaneously requires clear visualization guidance for safe navigation, meaning this additional step is unavoidable in clinical workflows.

1.5.2 Catheter Navigation Limitations

The pursuit of autonomy in catheter navigation and manipulation has led to significant interest in learning-based control strategies, particularly reinforcement learning (RL), imitation learning (IL), and inverse reinforcement learning (IRL). While these methods offer compelling opportunities for enabling autonomous or semi-autonomous interventions, each faces substantial limitations that hinder direct application in medical robotics.

Reinforcement Learning: Despite its promise, RL suffers from fundamental drawbacks in the context of catheterization. A primary concern is sample inefficiency. RL agents typically require extensive numbers of interactions to converge on stable policies. This requirement is impractical in medical robotics, where real procedures are costly, time-consuming, and inherently risky for patients. Furthermore, RL methods trained in simulation encounter a pronounced sim-to-real gap, as catheter-tissue dynamics, vessel elasticity, and imaging artifacts are difficult to model with high fidelity, causing trained policies to fail upon transfer to the clinical environment. Another persistent challenge lies in reward engineering. Designing clinically meaningful reward functions that ensure safe, efficient navigation and proper force control is complex. Poorly specified rewards may yield

suboptimal or unsafe strategies. Additionally, the exploratory nature of RL raises significant safety concerns, as the algorithm may generate hazardous actions (e.g., vessel perforation or catheter misplacement) during training, unacceptable in a clinical setting. Finally, the high dimensionality of the catheter’s continuous action space, including deflections, rotations, and applied forces, slows convergence and contributes to policy instability.

Imitation Learning: IL aims to bypass reward engineering by directly learning policies from expert demonstrations. However, its effectiveness is constrained by several limitations. First, IL exhibits a strong dependence on expert demonstrations, requiring large, high-quality datasets of surgical trajectories. Collecting such data in catheterization is both resource-intensive and non-standardized across institutions. Second, IL is vulnerable to compounding errors (covariate shift) — small prediction errors accumulate as the catheter deviates from expected states — potentially leading the policy into unseen and unsafe regions of the state space. Third, IL suffers from a lack of adaptability, as learned policies are largely restricted to mimicking the training distribution and often fail when encountering patient-specific anatomies, vessel geometries, or novel imaging conditions. In addition, ambiguity in demonstrations complicates the learning process — different surgeons may employ distinct strategies to achieve the same outcome — and IL struggles to reconcile such inconsistencies. Finally, IL methods do not explicitly optimize for safety or performance—they replicate observed behavior, which may not represent the optimal balance between minimizing tissue forces, maximizing efficiency, and ensuring patient safety.

Inverse Reinforcement Learning: IRL attempts to overcome reward specification challenges by inferring the underlying reward function from expert demonstrations. While conceptually attractive, IRL faces several limitations in catheter-based interventions. One critical issue is the ambiguity of rewards. Multiple candidate reward functions may equally explain the same expert behavior, making it difficult to identify the true clinical objectives. Moreover, IRL is computationally demanding and exhibits scalability issues, particularly when extended to high-dimensional catheter environments involving continuous control and complex vessel geometries. Finally, IRL methods often struggle with generalization, as reward functions learned from a fixed set of demonstrations may fail to adapt to new patient anatomies, unseen procedural variations, or differences in imaging modalities.

1.5.3 Research Objectives

The systematic review of existing methodologies and the critical analysis of their limitations reveal several fundamental knowledge gaps that directly motivate this doctoral research. In particular, unresolved challenges in force sensing, catheter visualization, and autonomous navigation continue to restrict progress toward safe, accurate, and efficient catheter-based interventions. Addressing these gaps is essential to advance both the technical and clinical frontiers of the field. Motivated by these shortcomings, the primary objective of this research is to propose a series of novel methodologies grounded in artificial intelligence and deep learning. These approaches are designed to overcome the limitations of existing solutions by enabling more accurate force estimation, improving the reliability and efficiency of catheter visualization, and developing robust learning-based frameworks for autonomous navigation that generalize across diverse patient anatomies while maintaining stringent safety requirements. These frameworks are intended as a step toward a deep learning foundation-model paradigm for catheter-based interventions, emphasizing the learning of transferable representations that can support generalization across anatomical variability, procedural settings, and downstream tasks. In light of the limitations identified in the literature and the corresponding research motivations, this doctoral research pursues two overarching objectives:

- Design and development of clinical force estimation frameworks for catheters using real-time, sensor-free and deep learning-based methodologies.
- Design and development of a deep learning model-based autonomous Robot-assisted Cardiovascular Intervention (RCI) system, integrating catheter perception, localization, path planning, and navigation.

1.5.4 Research Scopes

To achieve these objectives, the following specific criteria must be satisfied:

Force Estimation Framework:

- The estimator must remain compatible with standard clinical catheters (e.g., Boston Scientific Blazer II XP) without requiring structural modification.

- The approach must be independent of mechanistic modeling, ensuring robust performance under real clinical conditions.
- The design must avoid embedding physical sensors, thereby reducing implementation cost and preserving catheter integrity.
- The system must leverage existing imaging modalities already in use in the operating room (e.g., fluoroscopy for ablation procedures).
- The framework must estimate 3D contact forces along the x, y, and z axes.
- The system must simultaneously perform segmentation and force estimation in 3D, eliminating the need for additional preprocessing steps.
- The system must leverage a multitask stereo learning framework to segment the catheter from two complementary viewpoints, enabling consistent cross-view understanding.
- The model must demonstrate robustness to perturbations in imaging system position (e.g., translation, rotation, scaling), minimizing operator effort in calibration.
- As a data-driven system, the framework must allow fine-tuning across different catheter types, enabling adaptability to varying mechanical properties.

Autonomous Navigation Framework:

- The system must be capable of localizing the catheter within the vascular system exclusively from imaging data, without additional sensors.
- The framework must operate efficiently within a unified architecture, avoiding the incorporation of separate perception modules such as explicit detection or segmentation pipelines.
- The navigation policy must be explicitly grounded in task-level objectives, such as a designated target point, to ensure goal-directed behavior.
- The learning framework must minimize reliance on trial-and-error exploration by leveraging structured guidance or informed learning strategies.

- The system must be designed to function in a physical experimental setup, thereby avoiding dependency on simulation environments and mitigating sim-to-real transfer challenges.
- The perception modules must be built upon continuous image-based data streams derived from standard clinical imaging systems.
- The autonomous agent must learn catheter actuation policies and control the robotic manipulator without direct surgeon intervention during training.

1.6 Thesis Contributions

This thesis introduces a series of novel contributions that advance the state-of-the-art in sensor-free force estimation and autonomous catheter navigation. For the first time, a lightweight fusion-based Convolutional Neural Network (CNN) architecture is proposed that enables end-to-end mapping of catheter deflections directly from stereo images to 3D contact forces along the x , y , and z axes. Unlike prior methods that rely on handcrafted features, the fusion process inherently captures the variation of deflections in three dimensions without additional preprocessing, ensuring robustness against rotation and scaling. Building on this foundation, the thesis further presents an end-to-end multitask architecture that unifies 3D force estimation and stereo semantic segmentation. By jointly segmenting the catheter from two X-ray views and leveraging the decoded feature maps, the model achieves simultaneous 3D reconstruction and force estimation in a single lightweight network, effectively eliminating the need for segmentation as a separate preprocessing stage. This approach establishes a new state-of-the-art by balancing accuracy, memory efficiency, and latency in clinical settings.

Extending beyond CNN architecture, the thesis introduces a novel transformer-based multitask model that leverages a patch-based encoder–decoder built on Vision Transformers (ViTs) with cross-attention mechanisms. This architecture simultaneously performs stereo catheter segmentation and 3D force estimation, where the segmentation task guides the force estimation head to attend on catheter deflections while suppressing background artifacts, further pushing the boundaries of state-of-the-art performance. Finally, the thesis presents the first goal-conditioned, vision-based control model for fully autonomous catheter navigation. Inspired by expert surgical maneuvers, the

model fuses kinematic and visual cues in an end-to-end DINO-based framework, directly steering the catheter toward its target destination without relying on explicit object detection, localization, or segmentation. By conditioning navigation on the image of the final destination, the model demonstrates the feasibility of autonomous, perception-driven robotic catheterization, thus bridging the gap between image-based force estimation and closed-loop robotic control. Motivated by recent advances in world models and deep learning foundation models, the proposed architecture is positioned as a step toward such unified learning workflows, in which perception, representation, and control are jointly learned within a single, goal-conditioned framework.

1.7 Publications

The research conducted throughout this doctoral study has led to a series of publications that not only support the core contributions of this thesis in sensor-free force estimation, catheter segmentation, and autonomous robotic navigation, but also extend to related domains including haptics guidance, robotics, perception systems for self-driving vehicles, telesurgery, and surgical skill transfer. Collectively, these works reflect the breadth and depth of this doctoral research, advancing multiple areas at the intersection of artificial intelligence, deep learning-based computer vision, medical robotics and mechatronics, and autonomous systems. The publications are listed below:

Journals Papers

- (1) P. Fekri, J. Dargahi, and M. Zadeh, “Deep learning-based haptic guidance for surgical skills transfer,” *Frontiers in Robotics and AI*, vol. 7, 2021. [77]
- (2) 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. [78]
- (3) 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. [79]
- (4) P. Fekri, M. Zadeh, and J. Dargahi, “TransForSeg: A Multitask Stereo ViT for Joint Stereo

Segmentation and 3D Force Estimation in Catheterization,” submitted to Pattern Recognition (Elsevier). [80]

- (5) P. Fekri, M. Roshanfar, S. Barbeau, F. Famouri, T. Looi, D. Podolsky, M. Zadeh and J. Dargahi, “DINO-CVA: A Multimodal Goal-Conditioned Vision-to-Action Model Toward Autonomous Catheter Navigation,” submitted to IEEE Transactions on Medical Robotics and Bionics. [81]

Conferences Papers

- (1) 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. [82]
- (2) 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. [1]
- (3) 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,” in *IEEE 5th International Conference on Soft Robotics (RoboSoft)*, 2022. (Dual acceptance)
- (4) P. Fekri, M. Zadeh, and J. Dargahi, “Fs-net: An encoder-decoder architecture for catheter segmentation and contact forces estimation in intracardiac catheters,” in *2025 IEEE World Haptics Conference (WHC)*, 2025, pp. 123–132. [83]

Other Publications

- (1) 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. [63]

- (2) V. Abedi, J. Dargahi, M. Zadeh and P. Fekri, "Software Failures Prediction in Self-Driving Vehicles" in *IEEE 92nd Vehicular Technology Conference: IEEE VTC2020*, Victoria, BC, Canada, 2020. [84]
- (3) P. Fekri, Dargahi. J, and Zadeh. M , "On the Safety of Autonomous Driving: A Dynamic Deep Object Detection Approach," *SAE Technical Paper* 2019-01-1044, Detroit, MI, USA, 2019, <https://doi.org/10.4271/2019-01-1044>. [85]
- (4) Tavakolian. N, Fekri. P, Dargahi. J, Zadeh. M. "KDepthNet: Mono-Camera based Depth Estimation for Autonomous Driving" in *SAE Technical Paper*, e-ISSN: 2688-3627, Detroit, MI, USA, 2022 [86]
- (5) 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. [87]
- (6) S. Famouri, P. Fekri, and J. Dargahi, "Kinematics skill modeling of cardiac catheterization via deep learning method," in *2024 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE*, 2024, pp. 1–9. [88]
- (7) M. Roshanfar, P. Fekri, and J. Dargahi, "Toward autonomous cardiac catheterization through a parametric finite element simulation with experimental validation," *ICAS 2023*, p. 23, 3 2023. [89]
- (8) 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. [90]
- (9) S. Barbeau, P. Fekri, D. Osowiechi, A. Bahri, M. Yazdanpanah, M. Aminbeidokhti, and C. Desrosiers., "Cta: Cross-task alignment for better test time training", 2025. [91]
- (10) N. Mallouk, M. Roshanfar, P. Fekri, and J. Dargahi, "A brief review on recent advances in haptic technology for human–computer interaction: Force, tactile, and surface haptic feedback," in *Proc. 7th Int. Conf. Information and Computer Technologies (ICICT)*, 2024,

pp. 34–39, doi: 10.1109/ICICT62343.2024.00012. [92]

- (11) M. Roshanfar, P. Fekri, R. H. Nguyen, C. He, P. H. Kang, J. Drake, E. D. Diller, and T. Looi, “Learning-based tip contact force estimation for FBG-embedded continuum robots,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, Vienna, Austria, May 19–23, 2025. [93]

1.8 Thesis Layout

Chapter 1 introduced the background and motivation for this research. It began with a review of the clinical needs in catheter-based cardiovascular interventions and the limitations of existing solutions for force estimation, catheter visualization, and autonomous navigation. Through an extensive literature review, current learning-based methods were critically analyzed to expose their shortcomings in accuracy, generalizability, and clinical applicability. This analysis revealed the key knowledge gaps that motivated the research and led to the formulation of the objectives and scope of the thesis.

Chapter 2 presents the first proposed solution addressing the force sensing problem through a novel architecture called Y-Net. This model directly predicts 3D forces from RGB images captured under biplanar fluoroscopy. Y-Net is designed as a lightweight late-fusion architecture capable of learning catheter deflections without requiring physical sensors or explicit geometric modeling [78].

Chapter 3 presents H-Net, the first CNN-based multitask architecture for catheterization. Unlike prior methods, H-Net simultaneously segments the catheter and estimates 3D forces, thereby eliminating the need for a separate segmentation phase. To further validate performance, the chapter also proposes a synthetic X-ray generator and evaluates the network across multiple datasets, demonstrating improved accuracy and generalizability [79, 83].

Chapter 4 advances the state-of-the-art with a multi-input, multi-output Vision Transformer (ViT)-based model. This lightweight encoder–decoder architecture incorporates dual heads for stereo segmentation and 3D force estimation. By leveraging cross-attention mechanisms, the framework pushes beyond CNN-based approaches, offering both enhanced interpretability and efficiency [80].

Chapter 5 focuses on autonomous catheter steering through the introduction of the Dino-Action model. This multi-stage transformer framework projects sequential video frames into control actions, enabling perception-driven navigation of a catheter within vascular phantoms. The work is validated in a robotic experimental setup, where a catheter is steered by the model while a vision system captures real-time imaging feedback [81].

Chapter 6 concludes the thesis by summarizing the key findings and emphasizing their impact on sensor-free force estimation and autonomous robotic catheterization. The chapter also outlines several directions for future work, including model generalization to diverse clinical scenarios, integration with real-time robotic platforms, and translation toward in vivo experimentation and clinical deployment.

Chapter 2

Y-Net: A CNN for 3D Estimation of Contact Forces in Intracardiac Catheters

2.1 Introduction

Atrial Fibrillation (AF) is a cardiac arrhythmia characterized by impaired electrical signaling in the heart, resulting in an inability to maintain a regular sinus rhythm [94]. According to the Heart Disease and Stroke Statistics update from the American Heart Association, the prevalence of AF is projected to reach 12.1 million individuals in the United States by 2023 and 17.9 million in the European Union by 2060 [95]. Pharmacotherapy, electrical cardioversion, artificial pacemakers, and catheter ablation constitute the most commonly available treatment options for AF [96]. Although the first three approaches can be effective in certain cases, catheter ablation remains the predominant and well-established therapeutic procedure adopted in clinical practice [97]. In this minimally invasive procedure, a long, flexible, wire-like instrument known as a catheter is introduced into the cardiovascular system through vascular access points such as the groin, neck, or shoulder. The catheter is then navigated through the vascular network until its tip is positioned near arrhythmogenic regions of the heart [98]. In tendon-driven ablation catheters, approximately the

distal 10 cm of the flexible shaft—referred to as the distal shaft—can be actively deflected by the surgeon to target abnormal cardiac tissue [99]. Radiofrequency energy delivered at the catheter tip generates localized heating, which disrupts the bioelectrical activity of these regions and facilitates the restoration of normal cardiac rhythm. The entire procedure is typically performed manually under X-ray fluoroscopy guidance. Several interrelated factors, including ablation duration, delivered electrical power, and catheter–tissue contact force, have been shown to significantly influence the outcome and safety of the ablation procedure [100].

Within this focused scope of interventional cardiac catheterization, this Chapter concentrates on catheter-based treatment of atrial fibrillation as a representative and clinically significant application. In particular, it addresses one of the two central challenges outlined in Chapter 1.5.1, namely the accurate estimation of catheter–tissue contact forces in three dimensions. Despite its critical role in procedural safety and ablation efficacy, force sensing remains limited in many clinical workflows, motivating the development of sensor-free, image-based force estimation methods. The following sections therefore investigate this problem in detail, forming the foundation of the proposed approach.

2.1.1 Related Work

As discussed previously, this Chapter focuses specifically on the challenges associated with estimating contact forces at the tip of an ablation catheter. During catheter insertion and manipulation, surgeons lack direct tactile feedback, as it is not possible to physically sense the forces applied when the catheter interacts with anatomical lumens such as cardiac or vascular tissue. Both excessive and insufficient contact forces can lead to procedural failure, including tissue damage or incomplete ablation [101]. Consequently, accurate force sensing has the potential to substantially improve the safety and efficacy of catheterization procedures [44, 45]. It is therefore critical to detect and estimate applied contact forces and to provide timely feedback to surgeons in order to mitigate hazardous catheter–tissue interactions.

As reviewed in Chapter 1.4.1, catheter contact forces can be obtained using two principal approaches: (i) embedded force sensors integrated into the catheter tip and (ii) distal shaft shape analysis [46, 47]. Commercial systems such as the THERMOCOOL[®] catheter (Biosense Webster) and

the TactiCathTM contact force ablation catheter (Abbott, Abbott Park, IL, USA) incorporate force sensors directly within the catheter tip. Despite their clinical utility, such sensor-based catheters are considerably more expensive than conventional sensor-less catheters. Moreover, integrating force sensors typically requires substantial modifications to the catheter’s mechanical design and structural composition [51].

In contrast, sensor-less force estimation methods have attracted growing attention due to their ability to alleviate several limitations inherent to sensor-based approaches [47]. These methods generally infer contact forces by analyzing distal shaft deflections of standard catheters using image-based features. Broadly, existing sensor-less force estimation techniques can be categorized into two main classes: (i) model-based approaches and (ii) learning-based approaches.

In the model-based methods, the mechanical characteristics of the catheter are extracted from the input images using solutions such as Beam theory models, Cosserat-type rod theories, and multi-body dynamics. [52, 53, 54, 55, 56]. Nevertheless, uncertain model parameters still remain as a challenge for the model-based solution [61].

Conversely, learning-based systems constitute a sensor-less paradigm in which the relationship between catheter deflection patterns observed in images and the corresponding contact forces is learned through data-driven modeling. In this framework, the target surgical procedure (e.g., cardiac ablation) is reproduced using a purely mechanical setup, a fully virtual environment, or a hybrid physical–virtual simulation. Within these experimental settings, a specific catheterization task is designed to induce controlled catheter bending within a predefined curvature range. This design enables the systematic generation of diverse deflection profiles and their associated contact forces along multiple directions under controlled, non-clinical conditions. Subsequently, a dataset is constructed comprising catheter deflection images paired with measured force values. This dataset serves as the basis for training learning-based models that map observed deflections to the corresponding contact forces along the x, y, and z axes.

In a study, Khoshnam *et al.* proposed a method in which the variation of the distal shaft curvature is monitored in an image-based algorithm to estimate the applied force of the tip [62]. In [63], a machine vision–based feature extraction algorithm was proposed to transform catheter deflection images from the pixel domain into a compact feature vector space. The corresponding

contact force components along the x and y axes were then mapped to these feature vectors using an Artificial Neural Network (ANN) and Support Vector Regression (SVR). In contrast, [1] proposed a deep learning approach based on a ResNet architecture that directly performs regression from input images to output force values. The inputs consist of catheter deflection shapes acquired from an experimental setup, enabling end-to-end learning without explicit hand-crafted feature extraction.

2.1.2 Contributions

In most previously proposed methods, the force component along the z direction has not been explicitly estimated, as the input to the model is limited to a single two-dimensional image of the catheter. Consequently, out-of-plane catheter deflections cannot be reliably captured from a single 2D projection. To address these limitations, this Chapter introduces a novel deep convolutional neural network architecture designed to extract catheter deflection features from two distinct viewing orientations. By leveraging complementary information from multiple perspectives, the proposed framework enables the estimation of contact forces in three dimensions. The main contributions of this work are summarized as follows:

Inspired by twin-network architectures (e.g., Siamese neural networks), the proposed model processes paired images acquired from two different viewpoints of the scene [102]. However, unlike Siamese networks, the proposed architecture consists of two parallel processing branches, each with its own independent set of learnable parameters. Acting as a fusion-based network, the latent representation produced by each branch is combined at a later stage to form a unified feature vector. This fusion mechanism enables the proposed end-to-end architecture to capture variations in catheter deflections across three dimensions.

Consequently, the system is capable of directly mapping each pair of catheter deflection images to the corresponding contact force components along the x , y , and z axes. Notably, the proposed fusion framework estimates forces directly from raw image inputs without requiring additional hand-crafted feature extraction or traditional machine vision preprocessing. Despite its dual-branch design, the network remains lightweight and extensible, and demonstrates robustness under geometric transformations such as rotation and scaling.

As with most learning-based approaches, the proposed method is data-intensive, with performance and generalization improving as the size and diversity of the training dataset increase. Based on extensive experimental evaluations and comparative analysis with existing methods, the proposed approach achieves state-of-the-art performance in three-dimensional catheter force estimation. For clarity and simplicity, the proposed architecture is hereafter referred to as Y-Net, reflecting the characteristic “Y”-shaped structure of its schematic representation.

2.2 Deep Convolutional 3D Force estimator

The proposed deep learning-based framework aims to estimate catheter contact forces along the x , y , and z axes directly from image inputs. Because a single two-dimensional image is insufficient to capture out-of-plane deformations along the z axis, the proposed system employs a stereo vision configuration. Specifically, the model processes a pair of images acquired from two distinct viewing orientations rather than a single view. This multi-view input enables the extraction of catheter deformation features associated with depth variations that are otherwise unobservable in a single 2D projection.

Given the high sensitivity of the catheterization workspace to minute movements and scene variations, Y-Net is designed to operate on segmented catheter shapes as input. This design choice allows the network to focus explicitly on catheter deflections while suppressing interference from unrelated anatomical structures or surgical instruments present in X-ray images. Such segmentation can be achieved using semantic segmentation techniques, including established architectures such as U-Net [103, 104].

Accordingly, the proposed method constitutes a novel end-to-end Deep Convolutional Neural Network (DCNN) comprising two parallel image encoder streams, corresponding to side-view and top-view observations of the catheter. Features extracted from both views are subsequently fused and mapped to the associated three-dimensional contact force components. As the framework follows a learning-based paradigm, model training requires a dataset consisting of paired distal shaft deflection images from two views, together with their corresponding force measurements along the x , y , and z axes.

2.2.1 Experimental Setup

In cardiac catheterization procedures, it is common clinical practice to configure a pair of X-ray emitters—such as those in a biplane fluoroscopy system—to visualize the patient’s chest and monitor the catheter within the body from two distinct points of view. Real-time visualization from multiple viewing angles enables surgeons to develop a three-dimensional understanding of the catheter’s spatial position and its interaction with cardiac structures.

In advanced clinical systems, such as those reported in [105] and [106], three-dimensional reconstruction and mapping techniques are employed to provide enhanced visualization of the heart chambers and catheter trajectory. The proposed Y-Net framework is inspired by this multi-view imaging paradigm and leverages analogous visual information to enable three-dimensional force estimation. Accordingly, the experimental setup is equipped with a dual-view vision system designed to emulate a biplane fluoroscopy device. This setup represents an enhanced version of a previously developed platform by Khodashenas et al. [63] and is capable of simulating cardiac motion through controlled heart-beating dynamics.

As illustrated in Fig. 2.1, a mechanical experimental setup was developed to acquire catheter distal shaft images from two distinct points of view along with the corresponding contact force measurements. The setup comprises a custom-designed two-degree-of-freedom (2-DOF) motorized linear actuator, two cameras (Logitech C920, 640×480 resolution), a bi-directional ablation catheter (Boston Scientific Blazer™ II XP), a six-axis force sensor (ATI Mini40), a data acquisition unit (National Instruments USB-6210), a microcontroller (Arduino UNO), a stepper motor driver (HANPOSE TB660), a plastic sheath serving as a vascular phantom, and custom 3D-printed components. As shown in Fig. 2.1 (a) and (b), the ablation catheter consists of two primary segments: a flexible distal shaft that can be actuated and a stiff proximal shaft that remains uncontrollable. In clinical practice, the distal shaft curvature is adjusted by the surgeon using a control knob, as depicted in Fig. 2.1 (a). To reduce variability and simplify data acquisition in the present study, the knob position was fixed throughout the experiments. The catheter was introduced through a plastic sheath simulating a human vein, while the distal shaft was maintained in contact with the force sensor to record applied forces. Two cameras were positioned above and beside the distal shaft,

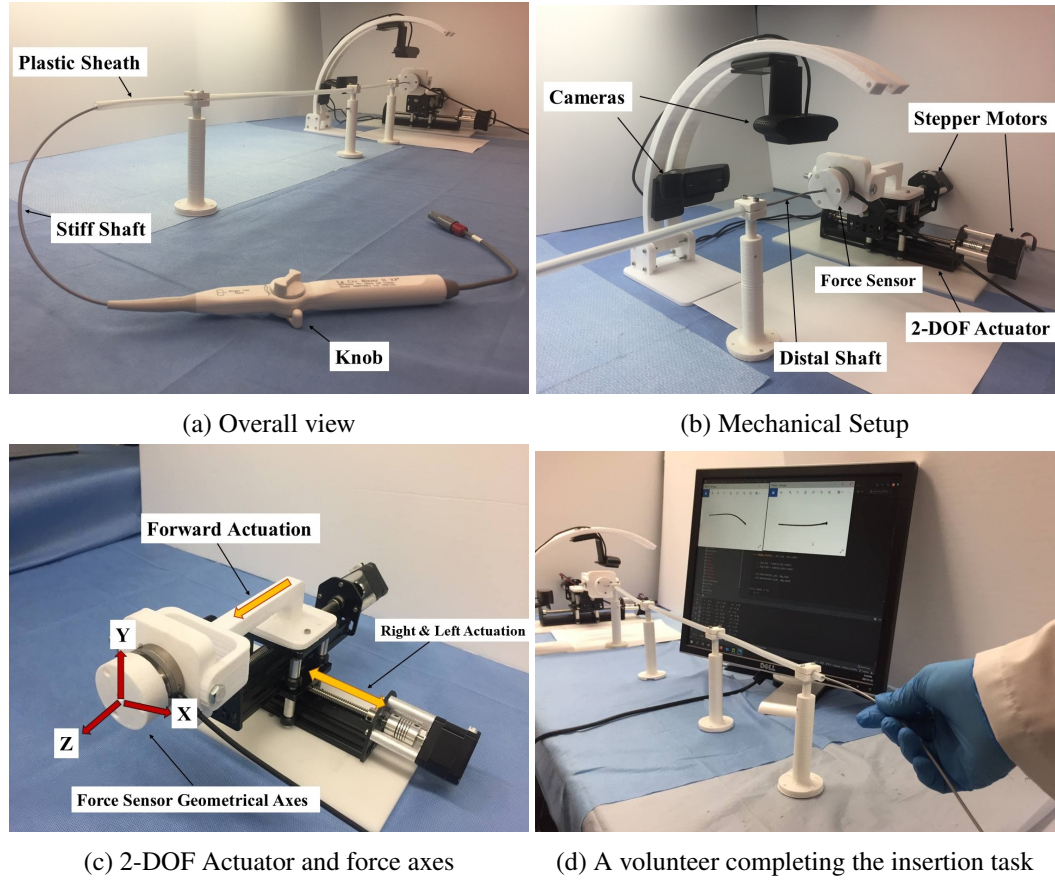


Figure 2.1: The designed experimental setup is shown in 4 subplots. This setup is equipped with a regular catheter, two cameras, two linear actuators and a force sensor to bend the catheter's distal shaft in different shapes.

emulating a biplane fluoroscopy configuration, to capture catheter deflections simultaneously from orthogonal viewpoints.

Using the aforementioned experimental setup, the catheter tip was pressed against the surface of the force sensor to generate a wide range of distal shaft deflection configurations, while maintaining contact forces within the range of 0–0.3 N along all three axes. The resulting bending shapes and their corresponding force measurements were recorded synchronously using the cameras and the force sensor, respectively. The catheter motion was governed by two linear actuators operating within a discrete movement space, as illustrated in Fig. 2.1(c). Specifically, the catheter displacement was discretized into incremental steps of 5 mm. At the start of each experiment, both actuators were initialized at a centered position. For each forward insertion step, lateral displacements were performed sequentially toward the left and right directions over 19 discrete steps. Beginning from

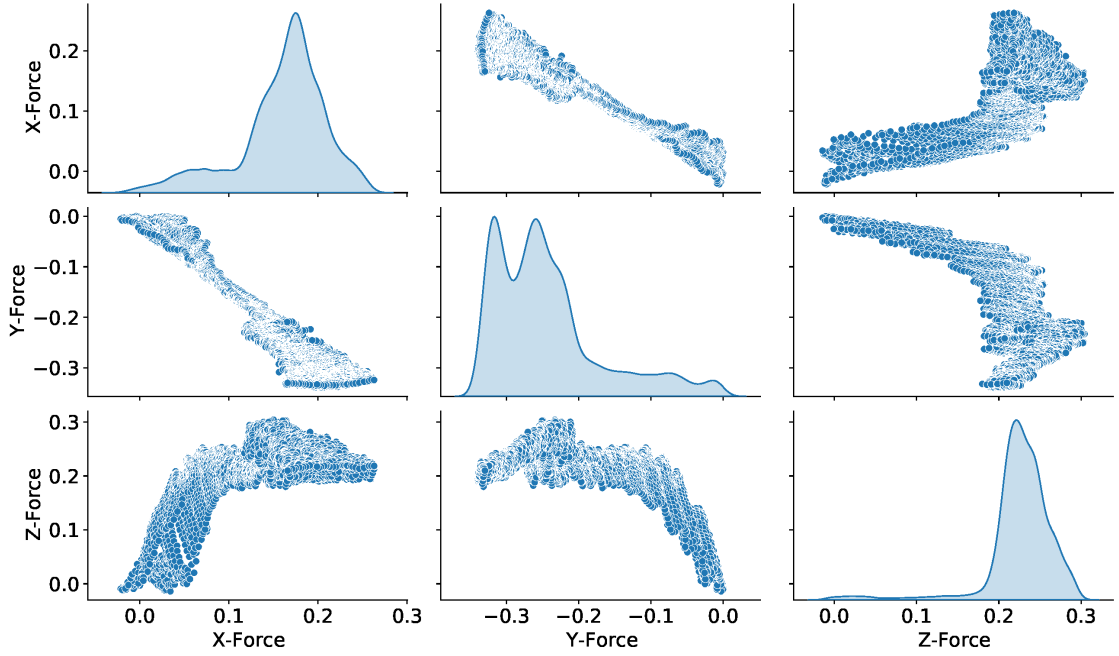


Figure 2.2: Joint distribution of the recorded contact force components along the x , y , and z axes. The diagonal subplots show the marginal histograms of each force component, while the off-diagonal scatter plots depict the pairwise relationships between the force components.

the centered position, the catheter was first displaced laterally to the left over 19 steps and subsequently returned to the center. The same procedure was then repeated toward the right direction. This lateral scanning process was executed at each forward insertion step until the catheter tip remained in contact with the force sensor throughout the full range of motion.

At each discrete position, control commands were transmitted from a Python-based controller to the Arduino microcontroller within approximately 500 ms, after which the actuators repositioned the catheter in an additional 500 ms. To ensure force measurement stability and suppress mechanical vibrations, the system paused for 2000 ms before data acquisition. Finally, synchronized image and force data were recorded over a 500 ms interval. In addition to the automated data collection process, Fig. 2.1(d) illustrates a manual experimental trial conducted using the same setup, which is further discussed in Section 2.3.2.

2.2.2 Dataset Compilation

Fig. 2.2 illustrates the joint distributions of the recorded contact forces using a pairwise scatter plot, providing insight into the variability of distal shaft deflections observed during the experiments. The diagonal subplots depict the marginal histograms of the force components along the x , y , and z axes, while the off-diagonal scatter plots illustrate the pairwise relationships between these force components. For example, the second subplot in the first row represents the variation of forces along the x axis with respect to those along the y axis. These distributions reflect the range of forces generated by the actuator-driven catheter motions throughout the data acquisition process. Using the experimental setup and data collection protocol described above, a dataset comprising 19,500 samples was compiled. Each sample consists of a pair of images with a resolution of 640×480 , corresponding to the top and side views of the catheter’s distal shaft. The associated ground-truth labels are the contact force components along the x , y , and z directions for each sample.

2.2.3 Methodology

With the rapid advancement of Deep Learning (DL) techniques, a wide range of application domains—including healthcare and the automotive industry—have benefited from state-of-the-art data-driven solutions. Deep learning represents a subclass of machine learning (ML) methods within the broader field of artificial intelligence (AI) and is primarily characterized by neural network architectures with multiple hierarchical layers capable of learning increasingly abstract representations of data. Among deep learning models, Convolutional Neural Networks (CNNs) have become the dominant paradigm for image-based analysis due to their ability to effectively model spatial hierarchies and local correlations within visual data. Unlike conventional machine learning approaches that rely on hand-crafted feature extraction, CNNs integrate feature learning directly into their network structure, enabling the automatic extraction of high-level representations as data propagates through successive layers.

CNNs have been widely adopted as backbone architectures for various computer vision tasks, including object detection [107, 108, 109], semantic segmentation [103, 104], and image classification and recognition [64, 110, 111]. The primary objective of CNN-based architectures is to

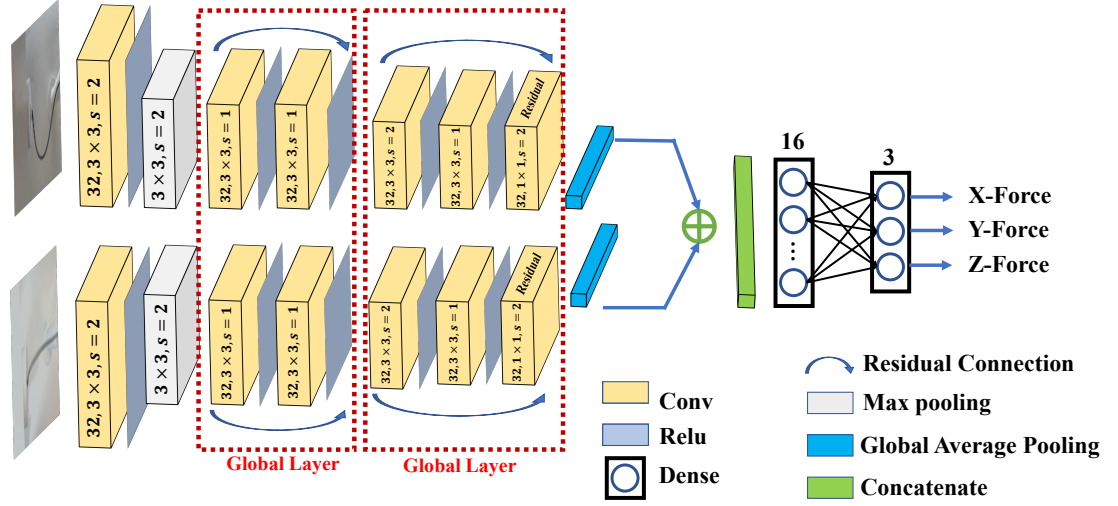


Figure 2.3: Architecture of the proposed Y-Net. The filter sizes and stride values are indicated on the corresponding layer blocks. The network consists of two parallel components that process image streams acquired from the top and side views, respectively. Each component has its own independent set of learnable parameters at every layer.

learn robust and compact feature representations that capture the essential characteristics of the input data while remaining invariant to irrelevant variations. Given that the data considered in this study consist of image-based observations of catheter deflections, the proposed methodology leverages convolutional neural networks as the core feature extraction mechanism for subsequent force estimation.

As discussed earlier, the objective of this Chapter is to estimate the contact forces applied at the tip of an ablation catheter by analyzing catheter deflection patterns observed in image data. Because out-of-plane deformations along the z axis play a critical role in determining the corresponding force component, the input to the problem consists of images acquired from two distinct viewing angles. This stereo vision configuration enables the inference of three-dimensional geometric information that cannot be recovered from a single two-dimensional projection. Accordingly, the force estimation task can be formulated as a regression problem in which a pair of input images is mapped to a continuous three-dimensional force vector.

To address this regression problem, Y-Net is proposed as a deep convolutional neural network that processes catheter images through two parallel input streams. Each stream is dedicated to extracting feature representations from a specific viewing orientation, allowing the network to capture

complementary geometric information. The features learned from both streams are subsequently fused to enable joint reasoning across views and to estimate the corresponding contact forces. Feature fusion architectures have been successfully employed in prior computer vision applications, particularly in the context of multi-sensor and multi-modal data integration [112, 113]. In a similar spirit, Y-Net adopts a feature-level fusion strategy while maintaining a feed-forward, end-to-end design, as the force estimation task does not involve temporal dependencies. As illustrated in Fig. 2.3, the network comprises two parallel branches—referred to as components—each with an independent set of learnable parameters. The formulation of the first layer in each component is described as follows:

$$A_{c,n}^l[h, w] = \sum_i \sum_j f_{c,n}^l[i, j] I_c^{l-1}[h - i, w - j] + b_n^l \quad (2.1)$$

where $A_{c,n}^l$ denotes the feature map produced by the n -th convolutional filter at layer l in component c of the network. Each component corresponds to one input stream (top or bottom view) and maintains an independent set of learnable parameters. The kernel $f_{c,n}^l \in \mathbb{R}^{k_h \times k_w}$ is convolved with the input I_c^{l-1} , and a bias term $b_n^l \in \mathbb{R}$ is added to the result. The output feature map $A_{c,n}^l \in \mathbb{R}^{h' \times w'}$. Afterwards, the outputs are passed through an activation function as follows:

$$\hat{A}_{c,n}^l[h, w] = \text{ReLU}\left(A_{c,n}^l[h, w]\right) = \max\left(0, A_{c,n}^l[h, w]\right) \quad (2.2)$$

where $\hat{A}_{c,n}^l$ denotes the output feature map obtained after applying the rectified linear unit (*ReLU*) activation function to $A_{c,n}^l$. The ReLU activation is applied element-wise at each spatial location (h, w) . The combination of convolutional layers and ReLU activations constitutes the fundamental building block of the proposed architecture. For both components ($c = 1$ and $c = 2$), a max-pooling operation is employed at layer $l = 1$ to reduce the spatial dimensionality of the input feature maps. In subsequent layers, spatial downsampling is achieved using strided convolutional layers rather than additional max-pooling operations. Based on the repetitive stacking of the convolutional and activation layers defined in (2.1) and (2.2), the overall model complexity can be adjusted by adding or removing higher-level layers, as illustrated in Fig. 2.3. To facilitate the training of deeper architectures and to alleviate the vanishing-gradient problem, residual connections are incorporated

[64]. Despite the lightweight design of the proposed model, residual connections are employed in selected layers and are defined as follows:

$$R(\hat{A}_{c,n}^{l-1}, A_{c,n}^l) = \text{ReLU}(\hat{A}_{c,n}^{l-1} + A_{c,n}^l) \quad (2.3)$$

where $R(\cdot)$ denotes the residual mapping function that combines the activated output of the previous layer $\hat{A}_{c,n}^{l-1}$ with the raw output of the current layer $A_{c,n}^l$ through element-wise addition, followed by a ReLU activation. The base graph of Y-Net (Fig. 2.3) employs the residual mapping defined in (2.3) in global layers 1 and 2 of both components ($c = 1$ and $c = 2$). When spatial downsampling is introduced via strided convolution at layer $l = 3$, a 1×1 convolutional projection is used in the residual path to ensure dimensional compatibility between feature maps.

Y-Net fuses the extracted features from both components (i.e., both image streams) using a late fusion strategy. Specifically, the feature maps from both streams are merged after the inputs have traversed all convolutional layers. The output of Y-Net at layer $l = 6$ consists of a set of two-dimensional feature maps. Rather than employing a flattening operation followed by fully connected layers, global average pooling is applied to each feature map to generate a one-dimensional feature vector $v_c \in \mathbb{R}^{n \times 1}$. Each element of the feature vector $v_c[n]$ is computed as follows:

$$v_c[n] = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W \hat{A}_{c,n}^l[h, w] \quad (2.4)$$

where H and W denote the spatial height and width of the feature maps at layer $l = 6$. Global average pooling is applied independently to the feature maps of both components, $c = 1$ and $c = 2$, corresponding to the top and bottom input streams, respectively. Given the resulting feature vectors, the fused feature vector $V \in \mathbb{R}^{2n \times 1}$ is obtained by concatenation:

$$V = v_1 \oplus v_2 \quad (2.5)$$

where $\oplus$ denotes the concatenation operation. The fused feature vector V aggregates complementary information from both image views and is subsequently mapped to the target force space, corresponding to three-dimensional contact forces. Since the force estimation task is formulated

as a regression problem, the fused feature vector V is processed by two successive fully connected layers. The operation of a fully connected layer is defined as:

$$\mathbf{z}^l = \sigma(W^l \mathbf{z}^{l-1} + b^l) \quad (2.6)$$

where $W^l \in \mathbb{R}^{m \times 2n}$ and $b^l \in \mathbb{R}^{m \times 1}$ denote the weight matrix and bias vector, respectively, and σ represents the activation function. For the first fully connected layer, the ReLU activation function is employed (Eq. 2.2), and the input is given by $\mathbf{z}^{l-1} = V$. In the output layer, $m = 3$ and a linear activation function is used, yielding the estimated force vector $\hat{\mathbf{F}} = \mathbf{z}^l \in \mathbb{R}^{3 \times 1}$.

During training, Y-Net minimizes the mean squared error (MSE) loss between the predicted force vector and the ground-truth force vector, defined as:

$$\mathcal{L}(\hat{\mathbf{F}}, \mathbf{T}) = \frac{1}{d} \sum_{k=1}^d (\hat{F}_k - T_k)^2 \quad (2.7)$$

where $\mathbf{T} \in \mathbb{R}^{3 \times 1}$ denotes the ground-truth force vector, $\hat{\mathbf{F}}$ represents the predicted force vector, and $d = 3$. The Root Mean Square Propagation (RMSprop) optimization algorithm is employed to minimize the loss function during training [114].

2.3 Evaluation and Discussion

This section presents the data preparation procedure and the configuration of the proposed model. Specifically, Section 2.3.1 provides a detailed description of the datasets used for training, validation, and testing. In addition, the training and inference configurations of Y-Net are described. Section 2.3.2 then evaluates the performance of Y-Net and compares it with three representative methods from the literature using a benchmark-based quantitative analysis.

2.3.1 Dataset Preparation and Model Configuration

Using the experimental setup described earlier, a dataset comprising 19,500 samples was collected and prepared for training, validation, and testing. Each sample consists of a pair of images capturing the catheter deflection from two viewpoints, along with a three-dimensional force vector

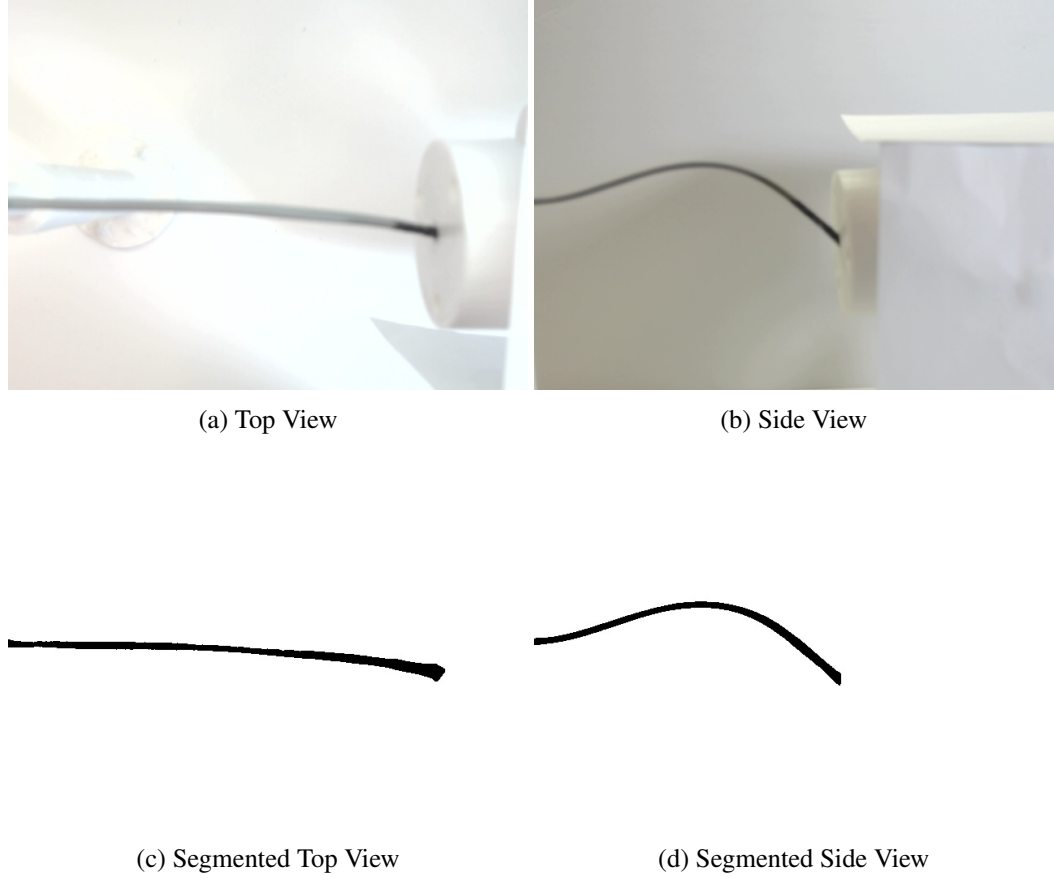


Figure 2.4: A sample of dataset records consisting of paired images acquired from the top and side views of an ablation catheter. The second row shows the corresponding images after the segmentation step performed using a thresholding algorithm.

representing the contact forces along the x , y , and z axes. As discussed previously, Y-Net operates on segmented catheter shapes obtained through a semantic segmentation process.

To emulate this segmentation stage, each input image from both views was first converted to grayscale, after which the catheter body was segmented using a binary thresholding operation. All images in the dataset were then resized to 224×224 pixels (single-channel) and normalized such that pixel intensities were scaled to the range $[0, 1]$. Fig. 2.4 illustrates a representative dataset sample, where the first row shows the raw images from the top and side views, and the second row presents the corresponding segmented outputs. In addition, data augmentation was applied independently to the images from each camera stream in order to improve generalization and evaluate the robustness of the model to camera misalignment. Specifically, each image was randomly rotated within $\pm 20^\circ$ and scaled by up to $\pm 20\%$ during the training and validation/testing phases, simulating

Table 2.1: The statistics of recorded forces (N) 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
z	19500	0.225902	0.041760	-0.01410	0.30412

(a) statistics of the main dataset

Force	Samples	Mean	std	Min	Max
x	200	0.00322	0.00333	-0.00406	0.01086
y	200	0.05842	0.02620	-0.00092	0.10967
z	200	0.22337	0.07950	-0.00902	0.36650

(b) statistics of the manual dataset

potential displacement of the vision system. Since the task is formulated as a regression problem, the force labels were left unchanged and did not undergo additional preprocessing. To provide further insight into the recorded force data, Table 2.1(a) reports the statistical distribution of the three-dimensional force components. The complete dataset was randomly shuffled and partitioned into three subsets. Specifically, 80% of the data (13,650 samples) was allocated to the training set, while the remaining 20% was evenly divided between the validation and test sets, resulting in 2,925 samples per subset.

With the dataset defined, the Y-Net architecture was configured as follows. Each network component receives a single-channel image as input. The input image is first processed by a convolutional layer with 32 filters of size 3×3 and a stride of 2. The resulting feature maps are passed through a ReLU activation function, followed by a max-pooling layer with stride 2 to reduce spatial resolution. Subsequently, the first global layer is introduced, consisting of two convolutional layers (the second and third layers) with 32 filters of size 3×3 and stride 1, each followed by ReLU activation. To enhance feature propagation and mitigate vanishing-gradient effects, a residual connection sums the output of the max-pooling layer with the output of the third convolutional layer [64].

From this point onward, the first convolutional layer of each global block (e.g., the fourth layer) employs a stride of 2, followed by a second convolutional layer (e.g., the fifth layer) with stride 1, while both layers maintain 32 filters of size 3×3 . To ensure dimensional compatibility for residual connections, a 1×1 convolution is applied where necessary to match feature map dimensions. This pattern can be repeated to increase the depth of the network through additional global layers. After feature extraction, the outputs of both components are processed using global average pooling and concatenated. The resulting fused feature vector is then passed through a dense layer with 16 neurons, followed by a final dense layer with three neurons to estimate the three-dimensional contact forces.

To analyze the impact of network depth on performance and generalization, three Y-Net configurations were evaluated, incorporating one, two, and three global layers, respectively. Each configuration was trained and tested under two conditions: with and without data augmentation. The learning rate was set to 0.003, and a batch size of 32 samples was used. All models were trained for 50 epochs, with performance monitored on the validation set after each epoch to enable early stopping and prevent overfitting. Final performance metrics were reported based on evaluation on the test set.

2.3.2 Results and Discussions

In this section, the performance of Y-Net under the configurations described above is evaluated. As a benchmark, Y-Net is compared against three recently proposed methods from the literature addressing image-based catheter force estimation. All trained models were assessed on unseen data, including test sets with and without data augmentation. Performance was quantified using the following evaluation metrics: Mean Absolute Error (*MAE*), Mean Squared Error (*MSE*), Root Mean Squared Error (*RMSE*), the coefficient of determination (R^2), and the ratio between RMSE and the average of the maximum force magnitudes across all directions (*R/M*).

Since the experimental setups employed in the works of Khodashenas *et al.* and Fekri *et al.* are analogous to the setup used in this study, a quantitative benchmark comparison is reported in

Table 2.2: The table reports the performance of the Y-Net comparing with the learning-based methods in the literature.

Method	FD	MSE	MAE	RMSE	R^2	R/M
ANN-based [63]	2	4.49e-05	0.0040	-	0.98	-
SVR-based [63]	2	6.27e-05	0.0046	-	0.98	-
ResNet [1, 64]	2	-	-	0.025	-	0.033
Y-Net (1- l)	3	1.17e-04	0.0085	0.010	0.95	0.052
Y-Net (2- l)	3	3.06e-05	0.0040	0.005	0.98	0.027
Y-Net (2- l + aug)	3	1.5e-04	0.0089	0.012	0.92	0.063
Y-Net (3- l)	3	2.83e-05	0.0039	0.005	0.98	0.026
Y-Net (3- l + aug)	3	1.0e-04	0.0073	0.010	0.95	0.052

Table 2.2 [63, 1]. The results presented in Table 2.2 were obtained from the inference stage of Y-Net under the three previously described configurations, evaluated on unseen test data with and without data augmentation. As the benchmarked methods report performance using different evaluation metrics, some entries in the table are marked with “-”, indicating that the corresponding metrics were not reported in the original studies.

The first and second benchmarked methods were proposed by Khodashenas *et al.*, where an image processing-based feature extraction pipeline is used to feed both an Artificial Neural Network (ANN) and a Support Vector Regression (SVR) model, without employing data augmentation or simulating camera displacement [63]. In contrast, Y-Net performs feature extraction internally as part of its end-to-end deep learning architecture. Furthermore, neither of the image processing-based approaches is capable of estimating the force component along the z direction. As shown in Table 2.2, Y-Net configurations with two and three global layers achieve improved prediction accuracy compared to these methods, despite reporting averaged errors over three force components instead of two. The Y-Net configuration with a single global layer, however, does not demonstrate superior performance, indicating insufficient network capacity for accurate force estimation.

Examining the ***R/M*** metric more closely, although the method proposed by Fekri *et al.* employs a deep learning model based on a ResNet architecture to estimate two-dimensional forces with an average maximum force of 0.73 N, Y-Net significantly outperforms this approach by accurately modeling three-dimensional forces with an average maximum force of 0.3 N. It is worth noting that the benchmarked methods remain incapable of estimating the force component along the *z* axis [1, 63].

The Y-Net configurations with two and three global layers were further evaluated under simulated camera displacement using augmented data. Although the loss values increased under these conditions, the training and validation loss curves shown in Fig. 2.5 indicate an appropriate model fit. In particular, training the three-global-layer Y-Net with data augmentation over additional epochs is expected to further improve performance. As illustrated in Fig. 2.5, the three-global-layer Y-Net exhibits no signs of overfitting in either the augmented or non-augmented settings, with validation loss closely following the training loss over 50 epochs and converging steadily. A similar trend is observed for the two-global-layer configuration without augmentation. In contrast, noticeable fluctuations in the validation loss for the augmented two-global-layer model and the single-global-layer configuration suggest insufficient model capacity for accurate generalization.

In a separate experiment, a manual data acquisition scenario was considered in which the motorized actuators were disabled, and a human subject—acting as a surrogate surgeon—manually manipulated the catheter and applied contact forces to the force sensor within the previously defined force range (Fig. 2.1(d)). During this experiment, force measurements were displayed in real time on a monitor, while catheter motion was restricted to insertion-only maneuvers, with rotational movements omitted. The trained three-global-layer Y-Net (without data augmentation) was used as the inference engine and evaluated on data acquired at a sampling frequency of 2 Hz. The recorded images and force measurements were preprocessed following the procedure described earlier. Table 2.1(b) summarizes the statistical properties of 200 samples collected during this manual experiment. The resulting prediction performance yielded $MSE = 3.56 \times 10^{-5}$, $MAE = 0.0044$, and $RMSE = 0.006$. Although the estimation accuracy remained high, a slight increase in error can be attributed to noise in the force sensor measurements, as the system was not paused at each position to allow mechanical disturbances to settle. Consequently, the reference force values may

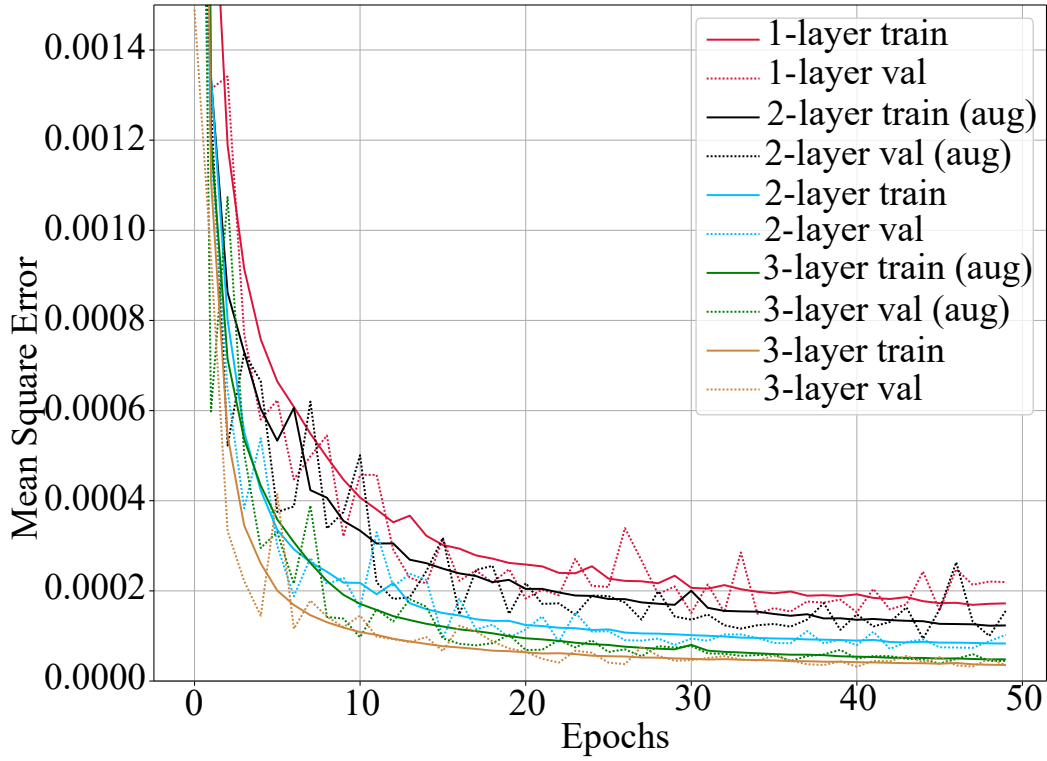


Figure 2.5: The diagram illustrates the training and validation loss curves for the different Y-Net configurations.

themselves contain measurement inaccuracies.

The distribution of prediction errors is illustrated in Fig. 2.6, where errors are computed as the difference between reference forces measured by the sensor and the predicted forces produced by the three-global-layer Y-Net on the test set. The results indicate that errors along the z direction are larger than those along the x and y axes, which may be attributed to camera calibration limitations. A potential solution to this issue involves incorporating a rotational camera mechanism that enables data acquisition from a wide range of viewing angles around the catheter. Such a setup would facilitate fine-tuning of the model to identify camera configurations with maximal discriminative power, thereby improving generalization and reducing sensitivity to camera placement.

Finally, to further illustrate the prediction accuracy of Y-Net, Fig. 2.7 presents representative prediction results obtained from the three-global-layer configuration on a subset of the test data. The three subplots depict the estimated forces and corresponding ground-truth values along the x , y , and z directions, demonstrating close agreement between predicted and reference forces. It is

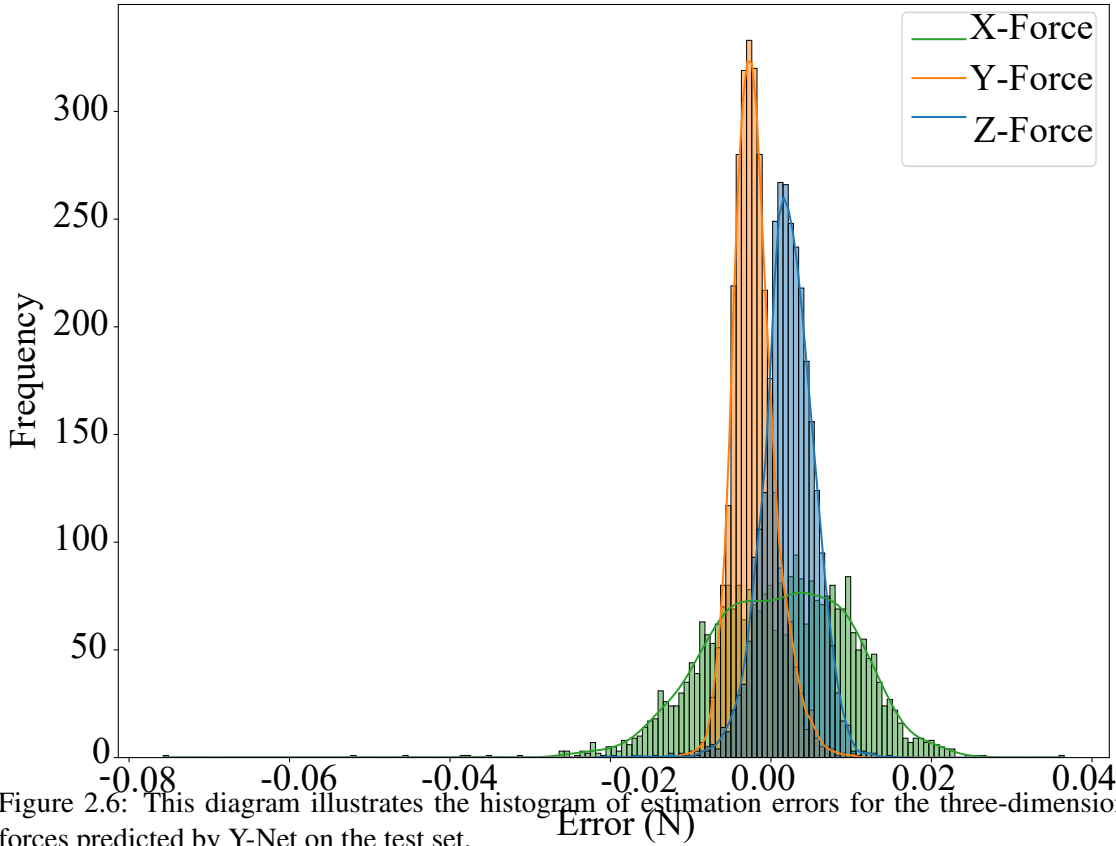


Figure 2.6: This diagram illustrates the histogram of estimation errors for the three-dimensional forces predicted by Y-Net on the test set.

worth noting that the proposed network was implemented in Python using TensorFlow version 2.4.0 [115]. The training phase and all experimental evaluations were conducted on a workstation running Ubuntu 20.04 and equipped with an NVIDIA RTX 2080 GPU. Overall, to the best of the authors' knowledge, this work presents the first dedicated deep learning architecture capable of directly and accurately estimating three-dimensional catheter–tissue contact forces from dual-view image inputs during an ablation procedure.

2.4 Conclusion

In this Chapter, a novel deep convolutional neural network architecture was proposed for estimating catheter–tissue contact forces during cardiac ablation procedures. By receiving a pair of images acquired from the top and side views of the catheter, the proposed method is capable of accurately estimating contact forces along the x , y , and z directions. To support this approach,

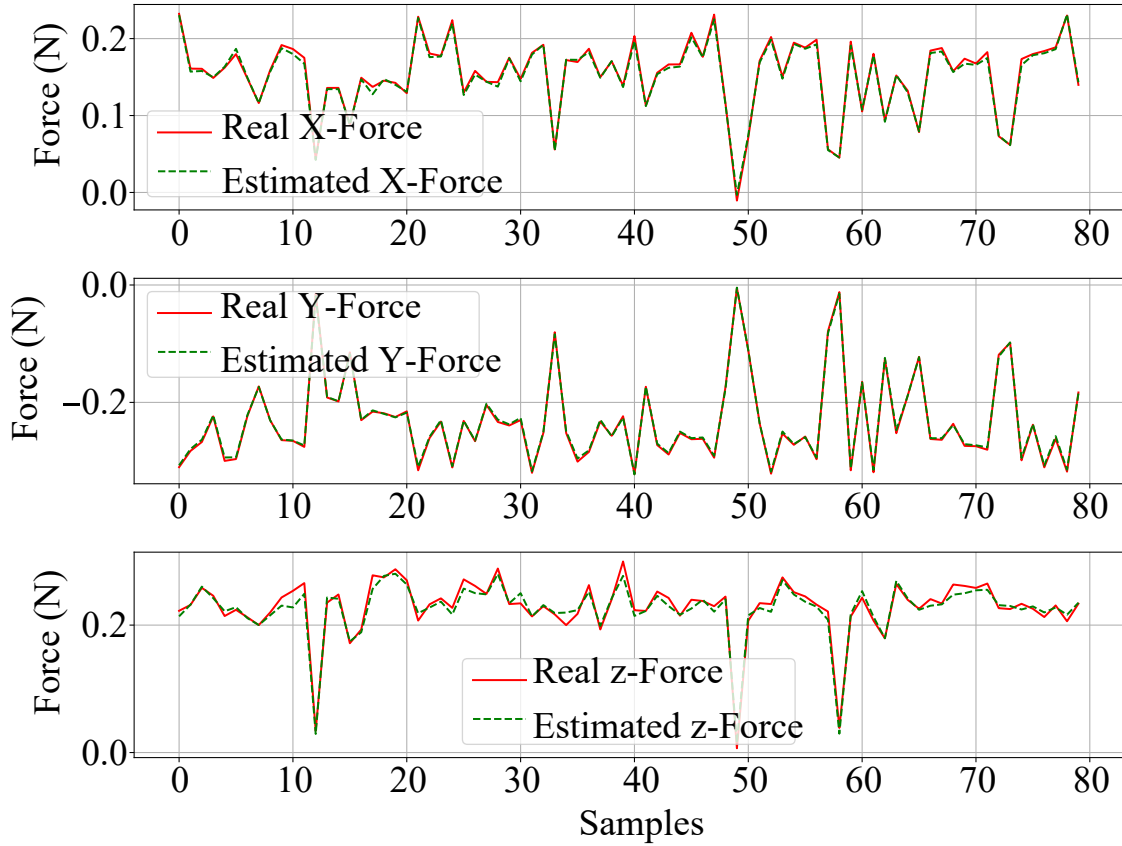


Figure 2.7: The diagram presents the predicted contact forces along the x , y , and z directions for 80 samples from the test set. Each sample consists of a paired set of images capturing the catheter deflection from two views.

a dedicated experimental platform was designed and implemented to realistically simulate an ablation procedure, enabling the collection of a comprehensive dataset for training, validation, and testing. Quantitative evaluations demonstrated that the proposed Y-Net architecture outperforms three recently published methods in the literature, particularly by enabling three-dimensional force estimation in a sensor-free and image-based framework. These results highlight the effectiveness of the proposed dual-view feature fusion strategy and end-to-end learning formulation for accurate force inference.

The Y-Net architecture demonstrates the effectiveness of 3D fusion-based learning for vision-driven contact force estimation by leveraging multi-view visual information. However, this formulation relies on segmented image inputs and therefore assumes access to clean, preprocessed visual

representations, limiting its applicability in scenarios where raw or cluttered images are encountered. Building on this observation, the subsequent chapters extend the problem formulation to a joint learning setting in which catheter segmentation and contact force estimation are addressed simultaneously and directly from image data. In particular, H-Net introduces a convolutional multi-input, multi-output architecture for coupled segmentation and 3D force estimation, followed by TransForSeg, which further advances this paradigm through an attention-based transformer architecture designed to jointly reason over multi-view visual inputs within a unified multi-task framework.

Chapter 3

H-Net: A Multitask Architecture for Simultaneous 3D Force Estimation and Stereo Semantic Segmentation in Intracardiac Catheters

3.1 Introduction

In a cardiac catheterization procedure, a surgeon navigates the cardiovascular system for a range of clinical purposes, including diagnostic assessment, measurement of pressure and oxygen levels, tissue biopsy, or therapeutic intervention. This is accomplished by steering a long, flexible tube—referred to as a catheter—into the patient’s vascular system, typically via femoral, jugular, or subclavian access routes, under the guidance of X-ray fluoroscopy imaging systems [116, 117]. Both biplane and monoplane fluoroscopy systems are utilized in such procedures. The choice of imaging modality directly affects catheter visualization accuracy, depth perception, and radiation exposure, and therefore plays a critical role in both procedural safety and the development of vision-based force estimation methods. Biplane fluoroscopy provides enhanced spatial information

by simultaneously capturing images from two different viewing angles, making it particularly valuable for complex interventions such as neurointerventional procedures and intricate cardiac ablations. However, monoplane fluoroscopy remains the most commonly employed modality in clinical practice due to its lower acquisition and maintenance costs, operational simplicity, and broader accessibility—especially in medical facilities with budgetary constraints. For many standard ablation procedures, the additional spatial information offered by biplane systems is not strictly required. When combined with advanced navigation technologies such as three-dimensional electroanatomical mapping systems, monoplane fluoroscopy can provide sufficient accuracy and clinical effectiveness while significantly reducing radiation exposure. These considerations make monoplane fluoroscopy a practical and efficient choice for a wide range of interventional procedures [118, 119]. Nevertheless, the limited depth perception inherent to monoplane imaging introduces challenges that must be addressed through complementary sensing or estimation strategies.

Despite the well-documented clinical efficacy of interventional catheterization procedures, several limitations and safety challenges persist. These challenges primarily arise from two key factors: the absence of tactile feedback at the catheter tip and difficulties in precise catheter localization within complex vascular environments. Conventional catheters do not provide surgeons with direct tactile feedback regarding interactions between the catheter tip and surrounding anatomical structures, such as myocardial or vascular tissues. As discussed in Chapter 1.3.1, effective ablation requires the application of carefully controlled contact forces—typically in the range of 0.1 N to 0.4 N—at the catheter tip to ensure adequate lesion formation while avoiding excessive tissue damage.

Furthermore, catheter insertion and manipulation are inherently susceptible to unintended motion due to the tortuous geometry of the vascular system and the flexibility of both the catheter shaft and vessel walls. Under limited visualization, such motion increases the risk of vessel wall perforation, endothelial injury, or abrasion, potentially leading to severe or life-threatening hemorrhagic complications. Enhanced visualization and accurate catheter localization are therefore essential for improving navigation precision, maintaining contact stability, and enabling safer force-controlled interactions with cardiac tissue [44, 45]. Accurate catheter visualization also constitutes a fundamental requirement for the development of autonomous or semi-autonomous robotic catheterization

systems. Robotic platforms equipped with reliable force estimation and precise positional awareness have the potential to achieve levels of accuracy and stability beyond those attainable through manual operation, thereby improving procedural consistency and clinical outcomes [77, 82, 101]. Finally, during catheter manipulation, there exists the additional risk of dislodging calcified plaques or thrombotic material from the vessel wall. Once released into the bloodstream, such embolic fragments may migrate to cerebral vessels, increasing the risk of ischemic stroke.

As discussed in Chapter 1.5 and Chapter 1.5.3, the importance of addressing both force estimation and catheter segmentation in the context of cardiac catheterization procedures is twofold. Improving force estimation and catheter segmentation offers two key benefits. First, it provides surgeons with critical information, helping them gauge the pressure applied by the catheter for precise and safe navigation in the vascular system and ensuring accurate positioning of the catheter, as seen in X-ray images. Second, these advancements are crucial for developing autonomous or semi-autonomous robotic systems for catheterization. Such systems, equipped with precise force and position controls, could perform with a level of accuracy and steadiness that enhances the outcomes of catheterization procedures.

3.1.1 Related Work

Regarding the measurement of applied forces, sensor-free approaches reported in the literature aim to reduce the production costs associated with catheters equipped with embedded force sensors (e.g., the TactiCath™ contact force ablation catheter by Abbott, Abbott Park, Illinois, United States). As previously reviewed, a sensor-free force estimation system typically estimates contact forces by analyzing images that depict the deflection of the catheter’s distal shaft, using either model-based or learning-based approaches. In model-based methods, the mechanical properties of the catheter are inferred from the input images by employing techniques such as beam theory models, Cosserat-type rod theories, and multi-body dynamics. However, a major limitation of these model-based solutions lies in the uncertainty of model parameters, which can negatively affect estimation accuracy and robustness. Consequently, these approaches are not the focus of this study [52, 53, 54, 55, 56, 61].

Learning-based approaches employ deep learning architectures to estimate contact forces from catheter deflections. The input to these models may consist of hand-crafted features extracted from

images, simulated data, or the raw images themselves [63, 89, 1, 90]. However, these methods face three major challenges: (1) hand-crafted features lack robustness to variations in imaging conditions [63]; (2) both learning-based and model-based feature extractors require an accurate segmentation of the catheter shape as a preprocessing step [1]; and (3) the models are unable to estimate forces along the z -axis [90]. The latter limitation arises from the fact that these approaches are designed for deployment on monoplane fluoroscopy, where out-of-plane catheter deflections cannot be recovered from a single two-dimensional image. To address this limitation, Y-Net was proposed in Chapter 2 to estimate three-dimensional contact forces by processing a pair of catheter images acquired using biplane fluoroscopy [78]. Although Y-Net overcomes the inability to estimate forces along the z -axis, it still relies on a catheter segmentation module in the preprocessing stage to provide a clean representation of the catheter shape, thereby preventing interference from surrounding anatomical structures or medical instruments during force estimation.

Furthermore, as a separate but closely related task, surgeons require reliable visualization methods to localize the catheter within X-ray images. This is typically achieved using image processing techniques such as thresholding or deep learning-based semantic segmentation models [120, 121, 122, 123, 124, 125]. In the majority of catheter segmentation studies, architectures such as U-Net, FCN, or HR-Net are commonly employed due to their simplicity and effectiveness. Experimental results reported in these works suggest that the data distribution in catheter segmentation is not sufficiently complex to necessitate highly sophisticated models. Nevertheless, it remains worthwhile to investigate alternative convolutional neural networks and open-set vision transformer-based architectures developed for semantic and instance segmentation tasks [126, 127, 128, 104, 129]. As discussed, catheter segmentation is a prerequisite for both visualization and force estimation in catheterization procedures. Ideally, two deep learning models would need to operate simultaneously to address these requirements, which in turn imposes significant computational demands.

3.1.2 Contributions

In response to the aforementioned issues, this Chapter presents a comprehensive solution encompassing catheter force estimation and segmentation for biplane fluoroscopy systems. We propose a novel multi-input, multi-output and multi-task architecture equipped with two parallel sub-networks, each consisting of an encoder and a decoder with the following contributions: This end-to-end architecture is designed to process two X-ray images of a catheter from two angles simultaneously. By fusing these images, the network can segment the catheter from both perspectives simultaneously through two separate segmentation heads, enabling 3D reconstruction and visualization. Leveraging the feature maps of both decoders to capture the catheter’s 3D variation, the force estimation head predicts the applied force at the catheter’s tip along the x and y axes, as well as the force along the z axis. Addressing the problem of catheter segmentation in the preprocessing phase, this approach is designed to ensure that the estimation head primarily focuses on the catheter, minimizing the influence of other object variations in the X-ray images.

Although the proposed architecture segments two X-ray images and estimates forces in 3D simultaneously, the network remains lightweight by sharing parameters between its sub-networks. This balances computational complexity, memory usage, and latency. Additionally, to assess the model’s performance on a dataset that closely mimics real-world conditions, we have designed a synthetic X-ray image generator. This generator superimposes images of actual catheter deflections onto a chest X-ray background, effectively simulating a catheterization procedure as observed under fluoroscopy. Leveraging this capability, we compiled two separate synthetic X-ray datasets, each with a different level of difficulty. The proposed model was trained and tested on these two synthetic datasets as well as an additional RGB dataset. Compared to other pioneering methods in the literature, for the first time, the proposed method accomplishes 3D force estimation and stereo segmentation tasks in an end-to-end architecture with superior performance, establishing it as the state-of-the-art.

3.2 Multi-Task Encoder-Decoder CNN

As reviewed in the previous section, the proposed architecture addresses the key limitations of existing learning-based catheter force estimation and segmentation methods through the following design choices:

- (1) It directly estimates contact forces from raw X-ray images, eliminating the need for a separate catheter segmentation engine as a preprocessing step.
- (2) It is capable of estimating applied forces along the x , y , and z directions.
- (3) It provides stereo catheter segmentation by processing biplane fluoroscopic images.
- (4) Designed as an end-to-end framework, it achieves these objectives without significantly increasing the overall computational complexity.

Inspired by both U-Net and Y-Net architectures, the proposed network processes two raw X-ray images through two parallel encoders with shared parameters [78, 104]. The feature maps extracted from both inputs are passed through corresponding decoders and subsequently routed to two segmentation heads and a shared force estimation head. The network outputs the segmented catheter shapes from two perspectives, along with the predicted contact forces along the x , y , and z axes. Architecturally, the model consists of two parallel sub-networks, each with its own input and output head, interconnected via a central force estimation module. This overall structure resembles the letter “H”; therefore, the proposed architecture is referred to as *H-Net* [79].

Moreover, FS-Net was developed as a direct extension of H-Net to address the constraints of monoplane fluoroscopy, which remains the most widely adopted imaging modality in clinical practice [83]. FS-Net (Force Segmentation Network) simplifies the original architecture by operating on a single X-ray image, focusing on two-dimensional force estimation and single-view catheter segmentation. The network employs a streamlined encoder–decoder design with one segmentation head and one force estimation head. By eliminating the need for dual-view inputs and three-dimensional reconstruction, FS-Net achieves improved computational efficiency while remaining suitable for real-time deployment in standard clinical environments. Despite this simplification, FS-Net preserves the advantages of an end-to-end learning framework by jointly performing force estimation

and catheter segmentation.

Consequently, the proposed H-Net and FS-Net architectures provide a unified learning-based solution capable of simultaneously addressing force estimation and catheter visualization in both biplane and monoplane fluoroscopy settings, while achieving performance comparable to or exceeding that of existing state-of-the-art methods [79, 83].

3.2.1 Synthetic Data Generation and Data Preparation

As noted earlier, H-Net is provided with two images of a catheter captured from two different viewing angles. Similar to Y-Net, this stereo vision configuration enables the network to infer depth information and three-dimensional deflection patterns of the catheter, thereby allowing force estimation along the z -axis. The data required to capture these characteristics were generated using a mechanical experimental setup, as described in Chapter 2.2.1 [78]. This setup emulates a biplane fluoroscopy system by employing two cameras positioned to observe the distal shaft of the catheter from two perspectives. The catheter tip is actuated by two motorized linear stages that press it against a force sensor, enabling controlled motion in multiple directions. Each data record produced by this setup consists of two images along with the corresponding measured forces along the x , y , and z axes. Since the Y-Net architecture operates on segmented catheter shapes, the background is removed using a thresholding-based segmentation method. Consequently, both segmented catheter images and the corresponding original RGB images are available from the previous work.

This dataset was subsequently modified in order to ensure compatibility with FS-Net by utilizing only a single RGB image per sample together with force measurements along the x and y axes. In this revised formulation, each RGB image is used as the network input, while the segmented catheter image and the corresponding two-dimensional force components serve as supervision targets during training.

Building upon this version of the dataset, the objective for H-Net is to further modify each data record such that it contains two input images of the catheter acquired from two distinct viewpoints. These images are mapped to their corresponding segmentation masks, along with the associated three-dimensional force measurements. Under this configuration, H-Net is trained and evaluated using two different types of input images.

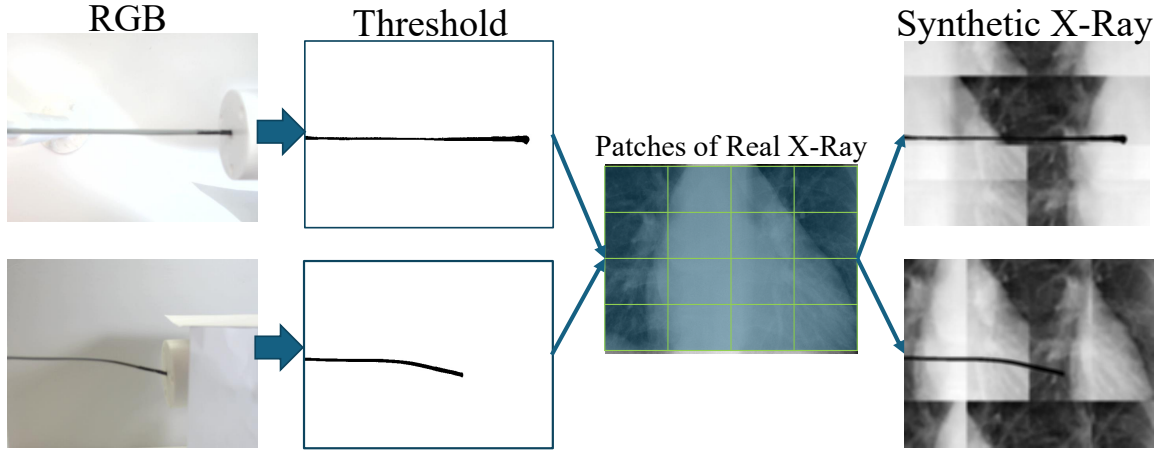


Figure 3.1: Overview of the synthetic X-ray image generation process using RGB images and their corresponding thresholded catheter masks.

The second type of the dataset for H-Net comprises synthetically generated X-ray images. Given that the ultimate goal is to deploy the proposed model in real operating rooms under fluoroscopic imaging, it is essential for H-Net to be exposed to data that closely resemble real X-ray scenarios. This strategy also facilitates a more informative comparison with RGB images, which typically exhibit less complex backgrounds than X-ray images. To this end, a synthetic data generation pipeline is proposed to convert the previously described dataset into synthetic X-ray images. First, a region containing the aorta, superior vena cava, and cardiac structures is cropped from a real chest X-ray image and selected as the background (see the background image under the grid in Fig. 3.1). These anatomical regions generally exhibit higher pixel intensities, resulting in relatively smooth background areas. However, radiolucent regions, which appear darker, may interfere with catheter visibility and thus create more challenging background conditions. Within the selected background region, a grid of 4×4 cells is defined, partitioning the image into 16 non-overlapping patches (refer to the patchifier in Fig. 3.1). These patches are later used to synthesize diverse background images.

For each dataset record, the catheter foreground is extracted from the thresholded (segmented) images. A blank background matrix with the same spatial dimensions as the RGB images is then initialized and divided according to the predefined grid layout. For each grid cell, one patch selected from the patchifier is randomly placed to construct a synthetic background. Finally, the extracted catheter foreground is superimposed onto the generated background, producing a synthetic X-ray

image, as illustrated in Fig. 3.1. This process is applied to all samples in the dataset. For both the synthetic X-ray dataset and the RGB dataset, the thresholded catheter images are used as segmentation ground truth during training.

3.2.2 Methodology

The tip of a catheter generates forces when it is pressed against the surface of lumens. These forces are correlated with the deflection shapes of the distal shaft, particularly when the catheter tendons are not engaged. In clinical practice, the catheter is visualized using continuous X-ray images provided by a fluoroscopy imaging system. Bearing this in mind, the objective is to segment the catheter and estimate the forces generated at the catheter tip directly from these X-ray images. This problem presents the following challenges: 1) the distal shaft of the catheter must be considered in three dimensions, as it can be deflected along all spatial axes; 2) other objects present in X-ray images, such as anatomical structures, may interfere with the visualization of catheter deflection shapes, potentially leading to inaccurate force estimation; and 3) in addition to force estimation, accurate catheter visualization is required to assist surgeons in precise measurement and steering. As mentioned earlier, these challenges are interdependent but have been addressed independently in prior studies [78, 124, 121, 130]. However, to the best of our knowledge, there currently exists no single network designed to address both force estimation and catheter segmentation simultaneously.

H-Net addresses the aforementioned challenges through a novel end-to-end deep learning architecture. The proposed architecture consists of two encoders and two decoders organized into two parallel sub-networks. As illustrated in Fig. 3.2, H-Net is a multi-input network capable of processing two X-ray images simultaneously, thereby providing stereo vision capability. In addition, it is a multi-output network comprising three distinct output heads. Two of these heads are dedicated to concurrently segmenting the two input X-ray images, each addressing a corresponding segmentation task. Utilizing feature representations extracted from both the bottleneck layers and the decoder stages, the third head is designed to estimate the applied forces along the x , y , and z axes.

To maintain low computational complexity and memory usage, the parameters of the encoder and decoder are shared between the two sub-networks. As a result, the lower sub-network can be

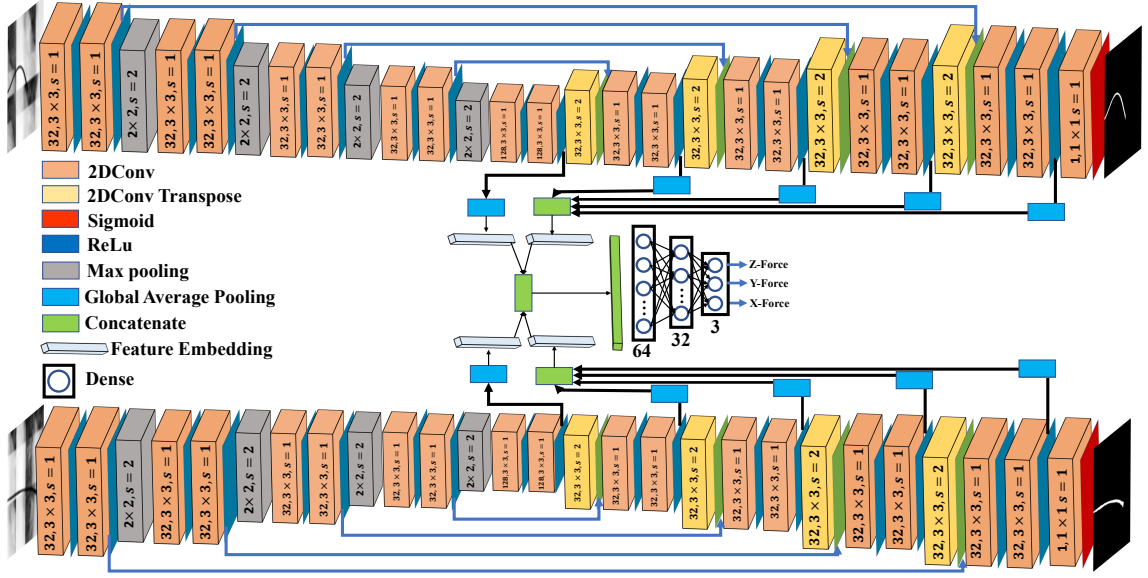


Figure 3.2: Overview of the H-Net architecture. The network consists of two parallel sub-networks with shared encoder–decoder parameters, two segmentation heads, and a central force estimation head. Given two input X-ray images, H-Net simultaneously produces two catheter segmentation maps and estimates the applied forces in three dimensions.

regarded as a mirrored version of the upper one. The inputs to the network are two X-ray images of size $I \in \mathbb{R}^{h \times w \times c}$, each depicting the catheter from a different viewing angle. The encoder of each sub-network consists of four convolutional blocks (b). Within each block, two consecutive two-dimensional convolutional layers with 3×3 kernels and ReLU activation functions are employed. The output of the second convolutional layer in each block is downsampled using a 2×2 max-pooling operation. Through this hierarchical structure, the encoder progressively extracts higher-level feature representations as the input propagates deeper into the network.

As noted previously, FS-Net can be regarded as an extension of H-Net that is specifically designed to meet the requirements of monoplane fluoroscopy [83]. While H-Net is formulated as a general multi-task framework capable of processing stereo (biplane) inputs, FS-Net adapts this design to the single-view setting that is more commonly encountered in clinical practice. To this end, FS-Net preserves the core architectural principles of H-Net while omitting the mirrored lower sub-network. This modification ensures compatibility with monoplane input images while maintaining computational efficiency. In this sense, H-Net serves as the foundational architecture for both models, with FS-Net representing its specialization for monoplane fluoroscopy. Accordingly,

the general structure of H-Net is introduced first, as it provides the basis for both architectures [79].

The output of each encoder block is denoted by $\hat{O}_{enc}^{sn,b}$, where sn indicates the sub-network and b denotes the block index within that sub-network. These feature maps are subsequently passed to the next encoder block, progressing toward the bottleneck layer. In addition, $\hat{O}_{copy}^{sn,b}$, which represents a copy of each encoder block's output prior to downsampling (max-pooling), is forwarded to the corresponding decoder block within the same sub-network. The spatial dimensions $h_b \times w_b$ of $\hat{O}_{copy}^{sn,b}$ match the input dimensions of the associated decoder block.

Within each encoder block of both sub-networks, the input feature maps are downsampled by a factor of two while higher-level features are simultaneously extracted. The encoder consists of four such blocks, after which the bottleneck block is applied. The bottleneck block comprises two two-dimensional convolutional layers with ReLU activation functions. This bottleneck serves two purposes: (1) it provides the input to the first decoder block of each sub-network, and (2) it produces an embedding that forms part of the input to the force estimation head, as described below:

$$v_{btl}^{sn}[k] = \text{GAP}(\hat{O}_{btl}^{sn})[k] = \frac{1}{H_{btl}W_{btl}} \sum_{h=1}^{H_{btl}} \sum_{w=1}^{W_{btl}} \hat{O}_{btl}^{sn}[h, w, k], \quad k = 1, \dots, n. \quad (3.1)$$

Here, (3.1) applies global average pooling channel-wise to the bottleneck feature maps $\hat{O}_{btl}^{sn} \in \mathbb{R}^{H_{btl} \times W_{btl} \times n}$, producing an embedding vector $\vec{v}_{btl}^{sn} \in \mathbb{R}^{n \times 1}$ for sub-network sn . In addition, $\hat{O}_{btl}^{sn}$ is provided as the input to the first block of the decoder. The decoder consists of four blocks, each of which upsamples the input feature maps by a factor of two. In each decoder block b of sub-network sn , a transposed convolution is first applied to increase the spatial resolution of the input feature maps. These upsampled features are then concatenated with $\hat{O}_{copy}^{sn,b}$, which is received from the corresponding encoder block via the skip connection. This concatenation is followed by two successive two-dimensional convolutional layers with 3×3 kernels and ReLU activation functions. The output of decoder block b in sub-network sn , denoted by $\hat{O}_{dec}^{sn,b}$, is produced by the second convolutional layer. In both sub-networks, this output is propagated to the subsequent decoder block and also contributes to the input embedding of the force estimation head, as described below:

$$\vec{v}_{dec}^{b,sn}[n] = GAP(\hat{O}_{dec}^{b,sn}) \quad (3.2)$$

Here (3.2), $\vec{v}$ denotes the embedding of decoder block b in sub-network sn , obtained via the global average pooling operation defined in (3.1). The fourth decoder block in each sub-network reconstructs the feature maps to match the spatial resolution of the input images. Subsequently, the output is passed through a two-dimensional convolutional layer with a single 1×1 kernel followed by a sigmoid activation function. This layer projects the reconstructed feature maps onto the segmentation plane, denoted by $\hat{O}^{sn}$. Consequently, catheter segmentation from the X-ray images is formulated as a classification problem at the output layer of each sub-network and is optimized using the following loss function:

$$L_{seg}^{sn}(\hat{O}^{sn}, t_{sn}) = -\frac{1}{N} \sum_{i=1}^N \left[t_{sn}^i \log(\hat{O}_i^{sn}) + (1 - t_{sn}^i) \log(1 - \hat{O}_i^{sn}) \right]. \quad (3.3)$$

Equation (3.3) corresponds to the binary cross-entropy loss, where $t_{sn} \in \{0, 1\}^N$ denotes the ground-truth label map associated with sub-network sn , $\hat{O}^{sn}$ represents the predicted segmentation map, and N denotes the total number of pixels in the image.

As previously discussed, the regression head receives input embeddings from the bottleneck layers, denoted by $\vec{v}_{btl}^{sn=1}$ and $\vec{v}_{btl}^{sn=2}$, as well as feature vectors provided by each decoder block in both sub-networks, denoted by $\vec{v}_{dec}^{b,sn}$ (see equations (3.1) and (3.2)). To this end, a portion of the final input embedding is constructed as follows:

$$\vec{V}_{sn} = \vec{v}_{btl}^{sn} \oplus [\|_1^b \vec{v}_{dec}^{b,sn}] \quad (3.4)$$

In (3.4), the operator ‘ $\|$ ’ denotes the concatenation of feature vectors generated by decoder blocks $b = 1$ to $b = 4$ (as defined in equation (3.2)) within each sub-network, denoted by sn . The resulting vector is subsequently concatenated ($\oplus$) with the bottleneck embedding $\vec{v}_{btl}^{sn}$. The bottleneck embedding captures the global characteristics of the input image, whereas the decoder embeddings

encode progressively refined spatial features while suppressing structures unrelated to the catheter. This combination enables the force estimation head to focus primarily on variations in the catheter shape, thereby modeling the applied forces while reducing sensitivity to background noise. As a result, a unified feature vector $\vec{V}_{sn}$ is constructed for each sub-network. Given these vectors, the input to the regression head is formed as follows:

$$\vec{V}_{reg} = \vec{V}_{sn=1} \oplus \vec{V}_{sn=2} \quad (3.5)$$

Here, $\vec{V}_{reg}$ incorporates the aforementioned feature representations from two distinct viewpoints, thereby providing the regression head with stereo vision capability. This input vector is passed through two fully connected hidden layers with 64 and 32 units, respectively. ReLU is used as the activation function for all layers except the output layer. The output layer consists of three units with a linear activation function, producing the estimated three-dimensional force vector $\hat{O}^{reg}$. The force estimation head addresses a regression task by minimizing the Mean Squared Error (MSE) loss function, defined as follows:

$$L_{reg}(\hat{O}^{reg}, t_{reg}) = \frac{\sum_{k=1}^d (\hat{O}_k^{reg} - t_{reg}^k)^2}{d} \quad (3.6)$$

Here, t denotes the target force vector, which consists of $d = 3$ elements corresponding to the forces along the x , y , and z directions. As discussed earlier, H-Net is an end-to-end architecture composed of two sub-networks, each equipped with a segmentation head, along with a shared force estimation head for the entire network. Consequently, the overall loss function of the network is defined as follows:

$$L_{total} = \beta_1 L_{seg}^{sn=1} + \beta_2 L_{seg}^{sn=2} + \beta_3 L_{reg} \quad (3.7)$$

In this context, $L_{seg}^{sn=1,2}$ is derived from equation (3.3), and L_{reg} is obtained from equation (3.6). In addition, the parameter β controls the relative contribution of each loss term within the overall objective function. The total loss function, defined in equation (3.7), is optimized using the Root

Mean Squared Propagation (RMSprop) optimizer to update the model parameters. It is worth noting that all layers and their associated parameters in both the encoder and decoder are shared between the two sub-networks, with the objective of maintaining computational complexity and memory usage at manageable levels.

3.3 Evaluation and Discussion

This section presents the data preparation procedures and model configurations used during the training and validation phases. In addition, it establishes a benchmarking framework for evaluating the performance of H-Net and its extension, FS-Net, in comparison with existing methods reported in the literature. The content of this section is organized into two subsections. The first one provides a detailed description of the data preparation process, followed by an explanation of the model configuration for training and inference. The second one subsequently evaluates the performance of the proposed models in both catheter segmentation and force estimation tasks through quantitative comparisons with state-of-the-art approaches.

3.3.1 Dataset Preparation and Model Configuration

H-Net utilizes three types of data, each of which is divided into training, validation, and test sets. The first data type consists of RGB images of a catheter that were previously used in the Y-Net study [63, 1, 78]. The second and third data types comprise synthetic X-ray images. As described, the difficulty level of the synthetic X-ray images is determined by the background regions, which influence the model’s ability to distinguish the catheter from the surrounding structures. Examples of these datasets are illustrated in Fig. 3.3. The XRay-1 dataset contains synthetic images with a lower difficulty level, characterized by relatively uniform intensity distributions in the selected background regions. In contrast, the XRay-2 dataset consists of synthetic images with mixed dark and bright pixel intensities that are randomly distributed across the background grid.

Each of the aforementioned datasets contains 19,500 samples. Each sample consists of two images with dimensions of $224 \times 224 \times 3$, which are normalized prior to being fed into the network. The targets for the segmentation head include two segmentation maps with dimensions of $224 \times$

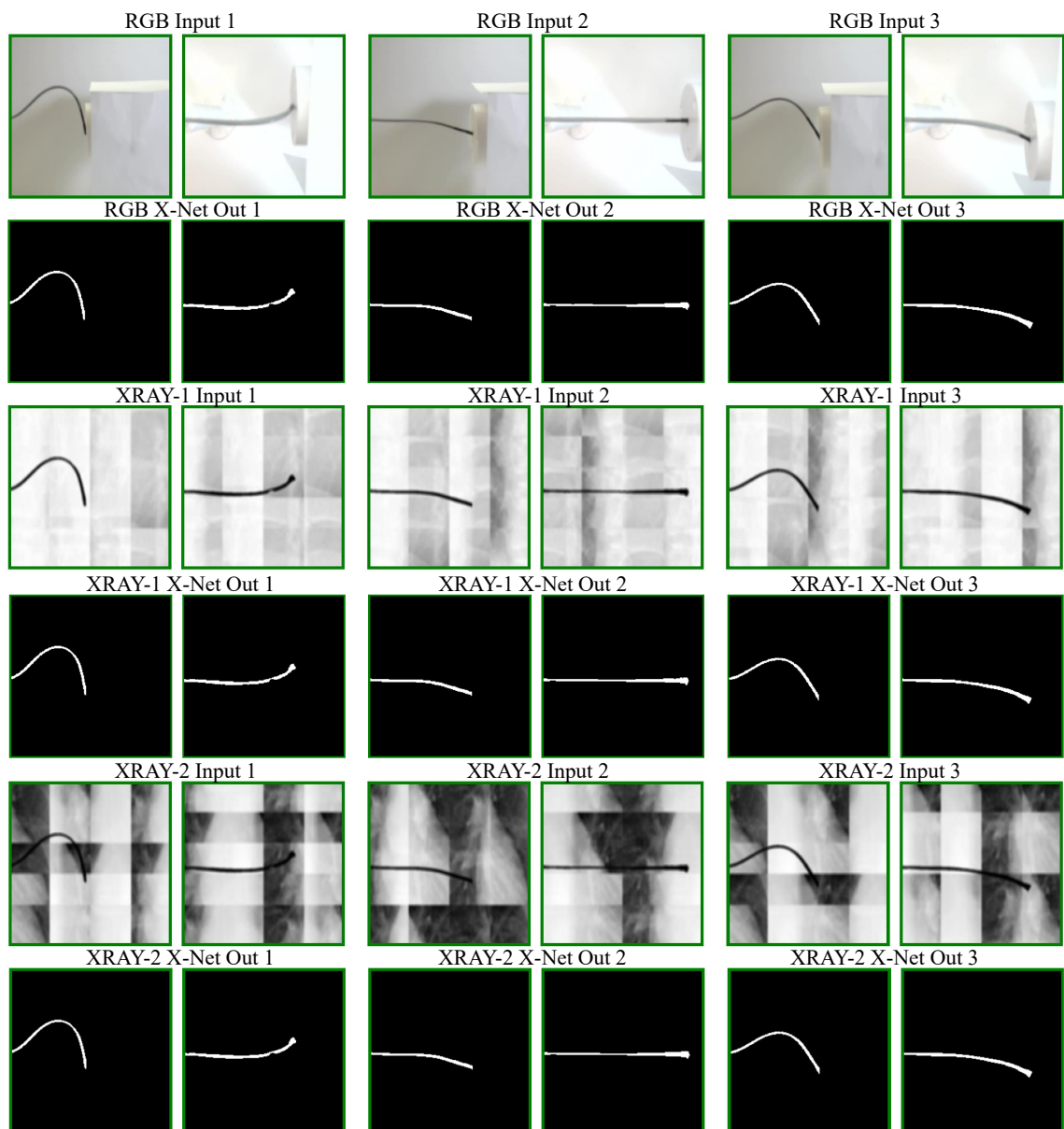


Figure 3.3: The diagram demonstrates two samples from RGB, synthetic XRay-1 and XRay-2 dataset. The outputs of H-Net for each sample are shown as well. For each dataset, H-Net was already trained on its train set.

224×1 , where each pixel is labeled as either catheter or background. For the force estimation head, the target associated with each sample is a three-dimensional force vector corresponding to the forces along the x , y , and z axes. Each dataset is randomly shuffled and partitioned into training, validation, and test sets. The training set comprises 80% of the data (13,650 samples), while the remaining 20% is evenly split between the validation and test sets, each containing 2,925 samples.

FS-Net, in contrast, relies exclusively on RGB images acquired from a single viewpoint. The primary objective of this extension is to demonstrate the model’s ability to adapt its architecture and training strategy while still achieving accurate force estimation and catheter segmentation. Unlike H-Net, FS-Net has not been subjected to the same level of extensive testing and validation, as the focus in this work is on establishing feasibility and illustrating its potential rather than providing a comprehensive performance assessment.

H-Net processes two input images, each of which is fed into the encoder of its corresponding sub-network. Each encoder consists of four blocks, where each block contains two successive two-dimensional convolutional layers with ReLU activation functions, followed by a max-pooling layer. The convolutional layers employ 32 filters of size 3×3 with a stride of 1. The max-pooling layers use a kernel size of 2×2 with a stride of 2. The feature maps produced by the encoders are subsequently processed by the bottleneck blocks in both sub-networks. These bottleneck blocks share a similar structure with the encoder blocks but do not include max-pooling layers.

In each sub-network of H-Net, the decoder receives the corresponding bottleneck output and upsamples the feature maps through four decoder blocks to generate segmentation maps for each input image at the segmentation heads. Each decoder block consists of a transposed convolutional layer followed by two standard convolutional layers. Both the transposed convolutional and convolutional layers use 32 filters of size 3×3 . The convolutional layers employ a stride of 1, while the transposed convolutional layers use a stride of 2. The final decoder block serves as the classification head and includes an additional two-dimensional convolutional layer with a single 1×1 filter, which produces the final segmentation map.

As described before, H-Net estimates the applied forces using a regression head that receives input embeddings from both bottleneck layers as well as from the decoder blocks of each sub-network. The regression head processes these embeddings through three fully connected layers

with 64, 32, and 3 units, respectively. The first two layers use ReLU activation functions, while the final layer employs a linear activation function to output the estimated forces. The loss weights associated with each head (β) are set to equal values. The configured H-Net is trained on the compiled training set using a batch size of 32, a learning rate of 1×10^{-4} , and a total of 100 epochs. To mitigate overfitting, the model’s performance is evaluated on the validation set at the end of each epoch, enabling the use of an early-stopping strategy.

As discussed previously, FS-Net can be regarded as a simplified extension of H-Net, in which the overall architecture and training strategy remain unchanged, but only a single sub-network is employed. Unlike H-Net, which processes two input images through parallel encoder–decoder pathways and fuses their representations within the regression head, FS-Net removes the lower sub-network and operates on a single input image. The encoder, bottleneck, and decoder of the retained sub-network preserve the same configuration of convolutional, pooling, and upsampling layers as those used in H-Net. The regression head also remains unchanged, except that it receives embeddings from only one bottleneck–decoder stream. In this manner, FS-Net demonstrates how the dual-stream architecture of H-Net can be reduced to a single-stream variant while still enabling end-to-end training for simultaneous catheter segmentation and force estimation.

3.3.2 Results and Discussions

Given that H-Net produces two catheter segmentation maps from two different viewing angles, in addition to estimating three-dimensional contact forces, the performance of these two tasks is evaluated separately. To assess H-Net’s capability in three-dimensional force estimation, the model is evaluated on the test set corresponding to the dataset on which it was trained. As previously described, H-Net is trained using three distinct datasets: RGB images, synthetic XRay-1 images, and synthetic XRay-2 images. Accordingly, consistent with the evaluation protocol adopted in Chapter 2.3.2, the following metrics are employed to quantify force estimation performance: Mean Absolute Error (**MAE**), Mean Squared Error (**MSE**), Root Mean Squared Error (**RMSE**), the coefficient of determination (**R**²), and the ratio between the RMSE (R) and the average of the maximum applied forces (M) across the available directions, denoted as **R/M**. Table 3.1 presents a quantitative comparison of H-Net’s force estimation performance against four state-of-the-art learning-based force

Table 3.1: The table reports the performance of the H-Net comparing with the learning-based methods in the literature.

Model	Inp/FD	MSE	MAE	RMSE	R^2	R/M
IP-MLP[63]	Seg/2	4.4e-05	0.0040	-	0.98	-
IP-SVR[63]	Seg/2	6.2e-05	0.0046	-	0.98	-
ResNet[1]	Seg/2	-	-	0.025	-	0.033
Y-Net [78]	Seg/3	2.8e-05	0.0039	0.005	0.98	0.026
FS-Net [83]	RGB	2.55e-05	0.0036	0.0047	0.99	0.036
H-Net [79]	RGB/3	3.6e-05	0.0039	0.005	0.98	0.026
H-Net [79]	XRay1/3	3.3e-05	0.0039	0.0049	0.98	0.026
H-Net [79]	XRay2/3	4.3e-05	0.0046	0.0058	0.97	0.03

estimation methods reported in the literature, using the aforementioned metrics. All five methods in this benchmark are evaluated on datasets generated using the same experimental setup. However, as discussed, the synthetic datasets used in this study are derived from the base dataset employed in the four reference methods.

The first two methods listed in Table 3.1 employ feature vectors that contain numerical descriptors of the catheter shape. These features are extracted using Image Processing (IP) algorithms applied to a thresholded and clean catheter segmentation obtained from a single-view input image. A MultiLayer Perceptron (MLP) and a Support Vector Regression (SVR) model are then used to map these feature vectors to two-dimensional force measurements [63]. In Table 3.1, the force dimension (FD) indicates whether the output force vector contains components along the x and y axes (FD = 2) or along the x , y , and z axes (FD = 3). In contrast, H-Net directly processes raw X-ray input images and is capable of predicting force components along all three spatial axes. When H-Net is trained on the RGB dataset, its performance in terms of MSE and MAE surpasses that of the IP-based models.

The third method listed in Table 3.1 employs a convolutional neural network based on the

ResNet architecture. This approach eliminates the explicit feature extraction stage by directly processing input images. However, it still requires a pre-segmented catheter shape as input, which necessitates an additional segmentation phase. Moreover, this method estimates forces only in two dimensions. In comparison with H-Net trained on the RGB dataset, the RMSE of the ResNet-based method is approximately five times higher [1].

FS-Net addresses the segmentation dependency observed in the previously discussed benchmark methods by introducing a unified end-to-end architecture that simultaneously performs catheter segmentation and force estimation. Similar to H-Net, FS-Net incorporates both a segmentation head and a force estimation head [83, 79]. Accordingly, FS-Net was trained on the RGB dataset and evaluated on the corresponding test set. When comparing H-Net and FS-Net trained on the same RGB data, H-Net exhibits slightly higher regression errors in terms of MSE, MAE, RMSE, and R^2 . It is important to note, however, that unlike FS-Net, H-Net predicts force vectors in three dimensions, which can reasonably account for the observed marginal increase in total error.

Y-Net, the fourth method reported in Table 3.1 which was introduced in Chapter 2.2.3, is a CNN-based architecture that addresses a major limitation of learning-based force estimation by enabling three-dimensional force prediction [78]. Similar to H-Net, Y-Net processes two catheter images simultaneously, capturing the catheter from two different viewpoints and thereby modeling variations along the z axis. Nevertheless, as mentioned, Y-Net relies on segmented catheter images as input rather than raw X-ray images. Since both H-Net and Y-Net estimate forces in three dimensions, a fair comparison can be established between the two approaches. When comparing H-Net trained on RGB data with Y-Net configured with three convolutional blocks, the results across all evaluation metrics indicate comparable performance.

The impact of synthetic X-ray images on the performance of H-Net was investigated using the XRay-1 and XRay-2 datasets. Compared to the RGB set, the model achieved similar results on the synthetic set with a smoother background (XRay-1). The MAE error of H-Net increased by approximately 17.9% when exposed to the more challenging synthetic XRay-2 dataset, compared to its performance on the XRay-1 and RGB datasets. However, the force estimation remains precise, aligning with state-of-the-art results. Fig. 3.4 displays the estimated forces alongside their corresponding actual forces for 80 samples from the XRay-2 test set, presented in the x , y , and z .

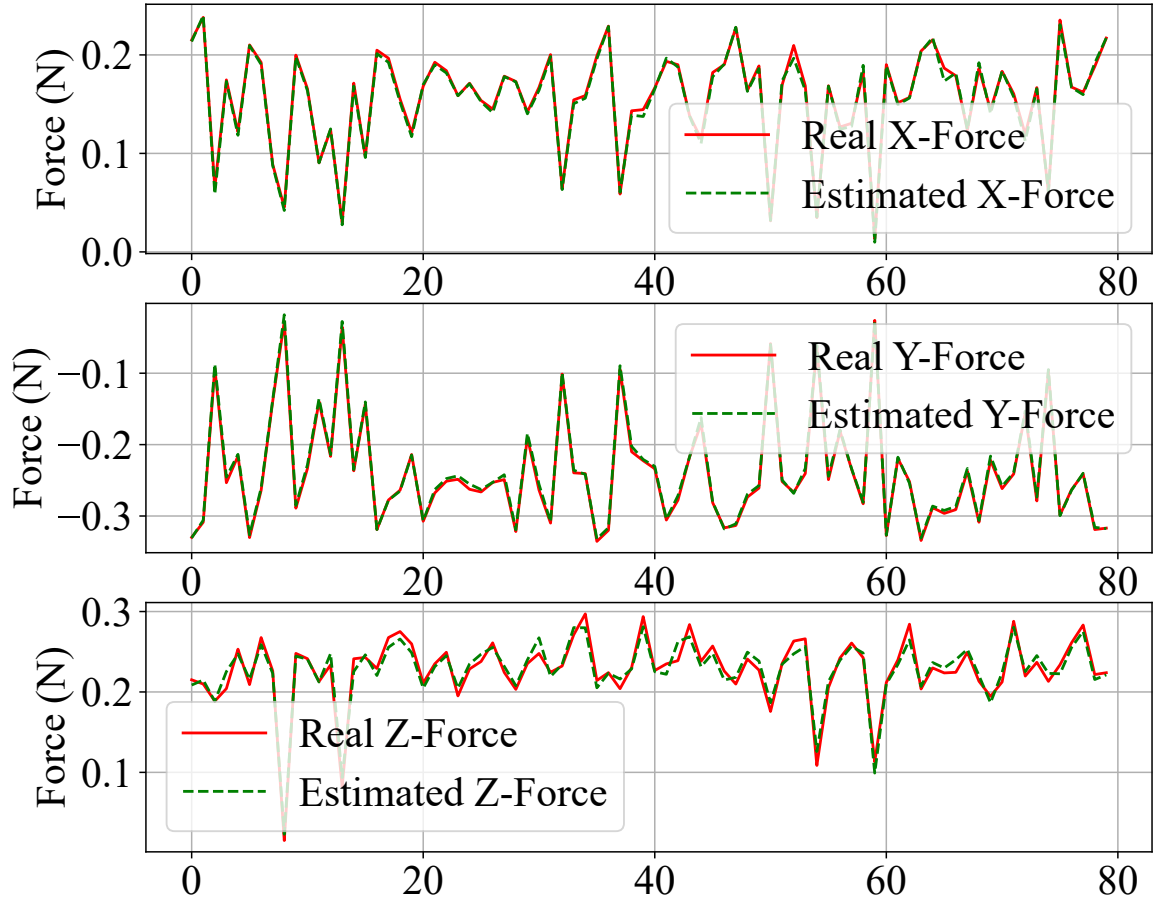


Figure 3.4: The diagram demonstrates the predicted forces in x , y and z , given 80 input samples of XRay-2's test set.

Obviously, every samples contains two images that are inputted into the model during the inference phase. The histogram of force errors for H-Net is depicted in Fig. 3.5. Comparing subplot (a) with subplots (b) and (c), it is noticeable that the error distribution for estimated forces along the z -axis for XRay-2 is more widely spread, while the errors for the other force axes are not as well centered around 0. As suggested in Chapter 2.3.2 for Y-Net [78], since the desired mean error value is above zero, a mechanism can evaluate the model's error at various camera positions relative to the catheter to identify the optimal position with the minimum prediction error. Additionally, compiling data from a rotating camera can also address this issue. With that in mind, H-Net is the only network that processes raw input images and accurately predicts forces in 3D across all datasets. This demonstrates its capability as a force estimator, particularly in extracting variations in the catheter's deflection from noisy images captured from two perspectives.

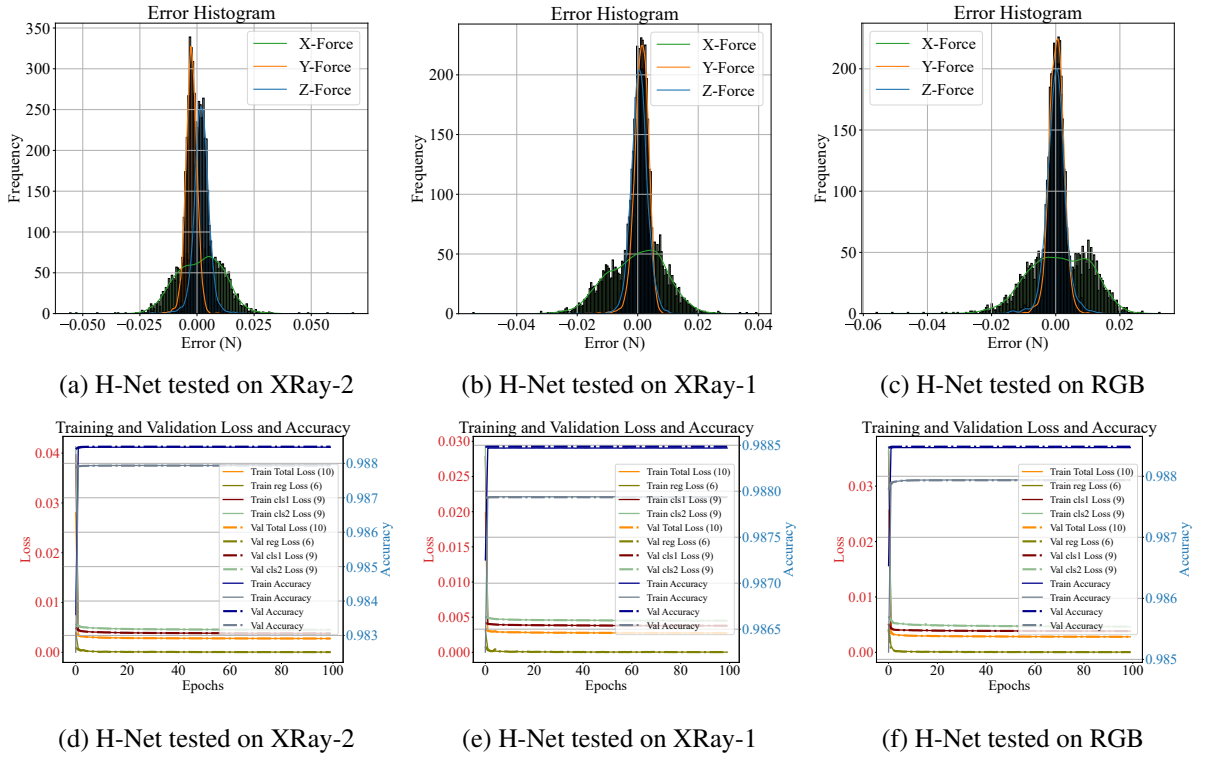


Figure 3.5: Histograms of force estimation errors generated by H-Net are shown for each dataset in subplots (a)–(c). Subplots (d)–(f) depict the corresponding training and validation loss and accuracy curves.

The segmentation heads of H-Net were evaluated separately on the test sets using accuracy (**acc**), mean Intersection over Union (**mIoU**), and mean Dice coefficient (**mDice**). In addition, the number of trainable parameters for each model is reported in millions, serving as an indicator of computational complexity. The segmentation head of FS-Net was also evaluated on the RGB test set. Table 3.2 presents a benchmark comparison between the two segmentation heads of H-Net, the single segmentation head of FS-Net, and three widely used semantic segmentation models in medical imaging applications, namely FCN, U-Net, and HR-Net [126, 125, 104, 130, 131, 124]. Due to computational constraints, the U-Net model included in this benchmark was a pruned variant, in which the number of feature channels was reduced to improve inference efficiency.

In this study, FCN and U-Net were randomly initialized and trained on the synthetic XRay-1 and XRay-2 datasets. In contrast, the HR-Net model was initialized using a pre-trained Cityscapes

Table 3.2: The table reports the performance of the H-Net’s segmentation heads comparing with the literature.

models / (datasets)	#params (m)	acc	mIoU	mDice
FCN (rgb)[126]	134	99.2	94.0	96.8
FCN (xray-1)[126]	134	99.2	94.1	96.8
FCN (xray-2)[126]	134	99.1	93.8	96.7
U-Net (rgb)[104]	1.1	99.8	95.7	98.0
U-Net (xray-1)[104]	1.1	99.7	95.7	98.0
U-Net (xray-2)[104]	1.1	99.7	95.4	97.8
Hr-Net (rgb) [131]	9.636	98.8	96.1	98.4
Hr-Net (xray-1) [131]	9.636	98.8	96.1	98.5
Hr-Net (xray-2) [131]	9.636	98.8	95.8	98.0
FS-Net (rgb) [83]	0.436	99.8	95.3	98.0
H-Net (rgb) [79]	0.460	98.8	95.7	98.0
H-Net (xray-1) [79]	0.460	98.8	95.7	98.0
H-Net (xray-2) [79]	0.460	98.8	95.5	97.8

checkpoint and subsequently fine-tuned on these datasets. Since all benchmarked models, except for H-Net, employ a single segmentation head, they were trained and evaluated using images from both viewing angles through the same head.

Although FCN contains the highest number of trainable parameters among the evaluated models, it consistently yields lower mIoU values across all datasets. HR-Net, which is the second largest model in the benchmark, demonstrates strong localization capability and achieves slightly higher mIoU values than H-Net. However, both H-Net and FS-Net are significantly more compact in terms of parameter count. The marginal improvement in HR-Net’s mIoU performance can be partially attributed to its initialization from a pre-trained Cityscapes model. U-Net, with approximately 1.1

million parameters, achieves mDice and mIoU values comparable to those of H-Net, despite H-Net attaining similar performance with more than two times fewer parameters.

H-Net is the only architecture in the benchmark capable of segmenting two images simultaneously. The mIoU and mDice values reported for H-Net represent the average performance across its two segmentation heads. Despite incorporating two encoders, two decoders, and an additional force estimation head, H-Net achieves the lowest total number of trainable parameters among the compared models due to extensive parameter sharing. Fig. 3.3 illustrates representative outputs from the segmentation heads of H-Net, showing two samples from each of the RGB, XRay-1, and XRay-2 datasets.

It is important to note that accuracy remains nearly constant across all models in the benchmark, whereas mIoU and mDice exhibit noticeable variation. This behavior is primarily due to the highly imbalanced nature of the binary segmentation task, where the number of pixels representing the catheter is substantially smaller than those corresponding to the background. Under such conditions, accuracy becomes a less informative metric. Furthermore, the Dice coefficient generally yields higher values than IoU for a given overlap, owing to its increased sensitivity to the intersection region, particularly when the segmented structures occupy a small area.

Finally, the second row of Fig. 3.5 presents the training and validation loss and accuracy curves of H-Net across all datasets over 100 epochs, demonstrating stable convergence without signs of overfitting. The labels *cls* and *reg* denote the classification (segmentation) and regression (force estimation) heads, respectively. All models included in the benchmark were implemented, trained, and evaluated on a system running Ubuntu 20.04 with an NVIDIA RTX 2080 GPU.

3.4 Conclusion

In this Chapter, we proposed a novel multi-task, multi-input, multi-output encoder-decoder architecture for simultaneous three-dimensional force estimation and semantic segmentation in cardiac ablation procedures. In addition, a synthetic X-ray image generator was developed to construct datasets that closely resemble real fluoroscopic images acquired from a biplane X-ray system. Given two input images capturing the catheter from two different viewing angles, the proposed network

processes them concurrently to segment the catheter in both images and estimate the applied forces along the x , y , and z directions. All tasks are performed within a unified end-to-end framework.

The proposed architecture was trained and evaluated on one RGB dataset and two synthetic X-ray datasets with different levels of background complexity. The performance of both the segmentation and force estimation heads was quantitatively compared against several state-of-the-art learning-based force estimators and semantic segmentation models. The results demonstrate that the proposed method achieves state-of-the-art performance across both tasks. To the best of our knowledge, this is the first end-to-end deep learning architecture capable of simultaneously performing catheter segmentation and three-dimensional force estimation directly from image data.

While H-Net demonstrates the effectiveness of convolutional encoder–decoder architectures for simultaneous catheter segmentation and 3D force estimation, the joint modeling of multiple views, tasks, and output modalities introduces increasingly complex dependencies between visual and force-related features. To more effectively capture these interactions and support scalable multi-input, multi-output learning, the next chapter investigates an attention-based transformer architecture designed for unified multi-task perception and force estimation. This architecture leverages self- and cross-attention mechanisms to jointly reason over stereo visual observations and force-related cues within a single end-to-end framework.

Chapter 4

TransForSeg: A Multitask Stereo ViT for Joint Stereo Segmentation and 3D Force Estimation in Catheterization

4.1 Introduction

As reviewed in Chapter 1.2, surgeons or robotic platforms employ catheters for diagnostic angiography and therapeutic interventions, including hemodynamic assessment, intravascular biopsy, percutaneous coronary intervention (PCI), and targeted drug delivery. To ensure accurate placement, the catheter is introduced through femoral, jugular, or subclavian access points and guided in real time using X-ray fluoroscopic imaging [116, 132]. Surgeons, as well as semi-autonomous and autonomous robotic platforms, require both visual and tactile sensory feedback to safely perform these procedures. However, the majority of these interventions are conducted using standard catheters without integrated force sensors or micro-cameras at the catheter tip, primarily due to cost constraints. In practice, surgeons rely on haptic perception to avoid applying excessive pressure on the vessel lumen, thereby reducing the risk of complications. Simultaneously, visual feedback—particularly catheter localization and segmentation—is essential for precise navigation through the vascular pathway [89, 133, 77].

Motivated by the limitations discussed in Chapter 1.3.1, this chapter aims to address the lack of reliable force sensing and robust catheter visualization in current catheter-based intervention systems. In addition to sensing and perception challenges, the computational complexity of existing methods poses a significant barrier to real-time deployment in both surgeon-assisted and robotic platforms. This chapter therefore focuses on developing methods that jointly tackle force estimation and catheter segmentation while explicitly considering computational efficiency as a core design constraint.

4.1.1 Related Work

To address the challenges associated with tactile and visual information, learning-based deep learning models have been introduced to extract these features directly from X-ray images. The two state-of-the-art models of this type were introduced in Chapter 2 and Chapter 3. These data-driven approaches mitigate the need for physical force sensors by inferring contact forces and catheter positioning directly from imaging data. Deep learning-based methods for contact force estimation can be broadly classified into 2D and 3D force estimators, while visual information is typically obtained through semantic segmentation models, enabling precise catheter localization and tracking [78, 83, 134]. A more recent class of methods integrates both catheter segmentation and force estimation into a unified framework, leveraging a multimodal end-to-end network to extract both visual and tactile information directly from X-ray images. As introduced in H-Net and FS-Net in Chapter 3, these models simultaneously localize the catheter and infer contact forces, eliminating the need for additional hardware sensors or two-stage processing pipelines with separate networks for each task, thereby improving efficiency and real-time applicability [79, 83].

The majority of existing catheter force estimation models rely on CNN-based (Convolutional Neural Network) encoders to capture the relationship between catheter deflections observed from two different viewpoints. These stereo encoders transform spatial pixel information into a latent feature space by progressively expanding the receptive field through image downsampling. Once semantically rich embeddings are extracted, the feature vectors corresponding to the two input views are fused and fed into an MLP regression head to estimate contact forces along the x , y , and z axes. Catheter segmentation models, on the other hand, are predominantly monocular, employing either

CNN-based encoder–decoder architectures or Vision Transformer (ViT)-based encoders for pixel-wise classification [135]. To date, the only existing multimodal architecture that performs both stereo catheter segmentation and force estimation simultaneously is H-Net (Chapter 3) which is based on a CNN-based encoder–decoder framework [79]. Moreover, there is currently no ViT-based architecture specifically designed for stereo catheter segmentation in this application domain. Related uses of Vision Transformers in stereo vision have primarily focused on depth estimation, where a ViT module is employed to estimate disparity from paired images [136, 137]. Consequently, the potential of a multimodal, patch-based architecture leveraging stereo Vision Transformers to simultaneously segment the catheter from two viewpoints and estimate contact forces at the catheter tip remains unexplored. To contextualize this research gap, the following sections review existing approaches to catheter force estimation and catheter segmentation, which form the foundational components of the proposed multimodal framework.

Vision-based force estimation: These approaches estimate the contact forces applied at the catheter tip directly from X-ray images [63, 90]. Specifically, input images are projected into a latent feature space, where catheter deflections are mapped under supervised learning into a vector representing the applied forces. In [1], a ResNet-based encoder is used as a feature extractor to map single-view catheter images into a latent feature space, followed by an MLP regression head that estimates 2D forces. As reviewed in Chapter 2 [78], a Y-shaped architecture is proposed to process a pair of clean images simultaneously, capturing the catheter from two different angles. This stereo processing enables the model to estimate forces not only along the x and y axes but also along the z axis. Although both models successfully project images into the force space, they rely on a separate segmentation model during the data preprocessing stage to extract the catheter shape and generate clean input images. This additional step increases computational cost during training by requiring an extra network to be executed in parallel and negatively impacts the overall efficiency of the pipeline at test time.

Catheter segmentation: Localizing the catheter in X-ray images is essential for enabling surgeons to accurately navigate it through the vascular system and serves as the perception module in semi-autonomous catheterization robotic platforms [138, 139, 140, 141]. This task is commonly

addressed using semantic segmentation techniques, with CNN-based encoder–decoder architectures being the most widely adopted in the literature. In these models, the encoder compresses the input image to extract semantically rich feature maps, while the decoder progressively upsamples these features to reconstruct spatial resolution and generate the final segmentation map [104, 123, 130, 126, 125, 131, 124, 142]. In parallel, transformer-based semantic segmentation models have recently gained significant attention, inspired by the success of attention-based architectures in Natural Language Processing (NLP) [143]. These patch-based models process image patches as token sequences and capture long-range dependencies through self-attention mechanisms. The resulting embeddings—either patch embeddings or mask/object queries depending on the architecture—are subsequently upsampled, typically using CNN-based decoders or interpolation techniques, to generate dense segmentation maps [144, 145, 129, 146]. While early ViT-based segmentation models lacked inherent multi-scale feature extraction, recent architectures have introduced hierarchical or multi-scale representations to better capture features across different spatial resolutions [147, 148].

Multitask architectures: As discussed above, catheter segmentation serves dual roles: as a preprocessing step for vision-based force estimation and as a perception module for catheter navigation. In Chapter 3 [79, 83], a CNN-based encoder–decoder architecture is proposed to address both tasks within a single end-to-end framework. The model receives a pair of X-ray images as input and processes them through a shared CNN-based encoder. The resulting bottleneck feature maps are then passed to an MLP regression head for 3D force estimation and to a convolutional decoder for segmentation map generation. The work presented in this Chapter is inspired by this multitask paradigm and shares the same objectives. However, it addresses the problem using a novel encoder–decoder design based on Vision Transformers, leveraging global attention mechanisms to achieve improved representation capacity and performance.

4.1.2 Contributions

In this Chapter, a novel multitask encoder–decoder Vision Transformer (ViT) architecture is proposed, featuring two segmentation heads and a regression head to enable the simultaneous segmentation of the catheter from two viewpoints and estimation of the applied contact force at the

catheter tip. The model adopts a symmetric ViT-based encoder–decoder design with shared weights, where patch sequences extracted from two X-ray images are processed concurrently. These patches are projected into semantically rich embeddings that capture the global context of both views.

The extracted embeddings from the two viewpoints are progressively upsampled through a shared CNN-based decoder to generate the corresponding segmentation maps. Given that contact forces along the x , y , and z axes can be inferred from catheter deflections observed across two views, a cross-attention mechanism is applied on top of the ViT decoder to capture dependencies between the encoder and decoder embeddings. Finally, the $[\text{CLS}]$ token from the decoder output, after cross-attention, is passed to an MLP regression head to predict the 3D contact force vector.

To ensure computational efficiency and a lightweight design, the ViT decoder shares its weights with the ViT encoder, effectively mirroring its architecture. In addition, the CNN-based upsampling module used for segmentation map reconstruction is shared between the encoder and decoder, further reducing model complexity and overall parameter count.

The main **contributions** of this work are summarized as follows:

- A novel multitask, patch-based encoder–decoder architecture built on Vision Transformers (ViT) is proposed for simultaneous stereo catheter segmentation and 3D contact force estimation from X-ray images.
- The proposed framework estimates contact forces directly from raw X-ray images, with the segmentation task guiding the network to focus on catheter deflection geometry rather than background variations.
- A shared-weight ViT encoder–decoder design, combined with a cross-attention mechanism, enables efficient fusion of information from two viewpoints for accurate 3D force prediction.
- The model processes paired X-ray images and produces three outputs across two modalities: two segmentation maps and a force vector corresponding to contact forces along the x , y , and z axes.
- A shared CNN-based segmentation head is employed to upsample transformer embeddings and generate dense segmentation maps, reducing redundancy and improving computational

efficiency.

- Extensive experimental evaluation and ablation studies demonstrate state-of-the-art performance in sensor-free, learning-based 3D catheter force estimation and catheter segmentation.

As the proposed framework is based on a Transformer architecture for both catheter segmentation and force estimation, it is referred to as *TransForSeg* for brevity.

4.2 Multi-Input Multi-Output Encoder–Decoder Transformer

TransForSeg is a novel encoder–decoder Vision Transformer (ViT) architecture designed for the simultaneous tasks of catheter segmentation and contact force estimation. Unlike H-Net (refers to Chapter 3), which relies on convolutional operations to progressively expand the receptive field, TransForSeg leverages the self-attention mechanism inherent to Transformers to model long-range dependencies both within and across input images. The model processes two X-ray images acquired from different viewpoints, representing each image as a sequence of non-overlapping patches. These patch sequences are embedded and subsequently processed by a shared-weight encoder–decoder Transformer architecture, enabling the modeling of global contextual relationships. A cross-attention mechanism facilitates interactions between the encoder and decoder token representations, allowing information from both viewpoints to be jointly integrated. The embeddings generated by the encoder and decoder are then passed to two shared segmentation heads to produce semantic maps of the catheter in both views. In parallel, a regression head operates on the decoder representations to estimate the applied contact forces at the catheter tip along the x , y , and z axes. By unifying stereo catheter segmentation and 3D force estimation within a single multitask framework, TransForSeg provides an alternative to conventional CNN-based architectures, with an explicit emphasis on global context modeling and multi-view information fusion.

4.2.1 Dataset

In this Chapter, the datasets introduced in Chapter 3.2.1 in the context of the H-Net model [79] are used. In summary, these datasets are based on the Y-Net experimental setup described in Chapter 2.2.1, which was designed to replicate the operation of a biplanar fluoroscopy system [78]. In

this setup, a standard catheter was pressed against a force sensor while two cameras captured its deflection from different viewpoints. This process resulted in a total of 19,500 samples, each consisting of two RGB images acquired from different angles and a corresponding 3D force vector (x, y, z) measured by the force sensor. However, the captured RGB images were not representative of real-world X-ray imagery. To address this limitation, the H-Net framework introduced a synthetic X-ray generator that transformed RGB images into synthetic X-ray images with two levels of background complexity [79]. In addition, two segmentation maps were generated for each sample. As a result, each sample in the extended dataset includes two synthetic X-ray images (XRay1 and XRay2), the original RGB images, a 3D force vector, and two corresponding segmentation maps. For all experiments presented in this Chapter, the same datasets and original data splits are preserved to ensure a fair benchmark comparison. Specifically, 80% of the data (13,650 samples) is used for training, while the remaining 20% is equally divided between the validation and test sets, with 2,925 samples in each subset.

4.2.2 Methodology

TransForSeg is a stereo vision model that processes two X-ray images of a catheter captured from different perspectives. By fusing visual information from both views, the model enhances its ability to capture 3D catheter deflections, leading to more accurate estimation of contact forces along the x , y , and z axes. Furthermore, TransForSeg functions as a multitask transformer, simultaneously segmenting the catheter and producing two segmentation maps corresponding to the input views, alongside a 3D force estimation output. This end-to-end, multi-input multi-output network is built upon an encoder–decoder ViT architecture. Fig. 4.1 shows the architecture of TransForSeg.

Patch Embeddings: During a single forward pass, the top-view X-ray image is divided into non-overlapping patches of size $(n \times n)$. This patch projection is implemented using a 2D convolution with k kernels of size $(n \times n)$ and a stride of n , consistent with the original Vision Transformer (ViT) design [135]. Since TransForSeg does not utilize overlapping patches, the stride matches the patch size. The projection yields a sequence of $p \times p$ patches, where each $n \times n$ patch is mapped to an embedding of dimension k . Unlike some segmentation and detection models, a learnable [CLS] token with the same dimensionality as the patch embeddings is appended to the sequence, resulting

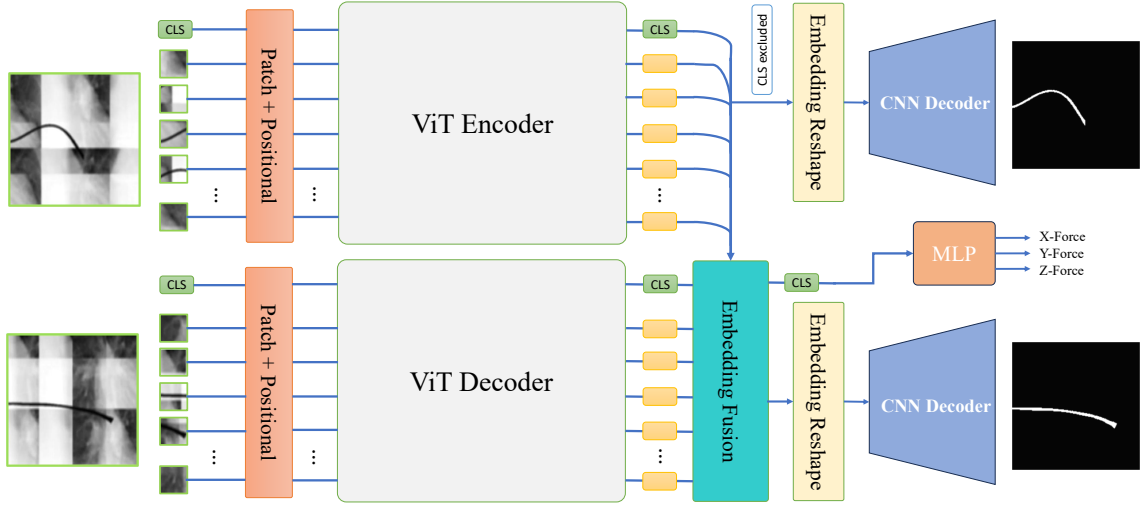


Figure 4.1: Overview of the TransForSeg architecture. The model consists of a Vision Transformer (ViT)-based encoder and decoder with shared weights. Two X-ray images are provided as input, one processed by the encoder and the other by the decoder. Each image is divided into non-overlapping patches, a $[CLS]$ token is appended, and positional embeddings are added to form the input token sequences. Both the encoder and decoder comprise T stacked Transformer blocks, with the final block producing the corresponding output embeddings. The decoder output embeddings are fused with the encoder embeddings through an embedding fusion (cross-attention) layer. The $[CLS]$ token from the decoder is passed to an MLP head for 3D contact force estimation, while the remaining token embeddings from both the encoder and decoder are reshaped and upsampled by a shared CNN-based decoder to generate the catheter segmentation maps.

in a total of $p^2 + 1$ tokens. This $[CLS]$ token is later used by the force estimation head, as described in the following section. Positional embeddings of shape $(p^2 + 1, k)$ are then added to inject spatial information into the sequence. The resulting input, a sequence of $p^2 + 1$ embeddings of dimension k , is passed to the transformer encoder.

ViT Encoder and Decoder: The encoder consists of a stack of T Transformer blocks. Each block begins with element-wise layer normalization applied to the input embeddings, followed by a Multi-Head Self-Attention (MHSA) mechanism with H attention heads. A residual connection is applied between the input tokens and the MHSA output. The result is then passed through a second layer normalization, followed by a Feed-Forward Network (FFN), typically implemented as a two-layer MLP. A second residual connection is applied between the MHSA output and the FFN output, which helps preserve gradient flow and improves training stability. A more detailed technical formulation is provided in the following section.

Unlike ViT-based object detection and segmentation models, the decoder in TransForSeg is fed with the side-view X-ray image. In models such as OWL-ViT, DETR, and SAM, the decoder typically receives object queries as input [149, 150, 129]. In contrast, the decoder in the proposed architecture directly processes patched X-ray images in the same manner as the encoder. The decoder shares the same architectural structure as the encoder, and for computational efficiency, weights are shared between them. The encoder output consists of embeddings corresponding to each input patch, which are then split into two branches. One branch is directed to the segmentation upsampler, while the other is passed to the decoder.

The encoder produces embeddings denoted as $x_t \in \mathbb{R}^{B \times (p^2+1) \times k}$, where B is the batch size, $p^2 + 1$ is the number of tokens (including the [CLS] token), and k is the embedding dimension. Similarly, the decoder produces embeddings from the side-view X-ray patches, denoted as x_s , with the same dimensionality as x_t . In TransForSeg, the decoder attends to the encoder output embeddings using Multi-Head Cross-Attention (MHCA) to capture long-range dependencies between the two X-ray views. This mechanism enables the model to focus on corresponding regions across both images where geometric variations contribute to differences in the estimated contact forces along the x , y , and z axes. More formally, the input embeddings $x \in \{x_t, x_s\}$ are first normalized using layer normalization:

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \quad (4.1)$$

$$norm(x) = \gamma \cdot \hat{x} + \beta \quad (4.2)$$

where μ and σ^2 denote the mean and variance computed across the embedding dimension. The term ε is a small constant (e.g., $\varepsilon = 10^{-6}$) introduced for numerical stability, and the operation in (4.2) represents an affine transformation with learnable parameters γ and β . The attention mechanism operates on the key (K), query (Q), and value (V) representations, while H denotes the number of attention heads. Each head attends to a distinct subspace of the embedding dimension. This is achieved by projecting Q , K , and V into head-specific representations as follows:

$$\begin{aligned}
Q_h &= x \cdot W_q^h + b_q^h \\
V_h &= x \cdot W_v^h + b_v^h \\
K_h &= x \cdot W_k^h + b_k^h
\end{aligned} \tag{4.3}$$

where all projection weight matrices $W \in \mathbb{R}^{k \times k_h}$, and $k_h = k/H$ denotes the dimensionality of each attention head. In (4.3), x denotes the source embedding of the attention operation. For multi-head self-attention (MHSA), $x = x_t$ in the encoder and $x = x_s$ in the decoder. For multi-head cross-attention (MHCA), the query projection is computed from the decoder embedding, i.e., $x = \text{emb}_s$, while the key and value projections are computed from the encoder embedding, i.e., $x = \text{emb}_t$. Given the projected key, query, and value representations, the attention score for each head is computed as follows:

$$A_h = \text{softmax}\left(\frac{Q_h \cdot K_h^T}{\sqrt{k_h}}\right) \cdot V_h \tag{4.4}$$

Next, the attention outputs A_h from all heads are concatenated to form the full attention representation A , which is then passed through a final linear projection followed by a residual connection and layer normalization:

$$A = \text{norm}(x + (A \cdot W_o + b_o)) \tag{4.5}$$

The embedding A is fed into a two-layer MLP, where the first layer expands the embedding dimension by a factor of F , and the second layer reduces it back to the original dimension. Finally, a residual connection is applied between the input embedding and the output of the FFN, followed by layer normalization.

Force Estimation Head: The force estimation head receives the [CLS] token from the decoder, which encodes the aggregated information from both X-ray views. The remaining patch embeddings are discarded, as the regression head does not require spatially resolved visual features. To ensure a fair comparison during benchmarking, the force estimation head adopts the same architecture as in H-Net, as described in Chapter 3.3.1, consisting of a three-layer MLP with 64, 32, and 3 units in the respective layers. The loss function used for force estimation is the mean squared error (MSE),

defined as follows:

$$\mathcal{L}_{\text{force}} = \frac{1}{FD} \|\hat{y} - y\|_2^2 \quad (4.6)$$

where FD denotes the force dimension equals 3 (i.e., along x , y and z).

Segmentation head: Both the ViT encoder and decoder output embeddings $\text{emb} \in \mathbb{R}^{B \times (p^2+1) \times k}$, representing patch tokens along with a [CLS] token for each input image. In the segmentation upsampling subnetwork, the [CLS] tokens are discarded, and the remaining patch embeddings are reshaped into feature maps $f(t, s) \in \mathbb{R}^{p \times p \times k}$, where t and s denote the top and side views, respectively. Several methods have been proposed in the literature for upsampling embedding feature maps, including bilinear interpolation, progressive upsampling using deconvolution, and direct patch-size mapping [144, 147, 146]. In TransForSeg, a deconvolution-based decoder, similar to the one used in H-Net, is employed to ensure consistency in benchmarking [79].

The embedding feature maps are upsampled to the original input image resolution through four stages. First, the embeddings pass through two successive 2D convolutional layers, which reduce the feature dimensionality by a factor of four while preserving the spatial resolution. Next, four upsampling blocks, in contrast to H-Net, each consisting of a deconvolution layer followed by a convolution layer, progressively increase the spatial resolution by a factor of two while reducing the number of feature channels. Finally, a sigmoid activation function is applied to generate a single-channel segmentation map. The segmentation upsampling module is shared between the ViT encoder and decoder and is trained using the Binary Cross Entropy (BCE) loss (3.3), denoted as $\mathcal{L}_{\text{seg}}$.

TransForSeg is an end-to-end multitask model optimized by minimizing the following total loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{force}} + \lambda_2 \mathcal{L}_{\text{seg}}^{\text{enc}} + \lambda_3 \mathcal{L}_{\text{seg}}^{\text{dec}} \quad (4.7)$$

where the coefficients $\lambda_1, \lambda_2, \lambda_3$ weight the individual components of the total loss: the mean squared error (MSE) for force estimation, and the binary cross-entropy (BCE) losses associated with the segmentation outputs from the encoder and decoder, respectively.

4.3 Results and Discussions

4.3.1 Implementation Details

The input to TransForSeg consists of two images of a catheter deflection captured simultaneously from two different viewpoints. For the synthetic X-ray dataset, the input images are denoted as $(x_s, x_t) \in \mathbb{R}^{224 \times 224 \times 1}$, while for the RGB dataset they are represented as $(x_s, x_t) \in \mathbb{R}^{224 \times 224 \times 3}$. TransForSeg is implemented in two configurations: ViT-Tiny and ViT-Small. In the Tiny variant, both the encoder and decoder follow the ViT-Tiny architecture. The input images are divided into patches using a 2D convolution with 192 kernels of size 16×16 , resulting in 197 tokens (including the [CLS] token), each with an embedding dimension of 192. The Transformer consists of 12 blocks, each with 3 attention heads, and the feed-forward network (FFN) expands the hidden dimension to 768. In the Small variant, each patch is projected into an embedding of dimension 384. Each Transformer block contains 6 attention heads, and the FFN expands the embedding dimension to 1536.

The embedding fusion module is implemented as a Transformer block with 8 cross-attention heads and an FFN expansion dimension of 2048. The [CLS] tokens are excluded from both the encoder and decoder outputs to form the fused feature maps $f(t, s) \in \mathbb{R}^{14 \times 14 \times n}$, where $n \in \{192, 384\}$, which serve as the input to the segmentation head. This shared segmentation head first reduces the feature dimensionality to 96 through two successive convolutional layers with 3×3 kernels. It then progressively upsamples the feature maps to the original image resolution of 224×224 using four deconvolution blocks to generate the final segmentation maps. Each block consists of a 2D convolutional layer with 3×3 kernels that reduces the channel dimensionality by a factor of two, followed by a transposed convolution layer with a 4×4 kernel and stride 2, which doubles the spatial resolution.

The force estimation head is implemented as a two-layer MLP that receives the [CLS] token from the ViT decoder as input. Depending on the model configuration, the input embedding has a dimensionality of either 192 (ViT-Tiny) or 384 (ViT-Small), which is projected to intermediate dimensions of 64 and 32, respectively. The output layer predicts the contact force vector as a three-dimensional output. Both ViT variants are pretrained on ImageNet. TransForSeg, in both

configurations, is trained on the RGB, X-ray1, and X-ray2 datasets using a fixed learning rate of 1×10^{-4} without learning rate scheduling, for 250 epochs, and with a batch size of 32 using the Adam optimizer. The entire network remains unfrozen during training. The loss weighting coefficients defined in (4.7) are empirically set to 1, assigning equal weight to each loss term.

For each dataset and model configuration, the training set is randomly shuffled and training is repeated three times. The best checkpoint from each run is selected based on the lowest mean squared error (MSE) on the unshuffled validation set. MSE is used as the checkpoint selection criterion due to the slower convergence of the force estimation head compared to the segmentation heads. All reported results in the following section represent the average performance of the three selected checkpoints for each model–dataset combination, evaluated on the corresponding test set. All experiments are conducted on an NVIDIA RTX 2080 GPU running Ubuntu 20.04.

4.3.2 Quantitative Results

Since TransForSeg is a multitask Transformer, each task is evaluated separately by comparison with state-of-the-art models reported in the literature. Similar to the evaluation protocol used for H-Net, as described in Chapter 3.3.1, the force estimation head is assessed using mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). The segmentation heads are evaluated using Accuracy (Acc), mean Intersection over Union (mIoU), and mean Dice coefficient (mDice). All benchmarked models from the literature, including those for both segmentation and force estimation, are trained, validated, and tested on the three datasets considered in this work: RGB, X-ray1, and X-ray2. All reported results represent the average performance over three independent runs to ensure consistency and reliability.

Force Estimation Table 4.1 presents a benchmark comparison in which TransForSeg-Small (S) and TransForSeg-Tiny (T) are evaluated against four learning-based catheter force estimation models from the literature. The experiments are conducted across four datasets: Segmented, RGB, Synthetic X-ray1, and Synthetic X-ray2. Each row in the table corresponds to the performance of the evaluated models on the test split of the respective dataset. All models are trained and validated using the corresponding training and validation splits to ensure a fair comparison.

In the segmented (Seg) dataset, models are provided with preprocessed images in which the

Table 4.1: Performance of Force Estimation Across Datasets

Data	type	Model	MSE	MAE	RMSE	R^2
Seg	NN	MLP [63]	4.4e-05	0.0040	-	0.98
	CNN	ResNet [1]	-	-	0.110	-
		Y-Net [78]	2.8e-05	0.0039	0.005	0.98
RGB	CNN	H-Net [79]	3.6e-05	0.0039	0.005	0.98
	ViT	TransForSeg (t)	1.77e-05	0.0027	0.0042	0.99
		TransForSeg (s)	1.81e-05	0.0028	0.0042	0.99
Xray1	CNN	H-Net [79]	3.3e-05	0.0039	0.0049	0.98
	ViT	TransForSeg (t)	2.04e-05	0.0029	0.0045	0.98
		TransForSeg (s)	2.79e-05	0.0034	0.0053	0.98
Xray2	CNN	H-Net [79]	4.3e-05	0.0046	0.0058	0.97
	ViT	TransForSeg (t)	3.43e-05	0.0038	0.0059	0.98
		TransForSeg (s)	3.34e-05	0.0038	0.0058	0.98

catheter has already been segmented and the background removed. The first two models (MLP and ResNet) estimate the applied force at the catheter tip only along the x and y axes, whereas Y-Net (refers to Chapter 2) performs full 3D force estimation. This dataset is considered the least challenging, as the foreground catheter is extracted using a thresholding-based method. However, this setup assumes the availability of an external segmentation module as a preprocessing step, which may negatively impact computational efficiency and real-time applicability in practical scenarios.

The second through fourth rows of Table 4.1 compare the proposed TransForSeg model against H-Net—introduced in Chapter 3—which is the only multitask architecture reported in the literature that jointly performs catheter segmentation and force estimation. Unlike the segmented dataset, these comparisons are conducted on unsegmented inputs, namely RGB, Synthetic X-ray1, and Synthetic X-ray2, listed in order of increasing complexity for both segmentation and force estimation. H-Net is based on a CNN encoder–decoder architecture, whereas TransForSeg employs a Vision Transformer (ViT) backbone. Both models estimate 3D contact forces.

As shown in the results, TransForSeg consistently outperforms H-Net across all datasets, achieving MSE improvements of 51% on RGB, 38% on X-ray1, and 22% on X-ray2. Furthermore, when compared with Y-Net evaluated on the segmented dataset, TransForSeg achieves superior performance on the RGB and X-ray1 datasets, with MSE improvements of 36% and 27%, respectively.

Table 4.2: Performance of Catheter Segmentation Across Datasets

Type	Model	#Params (m)	GFLOPs	RGB			X-Ray-1			X-Ray-2		
				Acc	mIoU	mDice	Acc	mIoU	mDice	Acc	mIoU	mDice
CNN	FCN [126]	134	42.4	99.2	94.0	96.8	99.2	94.1	96.8	99.1	93.8	96.7
	U-Net [104]	34	77.5	99.8	95.7	98.0	99.7	95.7	98.0	99.7	95.4	97.8
	Hr-Net [131]	9.6	3.5	98.8	96.1	98.4	98.8	96.1	98.5	98.8	95.8	98.0
	H-Net [79]	0.46	11.4	98.8	95.7	98.0	98.8	95.7	98.0	98.8	95.5	97.8
	DeepLabV3 [142]	13.8	11.5	99.9	97.0	98.4	99.9	97.2	98.5	99.9	96.9	98.4
Transformer	SegFormer [147]	3.7	1.3	99.9	98.6	98.9	98.8	97.6	98.7	99.9	97.5	98.7
	MaskFormer [146]	41.2	106.6	99.9	96.7	98.3	99.9	96.7	98.3	98.8	96.9	98.3
	TransForSeg (Tiny)	6.9	2.8	99.9	98.6	99.3	99.9	98.7	99.3	99.9	98.4	99.2
	TransForSeg (Small)	25.1	10.2	99.9	98.6	99.3	99.9	98.6	99.3	99.9	98.5	99.2

Regarding the two TransForSeg variants, the Tiny configuration yields better performance on the less complex datasets, while the Small configuration slightly outperforms it on the more challenging X-ray2 dataset. This behavior is attributed to the higher parameter capacity of the Small model, which enables better modeling of complex scenarios but may increase susceptibility to overfitting on simpler datasets.

Overall, the results demonstrate that TransForSeg achieves state-of-the-art performance in sensor-free 3D catheter force estimation, highlighting the effectiveness of Vision Transformer-based multitask architectures for this application.

Segmentation: Table 4.2 presents a comparison between TransForSeg and seven state-of-the-art catheter segmentation models from the literature. Each model included in the benchmark is independently trained on the three datasets—RGB, X-ray1, and X-ray2—and the reported results correspond to performance on the test split of each dataset. In addition to segmentation performance, the table reports the number of model parameters and Floating-Point Operations (FLOPs) to provide insight into the computational complexity of each approach.

All models considered in this comparison are single-task architectures, with the exception of H-Net and TransForSeg, which are multitask frameworks featuring two segmentation heads and one regression head. For these multitask models, the reported segmentation metrics are averaged across the two segmentation outputs. Unlike general-purpose multi-class segmentation problems, catheter segmentation represents a relatively constrained task, and most evaluated models demonstrate strong performance. Nevertheless, based on mIoU and mDice across all datasets, the FCN model exhibits the weakest performance among both CNN-based and Transformer-based approaches. In contrast,

DeepLabV3 achieves competitive results among the evaluated baselines and generally outperforms other CNN-based models.

Across all three datasets, TransForSeg achieves the highest mIoU and mDice scores, outperforming both CNN-based and Transformer-based segmentation baselines. While H-Net is also a dedicated multitask architecture designed for joint catheter segmentation and 3D force estimation, TransForSeg improves mIoU by approximately 3% on all datasets relative to H-Net. Despite its ViT-based encoder-decoder architecture, which incorporates an embedding fusion module, two segmentation heads, and a regression head, TransForSeg remains computationally efficient. This efficiency makes the model suitable for real-time deployment scenarios.

Moreover, integrating catheter segmentation with 3D force estimation within a single end-to-end framework eliminates the need for a separate segmentation model during data preprocessing for learning-based force estimation. At the same time, it enables the segmentation output to serve as a perception module for both surgeon-assisted and autonomous robotic catheterization systems. The reported FLOPs and parameter counts further demonstrate the ability of ViT-based architectures to achieve strong segmentation performance while maintaining relatively low computational complexity, owing to their capacity to generate semantically rich and spatially informative embeddings without reliance on skip connections during upsampling.

4.3.3 Ablation Study

As previously discussed, the segmentation heads in TransForSeg play a crucial role in enabling the model to estimate catheter contact forces directly from unsegmented X-ray images. The segmentation task guides the network to focus on catheter deflections, thereby reducing the influence of background variations during force estimation. To assess the contribution of the segmentation task, a modified version of the model, referred to as *TransForcer*, is introduced, in which the two segmentation heads are removed, converting the architecture into a single-task model.

Similar to TransForSeg, TransForcer processes input images acquired from two different viewpoints using its encoder and decoder. The embeddings produced by the encoder are fused within the decoder, and the [CLS] token from the decoder output is passed to a regression head for 3D force estimation. Unlike TransForSeg, TransForcer does not reshape the embeddings for segmentation,

Table 4.3: Ablation: Model with and without segmentation

Dataset	Model	MSE	MAE	RMSE	R^2
RGB	Tiny (with)	1.77e-05	0.0027	0.0042	0.99
	Tiny (without)	7.75e-06	0.0018	0.0028	0.99
	Base (with)	1.81e-05	0.0028	0.0042	0.99
	Base (without)	7.09e-06	0.0017	0.0027	0.99
X-Ray1	Tiny (with)	2.04e-05	0.0029	0.0045	0.98
	Tiny (without)	2.86e-05	0.0033	0.0054	0.98
	Base (with)	2.79e-05	0.0034	0.0053	0.98
	Base (without)	3.06e-05	0.0034	0.0055	0.98
X-Ray2	Tiny (with)	3.43e-05	0.0038	0.0059	0.98
	Tiny (without)	4.08e-05	0.0042	0.0064	0.97
	Base (with)	3.34e-05	0.0038	0.0058	0.98
	Base (without)	4.21e-05	0.0040	0.0065	0.97

as the segmentation heads are entirely omitted. Apart from the removal of the segmentation heads, the overall architecture of TransForcer remains identical to that of TransForSeg.

The TransForcer model is trained in both Tiny and Small configurations using the same training protocol described for TransForSeg. For each configuration, training is repeated three times on each dataset: RGB, X-ray1, and X-ray2. Instead of the multitask loss function defined in (4.7), TransForcer is optimized using the regression-only loss function defined in (4.6). Table 4.3 reports a comparison of force estimation performance between TransForSeg (denoted as *with* segmentation head) and TransForcer (denoted as *without* segmentation head) on the corresponding test sets.

Considering the X-ray2 dataset, which represents the most challenging scenario, the inclusion of the segmentation head improves the MSE by 15.9% for the Tiny configuration and 20.66% for the Small configuration of TransForSeg. A similar trend is observed on the X-ray1 dataset, where the segmentation head leads to MSE improvements of 28.6% and 8.8% for the Tiny and Small models, respectively. These results highlight the positive impact of integrating the segmentation task on the accuracy of the force estimation head when operating on synthetically generated X-ray images, in which the background is independent of the catheter deflection.

In contrast, this benefit is not observed on the RGB dataset, where models without the segmentation head outperform their multitask counterparts in both configurations. This behavior can be attributed to the fact that, in RGB images, background elements—such as the 3D-printed surface

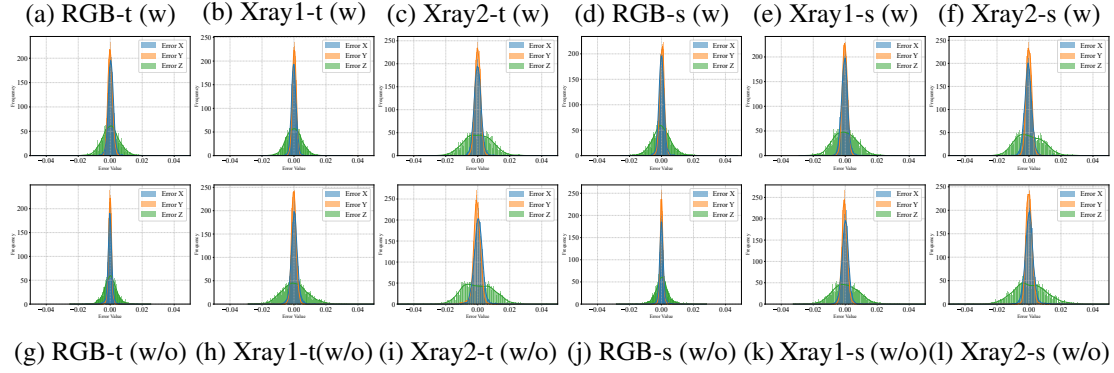


Figure 4.2: Error histograms of the force estimation head for TransForSeg (denoted as w , indicating models with segmentation heads) and TransForcer (denoted as w/o , indicating models without segmentation heads) across all evaluated datasets. In each subplot, t and s denote the Tiny and Small model configurations, respectively.

of the force sensor and associated shadows—are correlated with catheter motion, thereby providing indirect cues that facilitate force estimation even in the absence of explicit segmentation (see Fig.. 4.4).

Fig. 4.2 presents histograms of force prediction errors on the test sets for the RGB, X-ray1, and X-ray2 datasets, further illustrating the effect of the segmentation head on force estimation accuracy. The top row shows the error distributions for TransForSeg-Tiny and TransForSeg-Small, both equipped with segmentation heads, while the bottom row depicts the corresponding distributions for TransForcer-Tiny and TransForcer-Small, which do not include segmentation heads.

Fig. 4.3 illustrates the training and validation loss curves over epochs for both TransForSeg and TransForcer across all three datasets. Subplots (a)–(d) report the binary cross-entropy loss and mean Intersection over Union (mIoU) for the two segmentation heads in the Tiny and Small configurations of TransForSeg. These curves exhibit smooth convergence without observable signs of overfitting, indicating stable training behavior. Subplots (e)–(h) compare the regression loss curves of the force estimation head for TransForSeg and TransForcer in both model sizes. As observed, the inclusion of segmentation heads helps mitigate overfitting—particularly on the more challenging X-ray2 dataset—by providing additional supervision and promoting improved generalization.

Domain Shift and Generalization: Real-world deployment of segmentation and force estimation models often involves input data that deviate from the training distribution, underscoring the

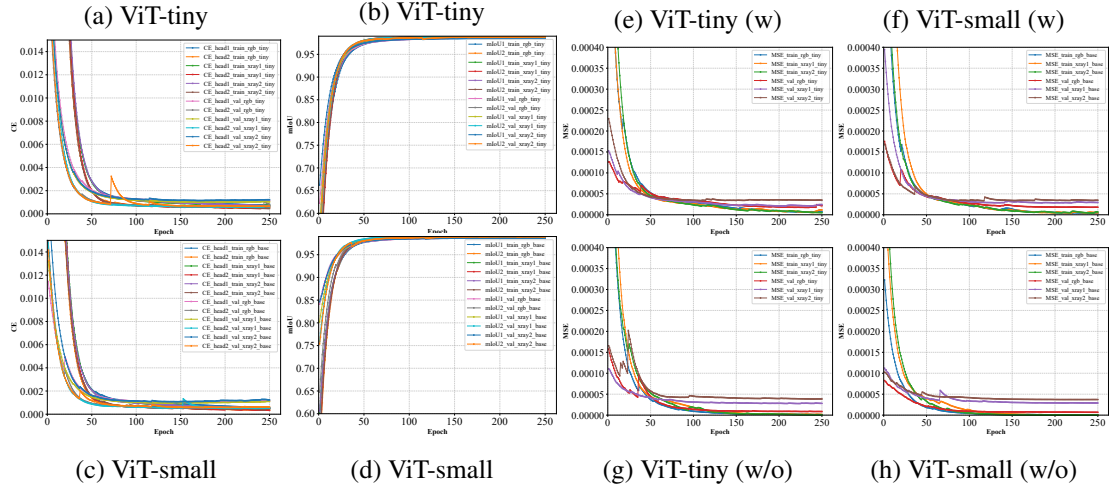


Figure 4.3: Training and validation loss curves. Subplots (a)–(d) show the segmentation loss for the Tiny and Small configurations of TransForSeg, while subplots (e)–(h) present the force estimation loss for TransForSeg (denoted as w , models with segmentation heads) and TransForcer (denoted as w/o , models without segmentation heads).

importance of evaluating robustness to domain shift and generalization. Having already assessed the performance of TransForSeg across different imaging modalities—namely RGB and synthetic X-ray datasets (X-ray1 and X-ray2)—its robustness is further investigated under input-level perturbations. To simulate realistic deployment conditions, six common noise types are applied to the test sets: Impulse noise, Gaussian noise, Poisson noise, motion blur, defocus blur, and stripe artifacts.

This extended evaluation enables a comprehensive assessment of the model’s robustness and generalization for both tasks—semantic segmentation and 3D force estimation—in the presence of real-world image corruptions. Table 4.4 reports the performance of TransForSeg Tiny and TransForSeg Small on these noisy test sets. In this context, the source domain corresponds to the original training datasets (RGB, X-ray1, and X-ray2), while the target domains consist of their respective corrupted test sets. It is important to emphasize that the models are not exposed to any of the noise types during training, and no domain generalization, domain adaptation, or data augmentation techniques are employed at either training or test time.

The statistics and parameters associated with each noise type are summarized in Table 4.4. For example, defocus blur is simulated using an isotropic Gaussian filter with a kernel size of $k = 10$ and a blur radius (standard deviation) of $\sigma = 2.0$. The isotropic nature of the blur indicates that

smoothing is applied uniformly in all directions. The kernel size k defines the spatial extent of the convolution operator, while σ controls the strength of the smoothing effect, with larger values producing stronger blur. This process emulates defocusing effects that may arise from depth-of-field limitations or optical imperfections in real-world medical imaging systems.

The results demonstrate that TransForSeg maintains robust performance on the X-ray1 and X-ray2 datasets when subjected to stripe artifacts, Poisson noise, and impulse noise, with no notable degradation observed in either segmentation accuracy or force estimation performance. In contrast, Gaussian noise, motion blur, and defocus blur have a more pronounced impact, resulting in a $1.5\times$ to $3\times$ increase in MSE and a 2% to 6% reduction in mIoU for the segmentation heads.

Despite the challenges introduced by domain shift and noise perturbations, TransForSeg continues to exhibit competitive performance relative to the methods reported in Tables 4.1 and 4.2, all of which are evaluated exclusively on clean, noise-free images. Among the evaluated datasets, the noisy RGB images prove to be the most challenging, particularly for the force estimation task. Stripe artifacts lead to a $12\times$ increase in MSE, while other noise types result in degradation factors ranging from $1.8\times$ to $4.8\times$. This sensitivity is attributed to the structured nature of RGB backgrounds—such as the force sensor and its surrounding environment—which are more susceptible to disruption by noise. In contrast, the synthetic X-ray backgrounds are randomly generated, encouraging the model to focus more directly on catheter geometry and improving tolerance to image variability.

Interestingly, segmentation performance on noisy RGB images remains relatively stable, following a degradation trend similar to that observed on the synthetic X-ray datasets. Fig. 4.4 illustrates qualitative segmentation results on the RGB, X-ray1, and X-ray2 datasets under both clean and noisy conditions. Finally, it is noted that test-time training and domain generalization techniques could be incorporated in real-world deployments to further enhance robustness without requiring additional domain-specific data collection or retraining [151, 152, 153].

Table 4.4: Performance of Models Across Different Noises

Noise	Noise Prop- erties	Dataset	Size	MSE	MAE	RMSE	Acc(%)	mIoU(%)	mDice(%)
Impulse	Salt (255) pepper (0) randomly applied to 20% of extreme pixels	RGB	Tiny Base	3.82e-05 (2.1x)	0.0042 (1.5x)	0.0061 (1.4x)	99 ($\approx$)	98 (-0.5)	99 (-0.2)
				3.29e-05 (1.8x)	0.0039 (1.4x)	0.0057 (1.3x)	99 ($\approx$)	98 (-0.4)	99 (-0.2)
		X-Ray1	Tiny Base	2.19e-05 ($\approx$)	0.0030 ($\approx$)	0.0046 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
				2.84e-05 ($\approx$)	0.0035 ($\approx$)	0.0053 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
		X-Ray2	Tiny Base	3.75e-05 (1.1x)	0.0039 ($\approx$)	0.0061 ($\approx$)	99 ($\approx$)	98 (-0.2)	99 (-0.1)
				3.56e-05 ($\approx$)	0.0039 ($\approx$)	0.0060 ($\approx$)	99 ($\approx$)	98 (-0.1)	99 ($\approx$)
Gaussian	Mean: $\mu = 0$ Std: $\sigma = 0.02$	RGB	Tiny Base	6.90e-05 (3.9x)	0.0059 (2.1x)	0.0083 (2.0x)	99 ($\approx$)	96 (-2.0)	98 (-1.0)
				5.77e-05 (3.2x)	0.0053 (1.9x)	0.0076 (1.8x)	99 ($\approx$)	97 (-1.2)	98 (-0.6)
		X-Ray1	Tiny Base	2.54e-05 (1.2x)	0.0033 (1.1x)	0.0050 (1.1x)	99 ($\approx$)	98 (-0.3)	99 (-0.1)
				3.019e-05 ($\approx$)	0.0036 ($\approx$)	0.0055 ($\approx$)	99 ($\approx$)	98 (-0.2)	99 (-0.1)
		X-Ray2	Tiny Base	5.42e-05 (1.6x)	0.0048 (1.3x)	0.0074 (1.2x)	99 ($\approx$)	97 (-1.2)	98 (-0.6)
				4.78e-05 (1.4x)	0.0046 (1.2x)	0.0069 (1.2x)	99 ($\approx$)	97 (-1.1)	98 (-0.5)
Poisson	Intensity- dependent ($\lambda = I$), Non- Gaussian noise, pixel-wise	RGB	Tiny Base	4.47e-05 (2.5x)	0.0046 (1.7x)	0.0067 (1.6x)	99 ($\approx$)	98 (-1.0)	99 (-0.5)
				3.74e-05 (2.0x)	0.0041 (1.2x)	0.0061 (1.1x)	99 ($\approx$)	97 (-0.1)	98 (-0.2)
		X-Ray1	Tiny Base	2.11e-05 ($\approx$)	0.0030 ($\approx$)	0.0045 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
				2.82e-05 ($\approx$)	0.0034 ($\approx$)	0.0053 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
		X-Ray2	Tiny Base	3.47e-05 ($\approx$)	0.0038 ($\approx$)	0.0059 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
				3.35e-05 ($\approx$)	0.0038 ($\approx$)	0.0058 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
Motion	Kernel (k=6), Motion Angle ($\theta = 20$), Linear Spread	RGB	Tiny Base	6.01e-05 (4.6x)	0.0053 (2.3x)	0.0078 (2.1x)	99 (-0.1)	92 (-6.4)	95 (-3.5)
				5.66e-05 (4.8x)	0.0051 (2.5x)	0.0075 (2.2x)	99 (-0.1)	92 (-6.1)	95 (-3.3)
		X-Ray1	Tiny Base	6.26e-05 (2.9x)	0.0054 (1.9x)	0.0079 (1.7x)	99 (-0.1)	92 (-6.4)	95 (-3.5)
				6.19e-05 (2.7x)	0.0054 (2.0x)	0.0079 (1.6x)	99 (-0.1)	92 (-6.4)	95 (-3.5)
		X-Ray2	Tiny Base	6.31e-05 (1.7x)	0.0054 (1.4x)	0.0079 (1.3x)	99 (-0.1)	92 (-6.4)	95 (-3.5)
				5.64e-05 (1.4x)	0.0052 (1.4x)	0.0075 (1.2x)	99 (-0.2)	92 (-6.2)	96 (-3.4)
Defocus	Kernel Size (k=10), Blur Radius ($\sigma=2.0$), Isotropic Blur	RGB	Tiny Base	5.79e-05 (3.2x)	0.0052 (1.9x)	0.0076 (1.8x)	99 ($\approx$)	97 (-1.7)	98 (-0.9)
				4.29e-05 (2.4x)	0.0044 (1.7x)	0.0065 (1.5x)	99 ($\approx$)	97 (-1.0)	98 (-0.5)
		X-Ray1	Tiny Base	6.74e-05 (3.3x)	0.0053 (1.8x)	0.0082 (1.8x)	99 ($\approx$)	96 (-2.6)	97 (-1.5)
				7.75e-05 (2.8x)	0.0055 (1.6x)	0.0087 (1.6x)	99 ($\approx$)	95 (-3.0)	97 (-1.6)
		X-Ray2	Tiny Base	7.85e-05 (2.3x)	0.0059 (1.6x)	0.0089 (1.5x)	99 ($\approx$)	95 (-3.3)	97 (-1.8)
				7.16e-05 (2.1x)	0.0057 (1.5x)	0.0085 (1.4x)	99 ($\approx$)	96 (-2.7)	97 (-1.4)
Stripe	Intensity $\alpha=0.1$, Vertical, Numbers (N=50), Max Width (w=10)	RGB	Tiny Base	2.13e-04 (12.0x)	0.0104 (3.8x)	0.0143 (3.4x)	99 (-0.2)	92 (-6.0)	96 (-3.3)
				2.18e-04 (12.0x)	0.0101 (3.6x)	0.0144 (3.4x)	99 (-0.6)	96 (-2.6)	98 (-1.3)
		X-Ray1	Tiny Base	2.39e-05 (1.2x)	0.0032 ($\approx$)	0.0049 ($\approx$)	99 (-0.9)	98 (-0.1)	99 ($\approx$)
				2.98e-05 ($\approx$)	0.0036 ($\approx$)	0.0055 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)
		X-Ray2	Tiny Base	3.48e-05 ($\approx$)	0.0038 ($\approx$)	0.0059 ($\approx$)	98 ($\approx$)	98 ($\approx$)	99 ($\approx$)
				3.35e-05 ($\approx$)	0.0038 ($\approx$)	0.0058 ($\approx$)	99 ($\approx$)	98 ($\approx$)	99 ($\approx$)

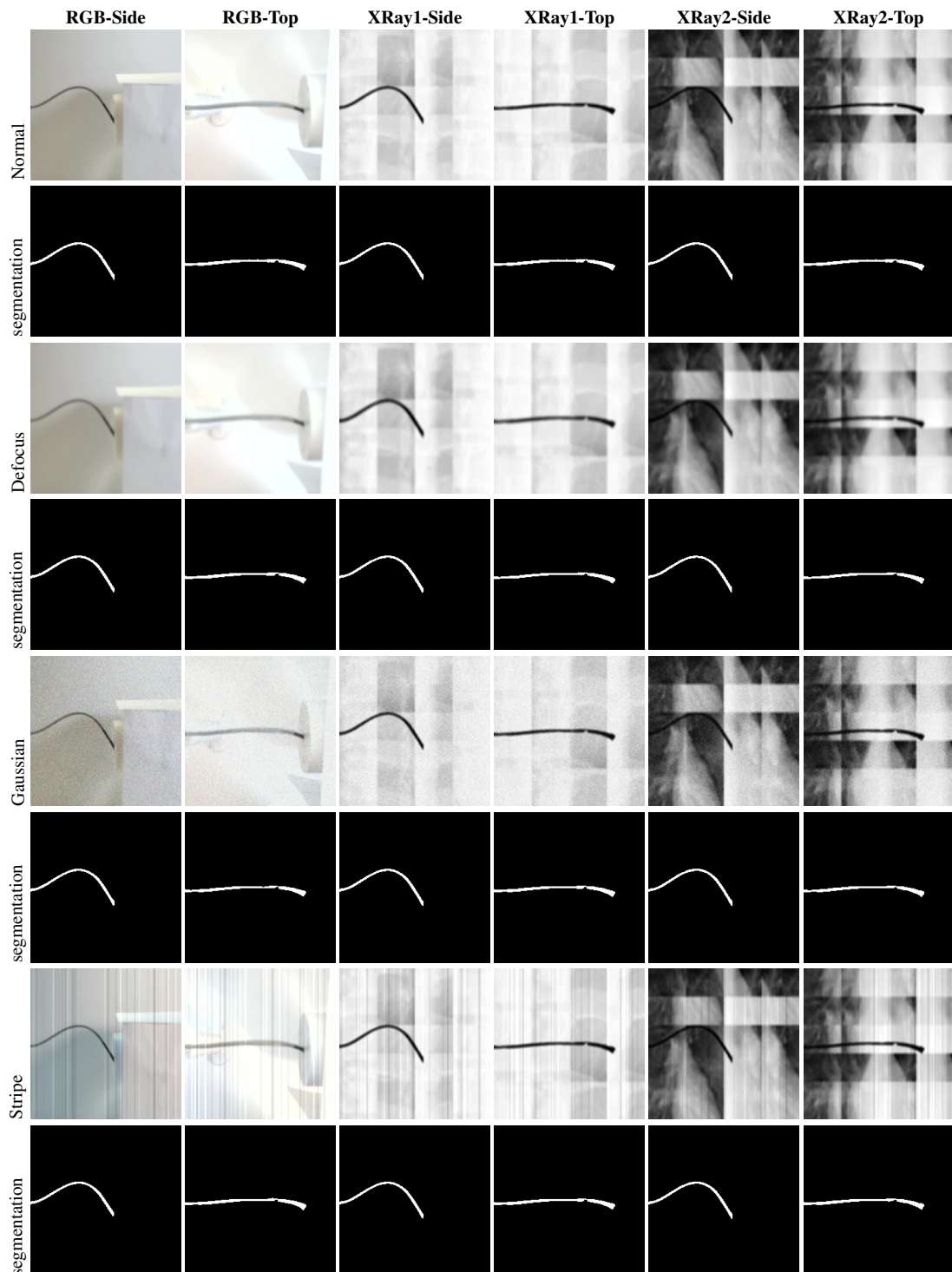


Figure 4.4: Qualitative results across 3 out of 6 noise conditions for 3 input samples. The first row displays the original (clean) samples, followed by the corresponding segmentation outputs produced by the model in the second row.

4.4 Conclusion

This Chapter presented TransForSeg, a novel multitask Vision Transformer–based architecture for the simultaneous tasks of catheter segmentation and 3D force estimation from stereo X-ray images. By integrating shared transformer components with task-specific heads, TransForSeg enables end-to-end learning without the need for additional preprocessing. Extensive evaluations on RGB and synthetic X-ray datasets demonstrate that TransForSeg achieves state-of-the-art performance in force estimation while maintaining computational efficiency. Ablation experiments using a single-task variant, referred to as TransForcer, confirm that the segmentation task plays a critical role in improving force prediction accuracy, particularly in more challenging domains. Furthermore, TransForSeg exhibits strong robustness to domain shift and input-level perturbations, maintaining reliable performance under a range of noise conditions. These results demonstrate that TransForSeg attains state-of-the-art performance among existing approaches, while remaining lightweight and generalizable for catheter-based interventions.

A compelling avenue for future work is to integrate recent medical foundation segmentation models into a teacher–student knowledge-distillation paradigm. The current literature shows rapid progress in promptable and foundation-based segmentation through SAM-derived medical models, including SAM-Med2D, MedSAM2, and VISTA3D, as well as parameter-efficient and domain-robust variants such as Med-SA, DeSAM, and SegMoTE [154, 155, 156, 157, 158, 159]. In parallel, open-world segmentation frameworks such as MedOpenSeg and MedSeg-R have introduced mechanisms for handling unseen or out-of-distribution anatomical structures [160, 161]. Motivated by these developments, a foundation teacher could be used in future work to provide dense and soft supervision for the anatomical scene, including vessels, adjacent organs, and uncertainty regions, while a compact task-specific student network, such as H-Net or TransforSeg, is trained to preserve catheter-specific accuracy and real-time feasibility. This formulation could improve the separation of the catheter from vessel branches by transferring higher-level anatomical priors and open-world awareness from large-scale foundation models into a deployment-oriented fluoroscopic segmentation model. Building upon the perception and force estimation framework established in this chapter, the next chapter shifts the focus to catheter navigation, examining learning-based

approaches for guiding catheter motion within vascular environments.

Chapter 5

DINO-CVA: A Multimodal Goal-Conditioned Vision-to-Action Model Toward Autonomous Catheter Navigation

5.1 Introduction

Minimally Invasive Surgery (MIS) was already reviewed in Chapter [1.1](#). This surgical method uses small incisions or natural orifices to access internal organs, avoiding the large openings of traditional open surgery. By reducing the extent of skin and tissue disruption, MIS techniques significantly lower the physiological burden on patients. These procedures typically rely on advanced instruments and imaging technologies, such as endoscopes and catheters, which enable surgeons to operate with precision through narrow “keyhole” entry points [\[2\]](#). Cardiac catheterization applies the principles of MIS by using a slender, flexible tube known as a catheter to traverse the vascular system and reach the heart. This approach enables both diagnostic and therapeutic procedures, such as angiography, electrophysiological mapping, and ablation, without the need for open-chest

surgery. Modern catheters are equipped with steerable distal segments that are actuated by pull-wires or tendon-based mechanisms, which provide precise navigation through complex vascular pathways [162]. This controllability allows operators to position the catheter tip accurately at targeted sites, including arrhythmogenic foci, where interventions such as stimulation or ablation can be carried out with high precision [8]. During cardiac catheterization, access is typically obtained through vascular entry sites such as the femoral or radial artery, and the catheter is advanced under continuous fluoroscopic monitoring. Fluoroscopy has become the gold standard for guiding these procedures, as it provides dynamic X-ray images that reveal the catheter’s position relative to surrounding anatomy. However, its routine use introduces important drawbacks. In the preceding Chapters, this thesis addressed several fundamental limitations of conventional cardiac catheterization by proposing learning-based solutions for force sensing and catheter visualization, with the aim of improving safety, accuracy, and operator awareness under fluoroscopic guidance. While these contributions mitigate important perceptual and sensing challenges inherent to current workflows, they do not fully resolve the broader limitations associated with manual catheter navigation, which were introduced in Chapter 1.5.2. The following discussion therefore shifts focus toward the navigation aspects of cardiac catheterization and the motivation for autonomous control strategies.

During catheterization procedure, the dependence on ionizing radiation creates health risks for both patients and clinical staff, with cumulative exposure linked to malignancies, cataract formation, and radiation-induced skin damage. These risks are particularly concerning in complex ablation cases, where extended imaging times substantially increase total radiation dose. Beyond radiation hazards, manual catheterization also places significant physical and cognitive demands on operators. Navigating the catheter through tortuous vasculature requires sustained attention, precise hand–eye coordination, and fine motor control, often maintained for hours at a time. These tasks are performed while wearing heavy lead aprons for radiation shielding, forcing clinicians to adopt static and ergonomically unfavorable positions. Over time, such working conditions contribute to operator fatigue and have been strongly associated with a high prevalence of musculoskeletal disorders among interventional cardiologists. Chronic back pain, neck strain, and joint injuries are frequently reported, underscoring the considerable occupational burden imposed by conventional catheterization practices [41, 42, 43].

As discussed in Chapter 1.2.2, robotic-assisted catheterization platforms have been developed to address limitations of manual procedures such as radiation exposure, operator fatigue, and limited precision [163]. Early systems like the Sensei Robotic System (Hansen Medical) and the Magellan Robotic System (Auris Health) enhanced catheter stability, reach, and steerability within tortuous vessels [19, 20, 21], while the CorPath GRX (Corindus/Siemens) enabled precise remote manipulation of guidewires, balloons, and stents during percutaneous interventions, reducing radiation exposure and demonstrating feasibility for tele-operated procedures [22, 23]. In parallel, magnetic navigation systems such as the Niobe (Stereotaxis) have leveraged externally generated magnetic fields to achieve superior catheter stability and mapping accuracy in electrophysiological ablation [24], whereas simpler designs like the Amigo Remote Catheter System (Catheter Precision) provided safe and effective remote manipulation of conventional catheters with reduced operator radiation exposure [25]. Recent research prototypes are further advancing the field by incorporating force feedback and semi-autonomous control, including magnetorheological-fluid-based haptic feedback and platforms such as MAGiC RMN, which improve contact force regulation and demonstrate high success rates in arrhythmia ablation [26, 27].

Despite these technological advances, current robotic catheterization platforms remain fundamentally limited by their follow–leader architectures, which require constant physician control. Although these systems reduce radiation exposure and improve catheter stability, they do not alleviate the cognitive and physical workload of the operator, nor do they provide intelligent capabilities such as adaptive decision-making or autonomous navigation in complex anatomies. To address these limitations, there is a growing need for methods that can be seamlessly integrated into existing robotic platforms to enable semi-autonomous or fully autonomous catheter manipulation.

This Chapter introduces an autonomous navigation system for cardiac catheterization, aimed towards reducing operator dependency and improving the reliability of catheter-based therapies. The proposed method is a multimodal vision–action model that learns from expert demonstrations in a learning-from-demonstration (LfD) framework by fusing visual observations with kinematic states. By moving beyond follow–leader configurations and incorporating intelligent navigation strategies, this approach represents an important step toward enabling autonomy that reflects expert practice. In the following section, related works on autonomous catheter navigation are reviewed,

after which the details of the proposed method and its key contributions are presented.

5.1.1 Related Work

The pursuit of autonomy in robotic catheterization has recently gained momentum, with research efforts primarily centered around three cutting-edge learning paradigms: Reinforcement Learning (RL), Inverse Reinforcement Learning (IRL), and Imitation Learning (IL). The fundamental these methods were reviewed extensively in Chapter 1.4.2. To maintain continuity and contextual clarity, the following sections present a concise recap of state-of-the-art methodologies within each paradigm as applied to robotic catheterization.

Reinforcement Learning (RL)

RL has been widely explored for autonomous navigation and skill transfer in vascular and catheter-based interventions, leveraging trial-and-error interactions in simulated or physical environments to develop adaptive decision-making policies [66, 67, 68, 69, 70, 28]. Studies have investigated diverse frameworks ranging from Deep Q-Network (DQN)-based navigation [66] and Soft Actor-Critic (SAC) methods for real-time guidewire control [67, 68] to Deep Deterministic Policy Gradient (DDPG)-driven virtual reality training platforms for skill transfer [69]. More recent work integrates image guidance and advanced path planning [70], while open-source simulators such as CathSim facilitate reproducible RL training with algorithms like SAC and Proximal Policy Optimization (PPO) [28].

Imitation Learning (IL)

IL has attracted growing interest in robotic catheterization and minimally invasive interventions as it enables systems to acquire expert-level behaviors directly from demonstrations (Learning from Demonstration "LfD"), bypassing the need for handcrafted models or purely trial-and-error learning [71, 72, 73, 74]. Early studies explored IL for tasks such as 3D/2D model-to-image registration in cardiac procedures, where neural networks learned to mimic expert alignment strategies to improve accuracy in image-guided interventions [71]. Generative Adversarial Imitation Learning

(GAIL) frameworks have further been used for autonomous catheter navigation and soft-tissue manipulation, often refined through Proximal Policy Optimization (PPO) to enhance precision, adaptability, and safety [72, 73]. More recent work integrates IL with reinforcement learning, combining expert demonstrations with policy optimization and curriculum learning to improve robustness, collision avoidance, and generalization across complex anatomical environments [74].

Inverse Reinforcement Learning (IRL)

IRL extends beyond standard imitation learning by inferring the latent reward functions underlying expert demonstrations, enabling the development of adaptive and generalizable control policies rather than simply replicating observed behaviors [75, 76]. In robotic catheterization, IRL has been applied to pre-operative path planning and autonomous navigation, where frameworks combining Behavioral Cloning (BC) and Generative Adversarial Imitation Learning (GAIL) inferred reward structures from expert trajectories to guide catheter movements in dynamic and patient-specific anatomies [75]. Other studies have used IRL to model expert decision-making in procedures such as mechanical thrombectomy, recovering reward functions from recorded state–action pairs and using them to train reinforcement learning policies via algorithms like Soft Actor-Critic (SAC), thereby enhancing precision, safety, and adaptability in complex surgical environments [76].

Challenges of Existing Paradigms

Despite significant progress, existing learning paradigms face major challenges in robotic catheterization. RL suffers from sample inefficiency, sim-to-real gaps, reward engineering complexity, safety concerns, and instability in high-dimensional control settings. IL mitigates reward design issues but depends heavily on large expert datasets, struggles with covariate shift, lacks adaptability to novel anatomies, faces ambiguity in demonstrations, and does not explicitly optimize for safety or performance; moreover, most IL methods rely on kinematic or kinesthetic demonstrations with limited visual perception of the surgical scene or require external detectors such as YOLOv5 for scene understanding [87, 88, 77, 82]. IRL attempts to infer latent rewards but encounters ambiguity in recovered objectives, computational burdens in high-dimensional spaces, and poor generalization to unseen clinical scenarios.

5.1.2 Contributions

Addressing the challenges outlined above, this Chapter presents a *toward-autonomy* approach for cardiac catheter navigation by introducing a *multimodal, goal-conditioned behavior cloning model*. This preliminary architecture serves as a foundational bridge toward overcoming these limitations, prioritizing supervised multimodal learning over full exploration-based autonomy in initial validations. Rather than aiming to fully solve autonomous catheterization, the proposed framework is intended as a foundational step toward learning-based multimodal navigation policies that integrate visual perception and control. The model is inspired by the architectural principles of DINO-WM and is designed to learn jointly from visual observations and the robot’s controller states (i.e., actions) over time [164]. Specifically, it operates as a multimodal vision–action model: given video frames of the catheterization task together with joystick states corresponding to the robot’s motors (translation, rotation, and knob), the model predicts the subsequent action in an autoregressive manner. Unlike prior approaches that rely on object detection for catheter or vessel localization or on pre-operative trajectory planning, the proposed ViT-based architecture learns navigation behavior by fusing visual cues and action states and processing them jointly over time with both temporal and spatial awareness.

The proposed model is trained in a supervised manner (autoregressive) using demonstration data collected from expert surgeons performing catheter insertion tasks [164]. To mitigate the challenges of sample inefficiency and reward design inherent to reinforcement learning, a learning-from-demonstration (LfD) framework is adopted. Unlike earlier LfD studies that primarily replicate kinematic or kinesthetic motions, the proposed approach integrates visual understanding with kinematic features to capture a more human-like representation of expert decision-making. A goal-conditioning mechanism is incorporated to improve adaptability and generalization beyond conventional behavior cloning. Specifically, the model is conditioned on a frame representing the target location of the catheter within the vascular system, enabling trajectory generation by projecting learned behaviors toward the surgical goal.

Also, imitation learning is favored over reinforcement learning due to the structural and procedural characteristics of cardiac catheterization. Unlike conventional reinforcement learning settings,

where exploration is essential for discovering optimal behaviors, the vascular anatomy in cardiac interventions is largely static and well-characterized prior to the procedure through pre-operative imaging, thereby reducing the necessity for extensive trial-and-error interaction with the environment. Moreover, expert demonstrations provide direct guidance for navigating the vascular network toward clinically relevant targets, allowing the model to learn goal-directed control in a safe and data-efficient manner. By employing a goal-conditioned imitation learning framework, the policy is explicitly conditioned on the desired anatomical endpoint, enabling reproduction of expert-like trajectories along vascular pathways while leveraging the static nature of the environment to promote safety and efficiency.

Furthermore, because the proposed architecture captures expert demonstrations in both visual and kinematic spaces, it establishes a basis for future extensions toward inverse reinforcement learning, where learned representations may assist in reward modeling and policy refinement. To support training and evaluation, a robotic platform equipped with a remote-control mechanism was designed and implemented, enabling operators to steer a catheter within an artificial vascular system. This setup facilitated the collection of expert demonstration data required for supervised training and empirical validation of the proposed model. To the best of the author’s knowledge, this work represents one of the earliest efforts to introduce a goal-conditioned, multimodal vision–action behavior cloning framework for autonomous catheter navigation, serving as an initial step toward more advanced multimodal and goal-driven robotic catheterization systems. For clarity and conciseness, the proposed framework is referred to as *DINO-CVA* (Dino Catheter Vision–Action model) throughout this chapter.

5.2 Experimental Setup and Data Compilation

Since commercial robotic catheterization platforms are not readily accessible for research purposes, a simplified yet representative robotic system was developed to support demonstration, data compilation, and model validation. The system was designed to emulate the follow–leader control paradigm commonly employed in clinical robotic catheterization platforms, while remaining modular and cost-effective for experimental use. A standard ablation catheter was integrated with a

custom robotic actuation unit and a teleoperation interface, enabling remote control of catheter motion by an operator. User commands were issued through a joystick-based controller and mapped directly to robotic actuators responsible for catheter translation, rotation, and tip steering. Real-time visual feedback was provided by a top-mounted camera, approximating the single-view perceptual constraint typically encountered during fluoroscopy-guided navigation. In this configuration, the operator interacted with the system by issuing joystick commands while relying exclusively on the camera feed for guidance, thereby reproducing key elements of the perception–action loop present in existing robotic catheterization workflows within a controlled experimental environment. The robotic platform, interface software, and data compilation procedure are described in detail in the following subsections.

5.2.1 Robotic Platform for Catheter Navigation

As shown in Fig. 5.1, a transparent and commercially available synthetic vessel phantom was used to approximate the vascular environment encountered during cardiac catheterization, enabling the insertion of an ablation catheter (Blazer XP, Boston Scientific, 110 cm length, outer diameter 2.5 mm). The phantom was constructed at a 1:1 scale based on a medically validated human anatomical model. It was mounted on a tabletop, while a top-mounted camera provided a fixed overhead view of the procedure, ensuring that both the phantom and the imaging setup remained stable throughout catheterization. The robotic platform provided three degrees of freedom (DoFs), each actuated by a NEMA 17 stepper motor and driven by an independent TB6600 microstepping driver. The platform base was securely mounted on an optical breadboard to minimize vibration and prevent undesired motion. The three DoFs included: (1) a linear rail with a travel range of 55 cm for catheter advancement and retraction, (2) a rotational stage directly coupled to the catheter steering knob for distal tip deflection, and (3) a base rotation module capable of rotating the entire catheter within a range of $\pm 90^\circ$. Structural components were fabricated using 3D printing to ensure secure catheter fixation, modularity, and smooth mechanical motion during operation.

A Logitech StreamCam (1080p resolution at 60 fps) was mounted above the phantom to capture real-time visual feedback of the catheter navigation process. To prevent catheter buckling during insertion, a hollow guiding rod was positioned at the center of the setup, through which the catheter

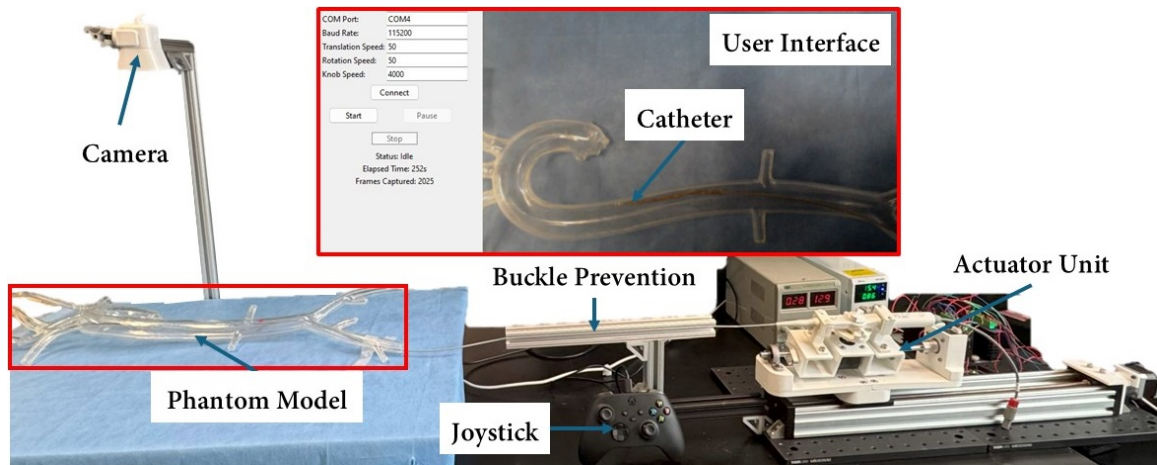


Figure 5.1: Overall experimental setup. The system includes a transparent vessel phantom positioned beneath a top-mounted camera, a robotic platform with an actuation unit mounted on a linear rail, and a joystick-based controller for teleoperation. A buckling-prevention guide ensures stable catheter insertion. The graphical user interface enables parameter configuration and provides real-time visualization of the catheter within the phantom.

was advanced. This configuration enabled precise, stable, and repeatable catheter control within the vessel phantom. For teleoperation, a joystick-based wired controller was employed to command the three robotic DoFs. The overall experimental setup is shown in Fig. 5.1, while a close-up view of the robotic platform is provided in Fig. 5.2.

5.2.2 Robot Interface Software

A custom software environment was implemented to support real-time teleoperation and synchronized data acquisition (see Fig. 5.1). The three robotic actuation modules—translation, distal tip steering, and base rotation—were assigned to corresponding control axes of a joystick-based controller, providing intuitive operation for the user. Controller inputs were transmitted to the robotic system through a microcontroller interface and converted into motor commands, enabling responsive execution during catheter navigation. A graphical user interface was developed to configure system parameters, including motor speed, communication settings, and calibration routines. The interface also displayed a continuous live video stream from the top-mounted camera, which served as the operator’s sole source of visual feedback. This integrated software framework enabled precise and consistent catheter navigation while ensuring synchronized recording of motor states and

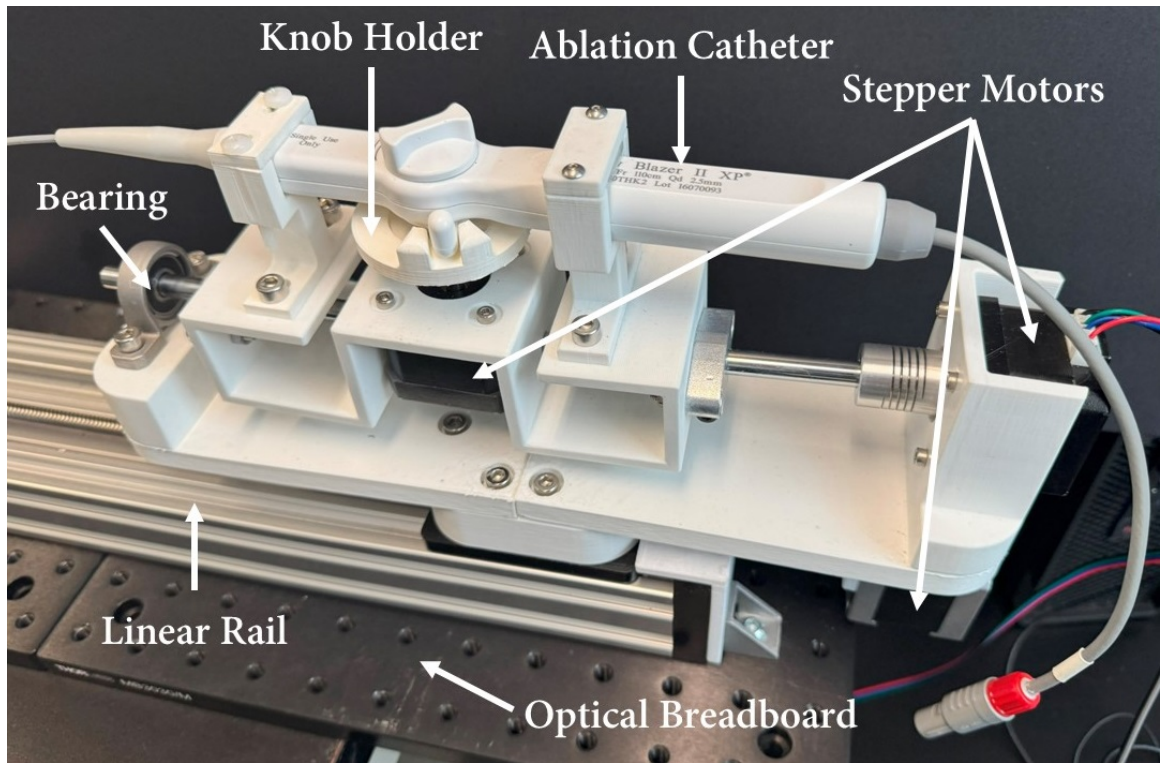


Figure 5.2: Close-up of the robotic platform. Three stepper motors, each driven by a microstepping driver, enable three degrees of freedom (DoFs): catheter translation, distal tip steering, and base rotation. The system is mounted on an optical breadboard, with 3D-printed fixtures securing the catheter during operation.

video frames for subsequent analysis and model training.

5.3 Dino-CVA: A Vision-Action model for Catheter Navigation

In general, autonomous catheterization systems can be understood as the integration of two tightly coupled components: a perception module and a control module. The perception module is responsible for real-time situational awareness, extracting information such as vascular geometry, catheter position, and tissue interactions from the operative scene. The control module then translates this information into actuation commands to enable safe and efficient navigation. Together, these modules support reliable catheter steering with the potential to shorten procedure times, reduce navigation errors, and enhance procedural efficiency.

Within the control module, two core functions are traditionally emphasized: path planning and

trajectory generation. Path planning refers to the process of identifying a feasible route for the catheter to travel from its current position to a target location while respecting anatomical constraints and avoiding collisions with vessel walls. Trajectory generation then produces a time-ordered sequence of states or control inputs required to execute the planned path in a stable and safe manner.

As an initial proof-of-concept toward full autonomy, this Chapter introduces a goal-conditioned behavior cloning framework that unifies perception and control within a single learning-based architecture. Rather than explicitly computing paths and generating trajectories, the proposed model learns navigation behavior directly from expert demonstrations by mapping visual and kinematic patterns to control actions. The proposed approach is designed to be compatible, in principle, with existing robotic-assisted catheterization platforms, such as Sensei or Magellan, by extending their conventional follow-leader functionality toward higher levels of autonomy. In the following sections, the proposed methodology and the associated dataset are described in detail.

5.3.1 Data Compilation

The dataset was constructed through a controlled experimental protocol in which the catheter was navigated toward nine predefined target points within a vessel phantom. Each trial began from the same femoral entry site to ensure consistent initial conditions across experiments. An expert operator provided demonstrations by guiding the distal catheter tip exclusively using the live camera feed. Each navigation sequence involved advancing the catheter toward a designated target point, and each experiment was repeated five times to reduce variability and improve repeatability. Only forward advancement phases were recorded, while retraction motions were excluded to avoid introducing heterogeneous control patterns that are not relevant to goal-directed navigation. During all procedures, synchronized video frames and corresponding robot state information were logged, resulting in a paired dataset of visual observations and control actions suitable for model training and evaluation. Further details regarding the dataset composition and usage are provided in [Section 5.4](#).

5.3.2 Methodology

The proposed model, referred to as DINO-CVA, is a supervised multimodal framework designed to steer a catheter toward a specified target by predicting controller signals for a robotic-assisted

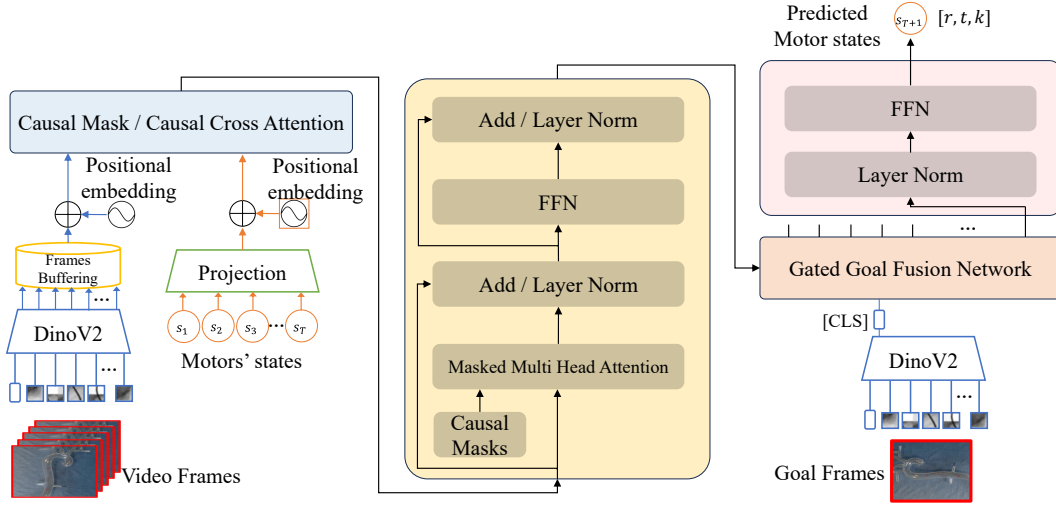


Figure 5.3: Architecture of the DINO-CVA model. The framework receives video frames of the catheterization task together with motor states (joystick inputs). The frames are encoded using a frozen DINOv2 backbone and organized into a temporal sequence. The motor states are linearly projected to match the visual embedding space. After positional embeddings are applied, cross-attention fuses the two modalities, while spatial and temporal dependencies are modeled by a causal transformer. The resulting fused representation is then conditioned on the target frame through a gated fusion network, and the action head predicts the next controller state s_{t+1} .

catheterization platform from combined visual and action inputs. Unlike kinesthetic or purely kinematic skill modeling approaches, DINO-CVA integrates visual observations and controller states over time, enabling control decisions to be generated based on temporal context rather than instantaneous inputs [77, 82, 87, 88]. As a behavior cloning model trained on expert demonstrations, the framework further improves generalization by conditioning its outputs on visual information corresponding to the final target location within the vessel. The overall architecture of DINO-CVA is illustrated in Fig. 5.3.

Formally, DINO-CVA is provided with an input sequence $X = \{F, S\}$ of length N , where F denotes the sequence of video frames capturing the catheterization scene and S represents the corresponding sequence of controller states for the robotic platform described above. Each video frame is represented as an image $I \in \mathbb{R}^{h \times w \times c}$, while each controller state is defined as a vector $s \in \mathbb{R}^3$, corresponding to translation, rotation, and steering knob actuation, respectively. The image frames are encoded using a pre-trained DINOv2 vision transformer, from which token embeddings are

extracted [165]. These visual tokens are then buffered as follows:

$$F_N = \{f_1, f_2, \dots, f_N\}, \quad (5.1)$$

$$f_i = \{e_{i1}, e_{i2}, \dots, e_{iP}\}, \quad (5.2)$$

where $F_N \in \mathbb{R}^{N \times P \times d}$ represents the sequence of visual token embeddings, with N denoting the number of frames in the sequence, P the number of patch tokens per frame, and d the embedding dimension of each token. Each embedding $e_{ip} \in \mathbb{R}^d$ corresponds to a patch-level representation extracted from frame i , including the [CLS] token. The corresponding motor state sequence is defined as $S_N \in \mathbb{R}^{N \times 3}$ and is projected to match the d -dimensional visual embedding space:

$$\hat{S}_N = \phi_s(S_N) = S_N W_s + b, \quad (5.3)$$

where $W_s \in \mathbb{R}^{3 \times d}$ and $b \in \mathbb{R}^d$ denote the learnable projection weights and bias, respectively, and $\hat{S}_N \in \mathbb{R}^{N \times d}$ represents the projected motor state sequence. Next, both modalities—the projected motor states $\hat{S}_N$ and the visual features F_N —are augmented with learnable positional embeddings to preserve temporal ordering and contextual alignment [135]:

$$\tilde{F}_N = F_N + E_F, \quad \tilde{S}_N = \hat{S}_N + E_S, \quad (5.4)$$

where $E_F \in \mathbb{R}^{N \times d}$ and $E_S \in \mathbb{R}^{N \times d}$ are two independent sets of learnable parameters. For the visual modality, the positional embedding E_F is applied uniformly across all patch embeddings within a given frame. That is, each frame in the buffered sequence receives a single temporal encoding that is shared by all of its P patch tokens, ensuring consistent temporal alignment while preserving the spatial structure within each frame.

Learnable positional embeddings are employed as the proposed model operates on fixed-length input sequences, enabling efficient position encoding while avoiding the additional computational overhead associated with dynamic or relative positional schemes. In contrast, sinusoidal positional

embeddings are typically preferred in variable-length settings, as they can generalize to arbitrary sequence lengths without retraining, whereas learnable embeddings are inherently tied to a predefined sequence length.

The temporally augmented modalities $\tilde{S}_N$ and $\tilde{F}_N$ are subsequently normalized using Layer Normalization (LN) and coupled through a masked causal cross-attention mechanism. Let $\tilde{S}_N \in \mathbb{R}^{N \times d}$ denote the query sequence and $\tilde{F}_N \in \mathbb{R}^{N \times P \times d}$ the visual token sequence. The visual tokens are flattened along the patch dimension to form the key-value sequence $K, V \in \mathbb{R}^{(NP) \times d}$. An additive causal attention mask $M \in \mathbb{R}^{N \times (NP)}$ enforces temporal causality by preventing queries at time index t from attending to tokens associated with future frames $t' > t$ [143]. Formally, the masked causal cross-attention operation is defined as:

$$\text{Attn}(Q, K, V; M) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right)V, \quad (5.5)$$

where $Q = \tilde{S}_N$ denotes the query sequence. The output of a multi-head attention (MHA) module with this attention mechanism is subsequently normalized using Layer Normalization, yielding the fused multimodal state representation:

$$H = \text{LN}\left(\text{MHA}(Q, K, V; M)\right), \quad (5.6)$$

where $H \in \mathbb{R}^{N \times d}$ denotes the embedding sequence obtained from the fusion of visual and motor-state modalities. This formulation allows controller-state queries at time index t to attend exclusively to visual patch tokens from the current and past frames, thereby preserving temporal causality while leveraging high-resolution spatial context.

The fused embeddings H are then processed by a causal transformer encoder to model temporal dependencies. The transformer consists of a stack of encoder layers, each comprising multi-head self-attention and feedforward submodules with *GELU* activations. To ensure autoregressive behavior, a temporal self-attention mask M^T is applied within each encoder layer (see Fig. 5.3). The output of this module, denoted as $\hat{H} \in \mathbb{R}^{N \times d}$, represents temporally contextualized embeddings that capture sequential structure in the catheter navigation process.

Next, the embeddings $\hat{H}$ are normalized and conditioned on the surgical goal. The goal representation is derived from a target image encoding the final catheter destination within the vascular system. This image is encoded using the pre-trained DINOv2 vision transformer, and only the [CLS] token is retained as the goal embedding $g \in \mathbb{R}^d$. The [CLS] token provides a compact global summary of the target image, avoiding patch-level redundancy while preserving semantic information relevant to the surgical objective.

The normalized temporal embeddings $\hat{H}$ are fused with the goal embedding using a gated goal fusion (GGF) mechanism. The goal embedding g is broadcast along the temporal dimension and concatenated with each timestep of $\hat{H}$. A gating function, parameterized by a linear projection followed by a sigmoid activation, regulates the relative contribution of temporal context and goal information:

$$G = \sigma([\hat{H} \parallel g]W + b), \quad (5.7)$$

where $[\hat{H} \parallel g] \in \mathbb{R}^{N \times 2d}$ denotes temporal concatenation, $W \in \mathbb{R}^{2d \times d}$ and $b \in \mathbb{R}^d$ are learnable parameters, and $\sigma(\cdot)$ is the element-wise sigmoid activation. The resulting fused representation $Z \in \mathbb{R}^{N \times d}$ is computed as

$$Z = G \odot \hat{H} + (1 - G) \odot g, \quad (5.8)$$

where $\odot$ denotes element-wise multiplication.

Finally, the goal-conditioned embeddings Z are passed to an action prediction head that maps the representation to controller commands for the robotic catheter platform. The action head is implemented as a multilayer perceptron (MLP) composed of successive linear layers with ReLU activations. To generate the output, only the final timestep of the sequence is used, yielding $\hat{a} = \text{MLP}(Z_N)$. This design reflects the autoregressive nature of the task, in which each prediction depends on accumulated temporal context and the surgical goal. The predicted action $\hat{a} \in \mathbb{R}^3$ corresponds to the controller commands for translation, rotation, and steering knob actuation. The model is trained in a supervised manner by minimizing the mean squared error (MSE) between the predicted action $\hat{a}$ and the ground truth controller state s_{t+1} .

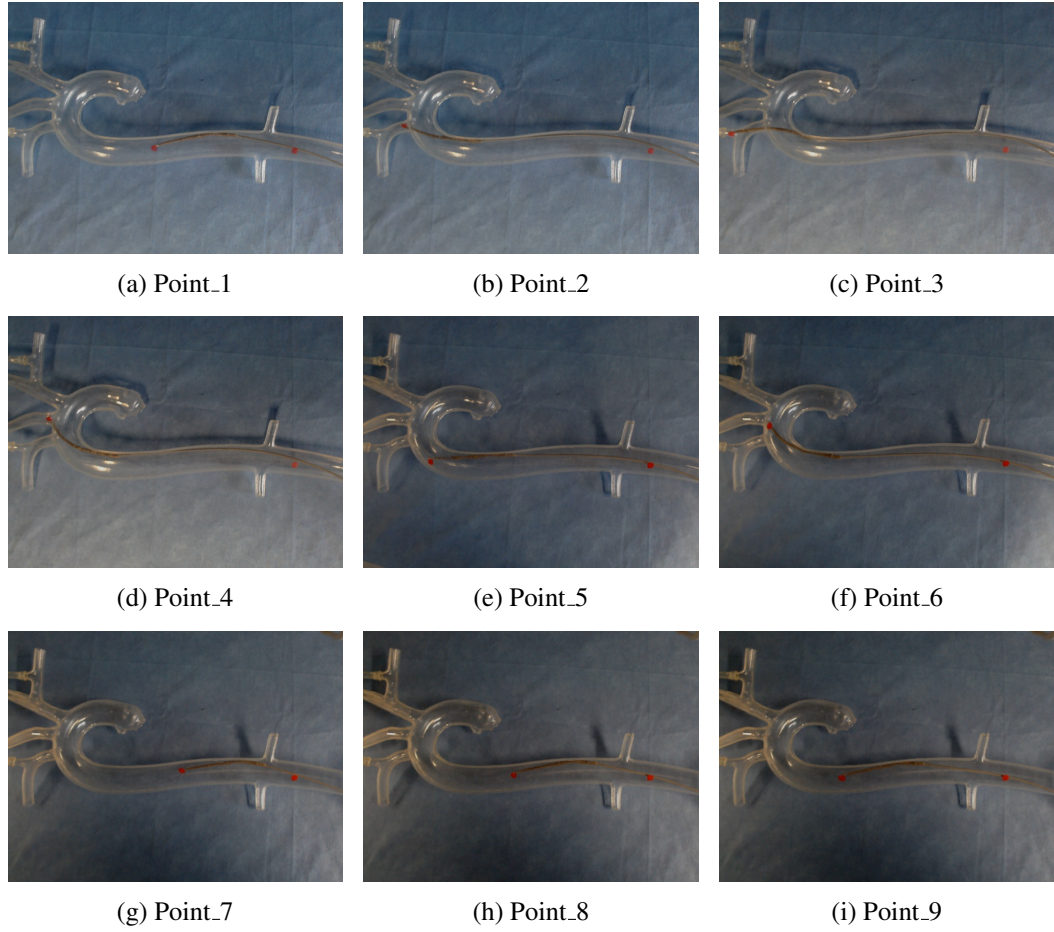


Figure 5.4: The figure shows nine predefined starting and ending points, indicated by red markers on the artificial vasculature. The images also illustrate representative visual inputs used to train and evaluate the DINO-CVA model. The starting location was kept identical across all experiments, while the target endpoints varied among the nine defined scenarios.

5.4 Results and Discussions

This section outlines the procedures employed for data preparation, including the construction of the training, validation, and test sets. It then describes the model configuration and training setup in detail. Finally, the evaluation results are presented and critically analyzed, with particular emphasis on ablation studies conducted to assess the contribution of individual components of the proposed approach.

5.4.1 Data Preparation and Model Configuration

Dataset Preparation: As outlined in Section 5.4.1, the datasets used in this study were compiled using the experimental setup described earlier, in which expert operators performed catheter steering tasks under teleoperation. Specifically, nine distinct tasks were defined, each characterized by a designated start and end point marked by two red reference points on the artificial vessel phantom (see Fig. 5.4). During each trial, the operator navigated the catheter while aiming to maintain the trajectory close to the vessel centerline and minimize collisions with the vessel walls.

The custom-developed interface software described in Section 5.2.2 simultaneously captured top-view video recordings and logged the corresponding joystick states. Each task was repeated five times, resulting in a dataset in which each data sample consists of an RGB image frame $I \in \mathbb{R}^{224 \times 224 \times 3}$ and an associated controller state vector $s \in \mathbb{R}^3$, representing translation, rotation, and steering knob actuation. For clarity, each defined task is referred to as a *scenario*, and each repetition of a scenario is referred to as an *episode*.

Based on the compiled dataset, two distinct training, validation, and testing protocols were established:

- **Episode-based:** In this setting, the training, validation, and test sets contain episodes drawn from all scenarios. As a result, the model is exposed to every scenario during training, while individual episodes differ across splits, providing variability in execution and operator behavior.
- **Scenario-based:** In this setting, the training, validation, and test sets are constructed such that no scenarios are shared across splits. Consequently, the model is evaluated on validation and test scenarios that are entirely unseen during training, enabling assessment of its generalization capability across distinct navigation tasks.

Table 5.1 summarizes the dataset statistics across motor states, scenarios, and episodes under the two partitioning strategies considered in this study. For example, under the scenario-based setting, the training set comprises 16,425 data points drawn from scenarios P1, P4, and P7, whereas in the episode-based setting, data points are sampled from all nine scenarios (P1–P9). Figure 5.5 illustrates the empirical distribution of motor states across both partitioning strategies, highlighting

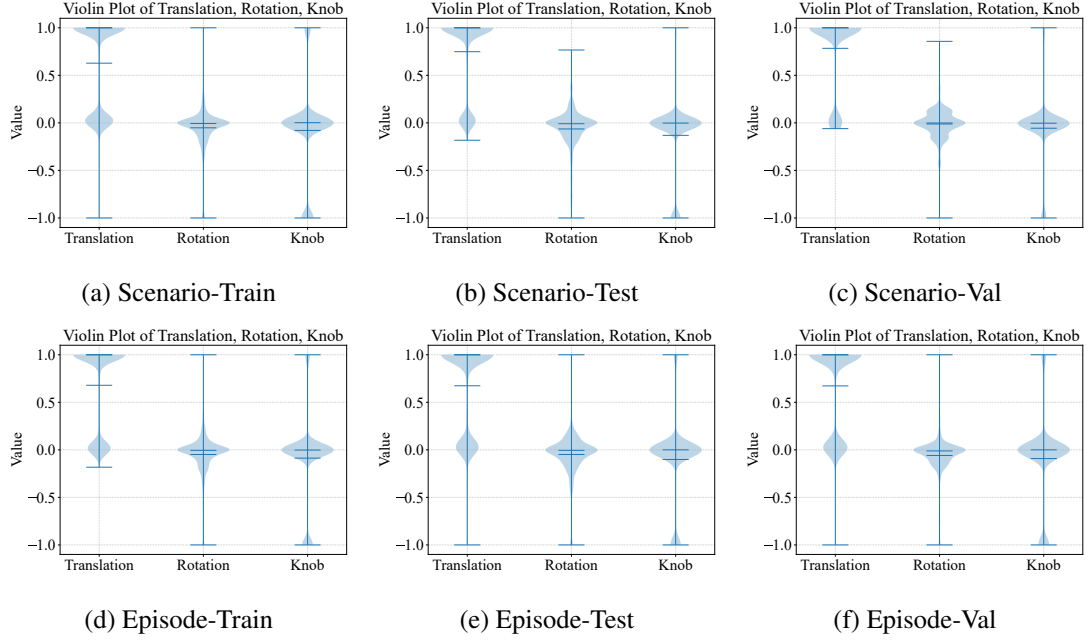


Figure 5.5: Violin plots illustrating the empirical distribution of motor states across both data-splitting protocols (episode-based and scenario-based) for the training, validation, and test sets. The plots highlight variability in catheter control inputs and reflect differences in control behavior across scenarios and execution episodes.

variability in catheter control inputs and corresponding motion patterns. To incorporate the temporal structure of the data, a sliding-window approach was employed. Each input sequence consisted of 50 consecutive time steps and was used to predict the subsequent controller state at the next timestep. The sequence length of 50 was selected as a practical trade-off, constrained primarily by available hardware resources. At each timestep t , the input RGB frame I_t is normalized using the mean and standard deviation statistics of the dataset on which DINOv2 was trained, ensuring compatibility with its pretrained feature representations [165]. The corresponding motor state vector s_t is standardized via z-scoring based on statistics computed from the training set. The prediction target s_{t+1} is retained in its original physical units to preserve interpretability and avoid post-hoc rescaling during inference.

Model and Training Configuration: DINO-CVA processes two input modalities: image sequences and joystick states. The image sequence F_N is encoded into $P = 257$ patch tokens per frame, each mapped to a $d = 384$ -dimensional embedding space using a frozen DINOv2-Small encoder. In parallel, the joystick state sequence S_N is linearly projected into the same 384-dimensional feature

Table 5.1: Data split statistics for scenario-based and episode-based datasets.

Type	Frames	Eps	Points	Var	Mean	Std	Min	Max
Train								
Scenario	16,425	25	P1-P4, P7	translation	0.63	0.46	-1.0	1.0
				rotation	-0.05	0.26	-1.0	1.0
				knob	-0.08	0.50	-1.0	1.0
Episode	15,705	27	P1–P9	translation	0.68	0.45	-0.18	1.0
				rotation	-0.05	0.24	-1.0	1.0
				knob	-0.09	0.47	-1.0	1.0
Val								
Scenario	2,680	5	P5	translation	0.78	0.40	-0.06	1.0
				rotation	-0.01	0.17	-1.0	0.85
				knob	-0.06	0.34	-1.0	1.0
Episode	5,276	9	P1–P9	translation	0.67	0.45	-1.0	1.0
				rotation	-0.06	0.22	-1.0	1.0
				knob	-0.09	0.47	-1.0	1.0
Test								
Scenario	7,139	15	P6, P8, P9	translation	0.75	0.42	-0.18	1.0
				rotation	-0.06	0.22	-1.0	0.77
				knob	-0.13	0.40	-1.0	1.0
Episode	5,263	9	P1–P9	translation	0.67	0.46	-1.0	1.0
				rotation	-0.05	0.26	-1.0	1.0
				knob	-0.10	0.45	-1.0	1.0

space, yielding the aligned representation $\hat{S}_N$. The input sequence length is set to $N = 50$, resulting in a total of $257 \times 50 = 12,850$ visual patch tokens per sequence.

To incorporate temporal dependencies, two distinct sets of learnable positional embeddings are defined: one for the video frames and one for the joystick states. For each frame f_i , the positional embedding corresponding to its temporal index is applied uniformly across all 257 patch tokens, thereby encoding temporal order while preserving the spatial structure captured by the vision transformer. A single-layer masked multi-head cross-attention module with 8 attention heads is then employed, where the encoded image features serve as keys and values, and the projected joystick states act as queries to enable multimodal fusion (see (5.5)). The resulting fused embeddings are subsequently passed to a causal transformer composed of four layers, each equipped with 8 attention heads and a feed-forward network of hidden dimension 1024. This configuration enables the model to capture long-range temporal dependencies while preserving the autoregressive structure required for sequence prediction.

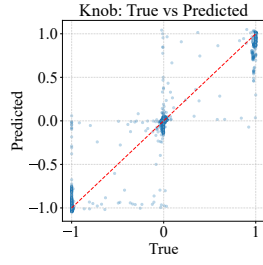
Subsequently, the Gated Goal Fusion (GGF) module incorporates goal information by conditioning the temporal representation $\hat{H}$ on the [CLS] token of the goal image, denoted as $g \in \mathbb{R}^{384}$,

Table 5.2: Performance comparison of the proposed model and LSTM baseline on episode-based and scenario-based datasets.

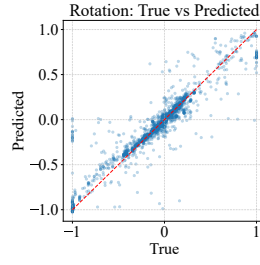
Episode-based dataset					
Model	Task	MSE	RMSE	MAE	R^2
DINO-CVA	Translation	0.0179	0.1338	0.0472	0.9145
	Rotation	0.0091	0.0955	0.0356	0.8676
	Knob	0.0094	0.0970	0.0279	0.9556
LSTM-based	Translation	0.0160	0.1264	0.0448	0.9237
	Rotation	0.0085	0.0922	0.0335	0.8765
	Knob	0.0115	0.1075	0.0311	0.9455
Scenario-based dataset					
Model	Task	MSE	RMSE	MAE	R^2
DINO-CVA	Translation	0.0116	0.1078	0.0369	0.9320
	Rotation	0.0050	0.0708	0.0354	0.8858
	Knob	0.0187	0.1367	0.0490	0.9116
LSTM-based	Translation	0.0125	0.1120	0.0416	0.9266
	Rotation	0.0050	0.0708	0.0311	0.8859
	Knob	0.0133	0.1154	0.0289	0.9158

extracted using the frozen DINOv2 encoder (see (5.7) and (5.8)). The final goal-conditioned representation Z_N is forwarded to an action prediction head implemented as a multilayer perceptron (MLP) consisting of four fully connected layers with output dimensions 512, 256, 128, and 3, respectively, to regress the next-step joystick state s_{t+1} .

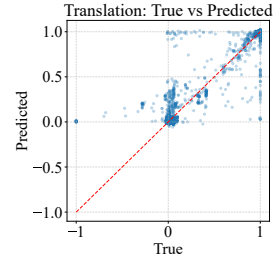
The network with the above configuration was trained separately under the episode-based and scenario-based dataset settings. Training was performed with a batch size of 16 for up to 50 epochs, with validation conducted after each epoch. The model was optimized using the Adam optimizer with an initial learning rate of 9.0×10^{-5} , regulated by a cosine annealing learning-rate scheduler. The mean squared error (MSE) loss function was employed, and early stopping based on validation performance was applied to mitigate overfitting. All experiments were implemented in PyTorch and executed on an NVIDIA RTX 2080 GPU with 8 GB of memory. The complete model contains 22.06 million parameters, of which 6.82 million are trainable.



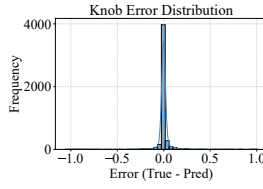
(a) Knob (True vs Predicted)



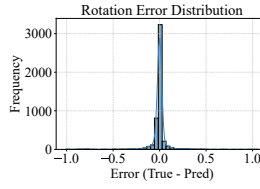
(b) Rotation (True vs Predicted)



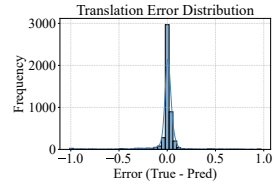
(c) Translation (True vs Predicted)



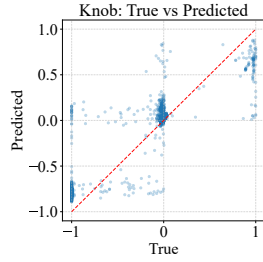
(d) Knob Error Distribution



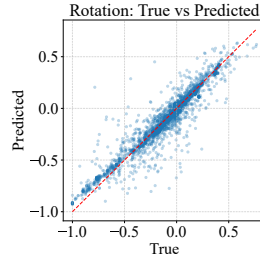
(e) Rotation Error Distribution



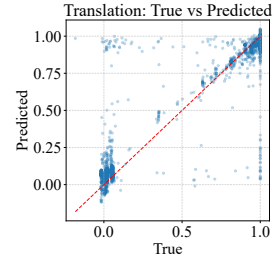
(f) Translation Error Distribution



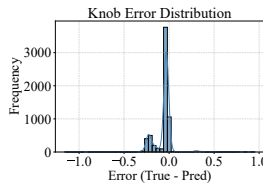
(g) Knob (True vs Predicted)



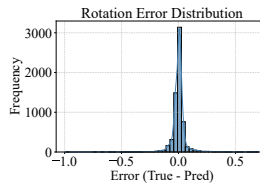
(h) Rotation (True vs Predicted)



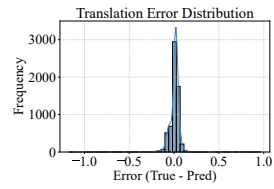
(i) Translation (True vs Predicted)



(j) Knob Error Distribution



(k) Rotation Error Distribution



(l) Translation Error Distribution

Figure 5.6: True–predicted plots and error histograms for knob, rotation, and translation control using DINO-CVA. Episode-based results are shown in (a–f): (a) knob true–predicted, (b) rotation true–predicted, (c) translation true–predicted, (d) knob error distribution, (e) rotation error distribution, and (f) translation error distribution. Scenario-based results are shown in (g–l): (g) knob true–predicted, (h) rotation true–predicted, (i) translation true–predicted, (j) knob error distribution, (k) rotation error distribution, and (l) translation error distribution. The true–predicted plots illustrate alignment between predicted and ground-truth motor states, while the histograms depict the distribution of prediction errors across tasks.

5.4.2 Results

As described earlier, DINO-CVA was trained under two distinct dataset partitioning strategies and evaluated on their corresponding test sets to assess its ability to steer the catheter by predicting joystick states toward partially unseen and entirely unseen target locations. Since the controller states s_{t+1} are generated in an autoregressive manner, model performance was evaluated using standard regression metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

To contextualize the results, a kinematics-only behavior cloning baseline based on a Long Short-Term Memory (LSTM) network—previously employed in related catheter navigation studies [87, 88, 77, 82]—was implemented. This baseline enables assessment of whether the proposed multimodal framework can achieve comparable predictive accuracy while additionally incorporating visual information. Table 5.2 summarizes the quantitative performance of DINO-CVA across both episode-based and scenario-based dataset partitions.

Across both evaluation settings, the results indicate consistently strong performance for both models. In the episode-based evaluation, translation and rotation exhibited low error values and high R^2 scores for both approaches, while steering knob prediction remained the most challenging degree of freedom but still achieved reliable accuracy. A similar trend was observed in the scenario-based evaluation, where both models maintained stable performance with only modest variations across metrics. These findings confirm that purely temporal modeling of kinematic sequences, as performed by the LSTM baseline, provides a strong reference for short-term joystick state prediction.

Importantly, the objective of the proposed model is not to surpass the LSTM baseline in terms of raw error values, but rather to demonstrate that comparable performance can be achieved while jointly modeling visual and kinematic information within a unified embedding space. Unlike the LSTM baseline, which relies solely on temporal continuity of joystick inputs, DINO-CVA learns a vision-aware and kinematics-aware policy that incorporates spatial context from the environment alongside control history. As a result, while the LSTM may accurately predict control commands

when kinematic patterns are consistent, it lacks explicit awareness of the catheter’s spatial relationship to the vasculature and target anatomy.

The central distinction of the proposed framework therefore lies in the representation it learns. By projecting visual observations and kinematic sequences into a joint embedding space, the model integrates spatial context and temporal dynamics to associate expert control behavior with the vascular environment in which it occurs. This capability is particularly important for autonomous navigation tasks, where successful control requires not only reproducing joystick sequences but also maintaining awareness of the catheter’s position and intended trajectory within the vasculature. Consequently, although the quantitative performance metrics appear similar, the proposed approach represents a necessary step toward embedding trajectory awareness into the control process—an aspect that purely kinematic models inherently lack.

Finally, Fig. 5.6 presents true-versus-predicted plots along with error histograms for translation, rotation, and steering knob states under both dataset settings. These visual analyses complement the quantitative results by illustrating the distribution and stability of prediction errors, further demonstrating that the proposed model produces reliable and well-bounded predictions. Also, these results validate DINO-CVA as a promising initial framework toward vision-guided autonomous navigation, matching kinematics-only baselines while introducing anatomical grounding—setting the stage for subsequent closed-loop integrations and clinical translations. In the following section, ablation study results are presented to examine the contribution of the goal-conditioning mechanism and multimodal fusion strategy, thereby clarifying their impact on overall performance.

5.4.3 Ablation Study

It is important to note that the proposed framework is grounded in imitation learning, specifically behavior cloning, which is inherently limited in its ability to extrapolate beyond the training distribution. While this limitation cannot be fully eliminated, the incorporation of explicit goal conditioning is intended to improve task-directed generalization by anchoring control predictions to a desired anatomical target.

To further investigate the contribution of goal conditioning and multimodal fusion, an ablation

Table 5.3: Ablation study on the scenario-based dataset. The table reports the performance of the model and its variants when goal or modality components are removed.

Condition	Task	MSE	RMSE	MAE	R^2
Baseline	Translation	0.0116	0.1078	0.0369	0.9320
	Rotation	0.0050	0.0708	0.0354	0.8858
	Knob	0.0187	0.1367	0.0490	0.9116
False goal	Translation	0.1134	0.3368	0.2996	0.3364
	Rotation	0.0113	0.1061	0.0637	0.7437
	Knob	0.0616	0.2483	0.1333	0.6102
No Goal	Translation	0.0321	0.1792	0.1461	0.8120
	Rotation	0.0078	0.0881	0.0669	0.8234
	Knob	0.0456	0.2136	0.1395	0.7114
No vision	Translation	0.6098	0.7809	0.6943	-2.5684
	Rotation	0.0447	0.2113	0.1322	-0.0163
	Knob	0.1890	0.4347	0.2044	-0.1951
No states	Translation	0.4273	0.6537	0.6108	-1.5007
	Rotation	0.0580	0.2408	0.1599	-0.3197
	Knob	0.1884	0.4341	0.1984	-0.1916

study was conducted on the scenario-based dataset, which represents the most challenging evaluation setting by requiring generalization to entirely unseen target destinations (Table 5.3). The first row reports the baseline performance of DINO-CVA, as previously presented in Table 5.2. The impact of goal conditioning was first examined. When an incorrect goal image—corresponding to a target inconsistent with the intended destination—was provided, performance degraded substantially across all control dimensions, with R^2 values dropping to 0.336 for translation, 0.744 for rotation, and 0.610 for steering knob control. This pronounced degradation highlights the central role of accurate goal specification in guiding the model’s predictions, as misleading goal embeddings disrupt the alignment between visual context and control actions.

When the goal-conditioning head was entirely removed, performance also deteriorated, though less severely than in the false-goal condition (e.g., $R^2 = 0.812$ for translation and $R^2 = 0.711$ for knob control). This result indicates that, in the absence of explicit goal information, the model partially relies on temporal patterns learned from demonstrations but lacks sufficient guidance to consistently align its predictions with unseen target destinations. The observed performance gap

confirms that goal conditioning contributes positively to trajectory planning, particularly for fine-grained control such as steering knob actuation.

The final two ablations evaluated the role of multimodal fusion by removing either the visual or the state modality. Eliminating visual input led to a severe collapse in performance, particularly for translation ($R^2 = -2.57$), indicating that the model’s predictions became substantially worse than a mean-value baseline. This result underscores the indispensability of visual feedback for spatial awareness and trajectory alignment. Similarly, removing motor state inputs resulted in large performance degradations across all tasks (e.g., $R^2 = -1.50$ for translation), demonstrating that kinematic information is also essential for effective control.

Overall, the ablation results confirm that the two modalities are both necessary and complementary: visual information provides global spatial context, while joystick states encode dynamic control history. Moreover, the goal-conditioning mechanism is shown to be critical for accurate navigation toward unseen destinations, with its removal or mis-specification leading to pronounced performance degradation. These findings collectively validate the architectural design of DINO-CVA and emphasize that the combination of multimodal fusion and goal conditioning is a key factor in enabling robust catheter steering under novel and previously unseen conditions.

5.4.4 Limitations

This study represents a step toward exploring the feasibility of multimodal, goal-conditioned architectures for autonomous catheter navigation, rather than presenting a definitive or clinically deployable solution. Several limitations should therefore be acknowledged.

First, the proposed model employs fixed-length, learnable positional embeddings, which constrains evaluation to sequences of predefined temporal length. While suitable for the current experimental setup, this design choice limits assessment of model behavior under variable-duration trajectories. Exploring alternative positional encoding strategies, such as sinusoidal or relative encodings, would be essential for evaluating robustness to temporal variations and differing procedural lengths.

Second, the goal representation is incorporated using the [CLS] token of the goal image, providing a compact global summary of the target location. Although this approach is effective, it

does not explicitly encode spatial relationships within the goal image. More expressive goal representations, such as spatially resolved embeddings or region-specific conditioning mechanisms, could further enhance the model’s ability to reason about fine-grained anatomical structure during navigation.

Third, the evaluation setup is limited by the available vascular phantom. While sufficient for controlled experimentation, the phantom represents a restricted anatomical configuration. The impact of visual awareness and goal conditioning should be examined in more diverse vascular geometries with increased branching complexity to better assess the model’s ability to disambiguate trajectories when multiple viable paths are present. Incorporating explicit destination cues—such as target coordinates or visual markers (e.g., highlighted regions or reference points in the goal image)—may further strengthen goal-directed navigation in such settings.

Fourth, the proposed framework is trained and evaluated under a teacher-forced imitation learning setting, in which model performance is assessed based on one-step-ahead prediction accuracy against expert demonstrations. While this evaluation protocol is standard in behavior cloning and facilitates stable and reproducible comparison across models, it does not explicitly capture the accumulation of prediction errors during long-horizon, closed-loop execution. In practical autonomous navigation, small per-step deviations may compound over time due to covariate shift between the training distribution and the states encountered during autonomous policy rollouts. Consequently, the reported metrics primarily reflect local prediction accuracy rather than long-term control stability. A systematic analysis of cumulative error propagation, closed-loop rollouts, and goal-reaching performance under autonomous execution was therefore beyond the scope of the present study and is left for future investigation.

Finally, training and evaluation were conducted using synthetic RGB data generated from the phantom setup, whereas clinical catheterization procedures are guided primarily by X-ray fluoroscopy. Extending the proposed framework to synthetic X-ray data would provide a more realistic representation of clinical conditions and constitute an important step toward bridging the gap between laboratory validation and real-world applicability.

5.5 Conclusion

This chapter introduced DINO-CVA, a multimodal, goal-conditioned behavior cloning framework for autonomous catheter navigation. By jointly embedding visual observations and kinematic control signals, the proposed model learns expert demonstrations in a manner that integrates both trajectory history and anatomical context. Training and evaluation on data collected from a robotic experimental setup with a synthetic vascular phantom demonstrated that the model achieves high predictive accuracy across both episode-based and scenario-based datasets. While exhibiting performance comparable to a kinematics-only LSTM baseline, DINO-CVA additionally grounds its predictions in visual context, enabling spatially informed decision-making.

Ablation studies further revealed that the combination of goal conditioning and multimodal representation learning plays a critical role in model behavior, with particularly pronounced benefits observed for fine-grained steering knob control. These results highlight the feasibility of multimodal, goal-conditioned architectures as an important step toward enhancing robustness and contextual awareness in catheter navigation, beyond purely kinematic sequence modeling.

Future work will focus on extending the framework to synthetic X-ray data, incorporating richer and more spatially expressive goal representations, evaluating performance in more complex vascular phantoms, and integrating simulation environments to complement the robotic platform for scalable training and systematic evaluation.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This dissertation addressed the challenges of perception, force estimation, and autonomy in cardiac catheter-based interventions through a series of learning-based, sensor-free frameworks. Motivated by the clinical importance of accurately sensing catheter-tissue interaction forces, reliably visualizing the catheter under fluoroscopic imaging, and enabling precise navigation within complex vascular environments, the overarching objective of this research was to establish unified methodologies that leverage standard clinical imaging without requiring specialized hardware or explicit mechanical modeling. The work was conducted as a multidisciplinary study at the intersection of artificial intelligence, deep learning-based computer vision, medical robotics, mechatronics, and autonomous systems.

The dissertation began by establishing the clinical and technical foundations of interventional cardiac catheterization. Chapter 1 reviewed minimally invasive surgery and its applications, with a focused discussion on catheter-based cardiovascular interventions. Through an extensive analysis of existing force sensing and navigation methods, this chapter identified fundamental limitations in current clinical workflows, including the absence of direct force feedback, reliance on handcrafted features, and limited integration between perception and control. These observations motivated the formulation of the central research objectives: to develop sensor-free force estimation frameworks compatible with standard catheters and to advance toward autonomous, vision-based robotic

catheter navigation.

The first technical contribution of the dissertation, presented in Chapter 2, investigated whether three-dimensional catheter–tissue contact forces could be inferred directly from visual information. This work introduced Y-Net, a lightweight late-fusion convolutional neural network designed to map stereo catheter images to contact forces along the x , y , and z axes. The primary objective of this chapter was to demonstrate that catheter deflections observed from multiple viewpoints encode sufficient information for accurate force estimation without relying on embedded sensors or explicit geometric modeling. Experimental validation confirmed that the proposed approach achieves reliable force estimation while remaining computationally efficient, establishing the feasibility of vision-based force sensing in catheterization.

Building upon this foundation, the dissertation next addressed the integration of catheter perception and force estimation within a single end-to-end framework in Chapter 3. This stage of the work introduced H-Net, the first convolutional multitask architecture capable of simultaneously performing catheter segmentation and three-dimensional force estimation. The objective of this chapter was to eliminate segmentation as a separate preprocessing step and to examine whether joint learning could improve force estimation performance. To support systematic evaluation, a synthetic X-ray image generator was developed to transform RGB images into fluoroscopy-like imagery with varying background complexity. The results demonstrated that multitask learning improves robustness and generalization across datasets while maintaining low computational overhead suitable for real-time applications.

The dissertation then advanced beyond convolutional architectures by introducing a transformer-based multitask framework in Chapter 4. This work culminated in TransForSeg, a multi-input, multi-output Vision Transformer encoder–decoder designed for joint stereo catheter segmentation and three-dimensional force estimation. The objective of this chapter was to exploit global self-attention and cross-attention mechanisms to capture long-range dependencies and complex catheter geometries that are difficult to model using convolutional operations alone. Extensive experimental evaluations and ablation studies demonstrated that TransForSeg achieves state-of-the-art performance while preserving efficiency, and that segmentation plays a critical role in guiding the force estimation process, particularly under domain shifts and imaging perturbations.

The final technical contribution of this dissertation focused on autonomous catheter navigation in Chapter 5. This work introduced DINO-CVA, a goal-conditioned, vision-based control framework that directly maps sequential visual observations and kinematic cues to catheter actuation commands. The objective of this chapter was to demonstrate that perception-driven navigation can be learned end-to-end without relying on explicit detection, localization, or path planning modules. Trained using imitation learning and validated in a physical robotic setup, the proposed framework successfully steered a catheter toward designated targets using only imaging feedback, representing an important step toward closed-loop autonomous catheterization.

Taken together, the contributions of this dissertation demonstrate that perception, force estimation, and navigation can be jointly learned within unified, sensor-free frameworks grounded in visual information. By progressively integrating convolutional neural networks, multitask learning, transformer-based architectures, and imitation learning, this work establishes a coherent pathway from image-based perception to autonomous robotic control in cardiac catheterization. The proposed methodologies advance the state-of-the-art in learning-based catheter force estimation and robotic navigation while remaining compatible with existing clinical imaging infrastructures.

6.2 Future Work

Several research directions naturally emerge from this dissertation, building upon the proposed learning-based frameworks for catheter perception, force estimation, and autonomous navigation. A primary avenue for future work involves extending the current force estimation models to more accurately capture the physical characteristics of real clinical catheters. In particular, incorporating the influence of catheter tendons represents an important step toward improved force modeling. Tendon-driven mechanisms introduce additional internal forces and constraints that significantly affect catheter deformation, stiffness, and tip behavior during steering and tissue interaction. Explicitly modeling tendon actuation within learning-based frameworks, either through augmented input representations or hybrid physics-informed architectures, could enhance force estimation accuracy and improve generalization across different catheter designs and manipulation strategies.

Beyond localized force estimation at the catheter tip, future research may extend the proposed

methods to estimate distributed contact forces along the catheter body. In many interventional scenarios, the catheter interacts with the vessel wall at multiple locations simultaneously, resulting in complex force distributions along both the shaft and the tip. Estimating multi-point contact forces along the catheter body would provide a more comprehensive understanding of catheter–tissue interactions and enable advanced safety-aware behaviors, such as contact-aware navigation and force-limited motion planning. Addressing this challenge may require spatiotemporal modeling of catheter geometry, sequential representations of catheter deformation, or graph-based formulations that capture continuous contact along the catheter length.

For autonomous catheter navigation, future work may investigate learning paradigms that extend beyond imitation learning toward higher levels of autonomy. In particular, the exploration of Vision–Language–Action (VLA) models represents a promising direction for integrating perception, reasoning, and control within a unified framework. Such models could leverage Multimodal Large Language Models (MLLMs) to interpret pre-operative imaging data, including X-ray or fluoroscopic images, and generate semantic descriptions of anatomical structures relevant to the intervention. In addition, patient-specific clinical history and laboratory measurements could be incorporated to contextualize the procedure, enabling personalized and risk-aware decision making. By fusing visual information with structured clinical data, MLLMs could generate high-level procedural plans and operation descriptions prior to intervention, which may then serve as structured inputs to VLA policies responsible for executing navigation and manipulation tasks during surgery.

Realizing this capability requires the development of a surgical world model that captures anatomical geometry, device dynamics, and procedural context across time. Pretraining such a world model on large-scale multimodal surgical data, encompassing imaging, kinematics, and clinical metadata, could enable predictive reasoning, long-horizon planning, and adaptive decision-making during interventions. This world model could support simulation-based pretraining of control policies, uncertainty-aware planning, and online adaptation during surgery.

In the longer term, integrating MLLM-driven planning, VLA-based execution, and force-aware

perception within a hierarchical control architecture may enable a new class of autonomous surgical systems. Such systems could generalize across multiple interventional tasks while maintaining safety and interpretability. When combined with humanoid robotic platforms, these capabilities may support flexible deployment in diverse clinical environments, extending autonomy beyond task-specific catheterization toward multipurpose surgical assistance. While substantial challenges remain, including safety validation, regulatory approval, and clinical acceptance, the methodologies developed in this dissertation provide a foundational step toward this broader vision of autonomous, intelligent surgical robotics.

Bibliography

- [1] Pedram Fekri, Hamid Reza Nourani, Masoud Razban, Javad Dargahi, Mehrdad Zadeh, and Ahmadreza Arshi. A deep learning force estimator system for intracardiac catheters. In *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE, 2021.
- [2] Kamran Mohiuddin and Scott J Swanson. Maximizing the benefit of minimally invasive surgery. *J. Surg. Oncol.*, 108(5):315–319, October 2013.
- [3] Sejal S Singh and Raju K Shinde. Minimally invasive gastrointestinal surgery: A review. *Cureus*, 15(11):e48864, November 2023.
- [4] Ibrahim Alkatout, Ulrich Mechler, Liselotte Mettler, Julian Pape, Nicolai Maass, Matthias Biebl, Georgios Gitas, Antonio Simone Laganà, and Damaris Freytag. The development of laparoscopy—a historical overview. *Frontiers in Surgery*, Volume 8 - 2021, 2021.
- [5] Tianci Chai, Yuhan Lin, Mingqiang Kang, and Jiangbo Lin. Thoracotomy versus video-assisted thoracoscopic resection of lung cancer: A protocol for a systematic review and meta-analysis. *Medicine*, 98(10), 2019.
- [6] Ilaria Bove, Domenico Solari, Teresa Somma, Riccardo Nocini, Carmela Palmiero, Felice Esposito, Paolo Cappabianca, and Luigi Maria Cavallo. Endoscopic endonasal approach for craniopharyngiomas with intraventricular extension: Anatomic-Clinical considerations and surgical outcomes in a series of 61 patients. *Operative Neurosurgery*, 9900.
- [7] Nathaniel B Langer and Michael Argenziano. Minimally invasive cardiovascular surgery: Incisions and approaches. *Methodist Debaquey Cardiovasc. J.*, 12(1):4–9, January 2016.

- [8] Giulia Guglielmi, Kaushiga Krishnathan, Andrew Constantine, and Konstantinos Dimopoulos. Cardiac catheterization in pulmonary arterial hypertension: Tips and tricks to enhance diagnosis and guide therapy. *International Journal of Cardiology Congenital Heart Disease*, 17:100527, 2024.
- [9] Del Rio-Pertuz, G Nugent, and K Argueta-Sosa. Right heart catheterization in clinical practice: a review of basic physiology and important issues relevant to interpretation. *American journal of cardiovascular disease*, 13(3):122–137, 2023.
- [10] Matias B. Yudi, Samin K. Sharma, Gilbert H.L. Tang, and Annapoorna Kini. Coronary angiography and percutaneous coronary intervention after transcatheter aortic valve replacement. *JACC*, 71(12):1360–1378, 2018.
- [11] Jeffrey Zhong, Nicholas Kamp, Agam Bansal, Ani Kumar, Rishi Puri, Amar Krishnaswamy, Samir Kapadia, and Grant W Reed. Balloon aortic valvuloplasty in the modern era: A review of outcomes, indications, and technical advances. *Journal of the Society for Cardiovascular Angiography Interventions*, 2(4), July 2023.
- [12] Michael L O’Byrne, Andrew C Glatz, and Matthew J Gillespie. Transcatheter device closure of atrial septal defects: more to think about than just closing the hole. *Curr. Opin. Cardiol.*, 33(1):108–116, January 2018.
- [13] Sripal Bangalore, Gregory W. Barsness, George D. Dangas, Morton J. Kern, Sunil V. Rao, Linda Shore-Lesserson, Jacqueline E. Tamis-Holland, Thrombosis on behalf of the American Heart Association Interventional Cardiovascular Care Committee of the Council on Clinical Cardiology; Council on Arteriosclerosis, Vascular Biology; Council on Cardiovascular, Stroke Nursing; Council on Cardiovascular Surgery, Anesthesia; Council on Lifestyle, and Cardiometabolic Health. Evidence-based practices in the cardiac catheterization laboratory: A scientific statement from the american heart association. *Circulation*, 144(5):e107–e119, 2021.
- [14] Jelena Kornej, Christin S. Börschel, Emelia J. Benjamin, and Renate B. Schnabel. Epidemiology of atrial fibrillation in the 21st century. *Circulation Research*, 127(1):4–20, 2020.

- [15] Gregory J. Dehmer, Douglas Weaver, Matthew T. Roe, Sarah Milford-Beland, Susan Fitzgerald, Anthony Hermann, John Messenger, Issam Moussa, Kirk Garratt, John Rumsfeld, and Ralph G. Brindis. A contemporary view of diagnostic cardiac catheterization and percutaneous coronary intervention in the united states: A report from the cathpci registry of the national cardiovascular data registry, 2010 through june 2011. *Journal of the American College of Cardiology*, 60(20):2017–2031, 2012.
- [16] Chi-Hsun Lien, Kung-Cheng Hsu, Yu-Kang Chang, Chien-Jung Chang, Ying-Ming Chiu, and Hung-Yi Hsu. Secular trends in the incidence of atrial fibrillation across different ages and sexes in taiwan from 2001 to 2021. *Heart Rhythm*, 22(8):e294–e300, 2025.
- [17] Kipp Slicker, Wesley G. Lane, Ola O. Oyetayo, Laurel A. Copeland, Eileen M. Stock, Jeffrey B. Michel, and John P. Erwin. Daily cardiac catheterization procedural volume and complications at an academic medical center. *Cardiovascular Diagnosis and Therapy*, 6(5), 2016.
- [18] Olurotimi J Badero, Bamikole Osibowale, Ayobami O Kuyoro, Adedeji Adebayo, Francis C Asogwa, Oyinkansola T Agaja, Loveth C Okonkwo, Olusegun D Alaga, and Olutomiwa Omokore. One-year review of cardiac catheterization procedures at a new private cardiac facility in nigeria: Indications, techniques, and outcomes. *Cureus*, 17(6):e86719, June 2025.
- [19] Alexander Rolls and Celia Riga. Endovascular robotics. *Ann. R. Coll. Surg. Engl.*, 100(Suppl 7):14–17, September 2018.
- [20] William Crinnion, Ben Jackson, Avnish Sood, Jeremy Lynch, Christos Bergeles, Hongbin Liu, Kawal Rhode, Vitor Mendes Pereira, and Thomas C Booth. Robotics in neurointerventional surgery: a systematic review of the literature. *Journal of NeuroInterventional Surgery*, 14(6):539–545, 2022.
- [21] Eyal Morag, Francois H. Cornelis, Giora Weisz, and Ripal Gandhi. Overcoming barriers and advancements in endovascular robotics: A review of systems and developments. *Techniques in Vascular and Interventional Radiology*, 26(3):100918, 2023.

- [22] Tejas M Patel, Sanjay C Shah, and Samir B Pancholy. Long distance Tele-Robotic-Assisted percutaneous coronary intervention: A report of First-in-Human experience. *eClinicalMedicine*, 14:53–58, September 2019.
- [23] Vitor Mendes Pereira, Hal Rice, Laetitia De Villiers, Nader Sourour, Frédéric Clarencon, Julian Spears, Alejandro Tomasello, David Hernandez, Nicole M Cancelliere, Xiao Yu Eileen Liu, Patrick Nicholson, Vincent Costalat, Gregory Gascou, Pasquale Mordasini, Jan Gralla, Mario Martínez-Galdámez, Jorge Galvan Fernandez, Monika Killer-Oberpfalzer, David S Liebeskind, Raymond D Turner, Raphael Blanc, and Michel Piotin. Evaluation of effectiveness and safety of the corpath grx robotic system in endovascular embolization procedures of cerebral aneurysms. *Journal of NeuroInterventional Surgery*, 16(4):405–411, 2024.
- [24] Carlo Pappone, Gabriele Vicedomini, Francesco Manguso, Filippo Gugliotta, Patrizio Mazzone, Simone Gulletta, Nicoleta Sora, Simone Sala, Alessandra Marzi, Giuseppe Augello, Laura Livolsi, Andreina Santagostino, and Vincenzo Santinelli. Robotic magnetic navigation for atrial fibrillation ablation. *JACC*, 47(7):1390–1400, 2006.
- [25] Zohaib A Shaikh, Michael F Eilenberg, and Todd J Cohen. The amigo™ remote catheter system: From concept to bedside. *J. Innov. Card. Rhythm Manag.*, 8(8):2795–2802, August 2017.
- [26] Linshuai Zhang, Jinshan Zuo, Ke Wang, Tao Jiang, Shuoxin Gu, Lin Xu, and Yujie Zhang. An advanced robotic system incorporating haptic feedback for precision cardiac ablation procedures. *Scientific Reports*, 15(1):6839, February 2025.
- [27] Gediminas Rackauskas, Ricardas Kundelis, Audrius Aidietis, Jurate Barysiene, Diana Sudaviciene, Justinas Bacevicius, Monika Kezeviciute, Germanas Marinskis, Niels Christian Foldager Sandgaard, Christian Jøns, Xu Chen, and Peter Karl Jacobsen. First in-human results of the MAGiC robotic magnetic navigation radiofrequency ablation catheter. *Journal of Interventional Cardiac Electrophysiology*, June 2025.
- [28] 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.
- [29] R West, G Ellis, N Brooks, and Joint Audit Committee of the British Cardiac Society and Royal College of Physicians of London. Complications of diagnostic cardiac catheterisation: results from a confidential inquiry into cardiac catheter complications. *Heart*, 92(6):810–814, June 2006.
- [30] Jacob A Doll, Ravi S Hira, Kathleen E Kearney, David E Kandzari, Robert F Riley, Steven P Marso, James A Grantham, Craig A Thompson, James M McCabe, Dimitrios Karpaliotis, Ajay J Kirtane, and William Lombardi. Management of percutaneous coronary intervention complications: Algorithms from the 2018 and 2019 seattle percutaneous coronary intervention complications conference. *Circ. Cardiovasc. Interv.*, 13(6):e008962, June 2020.
- [31] Claudio Godino and Antonio Colombo. Complications of percutaneous coronary intervention. In *Pan Vascular Medicine*, pages 2297–2322. Springer, Germany, 2002.
- [32] Ehtisham Mahmud, Ryan D. Madder, David H. Wohns, Jeffrey M. Schussler, Adam Salisbury, Paul Campbell, Tejas M. Patel, William L. Lombardi, William J. Nicholson, Manish A. Parikh, Natia Kelm, Ron Waksman, Alexandra J. Lansky, and Giora Weisz. Robotic-assisted percutaneous coronary intervention: Final results of the precision and precision grx studies. *Journal of the Society for Cardiovascular Angiography Interventions*, 4(7):103655, 2025.
- [33] C.B. Beaman, N. Kaneko, P.M. Meyers, and S. Tateshima. A review of robotic interventional neuroradiology. *American Journal of Neuroradiology*, 2021.
- [34] Saket Pradhan, Dennis Kundrat, and Giulio Dagnino. A ferrofluid-based haptic guidance system for robot-assisted endovascular procedures. *Journal of Robotic Surgery*, 19(1):226, May 2025.
- [35] Kriengsak Masnok and Nobuo Watanabe. Relationship of catheter contact angle and contact force with contact area on the surface of heart muscle tissue in cardiac catheter ablation. *Cardiovasc. Eng. Technol.*, 12(4):407–417, August 2021.

- [36] Eric Rasmussen, Daniel Guo, Vybhav Murthy, Rachit Mishra, Cameron Riviere, and Carmel Majidi. A soft resistive sensor with a semicircular cross-sectional channel for soft cardiac catheter ablation. *Sensors*, 21(12), 2021.
- [37] Anouchka Autar, Aladdin Taha, Richard van Duin, Ilona Krabbendam-Peters, Dirk J Duncker, Felix Zijlstra, and Heleen M M van Beusekom. Endovascular procedures cause transient endothelial injury but do not disrupt mature neointima in drug eluting stents. *Sci. Rep.*, 10(1):2173, February 2020.
- [38] Arun B Shivashankarappa, Nagesh C Mahadevappa, Anand Palakshachar, Prabhavathi Bhat, Ashita Barthur, Sripal Bangalore, Srinivas B Chikkaswamy, Rockey Katheria, and Manjunath C Nanjappa. Cerebrovascular events complicating cardiac catheterization - a tertiary care cardiac centre experience. *Heart Views*, 22(4):264–270, October 2021.
- [39] Eugenio Picano, Emanuela Piccaluga, Renato Padovani, Claudio Antonio Traino, Maria Grazia Andreassi, and Giulio Guagliumi. Risks related to fluoroscopy radiation associated with electrophysiology procedures. *J. Atr. Fibrillation*, 7(2):1044, August 2014.
- [40] Cosette M. Stahl, Quinn C. Meisinger, Michael P. Andre, Thomas B. Kinney, and Isabel G. Newton. Radiation risk to the fluoroscopy operator and staff. *American Journal of Roentgenology*, 207(4):737–744, 2016. PMID: 28829623.
- [41] Priya Bansal, Michael Luna, Margaret Dutton, Aneel Maini, Subhash Banerjee, Emmanouil S Brilakis, Brijeshwar Maini, and Houman Khalili. Enhancing occupational safety in the x-ray laboratory. *Coronary Artery Disease*, 33(4), 2022.
- [42] Ahmed Hassanin, Karim Al-Azizi, and Barry F. Uretsky. Occupational safety in the cardiac catheterization laboratory. *JACC: Cardiovascular Interventions*, 16(10):1236–1238, 2023.
- [43] Sylvia Marie R. Biso and Mladen I. Vidovich. Radiation protection in the cardiac catheterization laboratory. *Journal of Thoracic Disease*, 12(4), 2020.

- [44] Monica Y Lo, Prashanthan Sanders, Phillipp Sommer, Jonathan M Kalman, Usman R Siddiqui, Sri Sundaram, Christopher Piorkowski, Nicholas Olson, Steven M Madej, and Douglas N Gibson. Safety and effectiveness of a next-generation contact force catheter: Results of the tactisense trial. *JACC: Clinical Electrophysiology*, 2021.
- [45] Chao-Feng Chen, Xiao-Fei Gao, Mei-Jun Liu, Chao-Lun Jin, and Yi-Zhou Xu. Safety and efficacy of the thermocool smarttouch surroundflow catheter for atrial fibrillation ablation: A meta-analysis. *Clinical cardiology*, 43(3):267–274, 2020.
- [46] PJ French, D Tanase, and JFL Goosen. Sensors for catheter applications. *Sensors Update*, 13(1):107–153, 2003.
- [47] Kai Xu and Nabil Simaan. An investigation of the intrinsic force sensing capabilities of continuum robots. *IEEE Transactions on Robotics*, 24(3):576–587, 2008.
- [48] Naghmeh Bandari, Javad Dargahi, and Muthukumaran Packirisamy. Tactile sensors for minimally invasive surgery: A review of the state-of-the-art, applications, and perspectives. *IEEE Access*, 8:7682–7708, 2020.
- [49] Yohan Noh, Hongbin Liu, Sina Sareh, Damith Suresh Chathuranga, Helge Würdemann, Kawal Rhode, and Kaspar Althoefer. Image-based optical miniaturized three-axis force sensor for cardiac catheterization. *IEEE Sensors Journal*, 16(22):7924–7932, 2016.
- [50] Katsuaki Yokoyama, Hiroshi Nakagawa, Dipen C. Shah, Hendrik Lambert, Giovanni Leo, Nicolas Aeby, Atsushi Ikeda, Jan V. Pitha, Tushar Sharma, Ralph Lazzara, and Warren M. Jackman. Novel contact force sensor incorporated in irrigated radiofrequency ablation catheter predicts lesion size and incidence of steam pop and thrombus. *Circulation: Arrhythmia and Electrophysiology*, 1(5):354–362, 2008.
- [51] Liang Zou, Chang Ge, Z Jane Wang, Edmond Cretu, and Xiaou Li. Novel tactile sensor technology and smart tactile sensing systems: A review. *Sensors*, 17(11):2653, 2017.
- [52] Mahta Khoshnam, Mahdi Azizian, and Rajni V Patel. Modeling of a steerable catheter based

- on beam theory. In *2012 IEEE International Conference on Robotics and Automation*, pages 4681–4686. IEEE, 2012.
- [53] John Till, Vincent Aloï, and Caleb Rucker. Real-time dynamics of soft and continuum robots based on cosserat rod models. *The International Journal of Robotics Research*, 38(6):723–746, 2019.
- [54] Robert J Webster, Joseph M Romano, and Noah J Cowan. Mechanics of precurved-tube continuum robots. *IEEE Transactions on Robotics*, 25(1):67–78, 2008.
- [55] Minou Kouh Soltani, Sohrab Khanmohammadi, and Farzan Ghalichi. A three-dimensional shape-based force and stiffness-sensing platform for tendon-driven catheters. *Sensors*, 16(7):990, 2016.
- [56] Junghwan Back, Lukas Lindenroth, Kawal Rhode, and Hongbin Liu. Three dimensional force estimation for steerable catheters through bi-point tracking. *Sensors and Actuators A: Physical*, 279:404–415, 2018.
- [57] Amir Sayadi, Amir Hooshier, and Javad Dargahi. Impedance matching approach for robust force feedback rendering with application in robot-assisted interventions. In *2020 8th International Conference on Control, Mechatronics and Automation (ICCMA)*, pages 18–22, 2020.
- [58] Amir Hooshier, Amir Sayadi, Mohammad Jolaei, and Javad Dargahi. Accurate estimation of tip force on tendon-driven catheters using inverse cosserat rod model. In *2020 International Conference on Biomedical Innovations and Applications (BIA)*, pages 37–40, 2020.
- [59] Mohammad Jolaei, Amir Hooshier, Amir Sayadi, Javad Dargahi, and Muthukumaran Packirisamy. Sensor-free force control of tendon-driven ablation catheters through position control and contact modeling. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5248–5251, 2020.
- [60] Amir Hooshier, Amir Sayadi, Mohammad Jolaei, and Javad Dargahi. Analytical tip force estimation on tendon-driven catheters through inverse solution of cosserat rod model. In

- 2021 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1829–1834, 2021.
- [61] Shahir Hasanzadeh and Farrokh Janabi-Sharifi. Model-based force estimation for intracardiac catheters. *IEEE/ASME Transactions on Mechatronics*, 21(1):154–162, 2015.
 - [62] Mahta Khoshnam and Rajni V Patel. Estimating contact force for steerable ablation catheters based on shape analysis. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3509–3514. IEEE, 2014.
 - [63] Hamidreza Khodashenas, Pedram Fekri, Mehrdad Zadeh, and Javad Dargahi. A vision-based method for estimating contact forces in intracardiac catheters. In *IEEE 1st International Conference on Autonomous Systems, ICAS*. IEEE, 2021.
 - [64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
 - [65] Yan Zhao, Shuxiang Guo, Yuxin Wang, Jinxin Cui, Youchun Ma, Yuwen Zeng, Xinke Liu, Yuhua Jiang, Youxinag Li, Liwei Shi, and Nan Xiao. A cnn-based prototype method of unstructured surgical state perception and navigation for an endovascular surgery robot. *Medical & Biological Engineering & Computing*, 57(9):1875–1887, Sep 2019.
 - [66] Jihoon Kweon, Kyunghwan Kim, Chaehyuk Lee, Hwi Kwon, Jinwoo Park, Kyoseok Song, Young In Kim, Jeeone Park, Inwook Back, Jae-Hyung Roh, Youngjin Moon, Jaesoon Choi, and Young-Hak Kim. Deep reinforcement learning for guidewire navigation in coronary artery phantom. *IEEE Access*, 9:166409–166422, 2021.
 - [67] Glenn Jocher. YOLOv5 by Ultralytics, May 2020.
 - [68] Deai Yang, Jingzhou Song, and Yuhang Hu. Guidewire feeding method based on deep reinforcement learning for vascular intervention robot. In *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1287–1293, 2022.

- [69] Yu Wang, Juan Wang, Yaxin Li, Tao Yang, and Cheng Ren. The deep reinforcement learning-based vr training system with haptic guidance for catheterization skill transfer. *IEEE Sensors Journal*, 22(23):23356–23366, 2022.
- [70] Wentao Liu, Weijin Xu, Xiaochuan Li, Bowen Liang, Ziyang He, Mengke Zhu, Jingzhou Song, Huihua Yang, and Qingsheng Lu. Autonomous guidewire navigation for robot-assisted endovascular interventions: A knowledge-driven visual guidance approach, 2025.
- [71] Daniel Toth, Shun Miao, Tanja Kurzendorfer, Christopher A Rinaldi, Rui Liao, Tommaso Mansi, Kawal Rhode, and Peter Mountney. 3D/2D model-to-image registration by imitation learning for cardiac procedures. *Int. J. Comput. Assist. Radiol. Surg.*, 13(8):1141–1149, August 2018.
- [72] Wenqiang Chi, Giulio Dagnino, Trevor M. Y. Kwok, Anh Nguyen, Dennis Kundrat, Mohamed E. M. K. Abdelaziz, Celia Riga, Colin Bicknell, and Guang-Zhong Yang. Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2414–2420, 2020.
- [73] Ameya Pore, Eleonora Tagliabue, Marco Piccinelli, Diego Dall’Alba, Alicia Casals, and Paolo Fiorini. Learning from demonstrations for autonomous soft-tissue retraction. In *2021 International Symposium on Medical Robotics (ISMR)*, pages 1–7, 2021.
- [74] Zhen Li, Chiara Lambranzi, Di Wu, Alice Segato, Federico De Marco, Emmanuel Vander Poorten, Jenny Dankelman, and Elena De Momi. Robust path planning via learning from demonstrations for robotic catheters in deformable environments. *IEEE Transactions on Biomedical Engineering*, 72(1):324–336, 2025.
- [75] Andrea Fortuna, Mattia Magro, Angela Peloso, Alice Segato, and Elena De Momi. Inverse reinforcement learning algorithm for intra-vascular and intra-cardiac catheter’s navigation in minimally invasive surgery. EasyChair Preprint 8886, EasyChair, 2022.
- [76] Harry Robertshaw, Lennart Karstensen, Benjamin Jackson, Alejandro Granados, and Thomas C Booth. Autonomous navigation of catheters and guidewires in mechanical

- thrombectomy using inverse reinforcement learning. *Int. J. Comput. Assist. Radiol. Surg.*, 19(8):1569–1578, August 2024.
- [77] Pedram Fekri, Javad Dargahi, and Mehrdad Zadeh. Deep learning-based haptic guidance for surgical skills transfer. *Frontiers in Robotics and AI*, 7, 2021.
 - [78] Pedram Fekri, Hamidreza Khodashenas, Kevin Lachapelle, Renzo Cecere, Mehrdad Zadeh, and Javad Dargahi. Y-net: A deep convolutional architecture for 3d estimation of contact forces in intracardiac catheters. *IEEE Robotics and Automation Letters*, 7(2):3592–3599, 2022.
 - [79] Pedram Fekri, Mehrdad Zadeh, and Javad Dargahi. H-net: A multitask architecture for simultaneous 3d force estimation and stereo semantic segmentation in intracardiac catheters. *IEEE Robotics and Automation Letters*, 10(1):844–851, 2025.
 - [80] Pedram Fekri, Mehrdad Zadeh, and Javad Dargahi. Transforseg: A multitask stereo vit for joint stereo segmentation and 3d force estimation in catheterization, 2025.
 - [81] Pedram Fekri, Majid Roshanfar, Samuel Barbeau, Seyedfarzad Famouri, Thomas Looi, Dale Podolsky, Mehrdad Zadeh, and Javad Dargahi. Dino-cva: A multimodal goal-conditioned vision-to-action model for autonomous catheter navigation, 2025.
 - [82] Pedram Fekri, Peyman Setoodeh, Fariba Khosravian, AA Safavi, and Mehrdad H Zadeh. Towards deep secure tele-surgery. In *Proceedings of the International Conference on Scientific Computing (CSC)*, pages 81–86. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2018.
 - [83] Pedram Fekri, Mehrdad Zadeh, and Javad Dargahi. Fs-net: An encoder-decoder architecture for catheter segmentation and contact forces estimation in intracardiac catheters. In *2025 IEEE World Haptics Conference (WHC)*, pages 123–132, 2025.
 - [84] Vajiheh Abedi, Mehrdad H. Zadeh, Javad Dargahi, and Pedram Fekri. Software failures prediction in self-driving vehicles. In *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pages 1–5, 2020.

- [85] Pedram Fekri, Mehrdad Zadeh, and Javad Dargahi. On the safety of autonomous driving: A dynamic deep object detection approach. SAE Technical Paper 2019-01-1044, SAE International, 2019.
- [86] N. Tavakolian, P. Fekri, M. Zadeh, and J. Dargahi. Kdepthnet: Mono-camera based depth estimation for autonomous driving. SAE Technical Paper 2022-01-0082, SAE International, 2022.
- [87] Seyedfarzad Famouri, Pedram Fekri, Majid Roshanfar, and Javad 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)*, pages 216–221, 2023.
- [88] Seyedfarzad Famouri, Pedram Fekri, and Javad Dargahi. Kinematics skill modeling of cardiac catheterization via deep learning method. In *2024 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–9, 2024.
- [89] Majid Roshanfar, Pedram Fekri, and Javad Dargahi. Toward autonomous cardiac catheterization through a parametric finite element simulation with experimental validation. *ICAS 2023*, page 23, 3 2023.
- [90] Majid Roshanfar, Pedram Fekri, and Javad Dargahi. A deep learning model for tip force estimation on steerable catheters via learning-from-simulation. *EasyChair*, (10364), 6 2023.
- [91] Samuel Barbeau, Pedram Fekri, David Osowiechi, Ali Bahri, Moslem Yazdanpanah, Masih Aminbeidokhti, and Christian Desrosiers. Cta: Cross-task alignment for better test time training, 2025.
- [92] Nazih Mallouk, Majid Roshanfar, Pedram Fekri, and Javad Dargahi. A brief review on recent advances in haptic technology for human-computer interaction: Force, tactile, and surface haptic feedback. In *2024 7th International Conference on Information and Computer Technologies (ICICT)*, pages 34–39, 2024.
- [93] Majid Roshanfar, Pedram Fekri, Robert H. Nguyen, Changyan He, Paul H. Kang, James

- Drake, Eric Diller, and Thomas Looi. Learning-based tip contact force estimation for fbg-embedded continuum robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 844–850, 2025.
- [94] Jelena Kornej, Christin S Börschel, Emelia J Benjamin, and Renate B Schnabel. Epidemiology of atrial fibrillation in the 21st century: novel methods and new insights. *Circulation research*, 127(1):4–20, 2020.
- [95] Salim S Virani, Alvaro Alonso, Hugo J Aparicio, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Susan Cheng, Francesca N Delling, et al. Heart disease and stroke statistics—2021 update: a report from the american heart association. *Circulation*, 143(8):e254–e743, 2021.
- [96] Eric N Prystowsky, Benzy J Padanilam, and Richard I Fogel. Treatment of atrial fibrillation. *Jama*, 314(3):278–288, 2015.
- [97] Gjin Ndrepepa and Heidi Estner. Ablation of cardiac arrhythmias—energy sources and mechanisms of lesion formation. In *Catheter Ablation of Cardiac Arrhythmias*, pages 35–53. Springer, 2006.
- [98] Xiaohua Hu, Ang Chen, Yigang Luo, Chris Zhang, and Edwin Zhang. Steerable catheters for minimally invasive surgery: a review and future directions. *Computer Assisted Surgery*, 23(1):21–41, 2018.
- [99] Dieter Haemmerich. Biophysics of radiofrequency ablation. *Critical Reviews™ in Biomedical Engineering*, 38(1), 2010.
- [100] Luigi Di Biase, Andrea Natale, Conor Barrett, Carmela Tan, Claude S Elayi, Chi Keong Ching, Paul Wang, AMIN AL-AHMAD, Mauricio Arruda, J David Burkhardt, et al. Relationship between catheter forces, lesion characteristics, “popping,” and char formation: experience with robotic navigation system. *Journal of cardiovascular electrophysiology*, 20(4):436–440, 2009.

- [101] Dipen C Shah and Mehdi Namdar. Real-time contact force measurement: a key parameter for controlling lesion creation with radiofrequency energy. *Circulation: Arrhythmia and Electrophysiology*, 8(3):713–721, 2015.
- [102] Davide Chicco. *Siamese Neural Networks: An Overview*, pages 73–94. Springer US, New York, NY, 2021.
- [103] Pierre Ambrosini, Daniel Ruijters, Wiro J. Niessen, Adriaan Moelker, and Theo van Walsum. Fully automatic and real-time catheter segmentation in x-ray fluoroscopy. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*, pages 577–585, Cham, 2017. Springer International Publishing.
- [104] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).
- [105] Henri Veisterä and Jyrki Lötjönen. Reconstructing 3d boundary element heart models from 2d biplane fluoroscopy. In Toivo Katila, Jukka Nenonen, Isabelle E. Magnin, Patrick Clarysse, and Johan Montagnat, editors, *Functional Imaging and Modeling of the Heart*, pages 17–23, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.
- [106] FELIX BOURIER, TILKO REENTS, SONIA AMMAR-BUSCH, VERENA SEMMLER, MARTA TELISHEVSKA, MARC KOTTMAIER, CARSTEN LENNERZ, CHRISTIAN GREBMER, CHRISTOF KOLB, ISABEL DEISENHOFER, and GABRIELE HESSLING. Transseptal puncture guided by ct-derived 3d-augmented fluoroscopy. *Journal of Cardiovascular Electrophysiology*, 27(3):369–372, 2016.
- [107] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.
- [108] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

- [109] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.
- [110] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.
- [111] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
- [112] Luca Caltagirone, Mauro Bellone, Lennart Svensson, and Mattias Wahde. Lidar–camera fusion for road detection using fully convolutional neural networks. *Robotics and Autonomous Systems*, 111:125–131, 2019.
- [113] Joel Schlosser, Christopher K. Chow, and Zsolt Kira. Fusing lidar and images for pedestrian detection using convolutional neural networks. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2198–2205, 2016.
- [114] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- [115] Martín Abadi and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [116] Yugandhar R. Manda and Krishna M. Baradhi. *Cardiac Catheterization Risks and Complications*. StatPearls Publishing, Treasure Island (FL), 2021.
- [117] Rohit Mehta, Kyong-Jin Lee, Rajiv Chaturvedi, and Lee Benson. Complications of pediatric cardiac catheterization: A review in the current era. *Catheterization and Cardiovascular Interventions*, 72(2):278–285, 2008.
- [118] Chon-Seng Hong, Zhih-Cherng Chen, Kuo-Ting Tang, and Wei-Ting Chang. The effectiveness and safety between monoplane and biplane imaging during coronary angiographies. *Acta Cardiol Sin*, 36(2):105–110, March 2020.

- [119] David M Leistner, Lara S Schlender, Julia Steiner, Aslihan Erbay, Jens Klotsche, Patrick Schauerte, Arash Haghikia, Ursula Rauch-Kröhnert, David Sinning, Alexander Lauten, Hans-Christian Mochmann, Carsten Skurk, Ulf Landmesser, and Barbara E Stähli. A randomised comparison of monoplane versus biplane fluoroscopy in patients undergoing percutaneous coronary intervention: the RAMBO trial. *EuroIntervention*, 16(8):672–679, October 2020.
- [120] Marta Gherardini, Evangelos Mazomenos, Arianna Menciassi, and Danail Stoyanov. Catheter segmentation in x-ray fluoroscopy using synthetic data and transfer learning with light u-nets. *Computer Methods and Programs in Biomedicine*, 192:105420, 2020.
- [121] Anh Nguyen, Dennis Kundrat, Giulio Dagnino, Wenqiang Chi, Mohamed E. M. K. Abdelaziz, Yao Guo, YingLiang Ma, Trevor M. Y. Kwok, Celia Riga, and Guang-Zhong 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)*, pages 9967–9973, 2020.
- [122] Ibrahim Abdulhafiz and Farrokh Janabi-Sharifi. A hybrid approach to 3d shape estimation of catheters using ultrasound images. *IEEE Robotics and Automation Letters*, 8(4):1912–1919, 2023.
- [123] Shih-Sheng Chang, Ching-Ting Lin, Wei-Chun Wang, Kai-Cheng Hsu, Ya-Lun Wu, Chia-Hao Liu, and Yang C. Fann. Optimizing ensemble u-net architectures for robust coronary vessel segmentation in angiographic images. *Scientific Reports*, 14(1):6640, Mar 2024.
- [124] Yuhan Wang, Hak Keung Lam, Zeng-Guang Hou, Rui-Qi Li, Xiao-Liang Xie, and Shi-Qi Liu. High-resolution feature based central venous catheter tip detection network in x-ray images. *Med Image Anal*, 88:102876, June 2023.
- [125] Hongxu Yang, Caifeng Shan, Alexander F. Kolen, and Peter 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)*, pages 260–264, 2019.

- [126] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.
- [127] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.
- [128] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.
- [129] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
- [130] Amanda M. Aleong, Alejandro Berlin, Jette Borg, Joelle Helou, Akbar Beiki-Ardakani, Alexandra Rink, Srinivas Raman, Peter Chung, and Robert A. Weersink. Rapid multi-catheter segmentation for magnetic resonance image-guided catheter-based interventions. *Medical Physics*, 2024.
- [131] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019.
- [132] Yumiko Kanei, Tak Kwan, Navin C. Nakra, Michael Liou, Yili Huang, Lori L. Vales, John T. Fox, Jack P. Chen, and Shigeru Saito. Transradial cardiac catheterization: A review of access site complications. *Catheterization and Cardiovascular Interventions*, 78(6):840–846, 2011.
- [133] Harry Robertshaw, Benjamin Jackson, Jiaheng Wang, Hadi Sadati, Lennart Karstensen, Alejandro Granados, and Thomas C. Booth. Reinforcement learning for safe autonomous two-device navigation of cerebral vessels in mechanical thrombectomy. *International Journal of Computer Assisted Radiology and Surgery*, Apr 2025.
- [134] Anh Nguyen, Dennis Kundrat, Giulio Dagnino, Wenqiang Chi, Mohamed E. M. K. Abdelaziz, Yao Guo, YingLiang Ma, Trevor M. Y. Kwok, Celia Riga, and Guang-Zhong Yang.

End-to-end real-time catheter segmentation with optical flow-guided warping during endovascular intervention, 2020.

- [135] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.
- [136] Maximilian Durner, Wout Boerdijk, Martin Sundermeyer, Werner Friedl, Zoltan-Csaba Marton, and Rudolph Triebel. Unknown object segmentation from stereo images. *CoRR*, abs/2103.06796, 2021.
- [137] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6177–6186, 2021.
- [138] Tobias Behr, Tim Philipp Pusch, Marius Siegfarth, Dominik Hüsener, Tobias Mörschel, and Lennart Karstensen. Deep reinforcement learning for the navigation of neurovascular catheters. *Current Directions in Biomedical Engineering*, 5(1):5–8, 2019.
- [139] Olatunji Mumini Omisore, Toluwanimi Akinyemi, Wenke Duan, Wenjing Du, and Lei Wang. A novel sample-efficient deep reinforcement learning with episodic policy transfer for pid-based control in cardiac catheterization robots, 2021.
- [140] Hyeonseok You, EunKyung Bae, Youngjin Moon, Jihoon Kweon, and Jaesoon Choi. Automatic control of cardiac ablation catheter with deep reinforcement learning method. *Journal of Mechanical Science and Technology*, 33(11):5415–5423, Nov 2019.
- [141] S. Kevin Zhou, Hoang Ngan Le, Khoa Luu, Hien V Nguyen, and Nicholas Ayache. Deep reinforcement learning in medical imaging: A literature review. *Medical Image Analysis*, 73:102193, 2021.

- [142] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.
- [143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
- [144] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886, 2021.
- [145] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segformer: Transformer for semantic segmentation. *arXiv preprint arXiv:2105.05633*, 2021.
- [146] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv*, 2021.
- [147] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021.
- [148] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.
- [149] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.
- [150] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022.

- [151] Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Gustavo A. Vargas Hakim, David Osowiechi, Ismail Ben Ayed, and Christian Desrosiers. Tfs-vit: Token-level feature stylization for domain generalization. *Pattern Recognition*, 149:110213, 2024.
- [152] Marco Colussi, Sergio Mascetti, Jose Dolz, and Christian Desrosiers. Rec- ttt: Contrastive feature reconstruction for test-time training. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6699–6708, 2025.
- [153] David Osowiechi, Mehrdad Noori, Gustavo Adolfo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaei, Ismail Ben Ayed, and Christian Desrosiers. Watt: Weight average test-time adaptation of clip, 2024.
- [154] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyang Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d, 2023.
- [155] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos, 2025.
- [156] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, and Wenqi Li. Vista3d: A unified segmentation foundation model for 3d medical imaging, 2024.
- [157] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation, 2023.
- [158] Yifan Gao, Wei Xia, Dingdu Hu, Wenkui Wang, and Xin Gao. Desam: Decoupled segment anything model for generalizable medical image segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 509–519, Cham, 2024. Springer Nature Switzerland.

- [159] Yujie Lu, Jingwen Li, Sibojun, Yanzhou Su, he yao, Yisong Liu, Min Zhu, and Junlong Cheng. Segmote: Token-level mixture of experts for medical image segmentation, 2026.
- [160] Luisa Vargas, Eleonora Poeta, Tania Cerquitelli, Elena Baralis, and Maria A Zuluaga. Medopenseg: Open-world medical segmentation with memory-augmented transformers. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025.
- [161] Yu Huang, Zelin Peng, Yichen Zhao, Piao Yang, Xiaokang Yang, and Wei Shen. Medseg-r: Reasoning segmentation in medical images with multimodal large language models, 2025.
- [162] Sina Mohammadmahdi Keshavarz, Mohammad Khoobani, Rene Gilliland-Rocque, Mohammadmahdi Tahmasebi, Andrew Dueck, and M Ali Tavallaei. A self-expandable nitinol frame for cable-driven parallel mechanisms in minimally invasive cardiovascular interventions. *Journal of the Mechanical Behavior of Biomedical Materials*, 163:106889, 2025.
- [163] Majid Roshanfar, Mohammadhossein Salimi, Amir Hossein Kaboodrangi, Sun-Joo Jang, Albert J Sinusas, Shing-Chiu Wong, and Bobak Mosadegh. Advanced robotics for the next-generation of cardiac interventions. *Micromachines*, 16(4):363, 2025.
- [164] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning, 2025.
- [165] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.