Medical Instrument Detection in 3D Ultrasound for Intervention Guidance

Medical Instrument Detection in 3D Ultrasound for Intervention Guidance

Hongxu Yang

Medical Instrument Detection in 3D Ultrasound for Intervention Guidance

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. F.P.T. Baaijens, voor een
commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen
op donderdag 14 april 2022 om 16.00 uur

door

Hongxu Yang

geboren te Taiyuan, China

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ing. A.J.M. Pemen
promotor:	prof.dr.ir. P.H.N. de With
1e copromotor:	dr. A.F. Kolen (Philips Medical Systems)
2e copromotor:	prof.dr. C. Shan (Shandong Univ. Science & Technology)
leden:	prof.dr. H. Beerlage (UvA-Amsterdam UMC)
	prof.dr.ir. C.H. Slump (Universiteit Twente)
	prof.dr.ir. M. Mischi
adviseurs:	dr. R.A. Bouwman (Catharina Ziekenhuis Eindhoven)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

To my wife Qian, and parents Bin and Miao.

Medical Instrument Detection in 3D Ultrasound for Intervention Guidance

Hongxu Yang

Cover design: Xi Chen and Hongxu Yang

ISBN: 978-90-386-5488-1
NUR-code: 959

Summary

Cardiac interventions with a catheter and/or needle are commonly applied procedures with patients in hospitals. To guide medical interventions, advanced medical imaging systems such as ultrasound (US) and fluoroscopy are broadly applied and offer surgeons accurate visualization and measurement of anatomical structures, while displaying the interventional instrument and activities within or outside the operation regions. For efficiency reasons, a general trend in clinical environments appears towards a smooth and fast workflow of the patient treatment, where the emphasis is on shortening the procedure time, combined with improving the clinical outcome for a given intervention. For the intervention and surgical guidance, accurate instrument detection is important during the operation, but also forms a challenge with US imaging. First, an extensive multi-fold hand-eye coordination of the instrument and US transducer is required to continuously align the instrument within the beamforming region of the US transducer, even for a 3D US probe. Second, US imaging has intrinsic quality challenges like low signal-to-noise ratio, imaging artifacts and distorted appearance of the instrument, which complicates the interpretation of the image data. Existing technologies are neither accurate nor efficient enough for a real clinical application in 3D ultrasound. This is the field of research for this thesis, where the objective is to detect the medical instruments during the intervention in a live setting, where the patient is operated by clinicians and the intervention is imaged by ultrasound equipment. The imaging should support the live intervention by visualizing the instrument and enable to localize it inside the human body.

The thesis commences with an overview of existing computer vision methods for US interventional imaging. After initial experiments with machine learning in Chapter 3, the thesis proposes in succeeding Chapters 4-7 several efficient medical instrument segmentation and detection solutions in 3D US images, which are based on data-driven methods, such as deep learning. The image processing problem is approached in various ways using two-dimensional (2D) analysis of the 3D US output image (in Chapter 4), 3D processing of volumetric data containing the instrument (Chapters 5-6), and a hybrid combination of 2.5D processing (Chapter 7).

In the proposed methods, the majority of the designs, e.g. Chapters 4-6, are based on a coarse-to-fine framework, which facilitates both accuracy in segmentation and detection and efficiency in computation. In the framework, first, the regions containing the instrument are efficiently selected by a coarse locator (either in voxels or in image patches), which drastically reduces the computation cost for processing on whole image segmentation from several minutes to the level of seconds. Several novel coarse selection techniques are introduced in the thesis, such as a 3D Frangi model-based voxel pre-selection in Chapter 4, a 2D convolutional neural network-based coarse detector in Chapter 5 and a region-based reinforcement-learning detector in Chapter 6. These approaches all yield an efficient and lower computation cost than direct 3D processing of the volumetric data. The proposed detectors coarsely detect the instrument within a large 3D volumetric data with an execution speed of about 0.2-2 seconds, which generates the restricted regional volumes for the subsequent fine segmentation tasks. Then, as the second step, a fine 3D segmentation is applied based on the coarse results, such as CNN-based classification in Chapter 4 and a dimensional-fusion network in Chapter 5. This exploits the both 2D and 3D semantic information in the regional patches with higher accuracy than state-of-the-art methods, because of a better network architecture design and the coarse-to-fine strategy. More specifically in Chapter 5, a novel FuseNet is proposed to exploit the 2.5D and 3D information with an end-to-end training approach, which leads to better results than the conventional standard UNet. Using the proposed methods, the medical instrument can be segmented within 0.3-1.0 seconds for a 3D US image, which is at least 10 times faster than the existing published literature. More importantly, the segmentation can be achieved with about 70% Dice score in noisy and challenging US images, which leads to accurate detections with a localization error of about 1-2 mm only.

The second type of study focuses on annotation-efficient deep learning in Chapter 6, which reduces the initially required accurate annotation effort for deep neural network training. Especially for US imaging, the manual annotation of the instrument is costly and laborious. Specifically, a modified mean-teacher model, so-called Dual-UNet in Chapter 6, is proposed to exploit the prediction uncertainty on the unlabeled training images. This model aims at finding the unlabeled discriminative information in the images for performance improvement of the segmentation. The extensive validation shows the proposed uncertainty analysis method requires that only 30% of the images are annotated, while achieving comparable results to fully-supervised approaches with full annotations.

The third type of the study includes network architecture optimization in Chapter 7, which enhances the prediction efficiency by reducing the overall network architecture complexity. This method introduces a novel 3D-to-2D network projection module that projects the 3D volumes to 2D planes by an end-to-end learned network. The projection module drastically simplifies the overall network, so that the discriminative information is easier to learn and the overall ef-

ficiency is increased with at least a factor of 10. This results an execution speed of about 0.1-0.2 seconds for a standard volumetric image with a size of 160^3 voxels. Moreover, this speed improvement is obtained with limited automated detection errors, which are only about 2-3 voxels on different challenging datasets.

In conclusion, the developed techniques of this thesis can accurately and efficiently detect the medical instrument, such as a catheter and/or needle, in challenging 3D US images. The use of deep learning networks further generalizes the application of instruments in the US imaging experiments, such as catheters, guide-wires or needles. This finding is also supported by making no assumptions on the instrument's shape and thickness. The proposed algorithms already execute on existing computing platforms with a speed of 1-3 frames per second without any software optimization, which facilitate real-time applications in the near future. With accurate segmentation results, the algorithms can aid surgeons to find the instrument easily, which can contribute to a safer and more accurate instrument placement, thereby benefiting both patients and physicians.

US imaging is assumed to be widely applied in many different applications in regional or local hospitals to be supplementary to CT or X-Ray imaging, or even replace them for simple tasks, because it is less expensive and can be equipped with more advanced image processing software. With the trend of more affordable US machines and advanced US analysis in the future, a reduction of the hospital referring cost and better early outcomes for patients may be achieved. Future ultrasound systems with integration of the proposed deep learning algorithms will eventually improve the outcomes of image-guided interventions, and facilitate clinicians in training and operations. This will lead to a substantially broader usage of US-guided analysis, indicating a bright future for US-based image processing with AI techniques.

Samenvatting

Cardiale interventies met behulp van een katheter of een naald worden vaak uitgevoerd bij patiënten in ziekenhuizen. Geavanceerde medische beeldvormingssystemen, zoals echografie en fluoroscopie, worden daarbij gebruikt om de medische interventies te begeleiden. Zij bieden chirurgen de mogelijkheid anatomische structuren met hoge resolutie te visualiseren en te meten, terwijl het interventie-instrument en de activiteiten zowel binnen als buiten het operatiegebied in het lichaam worden gevisualiseerd. In de klinische omgeving is er een toenemende trend naar een soepele en snelle efficiënte workflow van de patiëntbehandeling, waarbij de nadruk ligt op het verkorten van de doorlooptijd en het verbeteren van het klinische resultaat van de geplande ingreep. Tijdens de operatie is nauwkeurige instrumentdetectie essentieel voor interventie en chirurgische begeleiding, maar dit vormt tevens een uitdaging wanneer echografie wordt gebruikt. Ten eerste is een nauwkeurige hand-oog coördinatie tussen het instrument en de ultrageluidsomvormer (*transducer*) nodig om het instrument voortdurend binnen het bundelvormingsgebied van de *transducer* te houden, zelfs met gebruik van een 3D-echografische sonde. Ten tweede heeft echografische beeldverwerking inherente kwaliteitsproblemen, zoals lage signaal-ruisverhoudingen, beeldverwerkingsartefacten en vervormde instrumentweergave, die de interpretatie van beelddata bemoeilijken. Bestaande technologieën zijn niet nauwkeurig of efficiënt genoeg voor de daadwerkelijke klinische toepassing van 3D-echografie. Dit is precies het onderzoeksgebied van dit proefschrift, dat gericht is op het detecteren van medische instrumenten tijdens interventies in een live omgeving, waarbij de patiënt wordt geopereerd door artsen en de interventie met beeldvorming wordt gevolgd met echografische apparatuur. De beeldvorming is bedoeld om een live-interventie te ondersteunen door het instrument te visualiseren en het mogelijk te maken dit instrument te lokaliseren in het menselijk lichaam.

De thesis begint met een overzicht van de meest gebruikte computervisiemethoden die worden toegepast bij echografische beeldvorming gedurende interventies. Na inleidende experimenten met conventionele *machine learning* in hoofdstuk 3, presenteert het proefschrift in de hoofdstukken 4-7 verscheidene efficiënte technieken voor medische instrumentsegmentatie en detectieoplossingen voor 3D echografie, gebaseerd op data-gedreven methoden zoals *deep*

learning. Het probleem van de echografische beeldverwerking wordt op verschillende manieren aangepakt, waaronder een tweedimensionale analyse van het 3D echografische uitgangsbeeld (hoofdstuk 4), driedimensionale verwerking van volumetrische data die het instrument bevatten (hoofdstukken 5-6) en een hybride combinatie van 2.5D beeldverwerking (hoofdstuk 7).

De meeste van de onderzochte methodes, zoals bijv. in de hoofdstukken 4-6, zijn gebaseerd op een grof-naar-fijn raamwerk dat zowel nauwkeurigheid in segmentatie en detectie geeft als efficiënte berekening hiervan mogelijk maakt. In het raamwerk worden eerst de regio's die het instrument bevatten efficiënt geselecteerd door een grove instrumentlocalisatie (hetzij in voxels of in beelddelen), hetgeen resulteert in een drastische reductie van de benodigde berekeningen voor de segmentatie van het gehele beeld (van minuten naar enkele seconden). Verschillende nieuwe methoden voor grove voxelselectie worden geïntroduceerd, zoals een 3D Frangi model-gebaseerde preselectie van voxels in hoofdstuk 4, een 2D convolutioneel neuraal netwerk-gebaseerde grove detector in hoofdstuk 5 en een op regio-gebaseerde *reinforcement learning* detector in hoofdstuk 6. Al deze benaderingen leveren efficiënte en verminderde berekeningskosten op dan directe 3D-volumetrische dataverwerking. De detectoren identificeren het instrument binnen een groot 3D-volumetrisch gebied met een uitvoeringssnelheid van ongeveer 0,2-2 seconden, waardoor begrensde regionale datavolumes worden geselecteerd die nodig zijn voor de daaropvolgende fijne segmentatie. Als een tweede stap wordt dan een fijne 3D-segmentatie toegepast op basis van de eerste grove resultaten, zoals een CNN-gebaseerde classificatie in hoofdstuk 4 en een multi-dimensionaal fusienetwerk in hoofdstuk 5. Dit maakt gebruik van zowel 2D als 3D semantische data in regionale beelddelen (*patches*) met een hogere nauwkeurigheid dan recente bestaande methoden, dankzij een betere netwerkarchitectuur en een grof-naar-fijn strategie. In hoofdstuk 5 wordt er meer specifiek ingegaan op een nieuw *FuseNet* dat gebruik maakt van 2.5D- en 3D-informatie met een *end-to-end* trainingsmethode die tot betere resultaten leidt dan een standaard UNet. Met de bovengenoemde methoden kan een medisch instrument in 0,3-1,0 seconde worden gesegmenteerd met een 3D-echografisch beeld, wat minstens 10 keer sneller is dan de reeds beschikbare methoden uit de literatuur. Nog belangrijker is dat de segmentatie kan worden bereikt met een Dice score van ten minste 70% met ruis en beperkte echografische beeldkwaliteit, wat desondanks resulteert in een nauwkeurige detectie met een lokalisatiefout van slechts 1-2 mm.

De tweede studie streeft naar annotatie-effectieve *deep learning* in hoofdstuk 6, dat het vereiste nauwkeurige annotatiewerk voor de training van het neurale netwerk vermindert. De handmatige annotatie van instrumenten in beelddata is duur en bewerkelijk, vooral voor echografische beeldvorming. Een gewijzigd leraar-student leermodel, het zogenoemde Dual-UNet in hoofdstuk 6, is ontwikkeld om de voorspellingsonzekerheid op ongelabelde trainingsbeelden te benutten. Dit model streeft naar het lokaliseren van ongelabelde karakte-

ristieke informatie in beelden om zo de segmentatiekwaliteit te verbeteren. De uitgebreide validatie toont aan dat de voorgestelde methode voor onzekerheidsanalyse slechts 30% van de beelden annoteert, terwijl dezelfde resultaatkwaliteit wordt bereikt als met getrainde systemen met volledig complete annotaties.

De derde studie is de optimalisatie van de netwerkarchitectuur in hoofdstuk 7, die de voorspellingsefficiëntie bevordert door de algehele complexiteit van de netwerkarchitectuur te verminderen. Deze techniek introduceert een nieuwe netwerkmodule voor 3D-naar-2D projectie, m.a.w. het projecteert 3D-volumes op 2D-vlakken via een end-to-end lerend netwerk. De voorgestelde projectiemodule vereenvoudigt drastisch het totale netwerk, waardoor onderscheidende informatie gemakkelijker te leren is en de algehele efficiëntie met ten minste een factor 10 toeneemt. Dit resulteert in een uitvoeringssnelheid van ongeveer 0,1-0,2 seconde voor een standaard volumetrisch beeld met een grootte van 160^3 voxels. Bovendien wordt deze snelheidsverbetering bereikt met slechts beperkte automatische detectiefouten, die slechts 2-3 voxels beïnvloeden voor verschillende uitdagende datasets.

Concluderend kan worden vastgesteld dat de ontwikkelde technieken in dit proefschrift het mogelijk maken om een medisch instrument, zoals een katheter of een naald, te herkennen in real-time 3D echografische beelden. Het gebruik van *deep learning* netwerken generaliseert en verbreedt dit tot algemeen instrumentengebruik in echografische beeldvorming, zoals katheters, geleidingsdraden en naalden. Deze uitkomst wordt ook ondersteund door het feit dat er geen aannamen zijn gemaakt over de vorm en grootte van het gebruikte instrument. De algoritmen zijn ontwikkeld op standaard computersystemen, die een verwerkingssnelheid leveren van 1-3 beelden/sec. zonder softwareoptimalisatie, waardoor real-time toepassingen in de nabije toekomst mogelijk zijn. De algoritmen kunnen met nauwkeurige segmentatieresultaten de chirurgen helpen bij het lokaliseren van het instrument, wat kan leiden tot een veiliger en preciezere plaatsing van het instrument, hetgeen ten goede komt aan zowel de patiënt als de arts. Het is aannemelijk dat echografie op grote schaal zal worden gebruikt in verschillende toepassingen in grote en regionale ziekenhuizen, als aanvulling op CT- of X-Ray beeldvorming, of zelfs deels vervangt in bepaalde procedures, omdat het minder kost en met geavanceerde beeldverwerkingssoftware kan worden uitgevoerd. Met de ontwikkeling van meer bereikbare echografie-apparatuur en geavanceerde echografie-analyses kan een verlaging van de diagnostische kosten in het ziekenhuis worden gerealiseerd naast een verbeterde vroegtijdige analyse voor de patiënt. Toekomstige echografiesystemen die gebruik maken van de eerder genoemde *deep learning* algoritmen zullen de uitkomsten van beeldgeleide interventies verbeteren en artsen helpen bij opleidingen en operaties. Deze toename biedt een gunstig perspectief voor het gebruik van echografie-gebaseerde beeldanalyse die gecombineerd wordt met artificiële intelligentie.

Contents

CHAPTER 1

Introduction

1.1 Image-guided Intervention

With the advancing development of imaging technologies, image-guided interventions on patients have become increasingly mature and robust for various interventions on organs and related applications. Examples of applications of such advances have been made for cardiac interventions and needle or catheter-based biopsy taking. Ultrasound (US) imaging facilitates a higher clinical outcome or robustness, since it is used as a supplementary imaging modality to another already existing imaging modality, such as X-ray imaging, to visualize the medical instrument that is inserted or part of the procedure. The primary imaging modality provides the overview of the operation area, whereas the ultrasound system is used to visualize the local navigation of the instrument. Besides this, with the growing quality of the ultrasound imaging system, it offers intervention physicians also visualization and measurement of anatomical structures. This means that the system displays the interventional activities within the region of interest. The imaging system also enables the guidance of medical instruments inside the patient body without making an incision. The latter is commonly known as minimally invasive image-guided intervention [1, 2, 3, 4]. This approach is being increasingly adopted to many surgical applications because of its lower risk of complications, shorter patient recovery time, and therefore overall lower cost of the intervention.

Ultrasound imaging is currently employed during the intervention procedures as a supplementary modality for other real-time imaging [1], such as 2D X-ray for cardiac catheterizations. In more detail, the US modality is only considered to visualize the instrument for the local region of interest, due to the limited field-of-view of the US probe, whereas fluoroscopic imaging generally provides a much larger visualization of the procedure to guide the instrument

inside the patient body [5]. Comparing US to X-ray imaging, US images can provide richer spatial information, such as 3D imaging, and better tissue characteristics, since the X-ray imaging cannot generate a clear boundary contrast of the organs, which mostly requires an additional contrast agent to be injected to guide the operation [6]. In current procedures, US imaging is commonly considered as a supplementary modality to visualize the instrument. However, on the longer term, it is foreseen that these two modalities will be jointly used, such as performed in 2D-3D image-stitching techniques for Virtual Reality, which will facilitate the operation with a better interpretation [7].

Ultrasound imaging is typically characterized as a noisy signal modality for clinical applications because it intrinsically has limitations of low contrast and low signal-to-noise ratio [8]. Nevertheless, due to the continuous developments of computer vision and image processing techniques, the imaging quality has been significantly improved, which ensures a better visualization of the instrument and besides this, also a partial insight into anatomical structure for navigation. Consequently, the visualization of the instrument in US has become possible by advanced image processing and/or computer vision techniques, and an complementary properly designed pre-processing, post-processing, and rendering visualization algorithms [8, 6]. For interventional applications, these techniques should ensure a real-time performance. Besides, the advanced hardware and image recording techniques provide the possibilities to obtain high-quality operation-related images in an efficient way [9], which enables the researchers to analyze the image data and provide solutions for interventional guidance.

This thesis concentrates on realizing medical instrument visibility in a continuous way, such that a cardiac catheter, guide-ware or anesthesia needle becomes always visible. Furthermore, the continuous visibility of the instrument by exploiting computer vision techniques, is facilitating the reliability of the intervention and giving a higher success rate. As a result, more accurate and easier interventions can be achieved for physicians, which potentially lead to faster operation and better outcomes. The proposed technical implementation of the above concepts should be based on commonly used surgical and imaging equipment.

1.2 Examples of Ultrasound-guided Instruments Visualization

This section describes a few key applications, where US imaging is used to navigate the instrument for the image-guided intervention in the clinical setting.

1.2.1 Needle-based Intervention

In clinical practice, needles that are inserted in patients, are used for administration of medication, performing biopsies or for treatments, such as e.g. a needle is used for biopsy taking of cancer tissue to refine the decision-making for further treatment or surgery. Such a procedure is demonstrated and shown in Fig. 1.1.

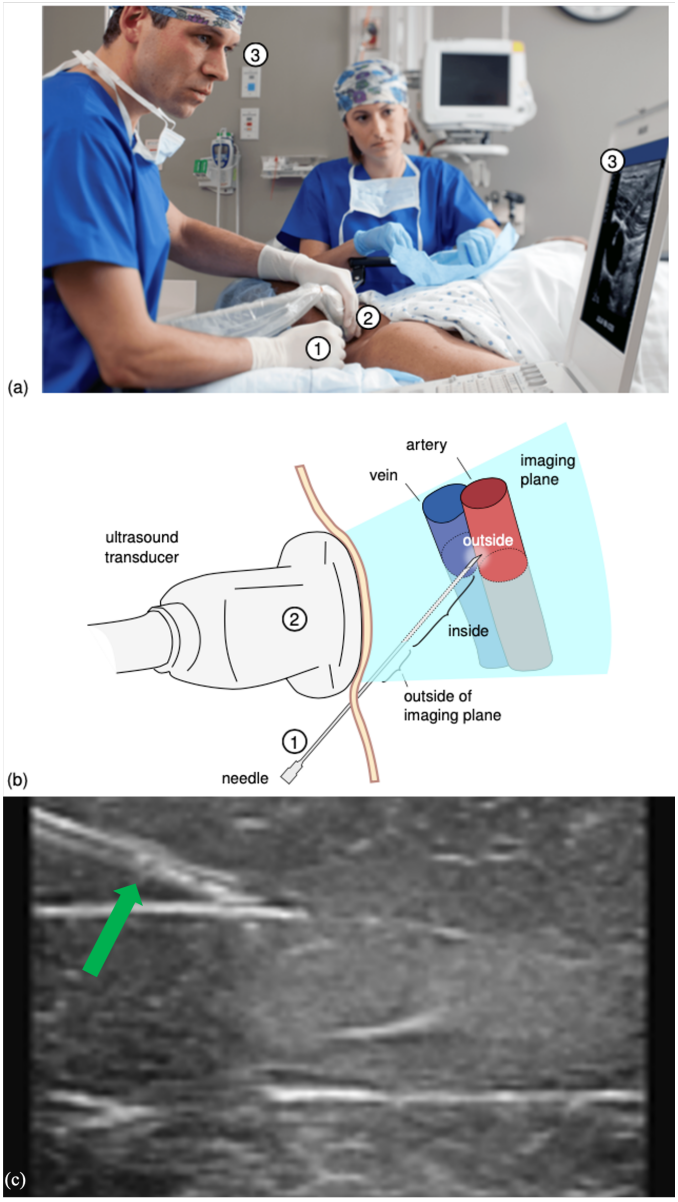


Figure 1.1 Ultrasound-guided needle biopsy taking (figure after [6]). (a) Clinical staff has to manage the multi-fold coordination of ① the needle, ② ultrasound transducer, while ③ looking at the US screen (Courtesy of Philips Ultrasound). (b) Schematic representation of guiding a needle using US imaging, depicting an example situation for vascular access. , where the needle tip is outside the imaging plane and is approaching an erroneous target area. (c) B-mode US volume slice containing the needle, pointed by a green arrow.

Fig. 1.1 (a) visualizes the physician at the left manipulating the needle with his right hand (in ①) and the US probe with his left hand (in ②). During the manipulation, the physician monitors the display of the US imaging system ③ at the right of the image. In the subfigure (b), the US transducer at the left (in ②) is radiating the sound pattern, indicated by the light blue color, into the patient tissue. The inserted needle is shown at the bottom in ① with an enlarged view in the middle subfigure, where it points to the artery at the right, inside the patient tissue. In the subfigure (c), which is commonly called a B-mode image, the resulting image of the US system is illustrated, where the needle is visible at the top-left corner, indicated by the green arrow (the horizontal white line indicates an anatomic structure of e.g. a muscle or similar element).

The difficulty of this type of imaging is that the physician has to look to the screen, while optimizing the visualization of the needle by manipulating the US transducer. More precisely, to visualize the instrument in US imaging, the 2D US plane needs to be oriented perfectly, such that it is fully parallelized with the needle orientation, and visualization of the US image coincides with the plane containing the full needle. Thus the clinical staff has to carefully align the coordinates between the needle and US probe, while looking at the US display for the operation. Besides the above orientation and alignment difficulties, it is still challenging for sonographers to distinguish the instrument from the background tissue in the B-mode images by human eyes. This aspect requires extra training and experience of the medical expert, in order to achieve a suitable instrument interpretation, such as in Fig. 1.1 (c).

1.2.2 Catheter-based Intervention

In catheter-based cardiac operation, a catheter is inserted into the patient heart chamber for tissue ablation or signal measurement. Such a procedure is demonstrated and shown in Fig. 1.2. Fig. 1.2 (a) depicts the overview of the capturing process of such US images. The US probe, commonly known as Transesophageal Echocardiogram (TEE) probe, is inserted into the patient body through the esophagus to reach the position close to the patient heart. In subfigure (b), the catheters reach the heart chambers via a patient vein or artery. The activities of the instrument are recorded by the TEE probe close to the heart. In subfigure (c), a captured 3D volumetric dataset is demonstrated by a standard rendering technique, which is difficult for interpretation. Finally in subfigure (d), the manually sliced B-mode image from (c) is visualized, which contains an instrument (indicated by the green arrow).

The difficulty of this type of imaging is similar to the needle-based operation, i.e. the multi-coordination manipulation required by the physician. By comparing catheter and needle-based interventions, the main difference in this section is coming from the imaging modality, where the 3D TEE probe generates the 3D images such that the careful alignment between instrument and US acoustic

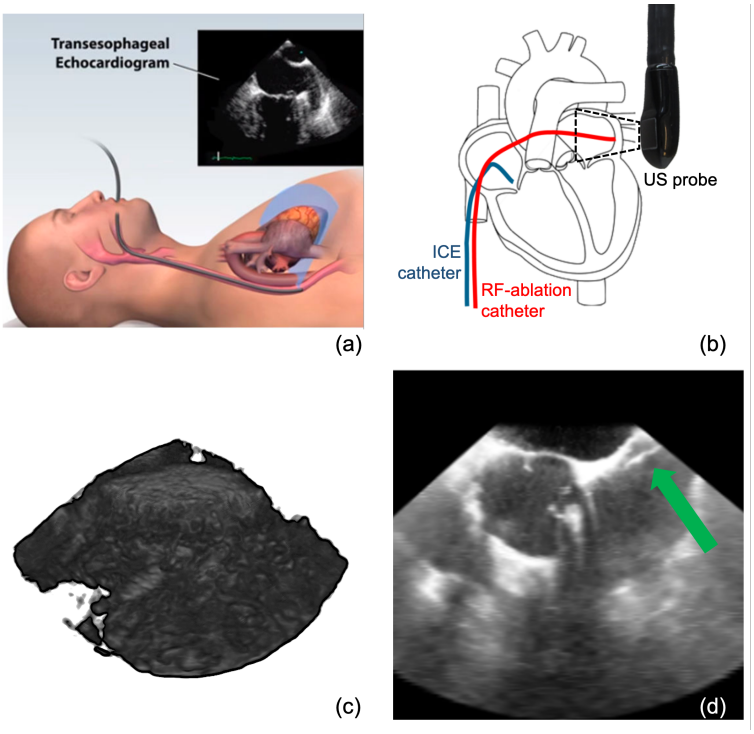


Figure 1.2 Example of 3D US imaging steps in cardiac operations. (a) The US probe is inserted into esophagus of the patient, which is passing by close to the patient heart [5]. (b) An example of cardiac catheterization in the heart with a phased array probe placed next to the heart chamber. Generally, catheters are inserted into the heart chambers. (c) The 3D volumetric dataset is rendered by standard techniques, where it is hard to interpret the image and localize the instrument in the volume. (d) The manually sliced B-mode image from the 3D volume, which contains the instrument (pointed by the green arrow).

plane is not required. However, due to the complex 3D US imaging with rendering techniques, it is still challenging for sonographers and physicians to tune the 3D US image such that the 2D plane or slice contains the instrument with good visibility. In addition, it is challenging to distinguish the instrument in B-mode images for human expert eyes, which requires extra training and experience from the medical expert.

1.3 Ultrasound Guidance of Instrument

Due to challenging conditions of the instrument visualization, such as limited field of view, complex anatomical structure, and low image contrast, extensive tuning and manipulation of instruments and US transducers are performed to continuously capture an instrument. This advanced manipulation and tuning

may lead to image interpretation errors during the operation and thereby lead to higher risks for patients.

A. Instrument Detection Techniques

To successfully find the medical instrument during the operation, instrument detection techniques have been studied in recent years, and various solutions have been developed to improve the manual alignment between the instrument and US transducers, which can facilitate the visualization of the instrument. These solutions employ various techniques, which can be classified into two categories. First, detection of the instrument by external or internal sensing devices (hardware-based), such as optical fiber sensing [10], electromagnetic tracking [11], and robotic-guided detection [12]. Second, the use of image-based approaches for instrument visualization, without employing any additional sensors. Although sensing-based methods have achieved promising results, the relatively high cost of their equipment and the involved sensors complicate the system set up in the operation room. Therefore, clinical users do not broadly accept the sensing-based approaches in practice. In contrast, image-based approaches have been proposed to detect the medical instrument in US images, since this solution is less obtrusive to the clinical user, but it is difficult to obtain a robust implementation. In addition, based on available equipment in the hospital, an image-based method provides a less costly and affordable solution. Therefore, we aim at image-based medical instrument detection for clinical applications to ensure the visualization of the instrument during surgery.

B. Potential of Image Processing

Image-based instrument detection is a promising technique for clinical practice because it minimizes the instrument-probe manipulation, while preserving the usage of conventional instruments and image recording systems [13]. This system approach is solely based on image processing techniques to improve the instrument representations. With advanced techniques of image processing and computer vision, the target instrument can be enhanced to obtain better visualization with improved contrast and visibility. Furthermore, with consistent detection of the instrument in US data, this approach will support the guidance of the intervention by enabling continuous visualization of the instrument. Therefore, with optimized viewing and image-based techniques, the instrument detection is considerably simplifying the placement of the instrument within the acquired field of view of the US probe.

To develop an image-based detection system, advanced image processing and computer vision techniques are investigated. Several studies have been conducted in the past two decades using image-based methods. However, these methods are lacking validation on realistic datasets or require specific recording conditions, such as recording images from phantoms or computer-generated

datasets. Despite the fact that these solutions have resulted in an acceptable performance for specific applications, the reported methods lack robustness in offering anatomical background when applied to real clinical situations. In recent years, the fast development of combining advanced image processing and artificial intelligence provides a new direction of designing an image-based medical instrument detection, which overcomes various challenges and yields novel and improved support for US-guided interventions.

1.4 Artificial Intelligence

The rapid development of digital computing has enabled computer systems to grow in an exponential rate, as predicted by Gordon Moore [14]. Digital signal processing techniques have been extensively applied in medical imaging, which can execute the complex and expensive computations with real-time operation, thereby increasing the clinical value of the different imaging modalities. Many signal processing components, such as de-noising and contrast enhancement, have been widely applied to imaging data, which has improved the image quality to a level that enables accurate image analysis. In addition, this rapid development in computation capabilities of processing units has resulted in a new period of artificial intelligence (AI), which is compounded with advanced sensing technologies. Nowadays, advanced (image) signal processing systems with AI, advanced sensors and smart devices have been deeply integrated into the daily life of millions of people in many different ways, such as the digital assistant on mobile phones, but also popular applications (Apps) like TikTok making use of AI.

The rapid development of AI has also a great impact in the healthcare area. Several essential domains have already experienced revolutionizing developments, ranging from cancer diagnosis, medicine development and computer-assisted intervention. These advanced techniques can help medical people and patients to perform a more efficient and effective healthcare treatment, by incorporating e.g. AI-based monitoring, or obtaining suggestions for personalized treatment. By employing AI in medical imaging systems, there will be a large impact on the outcomes of patients, by providing a more accurate diagnosis and treatment solution. Specifically, for US imaging empowered by AI, a faster and accurate image acquisition and interpretation can be achieved, which extends the usage of this modality to novice and non-expert practices. In addition, the advanced processing of US images during the intervention can enable a better interpretation of the image content and facilitate the operation with better understanding. Therefore, a safer and better intervention outcome for patients can be obtained at a faster pace.

1.5 Requirements and System Aspects of the Research

Advanced and robust processing solutions for accurate and efficient medical instrument detection for US-guided interventions is useful for clinical applications, since they provide potential improvement of the accessibility and suitability of minimally invasive procedures. Thanks to fast developments in image processing techniques, 3D US is gradually becoming feasible in clinical practice, because it provides richer spatial information and better instrument-related information capturing than conventional 2D US. Nevertheless, due to the large required space for volumetric data recording and limited image representations on a 2D monitor, it is challenging to interpret the 3D images with complex image manipulation, such as image-plane selection, probe rotation, etc. Therefore, it is time-consuming and inconvenient for sonographers or surgeons to manipulate the devices and instrument for maximum instrument visibility, rather than focusing on the operation itself. With the rapid development of AI-based techniques and advanced signal processing solutions, this limitation can be addressed, such that an optimized and automated visualization of the medical instrument in a 3D volume or in 2D planes is realized. This thesis focuses on developing robust and efficient image processing solutions for medical instrument detection, or more specifically, localization and segmentation in 3D US data volumes, enabling simplified manual coordination of imaging equipment and instrument.

When developing such intervention support system, both technical and clinical challenges should be identified and addressed. In the following, the most important considerations and aspects are summarized, which are essential for this study.

1.5.1 Data Acquisition and Standardization

An essential step for developing an AI algorithm is having access to large amounts of data. In the healthcare domain, there is a natural limitation to collect a large amount of clinical datasets. These limitations include technical limitations for advanced studies, but also ethical limitations (privacy rules and governmental regulations). Medical imaging systems are less available than commonly used video cameras, which limits the data collection in both efficiency and accuracy. In addition, the collection of datasets for development of novel techniques are commonly involving new and immature procedures, which are typically avoided for considerations of patient safety, ethical evaluation, and privacy protection. Moreover, the gold-standard data or so-called ground-truth data for AI development is obtained in practice from highly skilled medical experts, which are scarcely available and validated data are expensive to obtain.

All the above difficulties have made the medical datasets expensive and with limited availability. This has motivated the researchers to choose sometimes another experimental direction, such as simulation experiments using phantoms (*in-vitro*), either on animal tissue or cadaver (*ex-vivo*), and sometimes incidentally

on live animals and patients (*in-vivo*). Although most research studies from recent years are gradually concentrating on *ex-vivo* and *in-vivo* datasets, the limited amount of validation datasets still hampers the progress of the designing a reliable automated system.

1.5.2 Reliability and Decision-making Motivations

During the operation, it is important for surgeons to consider the uncertainty of a procedure for each decision, which ensures the treatment can be controlled to avoid any high-risk situation. By considering an automated detection system in clinical practice, it is important for physicians to understand the risk and uncertainty of the applied solution. Therefore, a reliable quality evaluation and an acceptable error margin of such a system should be provided to support the decision-making of operators, which ensures the surgeon can perform the operation with known risks. In addition to the above reliability of the system design, it is also important for clinical experts to understand the motivation for using such a system. When being able to explain the reason for each design step in the system, clinical specialists can gain trust in the detection results of the system, and know-how to integrate the system into operation procedures, thereby making every decision during the intervention responsible and accountable.

1.5.3 Applicability, Execution Speed and Accessibility

Clinical solutions designed for medical instrument detection and guidance are subject to strict requirements by the type of application. One of the major considerations in the intervention operation is to achieve the optimal procedure as fast as possible, to minimize the risk of complications. Therefore, the applicability of a detection system should not divert too much from the current work flow and should proceed without interruption of the normal decision-making and device-handling procedures. Moreover, the execution speed with real-time performance is also essential for the system design, since it avoids any delays in the operation stages. Moreover, the system should limit any additional complexity for the operation system, while ensuring that the end-users have know-how and control on the detection and guidance steps of the system during the intervention.

1.5.4 Evaluation Criteria

Automated assistance and guidance systems for healthcare are usually aiming at improving decision-making and treatment accuracy. Therefore, an accurate performance evaluation is important for the system development. Nevertheless, the direct contribution of such a system to patient health may not be always clear, not visible, and/or not measurable. Therefore, it is important to evaluate the system w.r.t. the intended application in clinical practice. In addition, the performance of such a system may be subject to variations due to the different handling of

the specialists. Therefore, the applied performance criteria for a clinical solution should be objective and quantitatively measurable. This approach has been followed also for this thesis, where in each chapter metrics for quality and accuracy are addressed.

1.6 Problem Statement and Research Questions

The presented technical and clinical challenges in developing a robust and efficient detection system for medical instruments result in the following problem statement of this thesis.

> **Problem Statement:** This thesis aims at the design and development of robust and efficient processing algorithms to detect and localize a medical instrument within the acquired 3D ultrasound volumetric data. This detection and localization is based on exploiting machine learning techniques, in order to properly present the medical instrument to the clinical expert and to support the interventional guidance.

An efficient automated detection is required to support the interventional guidance. As a result, a fully automated instrument detection algorithm should be exploited with at least near-real-time execution speed, e.g. 1-5 frames per second for the observed 3D volumetric data (based on algorithm complexity and image size).

Research Questions

From the above problem statement, specific research questions (RQ) are formulated below.

RQ1. Features and modeling of the instrument for an automated detection system

In order to design a system that can robustly detect the instrument in noisy acoustic image data with complex anatomical structures, it is necessary to design a full 3D feature description and modeling of the instrument and its background, which ensures a robust separation in 3D imaging. This leads to the following related research questions.

> **RQ1a.** What are good discriminative shape features for a medical instrument?
>
> **RQ1b.** Is it possible to model a curved instrument in 3D image data based on position information, from the initial classification of voxels?

RQ2. Pre-modeling and robustness of the detection system

The voxel-level detection may lead to redundancy and hamper the execution performance of the detection algorithm. In addition, increased robustness is needed to detect the instrument in complex anatomical US images. These topics lead to the following research questions.

RQ2a. For efficient removal of irrelevant voxels, is pre-modeling and selecting the voxel points in 3D volume data a feasible solution?

RQ2b. Can we directly describe the instrument within a deep neural network using (partially) 3D US data, to potentially improve the detection accuracy or improve the detection efficiency?

RQ3. Exploitation of the 3D context and near-real-time instrument detection

During the operation, the instrument only occupies a small volume of the acquired US data. Consequently, algorithms applied to the total US volume are less efficient in computation, thereby hampering execution speed. Exploiting the 3D complex context for the separation algorithm can improve the detection robustness. These aspects lead to the following research questions.

RQ3a. Can we implement an efficient and robust region-of-interest (ROI) instrument detection by means of a deep learning method?

RQ3b. Can the instrument be robustly segmented by deep learning after applying the ROI methods? Does this technique provide a more meaningful semantic model and improve the robustness against challenging volumes containing anatomical structures?

RQ4. Annotation-efficient training of a deep learning method for instrument detection

To train a successful deep learning model, a large amount of training data are required to ensure the robustness and accuracy of the prediction. Nevertheless, it is challenging to collect sufficient data with annotations for accurate model training purposes. To address this challenge, the following research questions are formulated.

RQ4a. Is it possible to train an efficient coarse localization method to find the sub-volume containing the instrument without requiring accurate voxel-level annotation?

RQ4b. How should the region of interest be segmented with a deep learning method, when the network is trained by only a small amount of annotated data (or even without) at voxel-level and a large amount of unannotated data?

RQ5. Real-time detection of medical instruments

A deep learning method for 3D US images involves a complex 3D model with large memory, which involves patch-based processing due to limited hardware capability. Therefore, a real-time performance is difficult to achieve and hampers the use of contextual information. These limitations lead to the research question stated below.

> **RQ5.** Is it possible to reduce the complexity of a 3D model by decreasing the number of dimensions in the neural network modeling, rather than reducing the image-scale/resolution, CNN filter sizes, etc.?

1.7 Scientific Contributions

The scientific and technical contributions of this thesis can be divided into 5 categories, which are summarized below.

1.7.1 Contributions to Feature Analysis and Instrument Model-fitting Algorithm

A two-stage algorithmic approach is proposed to detect an instrument in 3D US data. In the first stage, multi-scale features with multiple feature definitions are introduced to capture the discriminative information of the instrument to extract the corresponding instrument voxels from a complex 3D US volume. In the second stage, a specifically designed model-fitting algorithm is employed, which is able to localize the curvature instrument in 3D US images. Although the described two stages of the algorithm form a novel detection method for finding instrument in US data, they are based conventional computer vision techniques.

First, the classification of the proposed features using a non-linear classifier has achieved on average a Dice score of 83.7% on an *in-vitro* dataset, 55.2-67.0% on multiple *ex-vivo* datasets and 52.9% on an *in-vivo* dataset (Chapter 3). Second, based on the classified volumes, the position and orientation of the instrument are extracted after the specifically designed Sparse-plus-dense Random Samples Consensus (SPD-RANSAC) algorithm. The proposed system achieves an average localization error of 1.5 mm applied to an *in-vitro* dataset, 1.5-3.0 mm on multiple *ex-vivo* datasets and 1.9 mm on an *in-vivo* dataset (Chapter 3). The presented method is a first algorithm for catheter detection in 3D US images on various datasets. Although it has a decent score, it is not optimal and serves as a baseline for further algorithm development, where deep learning is extensively employed.

1.7.2 Contributions to Pre-modeling and Robust Voxel Classification

A voxel-of-interest deep learning-based classification framework is proposed to efficiently detect the instrument in complex 3D US images. First, a pre-filtering on the US images is applied to skip the non-interest voxels based on prior knowledge of the instrument shape. This shape is pre-analyzed by the Frangi vesselness filter [15], which efficiently reduces the computation load for complex deep learning classification. Second, a specifically designed convolutional neural network (CNN) is developed using a tri-planar approach for projecting 3D CNN input on these planes, which further reduces the computation load compared to straightforward processing on 3D US images.

Quantitative analysis of the proposed method shows the proposed voxel-of-interest pre-selection accelerates the overall procedure time from more than 100 seconds to approximately 10 seconds per volume on a standard PC with a Titan 1080ti GPU. The overall classification achieves about 53.7% recall and 59.1% precision on an *ex-vivo* dataset. The final localization error is about 1.7 mm for a catheter with a diameter of 2.3 mm. Therefore, the described system in Chapter 4 is able to improve the detection of the instrument, compared to the baseline algorithm with computer vision techniques, as presented in Chapter 3. Also, an ablation study shows a higher efficiency than a conventional exhaustive and iterative classification strategy.

1.7.3 Contributions to 3D Context and Semi-real-time Segmentation

A novel patch-of-interest deep learning method is proposed to semantically segment the instrument in 3D US images, which overcomes the limited use of context information in conventional voxel-wise classification methods. The proposed framework employs a fast region-of-interest selection method based on a 2D slice-based CNN with a large field-of-view, where the patches are processed by a regional patch-based semantic segmentation, using another novel 3D CNN trained by a contextual hybrid loss function.

The proposed system in Chapter 5 acts as a coarse-to-fine framework, which addresses the challenging class imbalance issues for 3D US images containing a small instrument such as a catheter. Moreover, it decreases the computational load for the 3D semantic network, thereby reducing the overall execution time. The proposed method achieves a Dice score of about 66.5 vs. 70.5% on *ex-vivo* and *in-vivo* datasets, respectively, with an execution time of about 1.3 seconds per volume on a standard PC with a Titan 1080ti GPU. This performance ensures a near real-time operation in clinical applications. Additionally, the proposed method with high Dice score ensures a successful instrument detection with a localization error of 2-3 mm only.

1.7.4 Contributions to Annotation-efficient Deep Learning Analysis

One challenging issue for a data-driven method following an AI approach, such as CNN for deep learning, is the large amount of labeled data that is required for training the model. Besides, this data is expensive and laborious to obtain. Therefore, a novel annotation-efficient approach is proposed to detect and segment the instrument in 3D US data, containing the following innovative aspects.

The proposed method involves a two-stage procedure. First, a deep reinforcement learning method is proposed to localize the region-of-interest containing the instrument, which does not require accurate voxel-level annotation of the instrument. Second, based on the candidate region, a novel semi-supervised trained CNN is applied to segment the instrument in the US images.

Quantitative analysis of the localization and segmentation in Chapter 6 shows that the proposed method achieves about 68.6 vs. 69.1% Dice score on *in-vivo* and *ex-vivo* datasets, respectively. The overall execution time is about 1 second per volume on a standard PC with a Titan 1080ti GPU, which is comparable to a fully supervised learning method. These results indicate that the proposed system is able to accurately detect the instrument with near real-time performance, but with much lower annotation effort (only about 30% training images require voxel-accurate annotations).

1.7.5 Contributions to Multi-dimensional Deep Learning for Real-time Detection

A novel multi-dimensional (3D-to-2D) transformation-based CNN architecture is proposed to efficiently localize the catheter and/or needle in large-volume US images. This method reduces the inherent 3D complexity of the CNN architecture, by introducing a dimensionality-reduction module, which effectively enables 2D processing. Therefore, the computation effort of the CNN model is dominantly reduced to 2D operations, yielding a near real-time execution time for detection.

Quantitative accuracy analyses of the efficient detection system in Chapter 7 on challenging *ex-vivo* datasets show that the proposed method achieves a detection error of 2.3-2.5 voxels with an execution time of 0.06-0.12 seconds per volume, which is much faster than the reported work in literature. The detection results can be efficiently visualized by the cross-section plane or 3D rendered volume, thereby creating a high clinical value for practical interventions, since it can efficiently indicate the location of the instrument and facilitate the operation procedures.

1.8 Outline and Scientific Background

This section presents an overview of the chapters in this thesis with the related publication background. As shown in Fig. 1.3, Chapter 2 provides a technical introduction of image-based medical instrument detection and related algorithmic components, which are employed in this thesis. Chapters 3-6 present various contributions and innovative algorithm designs. Finally, the conclusions and future work are elaborated in Chapter 8. The individual chapters are summarized below, including references to the corresponding publications.

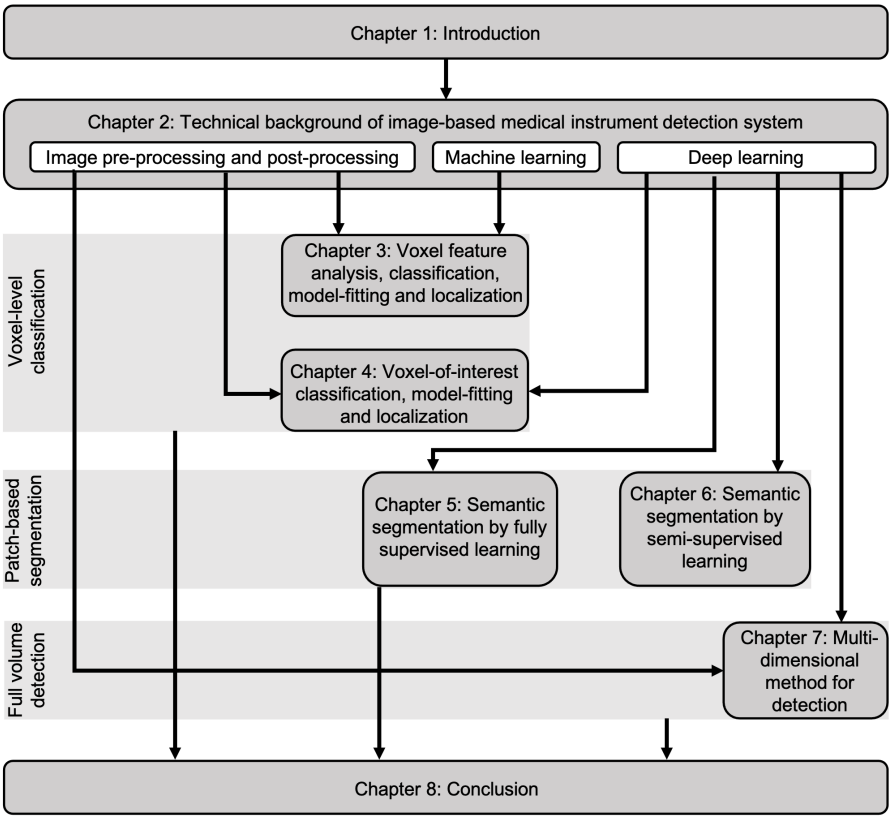


Figure 1.3 Structure of the chapter organization in the thesis.

Chapter 2

This chapter introduces a high-level overview of designing an image-based instrument detection system, based on conventional computer vision techniques. The detail of various algorithmic components are introduced. To design a medi-

cal instrument detection system based on computer vision, image pre-processing is first discussed. Then, feature analysis by exploring transformation techniques, feature design and extraction methods are used for developing feature vectors for further processing. Next, linear and non-linear classification techniques are explained for processing these feature vectors. Afterwards, a CNN is introduced as an alternative option to replace computer vision methods. Finally, post-processing, validation details and evaluation metrics are defined.

Chapter 3

This chapter presents a framework of automated medical instrument detection based on computer vision techniques, employing both feature analysis and model-fitting. The chapter is divided into two parts: 1) voxel-level classification and 2) model-fitting for localization of the instrument. Subsequently, the experimental validation is provided on phantoms, isolated tissue and animal data acquired during experiments.

The main topic of this chapter is the feature analysis and model-fitting algorithm design. This work was published in the SPIE Journal of Medical Imaging (2019) [J-7]. Moreover, the presented individual sub-systems were presented at the Int. Conf. SPIE Medical Imaging (2018) [C-8].

Chapter 4

This chapter presents a voxel-level coarse-to-fine classification system by investigating the combination of a vesselness filter and a CNN classification, which improves the overall localization efficiency. First, the vesselness filter is applied to coarsely select the voxels of interest. Second, this is followed by introducing a patch-based voxel-level CNN classification, which is applied to accurately segment the image. Subsequently, experiments with the proposed method are presented, as well as the final accuracy and efficiency of the proposed method.

The contribution of this chapter is published in a journal publication in IJCARS (2019) [J-6], which is based on a publication in the IEEE ICIP conference (2018) [C-7]. Besides this, the proposed method in the chapter has resulted in one patent application, which was published in 2020 [P2].

Chapter 5

This chapter describes the contribution of patch-based semantic segmentation by using a CNN, which involves a patch-of-interest coarse-to-fine strategy, aiming to achieve a high efficiency with high segmentation accuracy. For this purpose, a fast patch-extraction method is introduced with a 2D slice-based CNN. Next, a 3D regional CNN is applied to segment the instrument, which employs two individual sub-networks to fully exploit the spatial information, together with

a novel hybrid focal loss for learning. Subsequently, experimental results and analysis are presented on both challenging isolated heart tissue and patient data.

The contributions of this chapter were published in the Journal Medical Image Analysis (2021) [J-4]. Moreover, several international conference papers were published in MICCAI (2019) [C-5], IEEE ISBI (2019)[C-6] and IEEE ICIP (2019) [C-3, C-4].

Chapter 6

This chapter investigates the contribution of annotation-efficient segmentation by CNNs, aiming at reducing the laborious and expensive voxel-level annotation for semantic segmentation. To this end, a coarse deep reinforcement learning detection algorithm is introduced to find the potential region of the instrument. Then, a semi-supervised learning trained CNN is subsequently applied to segment the instrument, which is trained by a novel hybrid constraint to exploit the unlabeled information. The proposed method is validated in experiments and achieves state-of-the-art performance.

The contribution of this chapter is partly presented at the MICCAI Conference 2020 [C-1]. Moreover, its extension is reported in IEEE Journal of BHI [J-2].

Chapter 7

This chapter addresses a novel framework to reduce the computation cost in 3D CNN design for instrument detection. First, a multi-dimensional network is introduced with a detailed dimension-reduction module and image reconstruction steps. Second, validation and evaluation experiments are presented. The proposed method is explored for high efficiency and compared with the state-of-the-art methods.

The contributions of this chapter were published in the IEEE Trans. on BME [J-5], and also in one patent application in 2021 [P1].

Chapter 8

The final chapter summarizes the achieved results and addresses the research questions and answers found in the thesis work. In addition, this chapter concludes with a brief discussion on possible directions for the future and ways to achieve higher efficiency and accuracy.

CHAPTER 2

Overview of Image-based Instrument Detection Systems

2.1 Introduction

This chapter summarizes the state-of-the-art techniques at the time of starting research of the thesis and the latest developed image-based detection methods for medical instruments in 3D US volumetric data. As discussed in the first chapter, the focus is on image processing methods that are generic to different types of medical instruments, such as RF-ablation catheters, guide-wires and anesthesia needles, which have different appearances of both instrument and background tissue related to the intervention. These methods allow a seamless integration with existing US recording systems and introduce a minimal influence on the applied clinical protocol. When designing an instrument detection system, several techniques are exploited to identify the natural properties of instruments. For example, since the tubular shape of a catheter yields essential discriminative information, the majority of the techniques consider a vesselness filter to enhance and exploit such properties. Moreover, the echogenicity and linearity of the needle leads to the use of a parametrical transformation to exploit its discriminative information. Nevertheless, because clinical US data with complex anatomical structures introduce challenges of the intervention environment, additional preprocessing and post-processing methods are required to robustly distinguish the instrument from the anatomical environment.

In the field of image-based instrument detection, the majority techniques can be clustered into two groups, which are based on the addressed challenges in the clinical dataset and limitations of the assumed situations. These two groups are as follows.

1. *Transformation and model-fitting*: This forms the mostly studied method at the early stage, which gradually exploits the instrument-related information at different processing stages. These stages range from intensity information up to contextual model-fitting. The instrument can be detected with a carefully designed processing chain of various functions.

2. *Machine learning and prediction*: With more challenging US images containing complex anatomical environment, the straightforward definition and selection of such an intensity-based transformation is challenging. Therefore, more descriptive and discriminative information representations of data are modeled by machine learning techniques, based on a large size of image observations.

Conventional transformation and model-fitting methods have been proposed for instrument detection for a simplified and largely constrained dataset, which are unfortunately far from real-world clinical practice. In contrast, the proposed methodologies in this thesis provide an essential groundwork for a robust instrument detection system suited for clinical usage.

This chapter first presents a general architecture of the instrument detection system, which can address challenges appearing at several stages of the considered system. To this end, each stage of the system is briefly discussed, and some of the popular state-of-the-art techniques are introduced. Section 2.3 addresses pre-processing techniques to improve the quality of the US signal and approaches for selection the data of interest as the system input. Section 2.4 discusses the important stage of feature analysis, which aims at finding the characteristic properties of the instrument in the US data and creating an informative feature representation from those properties. Section 2.5 is dedicated to machine learning techniques employed for learning feature representation and deciding upon the presence of the instrument. Section 2.6 introduces the concept of convolutional neural networks (CNNs), which are able to learn and map the aforementioned processing stages automatically into a learned network. The various stages for the processing are discussed in the CNN context and design aspects. Section 2.7 briefly handles the post-processing and the concept of the model-fitting steps. Section 2.8 discusses the validation methods and appropriate metrics for objectively measuring the performance of the system. Section 2.9 concludes the chapter.

2.2 System Architecture

For a general image-based instrument detection system, discriminating information of the instrument or tissue are processed to enhance, detect, and/or visualize the target instrument in the US data. Since the complexity of clinical data introduces challenges for the instrument detection, auxiliary information from multiple information representations should be exploited to improve the performance

and robustness of the detection. This is commonly achieved by a learned model for high-level information processing, which performs better in multiple aspects than conventional image transformation, such as the intensity project [16]. The mentioned two groups from Section 2.1 form the essential stages for an automated instrument detection system, which are schematically depicted in Figure 2.1.

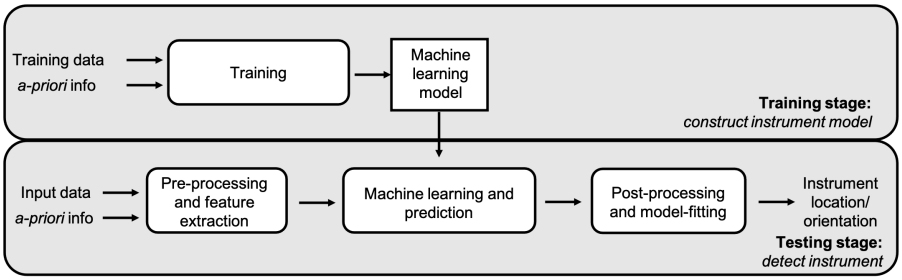


Figure 2.1 System architecture of a general image-based instrument detection system.

As described in Fig. 2.1, the key element of a general automated instrument detection system includes of a training stage, which constructs the model and decision criteria, based on available training data and *a-priori* knowledge. The training stage may be simplified to include a straightforward intensity-based threshold, or distinguish the properties of the instrument by shape analysis. In contrast, for a machine learning framework, the training stage includes exploiting a specific learning algorithm and parameters to achieve the highest and fastest detection results. Based on the model from the training stage, the testing stage of the architecture can be clustered into three steps, which are described as follows.

1. *Pre-processing and feature extraction*: The original US images are typically not suitable for automated detection of the instrument, i.e. mismatch in intensity dynamic range or too large image size. As a consequence, various steps such as image normalization, enhancement and the selection of a region-of-interest area (or voxel-of-interest) of the input image are required. Example steps of zero-mean normalization, discriminating representation extractions or region-of-interest extractions are applied to enhance the discriminating information and discard the redundant processing components.

2. *Machine learning and prediction*: The processed images are used for the instrument model derived from training stage, which can be a simple thresholding, or classifying the extracted feature vectors in a trained machine learning model.

3. *Post-processing and model-fitting*: The outcome of the machine learning and prediction step is further analyzed by morphology operations to omit outliers and noisy predictions, which are then processed by a model-fitting algorithm to localize the instrument in 3D space. The outcome results in the correct instrument voxel groups with their corresponding locations and orientations.

It is worthy to mention that the feature extraction and model training can be directly replaced by one of the latest end-to-end deep learning approaches for better information representation and learning. This approach is discussed in a later section of this chapter. The following sections will briefly introduce the commonly used techniques for each stage. Afterwards, a section is dedicated to novel methodologies of neural networks for developing the instrument detection system in a machine learning network.

2.3 Pre-processing

Pre-processing steps typically include two different techniques, such as image normalization and region-of-interest (ROI) pre-selection. Image normalization techniques are commonly applied to re-scale the voxel intensity range, which enforces the images from different recording sessions to obtain a similar response range. This transformation leads to processed images having similar appearance for the feature extraction stage, thereby achieving a higher performance (especially for deep learning applications). Moreover, the ROI pre-selection can reduce the computational load from the whole image level to a local area (e.g. window) of the image, which drastically improves the computational efficiency of the detection algorithm and leads to a real-time performance. The image normalization techniques and ROI selection methods are briefly introduced in this section, which form the basis of the system in the following chapters.

2.3.1 Image Normalization Techniques

Images for testing are commonly normalized to match the intensity range of the training images, which ensures that the constructed model processes the information within a fixed intensity distribution. Mostly, two different types of normalizations are applied for image processing: Z-score normalization and image subtraction.

A. Z-score Normalization

The first type of normalization is denoted as Z-score normalization and is defined for an image I by

$$I_{\text{Z-scored}} = \frac{I - \mu_I}{\sigma_I}, \tag{2.1}$$

where μ_I is the average of all the pixel/voxel intensity values and σ_I is the standard deviation of these intensity values. This transformation enforces the image to have a zero mean and unity standard deviation, which makes all the tested images to obtain a similar distribution at the image level. Alternatively, this transformation is also commonly applied without dividing by σ_I, i.e. only centering at zero value is performed. A typical example of this transformation is performed at the entrance stage of the VGG network [17].

B. Image Subtraction

Another image normalization technique for deep learning is defined as the subtraction of the averaged training samples. Specifically, for a training dataset, the averaged image $\bar{I}$ is obtained by adding all the observed training samples with the same coordinates, which is defined by

$$\bar{I}(x,y) = \frac{\sum_{i=1}^{N} I_i(x,y)}{N}, \tag{2.2}$$

where $I_i(x,y)$ is the training sample at position (x,y), the value N is the amount of training samples and index i is the index of the image. As a result, the normalization is defined as

$$I_{\text{normalized}}(x,y) = I(x,y) - \bar{I}(x,y), \tag{2.3}$$

where I is the test sample. By doing so, the test image obtains an attention region for the deep learning networks, because the static stationary structures are subtracted so that informative change points are notable for the networks. This improves the training efficiency and testing accuracy of the network. This technique is commonly applied to small patch-based segmentation or classification [18], since this pre-processing assumes the target objects are mostly at the center of the images. An example of image subtraction normalization is shown in Fig. 2.2, which is based on the method of voxel-based tri-planar classification for catheter detection [19]; this is addressed in detail in Chapter 4.

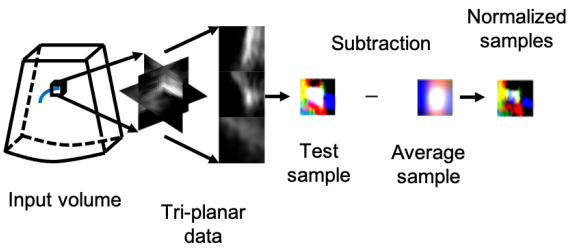


Figure 2.2 Visualization of image subtraction for voxel-based tri-planar classification.

2.3.2 ROI Pre-selection Techniques

With proper pre-processing techniques, the input images at the testing stage are commonly further processed by ROI pre-selection, which automatically selects a much smaller region-of-interest area containing the considered object. Since medical instruments are mapped into about 1/2,000-1/1,000 voxels of the whole image, the pre-selection can drastically reduce the overall computation cost and therefore improve the detection efficiency, which is essential for clinical practice with real-time performance requirements. Typically, there are two approaches to achieve the pre-selection purpose, known as voxel-of-interest (VOI) and region-of-interest (ROI) methods, as they are used for different purposes following these pre-selection methods.

A. VOI Pre-selection

As for VOI pre-selection, the candidate voxels belonging to the instrument are selected by voxel-level processing. Example VOI pre-selection is obtained by applying a Frangi vesselness filter on the volume, which is then thresholded to select the most confident voxels. Then, the selected voxels are classified by a machine learning classifier, such as Adaptive boosting or ConvNet (CNN), to further remove the non-instrument voxels. The overview illustration is depicted in Fig. 2.3. More details of the Frangi vesselness filter and adaptive thresholding are discussed in a later chapter.

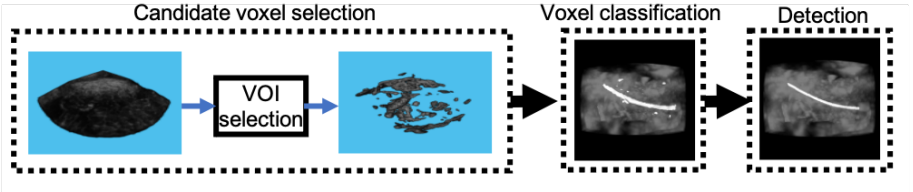


Figure 2.3 VOI pre-selection-based instrument detection. The voxels of interest are pre-selected to remove background and tissue voxels. Then, the remaining voxels are classified by the CNN for catheter segmentation.

B. ROI Pre-selection

In contrast to VOI pre-selection, which is processed by voxel-level classification, the ROI pre-selection is commonly succeeded by patch-based segmentation [20, 21, 22]. Specifically, this approach typically applies coarse and fast segmentation or detection techniques, to correctly localize the target location in the 3D volumetric data. Then, the local patches or regions containing the (parts of) target objects, are extracted for the second-stage processing, which may include Kalman filtering [20] or a more accurate semantic segmentation [21, 22]. With this

coarse-to-fine strategy, a more complex technique can be considered to provide an accurate classification result from the selected region. However, such processing is computationally expensive for full-image processing. An example of this framework is depicted in Fig. 2.4.

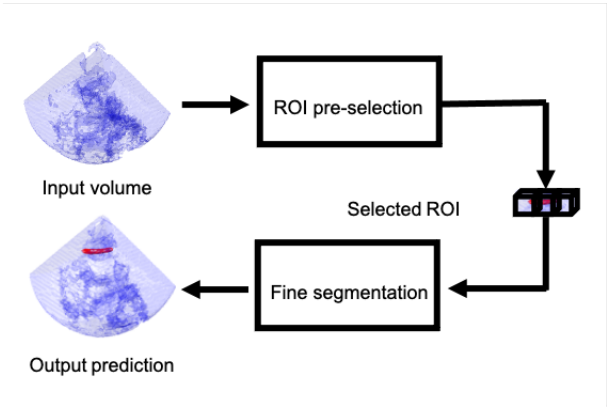


Figure 2.4 Visualization of the instrument detection using ROI pre-selection as the first pre-processing stage.

2.4 Feature Analysis

This section briefly introduces feature analysis and extraction techniques, which are specifically related to machine learning-based methods. In addition, to better model and represent the instrument in 3D US data, multiple measures are employed to form feature vectors that capture the shape and curvature information of the instruments. This procedure serves as the essential stage of the system discussed in Chapter 3.

A. Gabor Transformation

Gabor theory has been proposed to quantify the information capacity of the signal and gives the basis of signal representation by choosing an elementary function [23]. The Gabor elementary functions exploit time and frequency simultaneously with optimized resolution. In the Gabor expansion, the signal without infinite duration is defined with a certain inaccuracy, i.e. uncertainty relation, which is modulated by a Gaussian-shaped pulse having harmonic oscillations. Any signal can be formulated in terms of these elementary functions, which includes time analysis and Fourier analysis as extreme cases. This motivation of time analysis was described by Daugman [24] in an early study, where he exploited a neurophysiological plausible analysis of the uncertainty relation of the Gabor elementary functions in 2D space. Since then, vast applications of Gabor

filter have been proposed in the image processing domain, such as texture analysis and edge detection [25, 26]. Later on, Gabor transformation-based instrument detection techniques were proposed to exploit the instrument in US data by considering textual and edge information, which showed promising results in both 2D and 3D images [27, 28].

The conventional Gabor elementary function is expressed in Cartesian space in complex form, i.e. including real and imaginary parts. Nevertheless, this Gabor-based processing is influenced by a DC component of the signal [29]. To stabilize the performance with varying DC components and gray-value variations in US images, the log-Gabor function is commonly adopted. Specifically, the 3D log-Gabor elementary function $\mathcal{G}$ [29] in the frequency domain is specified by

$$\mathcal{G}(\omega,\phi,\theta) = \exp\left(\frac{\left(\log\frac{\omega}{\omega_0}\right)^2}{2\log\frac{B}{\omega_0}}\right) \times \exp\left(-\frac{\boldsymbol{\alpha}\left(\phi,\theta\right)^2}{2{\sigma_\alpha}^2}\right), \tag{2.4}$$

where B denotes the bandwidth of the filter in polar coordinates and ω_0 is the filter's central response frequency. The term B/ω_0 is set to a constant value to obtain constant-shape ratio filters. The direction of the filter is defined by azimuth angle ϕ and elevation angle θ. The position vector $\boldsymbol{\alpha}(\phi,\theta)$ at given frequency f point is defined as $\boldsymbol{\alpha}(\phi,\theta) = \arccos((f\cdot d)/|f|)$, where the unit direction vector is $\mathbf{d} = (\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi)$. The bandwidth of the angular direction is defined by σ_α. The discriminative information is extracted based on the real part of the frequency response, as it is focusing on the symmetric response on line detection. Specifically, the central frequency ω_0 of the Gabor function is characterized by the diameter of the instrument d_i, which is defined as $\omega_0 = 1/(2d_i)$ [30]. The resulting Gabor transformation is obtained by applying the convolution of the filter $\mathcal{G}(\omega,\phi,\theta)$ with the input volume.

B. Objectness Feature

Multi-dimensional objectness was first introduced by Antiga [31], who extended the traditional definition of the vesselness filter into the different shape descriptions for multi-dimensional images, see Fig. 2.5 as an example. For 3D images, the Hessian matrix is defined below, where f^σ is a Gaussian-filtered image with standard deviation σ, while $f_{xx}, ..., f_{zz}$ represent the second-order derivatives in the $x-$, $y-$, or $z-$directions. The Hessian matrix with these derivatives is specified in matrix form by

$$H_\sigma = \begin{bmatrix} f^\sigma_{xx} & f^\sigma_{xy} & f^\sigma_{xz} \\ f^\sigma_{yx} & f^\sigma_{yy} & f^\sigma_{yz} \\ f^\sigma_{zx} & f^\sigma_{zy} & f^\sigma_{zz} \end{bmatrix}. \tag{2.5}$$

From the Hessian matrix, the Eigenvalues are computed, which can be used to derive specific shape parameters R_A, R_B and S [31]. These parameters are defined by

$$
\begin{aligned}
R_A &= \Big(|\lambda_{M+1}|\Big) \Big/ \left(\prod_{i=M+2}^{3} |\lambda_i|^{1/(3-M-1)}\right), \\
R_B &= \Big(|\lambda_M|\Big) \Big/ \left(\prod_{i=M+1}^{3} |\lambda_i|^{1/(3-M)}\right), \\
S &= \sqrt{\sum_{j=1}^{3} \lambda_j^2}.
\end{aligned}
\tag{2.6}
$$

$$
O_\sigma^M = (1 - e^{-R_A^2/2\alpha^2}) \cdot e^{-R_B^2/2\beta^2} \cdot (1 - e^{-S^2/2\gamma^2}). \tag{2.7}
$$

The Eigenvalues of Eqn. (2.5) are ranked by $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$. Using these Eigenvalues, the M-dimensional ($M \leq 3$) shape structures are described by Eqn. (2.7) based on the parameters from Eqn. (2.6). Parameters R_A and R_B have the form of "normalized" Eigenvalues, which control the shape sensitivity of the Hessian matrix in two directions. This is implemented with parameter settings $M = 0$ for a blob, $M = 1$ for a vessel and $M = 2$ for a plate in shapes, while the Frangi vesselness filter equals to $M = 1$. Parameters R_A and R_B have two special cases. When $M = 2$, parameter $R_A = \infty$, and when $M = 0$, parameter R_B is set to zero.

As for Frangi's vesselness feature [15], parameter λ_j is set to be $\lambda_j < 0$ for $M < j \leq 3$. For the other cases of positive values of λ_j, the value of O_σ^M is set to be $O_\sigma^M = 0$. In addition, parameters α, β and γ are empirically determined, which define the sensitivity of the response [32].

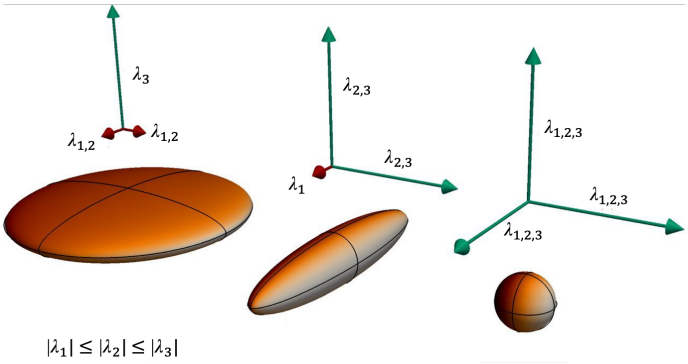


Figure 2.5 Objectness descriptors based on Eigenvalues of the Hessian matrix, showing different structures. Three coordinate systems indicate which Eigenvalues λ are dominant, leading to changes in the shape of the object. For example, in the second case at the middle, λ_1 is small, resulting into a tubular structure.

2.5 Machine Learning and Prediction

As described in the previous section, to enrich the description of the instrument in space, a feature vector is constructed such that it includes multiple descriptors, which is then classified to "instrument" or "no instrument". Typically, given a feature vector $\mathbf{x}$, the classification task is to assign it to one of K discrete classes $\mathcal{C}_k$, where $k = 0, 1, ..., K-1$. In case of classification, the assignment is unique that only one class is assigned for an input $\mathbf{x}$. As a result, the multiple inputs with different feature elements' values are divided into several regions by the classification decision boundary. In terms of instrument detection, this multi-classification is degraded to a two-class problem. In the following paragraph, two typical machine learning classification methods are introduced, which support the instrument voxel classification solutions proposed in this thesis.

2.5.1 Support Vector Machine

The support vector machine (SVM) is a popular classifier proposed by Cortes and Vapnik [33], which is based on empirical risk minimization. Given the input training feature vectors and corresponding binary label:

$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^d, \text{ with } y_i \in \{-1, +1\}\}_{i=1}^{N}, \tag{2.8}$$

where y_i denotes the class label of a feature vector $\mathbf{x}_i$ with d elements. The objective of SVM is to find the hyperplane $\mathbf{w}^\mathbf{T}\mathbf{x} + b = 0$ that maximally separates the two different classes. In the case the distributions are clearly non-linear and cannot be separated by a plane, a feature mapping $\phi : \mathbb{R}^d \to \mathcal{M}$ is applied to project the feature vectors into a high-dimensional space, which generally separates the classes more easily. The optimization based on the training dataset is achieved by the concept of *margin*, which is defined as the smallest distance between the decision boundary and training samples. Given the classification relationship between the estimated class label $\hat{y}_i$ and training feature vector $\mathbf{x}_i$, the decision procedure is defined as

$$\hat{y}_i(\mathbf{x}_i) = \text{sign}(\mathbf{w}^\mathbf{T}\phi(\mathbf{x}) + b), \tag{2.9}$$

where $\hat{\text{y}}_i = y_i$ for a perfect classification. The sign function produces a binary -1 or $+1$ output. As a result, the decision boundary of the optimized classification is to maximize the margin by optimizing the parameters $\mathbf{w}$ and b. Therefore, the maximum-margin solution is defined by solving:

$$\arg\max_{\mathbf{w},b}\{\frac{1}{||\mathbf{w}||}\min_i(y_i(\mathbf{w}^\mathbf{T}\phi(\mathbf{x}_i) + b)\}. \tag{2.10}$$

Because the distance from an arbitrary point to the decision surface is fixed after rescaling, the training points closest to the surface are satisfying

$$y_i(\mathbf{w}^\mathbf{T}\phi(\mathbf{x}_i) + b) = 1. \tag{2.11}$$

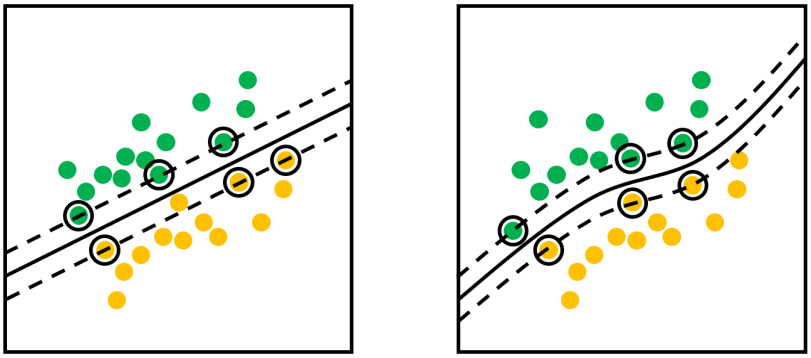


Figure 2.6 Visual illustration of SVM hyperplane separation in a 2D dataset. Left: Linear SVM, right: non-linear SVM. The support vectors are highlighted by a black circle, while the decision plane is depicted by a solid line. The margin is the distance between the two dashed lines in each figure.

This constraint leads to a case where the decision plane is only affected by the subset of the training samples that are closest to the hyperplane. Therefore, these samples are denoted as the support vectors, which are shown in Fig. 2.6. Because of this constraint, the optimization problem is simplified to maximize the distance of $1/||\mathbf{w}||$ with the constraint that the distance magnitudes at both sides of the plane are larger than unity.

When the SVM is constructed and learned, a new data point can be efficiently classified based on Eqn. (2.9), which simply determines at which side the data point is located. Because of this straightforward feature vector manipulation, SVM provides an efficient testing performance. Meanwhile, the trained parameters are consisting of simply the hyperplane information with only plane parameters and support vector points, which are used for model storage.

2.5.2 Adaptive Boosting

Adaptive boosting (AdaBoost) is an ensemble learning method, which uses an iterative approach to learn the discriminative information from the mistakes of weak learners, such as decision stumps, and turn them into a stronger classifier [34]. AdaBoost combines several base and simple classifiers, i.e. decision stumps, to form one optimized classifier. Based on several decision stumps, AdaBoost is constructed sequentially and the mistakes of previous models are learned by their successors. By doing so, the model dependencies are exploited based on the mislabeled examples. An example of AdaBoost and decision stumps are visualized in Fig 2.7, where the model is constructed based on several weak

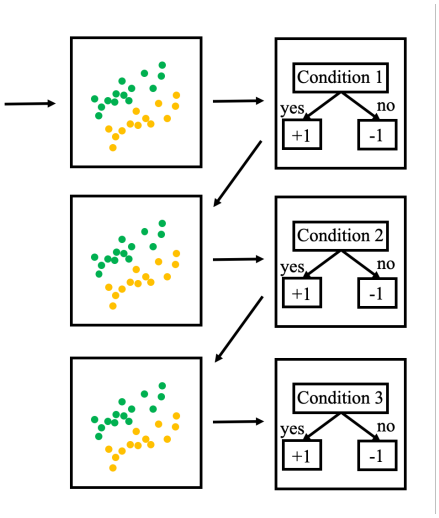


Figure 2.7 Overview of AdaBoost based on decision stumps. The model is constructed based on a sequence of weak classifiers, incorporating decisions of previous classifiers.

learners with sequential decisions. Given the input $\mathbf{x}$ and corresponding label y, the overall classifier $H(\mathbf{x})$ is defined as

$$H(\mathbf{x}) = \text{sign}\big(\sum_{t=1}^{T} w_t h_t(\mathbf{x}) \big), \tag{2.12}$$

where $h_t(\cdot)$ is the decision stump, while parameter w_t is the corresponding weight of the stump, and t is the index of T weak learners. Specifically, the weight w_t is initially defined as $w_t = 1/T$, while for each training iteration i, it is updated by the classification error rate E_t, so that the weight is updated by

$$w_t(i+1) = \frac{w_t(i) e^{-\alpha_t y h_t(\mathbf{x})}}{Z}, \tag{2.13}$$

where α_t is the classification influence of the decision stump, which is defined as $\alpha_t = \frac{1}{2}\ln((1-E_t)/E_t)$, with $\ln$ being the natural logarithm. Parameter Z is used to normalize the weight to confirm that the summation equals to unity. As for α, a larger value means an overconfidence of the weak classifier, so that the exponential will reduce its weight, since it is already performs well [35].

2.6 Convolutional Neural Networks (CNNs)

Besides the aforementioned classification methods, which employ a fixed transformation function between input feature vector and classification output, a flexible parametrical estimation of nonlinear transformations can be used for classification tasks. Specifically, instead of previous methods with experience-based

feature vector design, parameters of the nonlinear transformation can be directly obtained by adaptive learning from the original input images for both feature extraction and classification. The multi-layer perceptron (MLP) network in the early stage of neural networks is the most successful example of such a parametrical model. In the multi-layer perceptron network, discriminative information is learned by decomposing the input data into a multi-level representation, which is achieved by sequentially composing simple non-linear transformation modules. Therefore, the information contents are gradually extracted from low-level information to high and abstract level [36]. Based on the sequential structure of the network, the complex and abstract patterns can be constructed using low-level components. Moreover, because of the loss function, commonly denoted as objective function, the network tends to learn and amplify the information related to discriminative elements rather than the irrelevant variants of the input [37]. A typical system architecture for instrument detection by using a neural network is depicted in Fig. 2.8, which has a similar pipeline as in Fig. 2.1, but combines the feature extraction and task decision simultaneously.

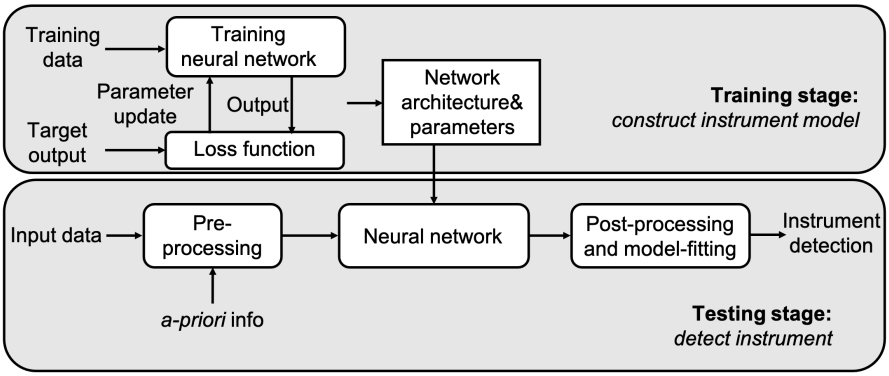


Figure 2.8 System architecture of instrument detection based on neural networks.

Neural networks (NNs) were initially considered to have limitations like lack of generalization and tend to overfit the training data with expensive computation in standard hardware. Since 2006, many papers have shown that the NN can achieve desirable results using a standard backpropagation with labeled data [38, 39]. NNs have become more attractive than in the past decades. Moreover, the improvement in computation ability of Graphics Processing Units (GPUs) and the vast amount of available data have enabled NNs to become one of the most popular techniques in machine learning and artificial intelligence. Specifically in computer vision and image processing communities, convolutional neural networks (CNNs) have achieved a dominant success because of their generalization and easy implementation properties [40].

The remainder of this section provides descriptions of neural networks, which are listed as follows. (1) A neural network, which formulates deep learning as the fundamental component. (2) Training of NNs by backpropagation. (3) Convolutional neural networks for different image detection tasks, which include classification, segmentation and reinforcement learning. (4) Commonly used architectures within the deep learning framework are briefly introduced.

2.6.1 Neural Networks

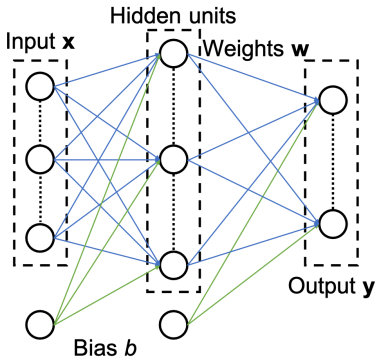


Figure 2.9 Network diagram for the two-layer multi-layer perceptron (MLP) network. The blue lines represent weight $\mathbf{w}$ while green lines mean bias b. The feedforward MLP network processes the input data from $\mathbf{x}$ to generate output $\mathbf{y}$.

The neural network typically consists of several non-linear layers, which include a linear transformation and a nonlinear operation. Specifically, for an NN with N layers, a common layer $i \in 1, 2, ..., N$ is defined as follows for input $\mathbf{x}$:

$$z_i = h(\mathbf{w}_i \cdot \mathbf{x} + b_i), \tag{2.14}$$

where $h(\cdot)$ is a differentiable non-linear activation operation. Parameter $\mathbf{w}_i$ denotes learnable weights while b indicates the bias. The commonly applied non-linear activations include the logistic sigmoid function, rectified linear unit (ReLU) or softmax function. As for a commonly used multi-layered perceptron (MLP), it is constructed by a sequence of layers. A typical two-layer network is formulated as:

$$\mathbf{y}(\mathbf{x}, \mathbf{w}, b) = h(\mathbf{w}_2 \cdot h(\mathbf{w}_1 \cdot \mathbf{x} + b_1) + b_2), \tag{2.15}$$

where parameters $\mathbf{w}$ and b include the weights and bias for all the layers. The corresponding diagram is shown in Fig. 2.9, which is a two-layer MLP network.

2.6.2 Network Training

With the defined network, and to determine its parameters, a common approach is to minimize the objective function, also known as loss function, by exploiting

input $\mathbf{x}$ and corresponding desired output $\mathbf{t}$. For instance, by a given training data set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, and their target $\{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$, the cost function can be straightforwardly defined as an error function between two terms:

$$E(\mathbf{w}, b) = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_i, \mathbf{w}, b))^2, \tag{2.16}$$

where $E(\cdot)$ is the commonly denoted mean squared error (MSE) loss function, which can be also replaced by cross-entropy [36]. Because of the nonlinear activations of the networks, most cost functions are not convex, and cannot be obtained by a closed form. Instead, the network parameters are commonly obtained by an iteratively gradient-based optimization approach, which minimizes the cost function during the training procedure. The optimization of a continuous nonlinear function can be obtained by iterative updating: a randomly initialized weight $\mathbf{w}^{(0)}$ is gradually moving to the desired state under the guidance of the negative gradient of the cost function. This is specified by

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)}), \tag{2.17}$$

where τ is the iteration count while η is a positive parameter known as the learning rate. With the updated parameters, the gradient is re-evaluated based on new input w.r.t. the target label, and then the parameter updating repeats itself until the end of the training procedures. More details of the derivation can be found in [41].

2.6.3 CNN for Image Processing

A convolutional neural network (CNN or ConvNet) is a class of neural networks for analyzing visual imagery or time series with grid-like topology [36]. The name of the CNN indicates that the network employs convolutional operations, which is a mathematical operation to correlate the information around the target point. The transformation weight $\mathbf{w}$ in a standard NN structure is replaced by a series of convolutional operations in a CNN, which drastically reduces the number of trainable parameters. Therefore, computing a CNN is affordable for a state-of-the-art GPU for training, which has resulted in a high popularity in applications and research of image processing and computer vision. Within a CNN architecture, three key components are essential to exploit the multi-scale and multi-level information of data.

1. ***Local receptive field.*** For any input image or processed result by a convolutional layer, all the resulting points are just a response to previous filters or point intensities. The local receptive field is only focusing on the neighborhood information of the point. Although the fully connected neurons in a standard NN can be adopted to learn the features, it is expensive and not practical to apply, as it leads to a high number of neurons. Alternatively, a local

receptive field extracts the regional features, e.g. edges, corners, etc., which are exploited by the sequential structure to find the contextual information.

2. *Weight sharing.* A standard image poses a large number of points, which leads to millions of weights in a standard NN. Furthermore, based on the concept of local receptive field, regional features should be invariant to location or translation. Therefore, sharing weights on the whole image can avoid overfitting of the location and achieve spatial invariant properties.

3. *Information aggregation.* With weight sharing and a local receptive field, neighborhood points will expect similar responses for the same stimulus, which unfortunately leads to spatial redundancy. Moreover, a regional receptive field only focuses on local information, while ignoring the contextual information. Therefore, in order to better exploit contextual information and reduce the spatial redundancy, information aggregation, typically known as *pooling*, is applied to concentrate neighboring information by max or mean operations. Moreover, with the combination of pooling and a local receptive field, high-level information can be processed for contextual analysis.

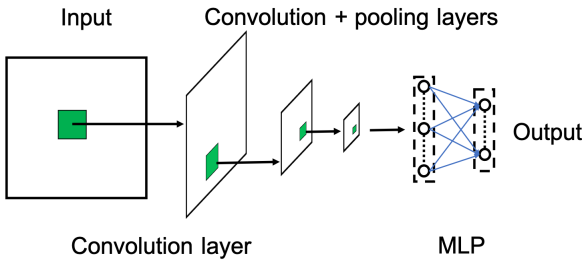


Figure 2.10 Diagram of a typical CNN architecture with convolutional layers and pooling layers, which are followed by fully-connected layers (also know as MLP). The green region in the input is local receptive field of the input convolutional filter.

A typical example of a CNN architecture for classification is shown in Fig. 2.10, which includes four convolutional layers with a multi-layer perceptron. The convolutional layers consist of several small-sized filters for local information processing. Based on the shared filters for the whole image, the spatial invariant information can be extracted and aggregated by the pooling layer. The resulting image has a compact and high-level information for succeeding layers. Finally, an MLP layer is introduced with dense connections for classification task. In this case, the non-shared weights exploit all the information from the input.

2.6.4 Deep Learning Networks and Application Tasks

As the depth of the neural network increases, the complexity and information extraction capability is augmenting. The deep neural networks are commonly

referred to after the name of the architecture, such as UNet [42], VGG [17] or ResNet [43], while sometimes the application is abbreviated and mixed into the name. The essential advantage of a deep learning network is to exploit hierarchical information from the training data involving resolution scaling, while the learning of the network is guided by a loss function that is typically minimized. Because of the data-driven optimization and automatic feature learning, more relevant information is captured from the input data, so that deep learning outperforms conventional handcrafted feature designing approaches.

Based on the popular applications of deep learning, there are two commonly considered tasks for deep learning networks: (1) image classification, and (2) image segmentation. Image classification is specifically defined as a procedure to classify the input image to a particular category by the deep neural network. A well-known example is to distinguish whether the input image includes a particular object (e.g. a cat or a dog). Alternatively, image segmentation is a more advanced task that not only distinguishes the image category, but also draws the boundary of the objects from the input image. Besides object classification and segmentation, deep networks are now also employed for motion analysis, behavior analysis, noise reduction, and other diversified tasks.

2.7 Post-processing

Based on the voxel-level classification or image segmentation, the outcome coarsely estimates the instrument location in the image data and its orientation. However, the results also include outliers or irregular structures from noisy and unseen data. Therefore, a proper post-processing is sometimes applied to refine the network outputs. A commonly used strategy includes cluster-connectivity analysis for outlier removal through binary morphological operations. Moreover, a model-fitting algorithm is commonly applied to localize the instrument by contextual level analysis. Specifically, the localization is achieved by random sample consensus, which is a typical technique for fitting a curved cylinder in a noisy dataset. As will be discussed in this thesis, instruments like a curved catheter or straight needle are localized in 3D space by sparse-plus-dense random sample consensus (SPD-RANSAC) algorithm. By removing outliers, the RANSAC algorithm acts as a filter for robust decision making.

2.8 Validation

Performance evaluation of the designed method is essential w.r.t. its designated tasks. For the medical domain, the performance evaluation with a good result not only provides detection capabilities and accuracy of that, but also implicitly indicates the stability and generalization of the learning approach. Therefore, the stability and accuracy can be obtained from the system when an unseen data is processed, which is required to assess whether the system is sufficiently powerful

for the clinical practice. A common way to evaluate a machine learning system is based on so-called ground-truth data, which is a pre-defined or measured type of information related to the input data and processing tasks. Therefore, the difference between the desired outcome and result from the system can be compared to qualify and quantify the performance.

In the machine learning area, training is defined as the procedure to generate a learned model based on the provided data, while testing indicates the evaluation step for the developed model. With a limited experimental dataset, appropriate selection and division is essential, which can reveal unbiased information for model generation, when supplying an unseen testing dataset. The key concept to split the dataset into training and testing parts without any overlapping can efficiently exploit the limited clinical dataset in practice. The most common techniques in machine learning for validation purposes are introduced below.

1. ***Leave-one-out cross-validation*** : The leave-one-out cross validation (LOOCV) is a commonly adopted technique for a limited dataset. More specifically, one sample is left out during the training, which is then used for the testing phase. For a dataset with N samples, this step is iteratively performed N times on each sample, so that the testing can be applied to all samples. The overall performance is obtained on the average of the N-times test. Because of using a single-data element for testing, a variability would be introduced from the occurring outliers. Moreover, it is expensive because the model is trained for N times.

2. ***K-fold cross-validation*** : Similar to LOOCV, instead of selecting one sample for testing, k-fold CV splits the dataset into k parts, and for each part, $k-1$ parts are used for training and the remaining one for testing. The experiment is performed k-times for overall performance evaluation. Compared to LOOCV, the performance variation is reduced, since more testing samples are included for each validation.

3. ***Dataset split*** : Another validation method is to split the dataset into training, validation and testing datasets, which is commonly applied now in deep learning methods. During training, besides the training samples for parameter optimization, a validation dataset is used to control the learning curve and evaluate the hyper-parameters that optimize the network to satisfy the loss function. Then the testing dataset is used to evaluate the final performance of the method. The validation dataset helps to provide an unbiased evaluation of the model fitting. Compared to the above cross-validation method, a dataset split provides less effort to train the model, but will overestimate the error rate at the testing phase for the model to fit on the whole dataset.

Based on the above evaluation approaches, evaluation metrics are required to quantify the performance of the designed system. Based on the tasks for classifi-

cation, segmentation and localization, different metrics w.r.t. those tasks are now introduced.

2.8.1 Voxel-level Classification and Segmentation Performance

In a binary classification system, appropriate evaluation metrics consider the relative correction of the decisions by defining the true positive, false positive, true negative and false negative results. In terms of these definitions, true and false refer to the decision agreement w.r.t. the ground-truth statement. The positive and negative labels are obtained based on the decisions from the automated classification algorithms. As a consequence, for a binary classification system for instrument voxel classification, the number of true positives (N_{TP}) is captured by the counted events of true occurrences of the desired class. A true positive indicates that the decision algorithm has correctly classified the involved instrument voxel as a positive occurrence. A false positive (where the count is the N_{FP}) indicates that a non-instrument voxels is wrongly classified as a positive class. Similarly, true negative (with total count N_{TN}) represents the non-instrument voxels are correctly classified, while false negative (which is counted into N_{FN}) indicates the instrument voxel is wrongly classified as a non-instrument region. These definitions are summarized in Fig. 2.11.

	Instrument voxel prediction	Non-instrument voxel prediction
Instrument voxel ground-truth	True Positive (*TP*)	False Negative (*FN*)
Non-instrument voxel ground-truth	False Positive (*FP*)	True Negative (*TN*)

Figure 2.11 Definition matrix of true and false positives and negatives in a binary classification problem.

Based on the above definitions, a high-level evaluation metric can be derived from these counted values, as specified in Fig. 2.11. These metrics are denoted as recall, precision, and specificity, which are defined as fractions in the unity interval based on the aforementioned counted numbers. The recall (R_c) of a binary classification algorithm indicates how many times an instrument voxel is correctly detected compared to the total amount of the instrument voxels. The precision (P_r) measures the how many times an instrument voxel detection is actually an instrument element compared to the total amount of predicted voxels. The specificity (S_p) refers to how many times a non-instrument voxel is correctly

detected when no instrument exists. These metrics are numerically computed as follows, and specified by:

$$R_c = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}, \tag{2.18}$$

$$P_r = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}}, \tag{2.19}$$

$$S_p = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FP}}}. \tag{2.20}$$

The ideal classification algorithm would have a value of unity or 100% score for the evaluation metrics. In real practice, the precision and specificity are closely related, such that a higher specificity always leads to a better precision. Nevertheless, for a highly imbalanced data distribution, for example of an instrument in 3D US, where the number of negative samples is several orders of magnitude larger than the number of instrument points, a high specificity cannot ensure a satisfactory classification result.

In case of image segmentation, a larger recall can be obtained by accepting a lower precision, and vice versa. Therefore, a typical evaluation metric for such task is considered by compounding them together, which is commonly denoted as the F_β-score. This score is specified by

$$F_\beta = (1 + \beta^2) \cdot \frac{R_c \cdot P_r}{\beta^2 \cdot P_r + R_c}, \tag{2.21}$$

where the parameter β controls the preference of a higher recall or precision performance in the results. Specifically, when $\beta = 1$, the metric is called F_1 score or Dice score, which is often practically applied for segmentation, because it focuses on the performance of true results.

Another type of evaluation metric for classification and segmentation results is to measure the volume similarity between the ground-truth volume and classified volume. Specifically, based on the definitions of N_{TP}, N_{FP}, etc., the Volumetric Similarity (VS) is defined as

$$\text{VS} = 1 - \frac{|N_{\text{FN}} - N_{\text{FP}}|}{(2 \cdot N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}})}. \tag{2.22}$$

The VS measures the overlapping regions between the ground-truth volume and classified volume, which is expected to be unity in the ideal case. In addition, the surface distance of the classified clusters in the 3D volume is measured by the Hausdorff Distance (HD), which is defined as

$$\text{HD}(A, B) = \max\big(d(A, B), d(B, A) \big). \tag{2.23}$$

In Eqn. (2.23), parameter $d(A, B)$ is the directed Hausdorff Distance, which is specified by:

$$d(A, B) = \max_{a \in A}\big(\min_{b \in B} ||a - b|| \big), \tag{2.24}$$

where the A and B denote voxel groups from the ground truth and the predicted results, respectively, and a and b are individual voxels from the corresponding groups, respectively. From this definition, the HD is commonly sensitive to outlier voxels, and can be smoothed by the Average Hausdorff Distance (AHD) and 95% Hausdorff Distance (95HD), which have different definitions of Eqns. (2.23) and (2.24). As for AHD, its distance $d(A, B)$ is alternatively defined as

$$d(A, B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} ||a - b||, \tag{2.25}$$

whereas for the 95HD metric, it is defined as

$$\text{95HD}(A, B) = Pct\big(\, d(A, B),\ d(B, A),\ 95\% \,\big), \tag{2.26}$$

where $d(A, B)$ follows the equation in Eqn.(2.24) and $Pct(\,\cdot\,, \cdot\,, 95\%)$ is the 95 percentile operation, which considers 95% voxels and excludes any noisy outlier information.

2.8.2 Instrument Localization Accuracy

The overall localization accuracy of a medical instrument detection system measures how close the estimated instrument coordinates are to the ground-truth annotation. Therefore, several instrument-level metrics are defined to asses the accuracy of the instrument position, which is visually defined in Fig. 2.12.

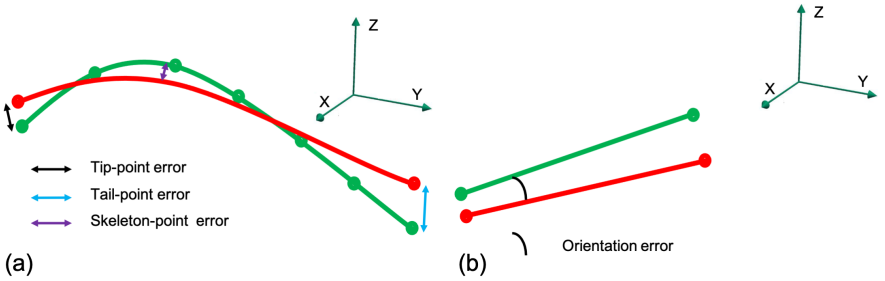


Figure 2.12 Examples of evaluation metrics for the instrument localization errors. (a) Distance error metrics: tip-point error, tail-point error and skeleton-point error. The red curve indicates the ground-truth skeleton, the green curve highlights the localized catheter skeleton. (b) Orientation metric, which measures the angle difference between the segmented skeleton of segmentation and the annotation skeleton.

As depicted in Fig. 2.12, three types of distance errors are defined: skeleton-point error, and two errors concerning the beginning and ending of the model, i.e. tip-point error and end-point error, respectively (the average of tip-point errors and tail-point errors). The latter two error types are defined as the distances between the localized point with the corresponding ground-truth point, either at

the tip or at the tail of the catheter. The skeleton-point error denotes the distance between the sampled points from the localized catheter to the ground-truth skeleton. All errors are measured visually in the images and are initially expressed in voxels, which can be translated to a distance using the voxel resolution. In addition, the orientation difference is also defined as the angle between the detected and ground-truth orientation vectors.

2.9 Conclusions

This chapter has presented an introductory overview of the common techniques for image-based instrument detection methods in 3D US data. Different algorithmic components have been presented, such as SVM, and CNN classification methods. In addition, several performance validation metrics of the development detection system have been presented in this chapter, which will be extensively used in the upcoming chapters of this thesis. Based on the objective of the instrument detection system, evaluation metrics for voxel-level classification, segmentation and instrument-level localization are reported.

The presented algorithmic components are used to design a robust and efficient image-based instrument detection system for 3D US data, aiming at giving guidance for challenging intervention tasks. The algorithmic components can be used to design and implement an instrument detection and segmentation framework, which is divided into two categories as follows.

1. ***Conventional instrument detection*** : The thesis starts with feature analysis and design to extract the discriminative information of the instrument, based on its local texture distribution and shape information. With the obtained features, SVM and other machine learning techniques are considered, as they can learn more complex features and combinations, which leads to successful voxel separation in complex 3D US volumes.

2. ***Deep learning-based instrument detection*** : Deep learning techniques are attractive for learning high-level representations of complicated structures in 3D images, which can separate the instrument voxels from the background by the fully automated learned discriminative features. With a tailored design of the CNN and its loss function, the network can extract specific characteristics of the instrument from the 3D data volume.

In Chapter 3, feature analysis and instrument model-fitting are exploited to classify the instrument, or more specifically cardiac catheters, in 3D US volumes. This setup serves as a baseline of the AI-based instrument detection in 3D US volumes for the subsequent chapters.

CHAPTER 3

Handcrafted Feature Analysis and Model-fitting for Catheter Detection

3.1 Introduction

The previous chapter has reviewed the fundamental technology and introduced a general state-of-the-art overview of techniques for image analysis. The chapter has also described an image-based automated detection system for medical instruments in 3D US data. The machine learning-based methods have been briefly overviewed for detection of the considered instrument by employing image segmentation and model fitting. The common pipeline for the detection system is referred to as a *segmentation model*, which includes instrument segmentation in 3D US and a model-fitting stage for the instrument. This approach can work properly only if the segmentation of the instrument voxels are accurately classified in the volumetric images.

However, the above assumption is usually not valid in clinical practice and data points in 3D US are complex in nature for ensuring accurate segmentation. The complexity of the data points are caused by anatomical structures, resulting from comparable spatial representations in the image, e.g. the heart chamber and heart valve have a similar appearance in cardiac US when compared to a catheter. An example of a US imaging setup are shown in Fig. 3.1. As can be observed, the RF-ablation catheter has almost the same appearance as the background, which makes it difficult to be identified by human eyes and this holds similarly for the segmentation algorithm. In order to create more stable segmentation, it is crucial to build 3D discriminating descriptors, based on different spatial descriptions and definitions. In this manner, challenging structures and anatomical backgrounds can be omitted by the segmentation algorithm, as they consist of different discriminative information. In addition, a stable model-fitting algorithm

for detection should be considered with high efficiency for complex segmentation output.

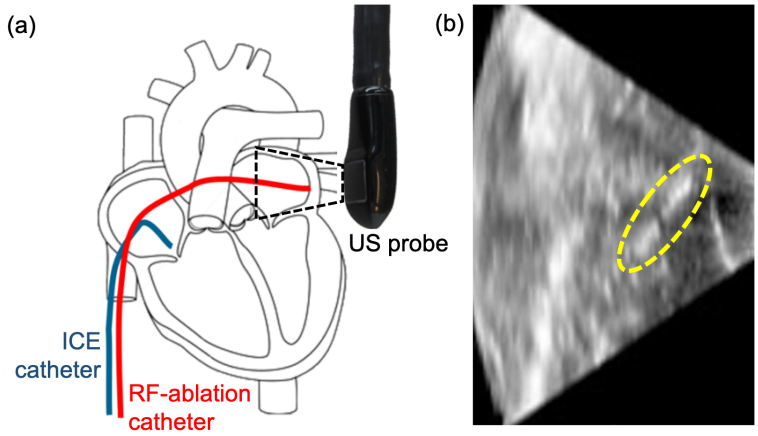


Figure 3.1 Ultrasound imaging of a cardiac intervention. (a) Example of cardiac intervention therapy with two catheters in the heart. (b) 2D image slice from a 3D image, where the RF-ablation catheter is located within the marked yellow ellipse.

3.1.1 Objective and General Challenges

This chapter aims to model the discriminating features of the medical instrument, especially for RF-ablation catheters, in 3D space by multi-definition features that capture the discriminative information with scale and definition sensitivity. As a result, each voxel is represented by its spatial frequencies, orientations and local statistical information, which jointly enable a robust detection of catheter voxels among other anatomical structures in the complex US volumetric data. With the obtained voxels, further analysis and elegant model-fitting is applied to localize the catheter with its orientation inside the US data. In more detail, the objective of this chapter is to design classification and model-fitting algorithms that are capable of: (1) robustly representing the voxel information in complex 3D space with anatomical structures, and (2) efficiently and robustly performing model-fitting to localize the catheter in the noisy segmentation output.

In order to achieve these objectives, we propose a specific solution to improve the detection performance for a robust detection and localization system. The proposed catheter detection system consists of two key steps, as depicted in Fig. 3.2. These steps are summarized as follows.

1. *Feature extraction and classification for voxels:* Handcrafted features of the catheter are designed to describe the information around the voxel. Supervised modeling and classification of the catheter voxels are applied prior to

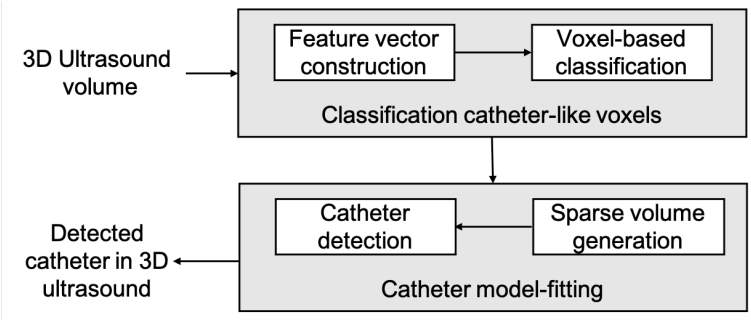


Figure 3.2 Block diagram of the proposed system for catheter detection.

the model-fitting stage, where both enclosed steps remove the non-catheter points in the volume. This will increase the detection accuracy and reduce the computation complexity.

2. *Catheter localization by model-fitting:* Catheter localization is applied based on the classified volume, which clusters the catheter-related voxels as a group with specific geometric properties, such as, e.g. a fat tube in an empty space. Nevertheless, some misclassified voxels are inevitable, which lead to difficulties in directly localizing the catheter in the classified volume. An advanced modeling step is proposed to localize the instrument in the full-3D US volume.

3.1.2 Specific Challenges for Feature Extraction and Voxel Classification

In order to robustly extract catheter points from the complex input US imaging, two essential algorithmic aspects, i.e. the use of multiple definitions and multiple scales, are proposed to perform information extraction in complex volumetric data. Each aspect aims at eliminating the background voxels that have a different intensity value, shape or structure from the catheter. By doing so, the overall localization robustness can be improved. We now briefly address the specific issues and objectives associated with multi-definition and multi-scale aspects.

1. *Multiple definitions of discriminative features:* The complex anatomical structure in 3D US imaging has comparable information and is yet different from the catheter. To remove these background voxels from the classification, a stable and discriminative feature should be considered. Nevertheless, feature extraction based on only one feature type (like Gabor feature) is considered too weak for catheter segmentation in 3D cardiac imaging. To better capture the discriminating information, we propose multiple 3D de-

scriptors for the US images, which better extract the spatial patterns of the catheter.

2. *Multiple scales of discriminative features:* Besides the above information extraction by the multiple features, discriminative features are also related to the feature scales. In previous studies, the papers are focusing on information extraction at specific scale, which only relate to the target. However, this approach ignores the information of the anatomical structure or background. Therefore, features are analyzed at multiple scales for each point, which is expected to improve the classification performance of the catheter in complex 3D US images.

3.1.3 Specific Challenges for Catheter Localization by Model-Fitting

Based on the image segmentation, it is important to localize the instrument by a model-fitting algorithm. Nevertheless, conventional model-fitting methods have limitations in both accuracy and efficiency for the catheter detection application. The involved issues are now briefly listed below.

1. *Coherency in the voxel labeling:* The voxel-based classification only categorizes the voxels independently. Therefore, for better labeling of the voxels, a joint labeling strategy is performed among the voxels, which removes the isolated small outliers and noise from the classified images. In this way, the segmented regions in the images are divided into sub-regions.

2. *Redundancy of clustered regions:* A typical model-fitting algorithm, such as RANSAC, is applied on the processed images after the group analysis to localize the instrument. Nevertheless, catheters are thicker than needles or thin instruments, leading to the issue that standard RANSAC-type of algorithms introduce redundancy for the model-fitting. Therefore, the accuracy and efficiency of such algorithms are degraded in practice. To address this, a concept is employed where only limited voxel data are allowed surrounding the catheter candidate voxels, to constrain the redundancy growth.

3. *Catheter localization in constrained 3D data:* Based on the constrained volume, the model fitting is performed with a reduced complexity of the surrounding background. This should enhance the catheter localization.

The sequel of this chapter is structured in the following way. Section 3.2 summarizes the related work. Sections 3.3 and 3.4 describe the proposed method in detail, including the various features for voxel classification and the proposed model-fitting algorithm. Section 3.5 discusses the collected datasets and experimental results. Finally, Section 3.6 concludes the chapter and presents some discussions on possible refinements.

3.2 Related Work

3.2.1 Non-machine Learning-based Detection

Many studies have recently focused on image-based medical instrument detection or localization in 3D US, but their approaches may be problematic for catheter detection in cardiac imaging. Methods like Principal Component Analysis (PCA) [44], Hough transformation [45] and Parallel Integral Projection transformation [16] were proposed to detect straight electrodes in 3D images. However, these transformation-based methods are not stable when the background includes high-intensity values, as is the case with instruments. This instability results from the fact that an image transformation cannot extract the discriminant shape information of tools to distinguish them from bright tissues or noise. Cao *et al.* [46] proposed a template-matching method to detect a catheter with *a-priori* knowledge of direction and diameter. Nevertheless, this method is still limited because it requires also by *a-priori* assumptions on the shape and orientation of the catheter. Besides, the carefully designed template is not only unstable to catheter appearance variations, but also lacking discriminating information.

3.2.2 Machine Learning-based Detection

Uherčík *et al.* [47, 48] applied Frangi [15] vesselness features to classify instrument voxels using supervised learning algorithms. The model-fitting based on RANdom SAmple Consensus (RANSAC) was employed to identify straight tubular-like instruments. Meanwhile, Zhao *et al.* [20] applied a similar method to track a needle on an ROI-based Kalman filter. Although the ROI-based algorithm decreases computation complexity, there are still some limitations. Firstly, the ROI-based algorithm requires a fixed view of images, which poses an extra limitation of avoiding ultrasound transducer movement during operation. Furthermore, both Uherčík [20] and Zhao [20] only considered a pre-defined Frangi feature as the discriminating key information, which is not only less robust to diameter variation, but also considers a small amount of information only, i.e. the information in the captured ultrasound volume is not fully exploited for discriminating classification. Recently, Pourtaherian *et al.* [49, 50, 28] have intensively studied needle detection algorithms based on 3D US. Their method detects the candidate needle-like voxels by incorporating a Gabor-based feature. This feature introduces more discriminating information on local orientation distribution, which is similar to the histogram of gradients. After the voxel-based classification, a two-point RANSAC algorithm is applied to estimate the axis of the needle. However, their proposed method is specifically designed for a thin needle with a large length versus diameter ratio in a high-quality US image, which does not apply to cardiac catheter detection. Although they did an experiment on catheter detection in an *in-vitro* dataset, their results showed that further studies on detecting the catheter on *ex-vivo* or *in-vivo* datasets were found necessary.

3.2.3 Challenges of the Current Methods

Although the methods discussed above have shown successful results in 3D US-based instrument localization, such as needle detection for anesthesia, there are still many challenges concerning cardiac intervention in both the segmentation and model-fitting stages.

As for the segmentation by non-machine learning-methods and already proven in the literature [48, 51], it is considered that the machine learning methods can generate more stable and accurate segmentation outcomes. Therefore, machine learning methods are superior over conventional methods in most cases, such as specific filtering techniques applied to the image followed by advanced thresholding. As shown in recent literature [28], machine learning methods can better exploit the discriminative information of the instrument voxels. In the article, Frangi filtering is compounded with a Gabor filterbank [28] to extract the discriminative feature vector, which is then classified by a learned SVM classifier. Nevertheless, only a preliminary study was performed on an *in-vitro* phantom dataset for catheter segmentation, which obtained insufficient performance. In addition, with a limited description of the catheter and non-catheter information, the discriminative features cannot be properly extracted for the supervised classification, which hampers catheter segmentation performance. As a result, the methods in current literature may not be sufficient and robust for catheter localization in cardiac interventional imaging. Motivated by the above discussion, we propose a novel feature extraction method for catheter segmentation, which includes multiple definitions and scales for a better feature representation and increased robustness.

As for the model-fitting methods, the current publications are focusing on straight-line instruments, such as a needle or an electrode. However, a catheter can be curved inside the volume, which makes the current fitting method less attractive and unsuccessful. In addition, the catheter has a thicker segmentation outcome than a needle which causes a conventional RANSAC algorithm to introduce computational redundancy, thereby hampering the fitting efficiency. To better fit the curved instrument properties and improve the efficiency, we will modify the RANSAC algorithm based on the concept of a sparse volume [52] and three-point fitting [48].

3.3 Method Part A: Feature Design and Voxel Classification

This section reviews and discusses several feature extraction techniques to evaluate them in the context of the catheter detection. To this end, these techniques will be compared in the same framework and input data to highlight their suitability for segmentation of the candidate catheter structure. The techniques will later be appended by a model-fitting procedure for further accurate catheter detection. This model-fitting will be discussed in the subsequent section.

The block diagram of the proposed catheter detection system is shown in Fig. 3.2. In the first step, the 3D volumetric image is processed to extract features from each voxel. The voxels are then classified by supervised learning methods into catheter-like voxels and non-catheter voxels. The procedure of catheter-like voxel classification consists of two steps. First, 3D discriminating features are extracted from each voxel in the 3D US image. Second, the supervised learning classifier is applied to classify the voxels. The discriminating features employed for voxel classification are described in the following subsections.

3.3.1 Objectness Feature

Multi-dimensional Objectness was first introduced by Antiga [31], who extended the traditional definition of the vesselness filter into the different shape descriptions for multi-dimensional images, see Fig. 3.3 as an example. For 3D images, the Hessian matrix is defined below, where f^{σ} is a Gaussian-filtered image with the standard deviation σ, while $f_{xx}, ..., f_{zz}$ represent the second-order derivatives in the $x-$, $y-$, or $z-$directions. The Hessian matrix with these derivatives is specified in matrix form by

$$H_{\sigma} = \begin{bmatrix} f^{\sigma}_{xx} & f^{\sigma}_{xy} & f^{\sigma}_{xz} \\ f^{\sigma}_{yx} & f^{\sigma}_{yy} & f^{\sigma}_{yz} \\ f^{\sigma}_{zx} & f^{\sigma}_{zy} & f^{\sigma}_{zz} \end{bmatrix}. \tag{3.1}$$

The Eigenvalues of Eqn. (3.1) are ranked by $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$. Using these Eigenvalues, the M-dimensional ($M \leq 3$) shape structures are described by Eqn. (3.3) based on the parameters from Eqn. (3.2). This is implemented with parameter settings $M = 0$ for a blob, $M = 1$ for a vessel and $M = 2$ for a flat surface in shapes, i.e. the Frangi vesselness filter equals to $M = 1$. Parameters R_A and R_B lead to two special cases. When $M = 2$, parameter $R_A = \infty$, and when $M = 0$, parameter R_B is set to zero. These shape parameters are derived from the computed Eigenvalues of the Hessian matrix, which are defined by

$$\begin{aligned} R_A &= \Big(|\lambda_{M+1}|\Big) \Big/ \left(\prod_{i=M+2}^{3} |\lambda_i|^{1/(3-M-1)}\right), \\ R_B &= \Big(|\lambda_M|\Big) \Big/ \left(\prod_{i=M+1}^{3} |\lambda_i|^{1/(3-M)}\right), \\ S &= \sqrt{\sum_{j=1}^{3} \lambda_j^2}. \end{aligned} \tag{3.2}$$

Similar to Frangi's vesselness feature [15], when $\lambda_j < 0$ for $M < j \leq 3$, the measurement of Objectness O^M_{σ} is defined by

$$O^M_{\sigma} = (1 - e^{-R_A^2/2\alpha^2}) \cdot e^{-R_B^2/2\beta^2} \cdot (1 - e^{-S^2/2\gamma^2}). \tag{3.3}$$

For the other cases of j, the value of $O_{\sigma}^{M} = 0$. The parameters α, β and γ are empirically determined, which defines the sensitivity of the response [32].

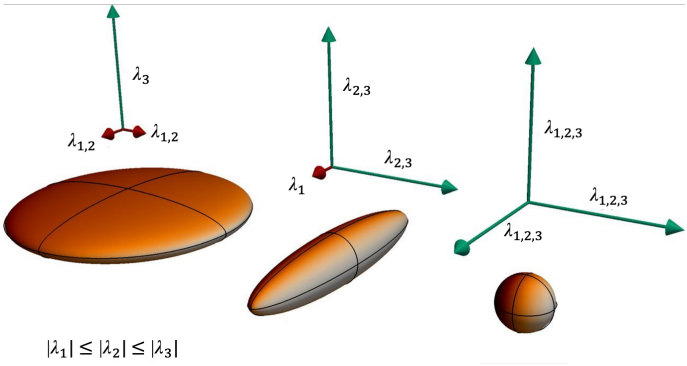


Figure 3.3 Objectness descriptors based on Eigenvalues of the Hessian matrix, showing different structures. Three coordinate systems indicates values of λ are dominant, leading to changes in the shape of the object. For example, in the second case at the middle, λ_1 is small, resulting into a tubular structure.

From the original definition, both Antiga and Frangi select the maximum response per pixel among a range of spatial scales, e.g. the maximum value among the scale range with $\sigma \in [1, 2, .., 5]$. However, this maximizing step loses scale-distribution information. Therefore, we propose to exploit all scale responses as features. Meanwhile, we calculate three different shape measurements, i.e. $M = 0, 1, 2$, instead of the tube descriptor used for needle detection in Uherčík [48]. Based on the above definitions, for each voxel v in the 3D volume V, the final feature vector is $\boldsymbol{\mathcal{O}}(v) = (O_{\sigma=1}^{M=0}(v), O_{\sigma=1}^{M=1}(v), O_{\sigma=1}^{M=2}(v), ..., O_{\sigma=3}^{M}(v), ...)^T$ in the multi-scale approach, where σ represents the standard deviation of a Gaussian filter and M denotes the type of Objectness feature, ranging from 0 to 2.

3.3.2 Hessian Features

Essentially, the Eigenvalue analysis in the Objectness feature space is to extract the directional information of edge distributions through the Hessian matrix and to remove the noise. However, the pre-defined descriptors in Objectness may lose some information because of (a) low signal-to-noise ratio and (b) the projection of nine features into three Eigenvalues. To preserve more information from a low-contrast image, we consider the elements of the Hessian matrix, as given in Eqn (3.1). Due to the symmetric structure of the Hessian matrix and to preserve the orientation response, we use six elements from the upper right of the Hessian matrix, and then shift the maximum response to the first position via circular shifting. As a result, the feature vector contains six elements for a specific scale (σ). The multi-scale Hessian feature $\boldsymbol{\mathcal{H}}(v)$ is denoted as

$\boldsymbol{\mathcal{H}}(v) = (\boldsymbol{\mathcal{H}}(v, \sigma = 1), ..., \boldsymbol{\mathcal{H}}(v, \sigma = 3), ...)^T$ with circularly shifted elements of Eqn. (3.4) in each scale. The feature vector per scale σ is specified by

$$\boldsymbol{\mathcal{H}}(v, \sigma) = (f^{\sigma}_{xx}, f^{\sigma}_{xy}, f^{\sigma}_{xz}, f^{\sigma}_{yz}, f^{\sigma}_{zz}, f^{\sigma}_{yy})^T. \tag{3.4}$$

3.3.3 log-Gabor Feature

Pourtaherian *et al.* [49, 50, 28] introduced Gabor features as an attractive discriminative feature for needle detection. The conventional Gabor-based features can be influenced by the DC components of the images [29]. To stabilize the performance for varying DC components and related gray-scale image variations in different US images, we adopt 3D log-Gabor features [29]. The 3D log-Gabor filter in the frequency domain is defined by:

$$\mathcal{G}(\omega, \phi, \theta) = \exp\left(\frac{\log^2\left(\frac{\omega}{\omega_0}\right)}{2\log\left(\frac{B}{\omega_0}\right)}\right) \times \exp\left(-\frac{\boldsymbol{\alpha}^2(\phi, \theta)}{2\,{\sigma_{bw}}^2}\right), \tag{3.5}$$

where B is the bandwidth of the filter in polar coordinates, and ω_0 is the central response frequency of the filter. The term B/ω_0 is set to a constant value to define constant shape-ratio filters. The direction of the filter is defined by the azimuth angle ϕ and elevation angle θ. The position vector $\boldsymbol{\alpha}(\phi, \theta)$ at frequency f is defined as $\boldsymbol{\alpha}(\phi, \theta) = \arccos((f \cdot \boldsymbol{d})/|f|)$, where the unit direction vector is $\boldsymbol{d} = (\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi)$. The polar coordinate system here is defined with an angle ϕ to the z-direction, and angle θ in the polar circle. The bandwidth of the angular direction is defined by σ_{bw}. Discriminative features are extracted using the real parts of the response in the spatial domain, due to their symmetry. The circular shifting operation is performed to shift the maximum response to the center and is denoted by $\boldsymbol{\mathcal{G}}(v, \omega)$ at specific frequency ω. The log-Gabor feature is then denoted by $\boldsymbol{\mathcal{G}}(v) = (\boldsymbol{\mathcal{G}}(v, \omega = 2\pi), ..., \boldsymbol{\mathcal{G}}(v, \omega = 6\pi), ...)^T$ for multiple harmonic frequencies with unit 2π. We have empirically chosen both angle parameters, i.e. ϕ and θ, to be set at $\{15°, 65°, 115°, 165°\}$. The amount and the interval between the angles is chosen such that the feature vector size is explicitly constrained.

3.3.4 Statistical Features

To extract more local information from a 3D cube, we propose to introduce a novel feature type, i.e. local statistical features. For a voxel v at the center point, we extract a 3D cube with specific sizes, such as $3 \times 3 \times 3$ voxels. The statistical features are obtained by calculating the mean, standard deviation, maximum, minimum value and local entropy of this cube. The statistical feature $\mathcal{S}(v)$ in the multi-scale case is denoted as

$$\boldsymbol{\mathcal{S}}(v) = (\mathrm{I}(v), \mathrm{mean}_{s=3}(v), \mathrm{std}_{s=3}(v), \mathrm{max}_{s=3}(v), \mathrm{min}_{s=3}(v), \mathrm{entr}_{s=3}(v), ...)^T, \tag{3.6}$$

where $\mathrm{I}(v)$ is the voxel intensity, parameter s is the size of the cube expressed in voxels, using v as the center point .

Table 3.1 lists the proposed 3D features with their symbols, scale variables and feature lengths for each scale. To enhance the performance of voxel classification, we apply a feature fusion strategy, which combines four different types of features in a multi-scale approach. The fused feature vector $\boldsymbol{\mathcal{C}}(v)$ is defined as $\boldsymbol{\mathcal{C}}(v) = (\boldsymbol{\mathcal{O}}(v), \boldsymbol{\mathcal{H}}(v), \boldsymbol{\mathcal{G}}(v), \boldsymbol{\mathcal{S}}(v))^T$, in which each component is developed in multi-scale fashion.

Table 3.1 Summary of 3D feature vectors

Name	Symbol	Scale param.	Size dim.
Objectness	$\boldsymbol{\mathcal{O}}$	σ	3
Hessian	$\boldsymbol{\mathcal{H}}$	σ	6
Log-Gabor	$\boldsymbol{\mathcal{G}}$	ω	16
Statistic	$\boldsymbol{\mathcal{S}}$	s	1+5

3.3.5 Supervised Classifiers for Voxel Classification

To achieve the best performance of the proposed features, we perform the classification under Linear Discriminant Analysis (LDA), Linear Support Vector Machine (LSVM), Random Forest (RF) and Adaptive Boosting (AdaBoost). Typically, the kernel-based SVM performs better than LSVM, but the fine-tuning of kernel parameters requires a large computation cost and empirical evidence shows that its performance is not better than RF and AdaBoost [53]. As a consequence, we consider only LSVM as the SVM classifier, which has a box constraint equal to unity. The RF is set to generate 50 trees. For the AdaBoost, weak learner is set to be decision stump with 50 learning cycles. During the training stage, due to an imbalanced class ratio, we down-sample the non-catheter voxels to have the same size as the catheter voxels. For testing, the voxels of whole captured volume is classified by the trained model, despite the fact that these voxels are imbalanced.

The section on feature extraction and classification is completed here. We have covered various feature extraction techniques, ranging from Objectness to Statistical features. Besides this, several classification approaches have been presented, which combine well with the presented feature extraction techniques. It is attractive to capture multiple elements in the feature vector and use also multiple scales for the feature representation. Therefore, the section has been completed

with an extended feature vector than combines several features into one larger feature vector with multi-scale representations within that vector. This concept will be tested later in this chapter in comparison with other techniques, in order to come to the best performance.

3.4 Method Part B: Catheter Model-fitting

This section develops a different part of the method, aiming at using the developed feature model for fitting that to a model for catheter detection. This model-fitting forms the second step for the identification of the catheters in the images.

In the second step, a modified RANSAC model-fitting is applied to localize the catheter in the noisy output of voxel classification. Given these noisy signal components, the choice of applying a RANSAC procedure is plausible. Misclassified voxels commonly occur, due to the complex local information from anatomical structures inside the heart and the non-perfect description of 3D features. As a result, after the voxel classification, there are multiple outlier blobs inside the US volumes. Fig. 3.4 shows example results from AdaBoost classification, where such outliers are clearly visible.

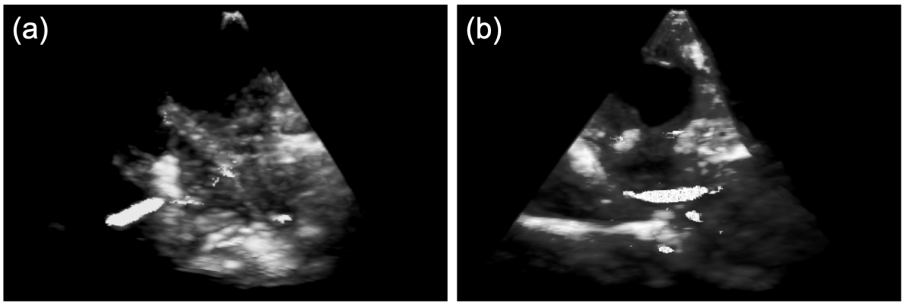


Figure 3.4 Examples of classified volumes in US imaging. The classified catheter voxels are highlighted in the images by bright intensity. Other bright areas may confuse the catheter detection.

To correctly localize the catheter in the noisy 3D image, we apply catheter model-fitting based on the *a-priori* knowledge that the shape of the catheter is a curved cylinder. The medical instrument model is conventionally reconstructed by fitting its skeleton together with instrument body voxels surrounding it[47, 28]. However, this method is not stable in detection and also inaccurate when assuming only a straight-line model in our challenging and noisy classified images. In our case, we have to assume that the instrument can also be curved and even can follow a more complicated 3D path. To localize a curved catheter in 3D US, a so-called Sparse-Plus-Dense-RANSAC (SPD-RANSAC)[52] has been reported earlier in literature. This concept is complex, so that we gradually explain and modify it for our purpose below. Meanwhile, we also modify the instrument

model into a 3-point curvature line to improve the localization accuracy. In the following paragraphs, we first describe the generation of a sparse volume, which reduces the complexity of the RANSAC algorithm. After this, a more complex catheter model is introduced to improve the detection accuracy, based on modified SPD-RANSAC.

3.4.1 Sparse Volume Generation for 3D US

After the voxel-level classification, the resulting binary image is called a dense volume V_d. Then, a connectivity analysis is applied to cluster the voxels, which are assumed to be part of a catheter-like shape, which can be either catheter or tissue voxels. The voxels from the same cluster are considered to belong to the same model. This means that the RANSAC algorithm includes many redundant sampling processes, if it is applied directly to dense data [52]. As a result, the centerline along the skeleton in each cluster is extracted to construct the sparse volume V_s, which reduces the model-fitting sampling iterations. The centerlines in the original SPD-RANSAC algorithm are generated directly by filtering the X-ray image, which benefits from using high-contrast imaging. However, in a coarsely classified 3D US image, the centerlines are difficult to extract directly and are not well-defined. As a result, we propose a new method to extract the centerline for each classified cluster in 3D US. This novel method for centerline extraction is described by pseudo-code in Algorithm 3.1, which is also leading to sparse volume generation. This terminology results from the fact that the algorithm locates a catheter cluster everywhere and then extracts centerlines, leading to sparse volume representations. Fig. 3.5 portrays an example of the obtained result when applying this sparse volume generation.

Algorithm 3.1: Sparse volume generation from a dense volume

```
Input: Dense volume V_d and empty V_s
  Find connected clusters in V_d
  for each cluster in V_d do
    Apply PCA analysis to find dominant axis among lat., az., and ax.
    for each 2D slice along dominant axis do
      Find connected 2D areas in the slice
      for each 2D area in the slice do
        Find center point of the area
        Project center point to V_s
      end for
    end for
  end for
Output: Sparse volume V_s
```

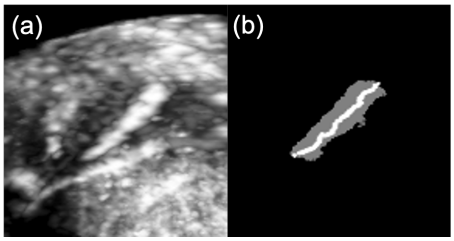


Figure 3.5 (a) Example of a catheter in 3D US imaging, (b) resulting dense cluster and sparse centerline, describing a sparse volume.

3.4.2 Model-fitting based on Sparse and Dense Volumes

In the proposed method, the catheter is modeled as a curved cylinder, which relies on fitting the centerline of the catheter [47]. Since we are looking for the catheter skeleton, the curved skeleton $\mathcal{K}$ can be described as:

$$\mathcal{K} = \{\boldsymbol{r} \in \mathcal{V}, \boldsymbol{r_0} \in \mathcal{V}, t \in \mathbb{R}, \boldsymbol{h_0}, \boldsymbol{h_1} \in \mathbb{R}^3 : \boldsymbol{r} = \boldsymbol{r_0} + t\boldsymbol{h_0} + t^2\boldsymbol{h_1}\}, \tag{3.7}$$

where $\mathcal{V}$ denotes the selected group of voxels from the 3D images, t is a real number and $\boldsymbol{h_0}$, $\boldsymbol{h_1}$ are vectors in 3D space. For catheter detection in 3D US, the model is fitted by a cubic spline interpolation, which is controlled by three control points [54]. For each RANSAC iteration, three control points are randomly selected from the sparse volume V_s, which are ranked by PCA analysis to define the interpolation order and to model the skeleton. The skeleton with the highest number of inliers in the dense volume V_d is chosen to be the catheter skeleton. The outliers are determined by computing their Euclidean distances to the skeleton. Finally, the inliers together with the skeleton in V_d are regarded as the localized catheter. Using the *a-priori* knowledge that the RF-ablation catheter cannot be heavily curved inside the heart chamber, we constrain the curvature by controlling the distance between the middle point to the straight line constructed from the endpoints. The maximum distance is empirically selected as 10 voxels in this chapter.

3.5 Experimental Results

For the experiments, the section starts with describing different datasets in Subsection 3.5.1. The evaluation metrics for classification and localization are introduced in Subsection 3.5.2. The results on the voxel classification using different features and classifiers are reported in Subsection 3.5.3. Subsection 3.5.4 shows the performance on catheter localization using the modified SPD-RANSAC algorithm.

3.5.1 Datasets

To validate the stability of our system, we have collected 3D US datasets under different recording conditions and performed the experiments on those data.

As for the *in-vitro* dataset, a Polyvinyl Alcohol (PVA) rubber heart was placed into a water tank. The images were captured by a 3D Transesophageal Echocardiography probe (TEE) while an RF-ablation catheter is inserted into it. Due to the less complex structure inside the rubber heart and the absence of anatomical material from a real heart, a clear contrast between catheter and background or phantom wall is shown.

For the *ex-vivo* datasets, porcine hearts were placed in several water tanks and images were captured through TEE, or a Transthoracic Echocardiogram probe (TTE). During the recording, the catheters were inserted into the ventricle or atrium. As for TEE-based images, although they were obtained from a different US system, we obtained a similar US quality because the same US probe was used. However, the dataset collected by employing a TTE probe, yielded noisy images with low-contrast appearance, due to a lower response at the low-frequency range.

Finally, we also collected an *in-vivo* dataset on a live porcine. During the recording, the TEE probe was placed next to the beating heart through the open chest, while the RF-ablation catheter was inserted through the vein to approach the heart. Because of challenging recording conditions and an unstable environment, the *in-vivo* dataset had the worst image quality.

More detailed meta-data about our datasets are presented in Table 3.2. All the datasets were manually annotated for catheter locations and confirmed by both medical and technical experts as the ground truth. In the following experiment, to fully exploit limited datasets, the Leave-One-Out Cross-Validation (LOOCV) is performed on each dataset. Some 2D slices from different datasets are shown in Fig 3.6.

3.5.2 Evaluation Metrics

Because of the class imbalance in the testing images, we use precision (P_r), recall (R_c), specificity (S_P) and F_1 score as evaluation metrics for classification performance after the supervised classification on each 3D US image. The definitions can be found in Chapter 2.

Our method starts with finding the voxels and identifying the catheter inside those voxels. The following major step is the previously discussed model-fitting to the classified voxels. The accuracy of the method can be defined as an absolute accuracy or a relative accuracy. The definition of absolute accuracy would require a completely calibrated physical setup with pre-defined phantoms or tissues and reference catheters. In our case, the accuracy is defined as the deviation of the visual ground truth, where the catheter is manually annotated within the image.

Table 3.2 Characterization of 3D ultrasound datasets for the experiments.

Dataset	Description
D1	*In-vitro* dataset from a PVA phantom heart with 2.3-mm RF-ablation catheter of 20 volumes, which was obtained by a 2-7 MHz phased-array transducer (TEE) using EPIQ7. The volume size ranges from $141 \times 168 \times 101$ to $145 \times 185 \times 101$ voxels (lat.$\times$ax.$\times$elev., which is the same for all datasets in this table) at a resolution of around 0.4 mm/voxel. The corresponding example is shown in Fig. 3.6 (a).
D2	*Ex-vivo* dataset from an isolated porcine heart with 2.3-mm RF-ablation catheter of 10 volumes, which was obtained by a 2-7 MHz phased-array transducer (TEE) using CX50. The volumes are resampled to a size of $179 \times 175 \times 92$ voxels at a resolution of around 0.4 mm/voxel. The corresponding example is shown in Fig. 3.6 (b).
D3	*Ex-vivo* dataset from an isolated porcine heart with 2.3-mm RF-ablation catheter of 10 volumes, which was obtained by a 2-7 MHz phased-array transducer (TEE) using EPIQ7. The volume size ranges from $120 \times 69 \times 92$ voxels to $193 \times 284 \times 190$ voxels at a resolution of around 0.6 mm/voxel. The corresponding example is shown in Fig. 3.6 (c).
D4	*Ex-vivo* dataset from an isolated porcine heart with 2.3-mm RF-ablation catheter of 12 volumes, which was obtained by a 1-5 MHz phased-array transducer (TTE) using EPIQ7. The volumes are resampled to a size of $137 \times 130 \times 112$ voxels at a resolution of around 0.7 mm/voxel. The corresponding example is shown in Fig. 3.6 (d).
D5	*In-vivo* dataset from a live porcine heart with 2.3-mm RF-ablation catheter of 8 volumes, which was obtained by a 2-7 MHz phased-array transducer (TEE) using EPIQ7. The volume size ranges from $146 \times 76 \times 153$ voxels to $172 \times 88 \times 178$ voxels at a resolution of around 0.4 mm/voxel. The corresponding example is shown in Fig. 3.6 (e).

In order to define the deviation as a distance, we define the skeleton of a catheter in the form of the center line of the shape. The deviation is then the distance between the annotated center line and the center line of model-fitted catheter. From the model, we obtain a limited set of key points, so that a spline function is used to construct a smooth curve going through the key points. This approach makes the model-fitted catheter well defined between the end points.

For our case, three types of errors are defined: skeleton-point error, and two errors concerning the beginning and ending of the model, i.e. tip-point error and end-point error (the average of tip-point error and tail-point error). These errors are visualized in Fig. 3.7. The latter two errors are defined as the distance between the localized point and the corresponding ground-truth point, either at the tip or at the tail of the catheter. The skeleton-point error is the distance between the sampled points from the localized catheter to the ground-truth skeleton. All errors are measured visually in the images and are initially expressed in voxels,

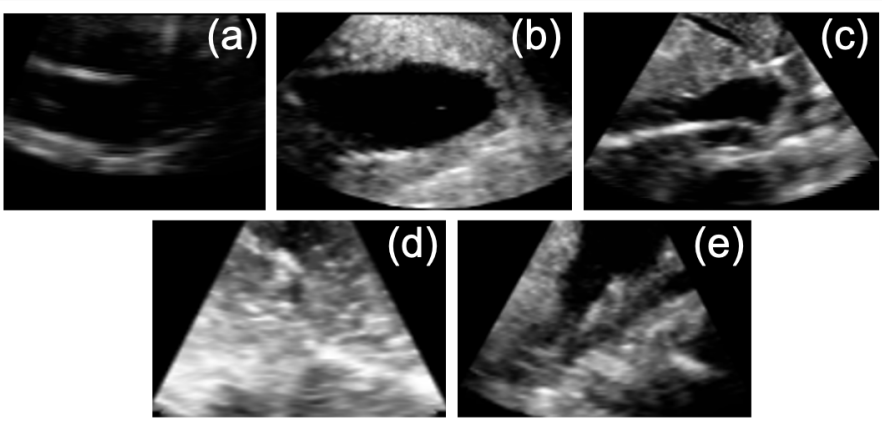


Figure 3.6 Examples of 2D slices from different datasets, which are corresponding to Table 3.2. (a) Dataset D1, (b) Dataset D2, (c) Dataset D3, (d) Dataset D4, (e) Dataset D5.

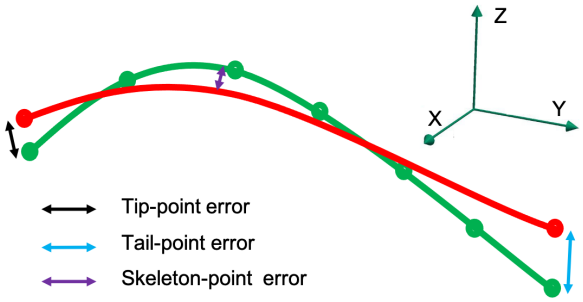


Figure 3.7 Example of three error metrics: tip-point error, tail-point error and skeleton-point error. The red curve is the ground-truth skeleton, the green curve represents the localized catheter skeleton.

which can be translated to distance using the voxel resolution. Further details and outcomes can be found in the experiments.

3.5.3 Voxel-based Classification

For the voxel classification, both feature and classifier can influence the performance of candidate voxel detection. To evaluate the discriminative power of the proposed features, we exploit their performances applying both a *single-scale* approach and a *multi-scale* approach. Conventional methods, e.g. needle detection in 3D US [48, 20], only consider a pre-defined scale size, i.e. *single-scale* and are denoted by SS-N for a pre-defined single-scale size N, based on *a-priori*

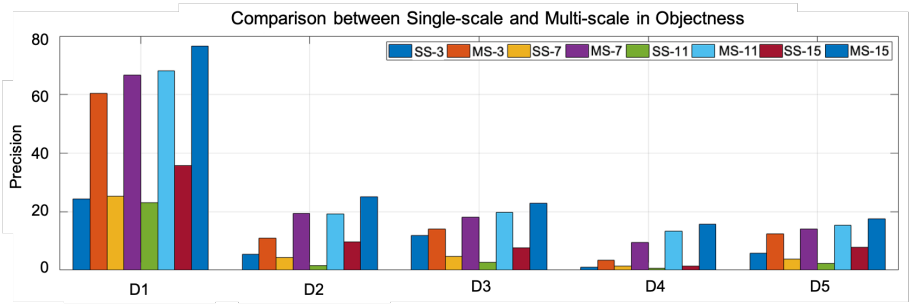


Figure 3.8 Average precision of single-scale (SS) and multi-scale (MS) Objectness features, depending on the applied datasets, D1 to D5.

knowledge of the instrument diameter. However, these pre-defined scales only extract discriminating information of tools while ignoring the information from the anatomical background, such as a heart wall or microvalve inside the heart. To extract more discriminative information for a better and stable classification, we also employ a *multi-scale* approach, which involves different scales simultaneously, e.g. the scale ranges from 1 to N, denoted as MS-N. In the following section, all comparisons are based on AdaBoost classification, due to its optimized performance which is shown in Fig. 3.12.

A. Single-scale vs. Multi-scale Feature Vector

Using the Objectness ($\mathcal{O}$) and Hessian ($\mathcal{H}$) features, we have performed experiments with σ ranging from 3 to 15 and a step size of 4. To measure the scale influence on the features in a simple way, we only employ the precision (P) as a metric, while fixing recall (R) at 75% in each volume. The experimental results are shown in Fig. 3.8 and Fig. 3.9 separately. These experiments lead to the following conclusions. (1) The multi-scale approach with the Objectness feature achieves a higher performance, because different shape information is contained for different scale sizes. When considering more scales, the features become more discriminating. (2) When comparing Frangi and Objectness features with Hessian features, the latter one has better performance, due to preserving more spatial information without PCA analysis. However, with the dataset D1, the Objectness gives a higher precision, which can be explained by the high-contrast image quality when compared with real tissue. Meanwhile, in all cases, the Frangi feature achieves a lower precision than Objectness.[32]

For features like the statistical feature $\mathcal{S}$ and log-Gabor feature $\mathcal{G}$, similar results are obtained, i.e. when the multi-scale range is increasing, the classification performance improves and the multi-scale approach achieves a higher performance than single-scale operation. We have performed the experiments with statistic feature $\mathcal{S}$ with scale ranging from 4 to 12 and step size 4. The experi-

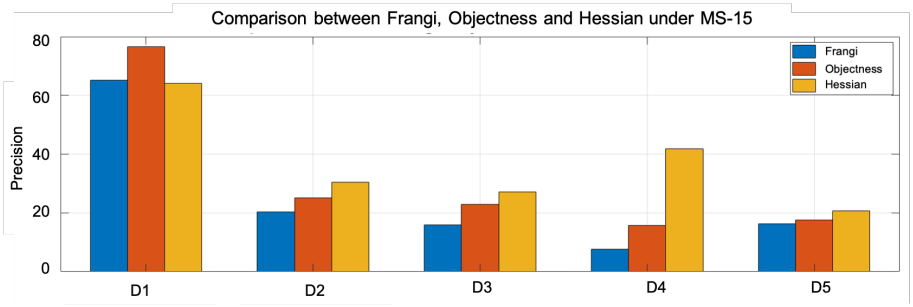


Figure 3.9 Average precision of Frangi, Objectness and Hessian features in multi-scale (MS) cases, depending on the applied datasets, D1 to D5.

ments based on the log-Gabor feature $\mathcal{G}$ are performed with scale ranging from 4 to 10 and step size 3. Experimental results are shown in Fig. 3.10. From the experiments, we conclude that the single-scale approach of the Gabor feature in needle detection [28] does not offer sufficient performance for our catheter detection in tissue-based images.

Based on the comparison between single-scale and multi-scale processing in different feature types, we have fixed the scale range to MS-15 for Objectness and Hessian features (see Fig. 3.8, the dark blue has the highest performance, and also Fig. 3.9). For the statistical features and log-Gabor features, MS-12 and MS-10 are the best choice, respectively.

B. Feature Comparison and Fusion

Based on the multi-scale approach in different features, their individual and fusion performances in each dataset are shown in Fig. 3.11. The results are demonstrated under AdaBoost, because it achieves the best performance when compared with other classifiers (under $\mathcal{C}$ and are shown in Fig. 3.12). All the results are obtained by LOOCV and thresholds are tuned to achieve the best F_1 score on the average. More detailed performance information can be referred to Table. 3.3.

From the performances in the table and figures, some observations are made. For the Phantom dataset, having less complexity and higher image contrast, the Objectness feature is able to achieve a promising result with *a-priori* defined descriptors. For *ex-vivo* datasets using different recording probes and US machines, the complex anatomical structure, which has a similar appearance as catheters, makes it difficult for the Objectness feature to describe the 3D space information. Moreover, when PCA is introduced, more spatial details are lost. Both Hessian features and log-Gabor features perform similarly in *ex-vivo* datasets, which may be explained by exploiting orientation and scale-sensitive features to describe the spatial information. For the Statistic feature, although it can extract 3D local intensity distribution information, the performance has less stability when

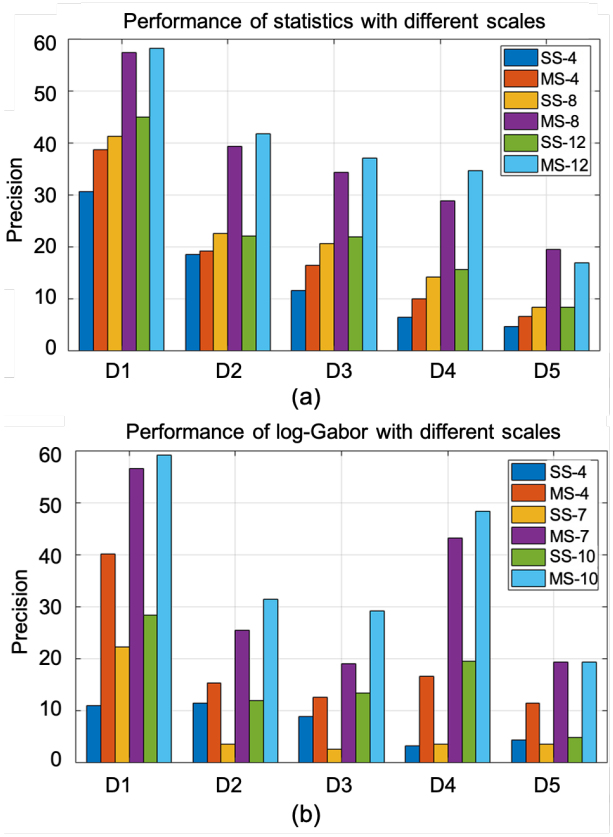


Figure 3.10 Average precision of (a) statistic features and (b) log-Gabor features with different scales, depending on the applied datasets, D1 to D5.

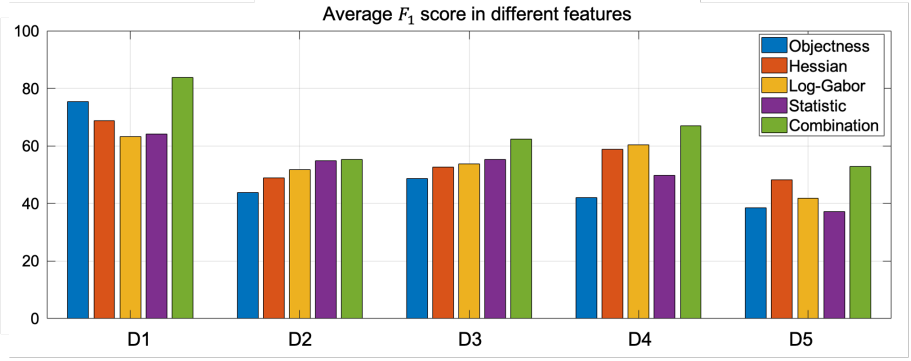


Figure 3.11 Optimizing the F_1 scores when tuning the thresholds, depending on the applied datasets, D1 to D5. The feature combination is the best choice, which corresponds to Table 3.3.

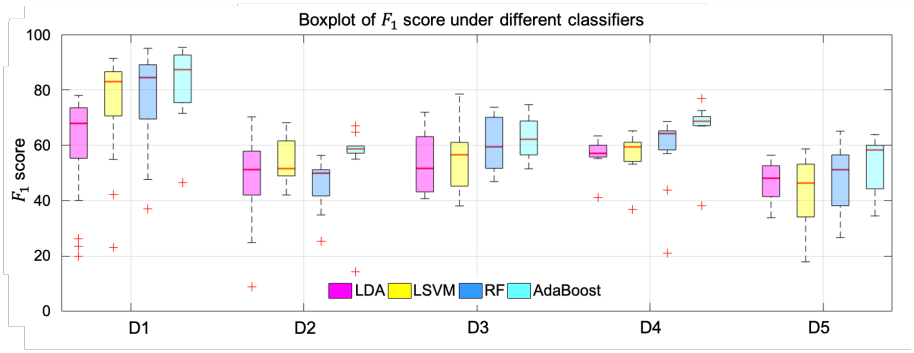


Figure 3.12 Distributions of F_1 score for different classifiers (LDA, LSVM, RF and AdaBoost), depending on the applied datasets, D1 to D5.

compared with Hessian and log-Gabor features. For the *in-vivo* dataset, due to challenging recording conditions and low-contrast image quality from real blood in the blood pool, the performances of all features are decreased. Although the log-Gabor feature introduces more orientation information, due to the low image contrast and the blurry boundary of the catheter, the orientation information cannot improve the classification performance. In all datasets, the feature combination is able to further improve the classification performance and appears to be the best choice.

3.5.4 Catheter Localization by Model-Fitting

After voxel-based classification, the SPD-RANSAC algorithm is applied to the binary images to localize the catheter in the noisy segmented images. The RANSAC algorithm generates the end-points and the skeleton of the catheter, which is used to analyze the localization error when compared to the ground truth. To evaluate the localization accuracy, we consider three types of errors: tip-point error (TE), end-point error (the average of tip-point error and tail-point error, EE) and skeleton-point error (SE). As in common practice, we regard the farthest point from the image border between the two end-points of the catheter as the tip. [55] The skeleton error is the average distance of five equally-sampled points (except the endpoints) on the localized skeleton to the annotated skeleton. For each sampled point, its distance to the ground truth center line is measured. An example of the three different error types is depicted in Fig. 3.7.

The localization performances from Table 3.4 are expressed in millimeters (mm) and involve three different model-fitting methods: 1) RANSAC with the two-point catheter model (R-2), 2) RANSAC with the three-point model (R-3) and 3) SPD-RANSAC with the three-point model (SR-3). Several slices of the tissue images are visualized in Fig. 3.13. The localized catheters are overlaid with colored annotations. To directly visualize in a 3D volume, the correspond-

Table 3.3 Average classification performances under best F_1 Score achieved per individual dataset. Numbers are mean and (standard deviation). Parameters $\mathcal{O}$ is Objectness features, $\mathcal{H}$ is Hessian features, $\mathcal{G}$ is log-Gabor features, $\mathcal{S}$ is statistics feature, and $\mathcal{C}$ is the feature combination. The notation of 1e-4 means $1 \cdot 10^{-4}$.

Dataset		Adaptive Boosting				
		$\mathcal{O}$	$\mathcal{H}$	$\mathcal{G}$	$\mathcal{S}$	$\mathcal{C}$
D1 (*in-vitro*)	Recall	80.06 (11.70)	71.05(24.94)	66.82(29.79)	73.16(22.78)	**89.77(12.51)**
	Precision	75.70 (19.39)	74.25(12.51)	66.90(13.64)	63.48(15.89)	**81.58(16.10)**
	F_1 Score	75.44 (12.54)	68.78(15.39)	63.35(21.30)	64.07(14.44)	**83.70(11.85)**
	Specificity	99.98 (3.3e-4)	99.98(1.6e-4)	99.98(1.5e-4)	99.96(2.7e-4)	**99.98(2.0e-4)**
D2 (*ex-vivo*)	Recall	50.31(25.73)	56.40(24.91)	58.46(25.30)	66.36(18.43)	**64.12(25.17)**
	Precision	48.55(18.92)	51.01(11.77)	53.18(9.15)	50.22(11.16)	**53.76(16.28)**
	F_1 Score	43.80(14.34)	48.87(11.01)	51.84(11.84)	54.80(7.45)	**55.24(14.82)**
	Specificity	99.93(6.5e-4)	99.94(5.2e-4)	99.95(4.04-4)	99.93(4.4e-4)	**99.94(4.3e-4)**
D3 (*ex-vivo*)	Recall	48.97(9.87)	55.15(16.27)	58.79(8.84)	60.53(10.62)	**70.62(11.70)**
	Precision	51.35(15.75)	52.87(14.92)	50.47(12.72)	51.91(9.22)	**57.96(11.34)**
	F_1 Score	48.66(10.03)	52.72(13.39)	53.75(10.24)	55.38(7.75)	**62.45(7.55)**
	Specificity	99.91(9.1e-4)	99.94(2.1e-4)	99.92(3.5e-4)	99.92(3.7e-4)	**99.93(3.2e-4)**
D4 (*ex-vivo*)	Recall	47.65(11.05)	71.58(19.07)	69.59(19.67)	66.35(11.63)	**75.59(17.75)**
	Precision	38.27(5.59)	51.14(8.88)	56.49(7.83)	40.75(6.78)	**63.79(7.77)**
	F_1 Score	41.99(6.93)	58.78(12.34)	60.47(12.40)	49.72(5.71)	**66.95(9.48)**
	Specificity	99.95(1.2e-4)	99.96(1.3e-4)	99.97(1.5e-4)	99.94(2.0e-4)	**99.97(1.2e-4)**
D5 (*in-vivo*)	Recall	37.67(19.21)	51.22(17.96)	43.25(13.63)	43.82(16.21)	**60.64(23.76)**
	Precision	43.24(23.54)	47.43(11.44)	41.99(13.43)	32.84(14.32)	**52.86(9.42)**
	F_1 Score	38.52(17.96)	48.32(14.72)	41.91(11.92)	37.27(14.77)	**52.93(10.65)**
	Specificity	99.95(4.0e-4)	99.95(1.4e-4)	99.94(3.7e-4)	99.92(2.3e-4)	**99.94(3.6e-4)**

ing 3D images are shown in Fig. 3.14. Furthermore, Fig. 3.15 shows an example of comparing our three-point SPD-RANSAC with a two-point RANSAC model-fitting [47, 28].

Table 3.4 Obtained average errors in catheter localization, expressed in mean±std.(mm). TE: tip-point error, EE: end-point error, SE: skeleton-point error. R-2: two-point RANSAC, R-3: three-point RANSAC, SR-3: three-point SPD-RANSAC. Numbers in bold are the best results.

Dataset	R-2		
	TE	EE	SE
D1	4.0±2.6	3.9±1.9	3.0±1.5
D2	4.0±1.8	4.8±0.9	3.1±0.7
D3	9.6±6.0	10.7±6.1	6.7±6.4
D4	4.0±1.3	4.3±1.2	2.9±0.6
D5	3.9±2.8	5.4±1.2	3.7±0.6
Average Error	5.0±3.7	5.5±3.6	3.7±2.2
Dataset	R-3		
	EE	SE	TE
D1	1.8±0.6	2.1±0.5	1.8±0.4
D2	1.9±0.4	2.5±1.3	2.1±1.1
D3	3.3±1.3	3.5±1.6	3.1±1.4
D4	**2.1±0.4**	2.2±0.3	2.0±0.1
D5	**2.0±0.9**	**2.0±0.5**	**1.7±0.3**
Average Error	2.1±0.9	2.4±1.0	2.1±0.9
Dataset	SR-3		
	EE	SE	TE
D1	**1.4±0.8**	**1.4±0.6**	**1.5±0.5**
D2	**1.2±0.3**	**1.7±1.0**	**1.5±0.6**
D3	**3.0±1.6**	**3.3±1.8**	**3.0±1.8**
D4	2.1±0.5	**1.9±0.4**	**1.8±0.2**
D5	2.4±2.8	2.4±1.4	1.9±0.8
Average Error	**1.9±1.4**	**2.0±1.2**	**1.9±1.0**

As shown in Table 3.4, the three-point models (R-3 and SR-3) are able to localize the catheters accurately when compared with the two-point methods (R-2). This is evident because almost every catheter is curved in the image, even a slightly curved catheter occurs when compared with the needle detection in the image. Meanwhile, the SPD-RANSAC algorithm is able to improve the localization accuracy when compared with R-3, which directly applies the model-fitting to the classified volume. As a result, our three-point SPD-RANSAC can achieve a higher localization performance giving an average localization tip-point error of only 1.9 mm.

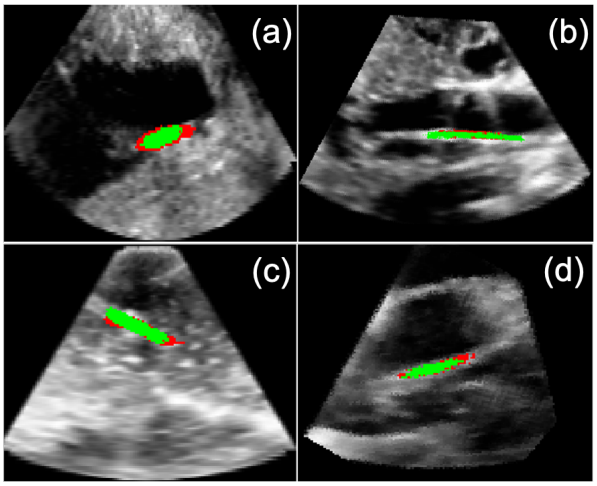


Figure 3.13 Slices (cropped) from real heart volumes, red color represents annotation and green color represents fitted catheter. (a) Dataset D2, (b) Dataset D3, (c) Dataset D4 and (d) Dataset D5.

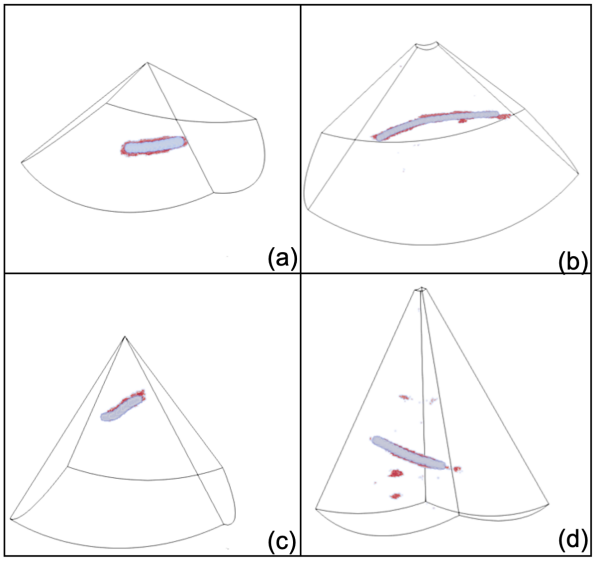


Figure 3.14 Classification results in tissue volumes, prediction (red color voxels) vs. annotation (gray color voxels). (a) Dataset D2, (b) Dataset D3, (c) Dataset D4, (d) Dataset D5.

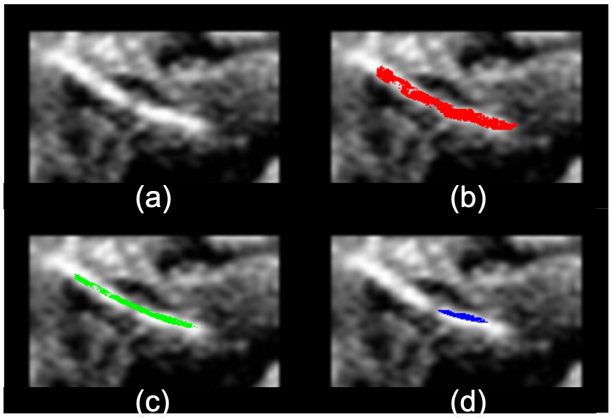


Figure 3.15 Comparison between SPD-RANSAC and a simple model-based RANSAC. (a) Original image. (b) Annotated catheter (red color). (c) SPD-RANSAC fitted catheter (green color). (d) Two-point RANSAC fitted catheter (blue color).

From the results in classification performance, multi-scale processing together with feature fusion are robust to classify the catheter voxels using AdaBoost classifiers. Although the classified volumes include some false positives (as shown in Fig. 3.14), *a-priori* knowledge of the catheter shape leads to a correct localization result. When an optimized 3D view generation would be implemented and added to our algorithms or alternatively, a 2D view on slices would be created, the catheter can be easily found and annotated for surgeons such that the cardiac intervention becomes easier and obtains a higher safety.

3.6 Conclusions

In this chapter, we have developed a robust and generic method to detect the catheter position and orientation in noisy 3D US volumetric data for challenging cardiac interventions. The automated detection of the instrument is aiming to distinguish the instrument from other anatomical structures in the obtained data, where the detection is characterized by voxel-based machine learning classification and model-fitting optimization. In the proposed framework, we have introduced a novel design and usage of the multi-scale and definition features as informative descriptors to better distinguish the catheter voxels from anatomical structures having similar appearance characteristics.

The framework of our method includes two essential stages: (1) classification of the catheter voxels and (2) catheter localization. In the first step, catheter voxel-level classification of the input image is based on the discriminative features, which achieves an F_1 score of 52%-83% on different experimental datasets.

Based on the proposed novel discriminative features, the overall performance of our method is much higher than the state-of-the-art techniques, which is because of a richer feature definition and multi-frequency exploration. The second stage of catheter localization includes a catheter-skeleton estimation by a Sparse-Plus-Dense-RANSAC (SPD-RANSAC) model-fitting algorithm. The positions of the catheter skeleton are optimized, which achieves a detection error of about 2 mm in the classified volumetric data. Since the detection error of the skeleton is similar to the catheter diameter, the instrument can be visualized by 3D in-volume rendering or 2D cross-section slicing. Therefore, a high clinical value can be achieved such that the sonographer can always locate the instrument during the operation.

The main contribution of this chapter can be summarized in two aspects: (1) a novel design of the 3D discriminative feature, which extracts the catheter voxels in the volumetric data, and (2) a novel design of RANSAC model-fitting, which improves the efficiency and accuracy of the instrument detection for better results. These aspects are elaborated below.

- *Novel discriminative feature:* To classify the catheter voxels, novel multi-scale and definition features are proposed, which describe the spatial and frequency information in different aspects. A thorough comparison is made for the classification in all datasets at the voxel level, which shows the proposed features are powerful and obtain higher performance than the state-of-the-art methods.

- *Modified model-fitting:* A modified model-fitting algorithm is proposed to accelerate the fitting efficiency and accuracy. Detailed experimental comparisons demonstrate the proposed method achieves faster and more accurate detection results, which is promising for clinical practice.

Further improvements are possible for higher detection performance and creating a real-time application. For example, tuning the US system to address varying recording conditions, e.g. adapting image gain or focal depth of the US array, may lead to better detection performance and higher robustness. Moreover, for different US resolutions and catheter appearances, the multi-scale processing with feature fusion (e.g. more features [56]) approach may be simplified or extended to achieve a better and robust detection accuracy.

With respect to the real-time application, the main challenge is coming from complex feature extraction during the voxel-level classification, which takes more than 85% of the whole processing time. Some possible solutions for enhancing the processing speed are: (1) embedding the feature extraction steps in a parallel manner on a GPU, which accelerates the computation efficiency, and (2) performing classification at voxel level can be accelerated by a coarse-to-fine strategy to reduce the calculation complexity.

The latter topic is explored in the next chapter (Chapter 4), where we will investigate an alternative approach, based on a coarse-to-fine strategy. Besides this, a novel deep learning method is introduced to replace the extensive feature extraction and classification. The coarse-to-fine strategy and the deep learning will jointly improve the detection efficiency and accuracy in challenging 3D cardiac US imaging.

CHAPTER 4

Catheter Detection by Voxel-of-interest-based CNN Classification

4.1 Introduction

The previous chapter (Chapter 3) has introduced a fundamental framework to create an image-based automated system, which distinguishes the catheter in 3D US data. For this purpose, a dedicated voxel-based classification and a model-fitting algorithm have been designed and evaluated for detecting the position and orientation in ultrasound images. It has been shown that the dedicated multi-scale and definition features generate catheter-like voxel classification results. This approach can work perfectly only if the spatial features of the catheter and surrounding tissue are following *a-priori* knowledge of the handcrafted feature design and corresponding learning. However, this can be challenged in a case that *a-priori* knowledge of the handcrafted features cannot properly extract the discriminative information, so that the classification results are not stable enough for clinical practice. In addition, the expensive computation and long-time execution hampers the application value, which is a real drawback for algorithm usage during the operation. Therefore, these two technical challenges raised during the algorithm design are further addressed in the following subsections.

4.1.1 Objective and Brief System Outline

In this chapter, we aim to model the automated medical instrument detection, especially for RF-ablation catheters, in 3D space by a voxel-level coarse-to-fine strategy. This method involves a coarse candidate voxel selection and a fine

catheter-like voxel classification. As a result, the catheter in the 3D US data can be segmented efficiently. The concept based on a coarse-to-fine strategy is adopted from various applications [20].

With the obtained voxels and following the general *segmentation-modeling* pipeline, the SPD-RANSAC model-fitting algorithm is applied to localize the catheter with its orientation inside the US data. In more detail, the objective of this chapter is to design a coarse-to-fine classification, leading to catheter segmentation in 3D US images, that is capable of: (1) efficiently and coarsely selecting the candidate voxels in complex 3D space for fine classification, and (2) accurately and robustly classifying the remaining voxels to further refine the segmentation results.

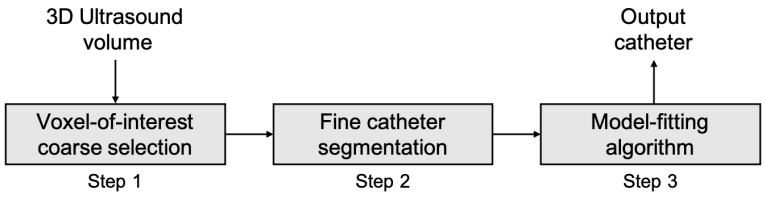


Figure 4.1 Diagram of the coarse-to-fine catheter detection and segmentation system.

In order to achieve these objectives, we propose a specific design to improve the efficiency and accuracy of the framework. The design is based on three basic steps, as shown in Fig. 4.1. First, from the voxel data, voxels are coarsely selected that could be candidate for the catheter segmentation. Second, the processing refines the coarse description for a fine catheter segmentation and classifies this description as being a catheter. Third, the model-fitting algorithm is applied to localize the instrument and confirm that a catheter is found.

Based on the above design outline, the key challenges originate from the first two steps in the method: coarse candidate voxel selection and fine catheter voxel classification. More detailed challenges and corresponding solutions are elaborated in the following subsections.

4.1.2 Specific Challenges for Candidate Voxel Selection

In order to coarsely select the interested voxels with high efficiency from the complex input US data, two essential aspects are considered for the coarse selection: (1) selecting the voxel-of-interest (VOI) efficiently, and (2) preserving as many catheter voxels as possible. Each aspect aims at improving the overall catheter segmentation result in both efficiency and accuracy. We now briefly address the specific issues and objectives associated with VOI selection and voxel preservation.

1. *VOI filtering:* A straightforward voxel-of-interest selection is to apply a machine learning classification on the voxels, such as a simplified classifier based on the Chapter 3. However, the feature extraction and classification is expensive and time-consuming, which is not suitable for the coarse VOI selection. To efficiently select the voxels, a promising choice is to simply apply a filter on the 3D images without considering a classification algorithm. Therefore, we propose to consider a Frangi filter as the VOI-selection approach. Specifically, the Frangi filter is able to efficiently extract the catheter-like voxels in the 3D images by *a-priori* knowledge, which has been proven successful in the past literature [48, 20].

2. *Voxel selection:* Based on the Frangi filtering, the output indicates the possible catheter voxels in the 3D space. However, a fixed and empirically determined threshold on the output can lead to unsatisfactory VOI selection results, which omits most of the background voxels, but also removes most of the catheter voxels due to appearance variations between the images. To better preserve the catheter voxels for fine classification, we propose to apply an adaptive threshold method for each individual image.

4.1.3 Specific Challenges for Voxel Classification

This part aims at performing the second step for which we will explore deep learning. With the obtained voxels from VOI selection, it is important to classify the residual voxels by a voxel-based classification method. Nevertheless, a conventional handcrafted design method has limitations in accuracy due to limited discriminative information capacity. The involved issues are now briefly listed below.

3. *Voxel-level classification:* The voxel-level classification can be achieved by the conventional handcrafted feature design and classification, as discussed in Chapter 3. This method may be not good enough due to the feature design being purely based on the experience or *a-priori* knowledge. In recent years, the deep learning methods, such as CNN-based classification, have proven to offer a superior performance in the classification task compare to the convention handcrafted design, which learns the discriminative information by a data-driven approach. Therefore, a CNN-based classification is considered to replace the conventional feature extraction and classification for a better voxel-level classification method.

4. *3D processing complexity:* With a CNN method, another challenge is the computation complexity in 3D images. In this chapter, the CNN is used to classify the voxels by employing the voxel's regional information, using a so-called 3D patch. Nevertheless, due to the complex CNN design with 3D filtering on the patch, it is computationally expensive to consider the voxel-by-voxel classification on the images, even when a coarse selection is

applied. The challenge is perform the classification in a 3D data patch with reduced computation cost and sufficient classification accuracy.

5. *Imbalanced class distribution:* For a common US dataset containing a catheter, the instrument occupies a small portion of the voxels, which typically amounts to 1/1000-1/2000. When deep learning is adopted as a starting point, this means that the catheter class is much smaller than the surrounding tissue and background voxels. This poses an imbalanced class distribution problem for CNN learning, so that the CNN could focus on the majority class instead of the catheter. In such a case, the network would misclassify the voxels, which omits the most of the catheter voxels, leading to unsatisfactory results. To address this, a two-stage training scheme is proposed to overcome the training bias and therefore improve the performance.

The sequel of this chapter is organized in the following way. Section 4.2 summarizes the related work in this field. Section 4.3 describes the proposed method in detail, including every step of the coarse-to-fine classification. Section 4.4 demonstrates the considered dataset and experimental results. Finally, Section 4.5 concludes the chapter and presents some discussions on possible refinements.

4.2 Related Work

4.2.1 Recent Methods for Instrument Detection

Medical instrument localization in US imaging is achieved by classifying the US voxels. Uherčík *et al.* [48, 15] have combined the image intensity with a Frangi filter response as a discriminating feature for voxel classification in needle localization. A recent study combined Gabor features with Frangi features to localize the catheter in a phantom heart [28]. Yang *et al.* [32] have used extended discriminating features within a multi-definition and multi-scale approach for catheter segmentation on *ex-vivo* datasets. However, these methods are less robust and less efficient when the US image has large variations in complex anatomical content. Recently, deep learning, e.g., convolutional neural networks (CNNs), have shown significant performance improvement in medical image analysis [9]. For US imaging, a CNN has been commonly used to classify voxels into different categories. Two different approaches for categorization exist: voxel-based CNN and semantic-based CNN. The first approach classifies individual voxels one-by-one through regional information of the voxels [57, 58, 59, 60]. The second approach of the semantic segmentation employs fully convolutional networks (FCNs) to predict segmentation masks directly [61]. Although this has obtained promising results by making use of the contextual information, the semantic segmentation

approach requires a large number of training data and has high computational complexity, which needs a careful trade-off during the algorithm design.

4.2.2 Direction of Our Method with Potential Improvements

As depicted in Fig.4.1, our method consists of three main steps.

- *Candidate voxel selection*: In the previous section, the pre-selection challenge has indicated a VOI procedure for finding candidates. The purpose of a VOI pre-selection procedure is to reduce the number of voxels to be processed by the second fine classification stage. To address the inconsistent response distribution in the filtered image, resulting from variations in imaging recording conditions and catheter appearance [32], an adaptive thresholding method is required for each image. This allows to adaptively preserve the majority catheter voxels, while omitting most non-catheter voxels based on the received image.
- *Voxel classification through CNN*: With the obtained candidate voxels, a dedicated voxel classification method is proposed, which outperforms the handcrafted feature design method by employing a deep learning approach. To further reduce the computation cost and improve the classification efficiency, a special tri-planar strategy is applied to the voxels in the 3D space prior to performing the deep learning method for limiting network complexity.
- *Catheter localization*: For the fine classified voxels, a cubic spline-based catheter model is fitted to localize the catheter by re-using the model-fitting algorithm from Chapter 3.

Compared to the recent methods from literature, our contributions can be summarized as follows. First, we employ a vesselness-based filter to coarsely select the candidate voxels to reduce the computation load for the CNN. We have designed a refined algorithm to improve the pre-selection results. Second, we propose a specific CNN for voxel classification in 3D US, which is in-depth compared with the existing methods. As a result, based on the proposed method, the catheter can be automatically detected with higher efficiency and accuracy than state-of-the-art methods.

4.3 Methods

As shown in Fig. 4.1, the proposed coarse-to-fine catheter segmentation and detection method is briefly summarized into three steps, which consists of coarse voxel-of-interest selection, fine catheter segmentation and a model-fitting algorithm. Each step is described in more detail below, which corresponds with visual diagram depicted in Fig. 4.2.

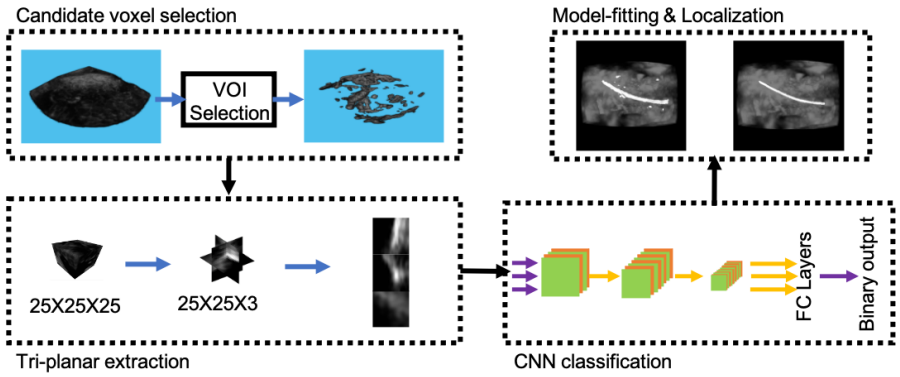


Figure 4.2 Diagram of the coarse-to-fine catheter detection and segmentation system.

1. *Candidate voxel selection:* VOI selection is applied on the input volume to coarsely select the voxels of interest for fine CNN-based classification and catheter segmentation. To achieve this goal, the Frangi vesselness filter [15] and a specifically designed adaptive thresholding step are applied. More details are presented in Section 4.3.1.

2. *Voxel-level catheter-like classification:* For each candidate voxel, a 3D regional patch is obtained, and three orthogonal planes are extracted and processed by the CNN for voxel classification. In particular, we propose a simplified tri-planar-based CNN, called *Share-CNN*, which reduces the computation complexity by sharing a single CNN for all orthogonal slices. Further details are presented in Section 4.3.2.

3. *Catheter localization and model-fitting:* Following the general computation vision processing of the *Segmentation-Modeling* as discussed in Chapter 2, the SPD-RANSAC algorithm from Section 3.4 is applied to finally localize the target catheter in complex 3D US images, as briefly addressed in Section 4.3.3.

4.3.1 Candidate Voxels Selection

The proposed method uses Frangi vesselness filtering to select the candidate catheter voxels from 3D US, which enables to dramatically reduce the number of samples to be classified by the CNN (typically a reduction from $\sim 10^6$ to $\sim 10^4$). Nevertheless, this simple selection results into a high false positive rate because of the weak voxel discrimination in noisy and low-quality cardiac 3D images [32]. To address this issue, we introduce an adaptive thresholding method for the VOI selection.

The 3D US image is first filtered by a Frangi filter with a pre-defined scale and re-scaled to the unity interval $[0; 1]$, leading to a normalized frame, called $\mathcal{V}_F$.

After the filtering, we apply an adaptive thresholding method to $\mathcal{V}_F$ to coarsely select N voxels with the highest vesselness response. The thresholding method is trying to find out the top N possible voxels in $\mathcal{V}_F$. Since the filter response has a large variance in different images, the adaptive tuning of the threshold can gradually select approximately N voxels, by iteratively increasing or decreasing the threshold T, depending on the image itself. The pseudo-code is described by Algorithm 4.1. Based on the coarsely pre-selected voxels in 3D US, which form about N candidates voxels, the 3D patches are extracted and processed to generate three orthogonal slices of each voxel for the CNN. In our experiment, the initial threshold is empirically set to $T = 0.3$. Value N is empirically selected to balance and trade-off the efficiency of CNN classification and classification performance.

Algorithm 4.1: Adaptive thresholding for candidate voxel selection

Input: filtered volume $\mathcal{V}_F$, required voxel Num. N and initial threshold T
Apply threshold to $\mathcal{V}_F$ by the initial threshold value T. Find the remaining voxels with amount K, which is larger than T.
if $K < N$ **then**

  while $K < N$ **do**
    $T = T - 0.01$. Apply thresholding to $\mathcal{V}_F$ by T, find No. of voxels K larger than T.
  end while
else if $K > N$ **then**

  while $K > N$ **do**
    $T = T + 0.01$. Do thresholding on $\mathcal{V}_F$ by T, find No. of voxels K larger than T.
  end while
end if
return The set of approximately N voxels with response larger than adapted threshold T.

4.3.2 Voxel Classification by CNN

For voxel-wise classification of volumetric data, the 3D regional information is processed by the CNN to classify the voxels. A straightforward way is to classify the voxel based on its 3D neighborhoods. For each candidate voxel located at the center of a 3D cube, the cube is processed by a 3D-CNN [57], as shown in Fig. 4.3 (a). However, when using a 3D data cube as input, this approach includes too many parameters in the network, which hampers the efficiency of the voxel-wise classification in 3D US volumes. To preserve the 3D information and yet reduce the convolution operations, a multi-slice-based method is proposed in [18],

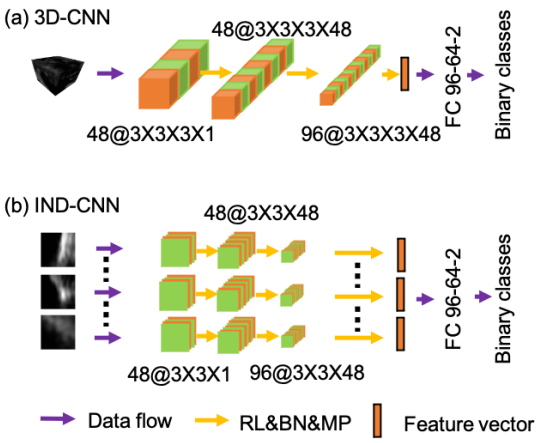


Figure 4.3 Two configurations of commonly used CNNs. (a) 3D-CNN. (b) IND-CNN. Note that the network IND-CNN can have more than three branches. The abbreviation RL&BN&MP stands for ReLu+Batch Normalization+Max pooling.

which effectively reduces the 3D operation to 2D processing. To preserve the 3D structure information, the authors of [18] employed a multi-view cross-section method that extracts slices from the 3D cube through different angles. Then each slice will be processed by an individual CNN using 2D processing. An example of this method is shown in Fig. 4.3 (b), which is called IND-CNN. The extracted feature vectors from the slices are concatenated to supply them into fully-connected layers (FCs). The 3D-CNN processes the information using 3D operations, which leads to too many computations and large execution times. Instead, although the IND-CNN keeps 3D information by a slicing approach, multiple individual CNN branches lead to redundancy in the network, which results from using a CNN for each slice. Because of these redundancies in the networks, 3D-CNN and IND-CNN are both sub-optimal choices in terms of application and computation time. This has motivated our research into an alternative solution with the aim to achieve a higher efficiency.

Method – proposed network architecture: This chapter proposes a simplified method to classify the voxels. To this end, we adopt the slice-based strategy, which is a good start to reduce the complexity of 3D processing. However, instead of training a CNN for each slice, we propose to train one shared CNN for all slices. All feature vectors from the slices are concatenated to form a longer feature vector for classification. This single network is called Share-CNN, which is depicted in Fig. 4.4 (b). There is a similar structure called RGB-CNN, reported in [58], which is shown in Fig. 4.4 (a). It extracts three orthogonal slices from

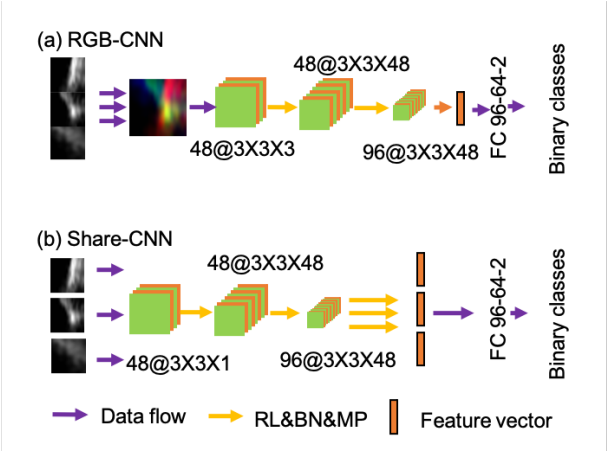


Figure 4.4 Alternative solutions with simplified CNNs for the network architecture of Fig. 4.3. (a) RGB-CNN. (b) Share-CNN. The abbreviation RL&BN&MP stands for ReLu+Batch Normalization+Max pooling.

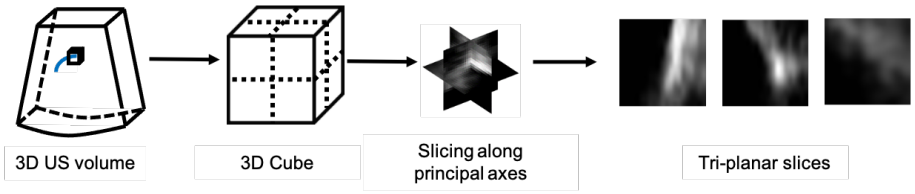


Figure 4.5 The steps to extract tri-planar slices from the obtained 3D cube of the interested voxel. The slices are obtained by extracting the planes passing through the center point of the cube and following the direction of the principal axes of the coordinate system.

the principal directions of the 3D cube, which are then re-organized into RGB channels. However, this introduces a limitation: the spatial information between each slice is processed rigidly by convolutional filters at the first stage of the network. The input stage of the CNN involves only shallow processing, where low-level features are processed. This simple strategy cannot fully exploit the spatial relationships between the slices. Alternatively, the proposed Share-CNN can exploit the spatial correlation at a high-level feature space. Based on the binary selection of candidate voxels during the pre-selection, a 3D cube is obtained for each candidate voxel located at the center of the cube. We extract a cube of size $25 \times 25 \times 25$ voxels, which is larger than a typical catheter diameter of 4-6 voxels in 3D cardiac US. Then, three orthogonal planes passing through the center point of the cube are sliced as the input for the CNN. More detailed demonstration is shown in Fig. 4.5.

Method – two-stage training: For training with medical images, the class imbalance is the most challenging issue. In our case, the ratio of catheter voxels vs. non-catheter voxels is commonly less than 1/1000. As a consequence and to fully exploit image information, we therefore perform a two-step training procedure when training the CNNs. First, the number of imbalanced voxels in training images are resampled on non-catheter voxels to obtain the same amount as catheter voxels. These balanced samples are used to train the CNNs. Second, the training images are validated on the trained models to select the falsely classified voxels, which are used to update the networks for fine-tuned optimization [59][60]. Specifically, unlike the diagram in Fig. 4.2, the training process is applied in the whole US image rather than the VOI processed one. This update step reduces the class imbalance, by removing the easiest sample points (so-called two-stage training).

Method – loss function adaptation: The parameters of networks are learned by minimizing the cross-entropy, using the Adam optimizer for faster convergence. During the two-step training, the cross-entropy is characterized into a different form to balance the class distribution. In the first training stage, the cross-entropy is characterized in a standard format. However, during the updating, the function is redefined as weighted cross-entropy. The difference between the cross-entropies avoid the bias in the updating stage, which occurs due to the number of false positives, being usually 5-10 times larger than the positive training samples in the second stage. As a result of the weighted cross-entropy, the networks tend to preserve more catheter voxels than discarding them after the classification. The weighted cross-entropy loss is formulated by

$$\text{Loss}_{\text{wCE}}(y, \hat{p}) = -(1-w)y \cdot \log(\hat{p}) - w(1-y) \cdot \log(1-\hat{p}), \tag{4.1}$$

where the parameter y indicates the label of the sample, $\hat{p}$ represents the class probability of the sample, and parameter w is the sample class ratio among the training samples.

Method – training details: During the training, the dropout is used to avoid overfitting with 50% probability in the FC layers, together with an L_2 regularization with 10^{-5} strength. The initial learning rate is set to 0.001 and re-scaled by a factor 0.2 after every 5 epochs. Meanwhile, to generalize the network in orientation and image-intensity variation, techniques for data augmentation like rotation, mirroring, contrast and brightness transformations are additionally applied. The mini-batch size is 128, and the total training epoch is 20, which involves about 25k steps in the first training, and iterations in the second training stage require about 100k cycles.

4.3.3 Catheter Localization

The classified volume may include some outliers, which are generated from the blurry tissue boundaries or catheter-like anatomical structures. To robustly localize the catheter, we employ the SPD-RANSAC method to fit a pre-defined catheter model. To robustly localize the catheter, the classified volume, so-called dense volume, is processed by connectivity analysis to generate clusters. Then, the cluster skeletons are extracted to generate the sparse volume. During the fitting stage, three control points are automatically and randomly selected from the sparse domain and ordered in orientation by principal component analysis. The re-ordered points ensure the cubic spline-fitting passes the points in sequential order, which generates the catheter-model skeleton. The localized skeleton with the highest number of inliers in the dense volume is adopted as the fitted catheter. More details are explicitly presented in Section 3.4.

4.4 Experimental Results

4.4.1 Datasets

In this study, we have collected a challenging *ex-vivo* dataset from 4 isolated pig hearts, which includes 65 volumes. During the recording, the hearts were placed in a water tank with an RF-ablation catheter (7 French (Fr) $\approx$ 2.3 mm) inside the heart chambers. Moreover, to ensure that the images in each heart are independent from each other, we changed the relative position between the heart and US probe to obtain a different appearance of the heart in each captured image. Furthermore, we extracted the catheter and re-inserted it into the heart chambers to make the images independent, i.e. 1 session for 1 image.

The dataset includes volumes of size ranging from $120 \times 69 \times 92$ to $179 \times 179 \times 202$ voxels, in which the voxel size was isotropically resampled to the range of 0.4-0.7 mm. The datasets were manually annotated by clinical experts to generate the binary annotation mask as the ground truth. Examples of three cases are shown in Fig. 4.6, which visually compares the recordings on phantom heart, pig heart and human heart. Compared to the phantom heart and human heart, the captured pig-heart images are more complex. Compared to the phantom data, real pig tissue has more complex anatomical structures, which makes it hard to distinguish the catheter from tissue. When compared to the real human heart, the chambers of the pig heart are collapsed due to the dead tissue, which leads to a small free space within the heart. Moreover, the human-heart image, which is shown in the figure, has a larger field of view than the pig-heart recordings, as the data was collected for the Transcatheter Aortic Valve Implantation (TAVI) operation. To fully make use of the limited datasets for deep learning, we perform threefold cross-validation on all collected images.

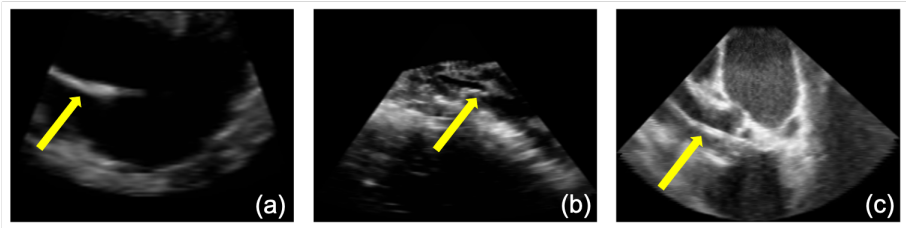


Figure 4.6 Visual appearances within different datasets. (a) Phantom US. (b) Pig heart US. (c) Human heart US. The arrows are pointing to the catheters.

4.4.2 Voxel-of-Interest Selection

To reduce the number of voxels for classification, we apply the Frangi vesselness filter to select the candidate voxels. However, it cannot selectively filter out the catheter voxels from tissue and background with a pre-defined scale, due to too many false positives [32]. In our method, we first apply the Frangi filter with scale size equal to 2.5 voxels to filter out most of the tubular-like structures. Then Frangi responses are re-scaled to the unity interval, which maps response into a probability-like range.

To evaluate the performance of thresholding, we consider three metrics: recall R_c (the remaining catheter voxels versus ground-truth catheter voxels), Ratio (thresholded voxels versus all voxels, to evaluate the voxel preserving ability), and their fusion score (this mimics the F_1 score by replacing precision by Ratio to evaluate a joint threshold performance), which enables to show the preservation performance of catheter voxels and removes non-catheter voxels. The metrics of Ratio and Score are defined in Eqn. (4.2), where TV denotes the remaining number of voxels after applying the threshold, while AV represents the number of all voxels. These metrics are specified by

$$\text{Ratio} = \frac{\text{TV}}{\text{AV}}, \quad \text{Score} = \frac{2 \cdot R_c \cdot (1 - \text{Ratio})}{R_c + (1 - \text{Ratio})}. \tag{4.2}$$

The performances of adaptive thresholding are shown in Fig. 4.7, where the threshold is chosen to be the required number of voxels N, which ranges from 10k to 190k voxels with a step size of 10k. The metric values are obtained by averaging the results of all the testing volumes, using threefold cross-validation. It can be observed that the proposed adaptive thresholding method provides a more stable voxel distribution, i.e., a smaller fraction of the whole pyramid area while keeping a higher recall. As a result, the proposed thresholding method provides a better selection for the voxels of interest. However, this pre-selection leads to a drop in the recall value. As a consequence, in the following step, a CNN with high recall for voxel classification is employed to improve the overall performance of finding catheter voxels.

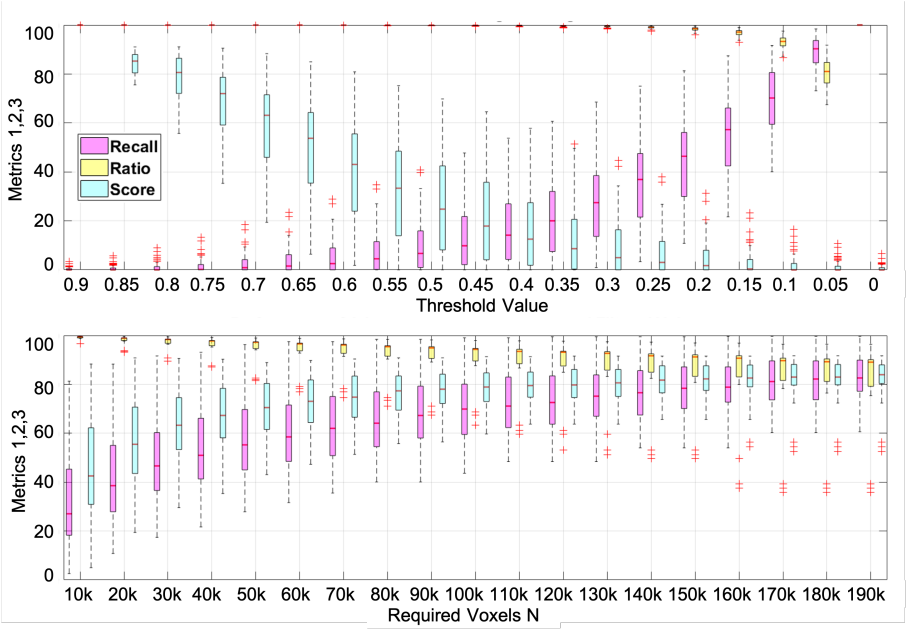


Figure 4.7 Performance of applying direct thresholding (top) and adaptive thresholding (bottom) on Frangi filtered images. Metric 1 is Recall with pink bars, Metric 2 is 1-Ratio in yellow bars, and Metric 3 is the Score in blue bars, see Eq. (4.2).

4.4.3 Voxel Classification

A. Comparison with existing methods

In the following experiments, three metrics, recall, precision and F_2 score are used for voxel classification at the image level. Specifically, the F_2 score is defined as

$$F_2 = \frac{5 \cdot R_c \cdot P_r}{4 \cdot P_r + R_c}. \tag{4.3}$$

We first compare the Share-CNN of the refinement step with the start-of-the-art methods. Two methods using handcrafted features, the Gabor feature with SVM (GF-SVM) [28] and the multi-scale and multi-definition features with Adaboosting (MF-AdaB) [32], are considered as baseline. For comparison, we also consider the semantic segmentation method 3D UNet [61]. The performances are listed in Table 4.1. We can observe that the Share-CNN outperforms conventional methods with handcrafted features. The standard 3D UNet produces the worst performance on our challenging data. This may be explained by the high complexity of 3D UNet in their design, resulting in over-fitting. Fig. 4.8 illustrates some example results obtained with 3D UNet.

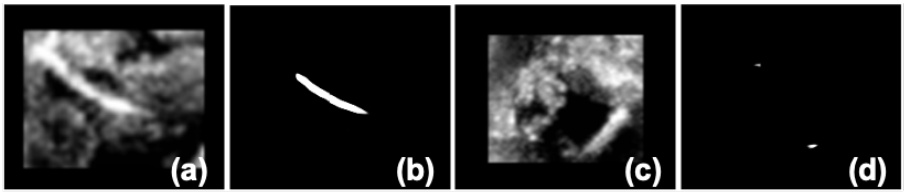


Figure 4.8 Segmentation results from 3D UNet. (a) Original image, and (b) successful segmentation. (c) Original image, and (d) segmentation failure.

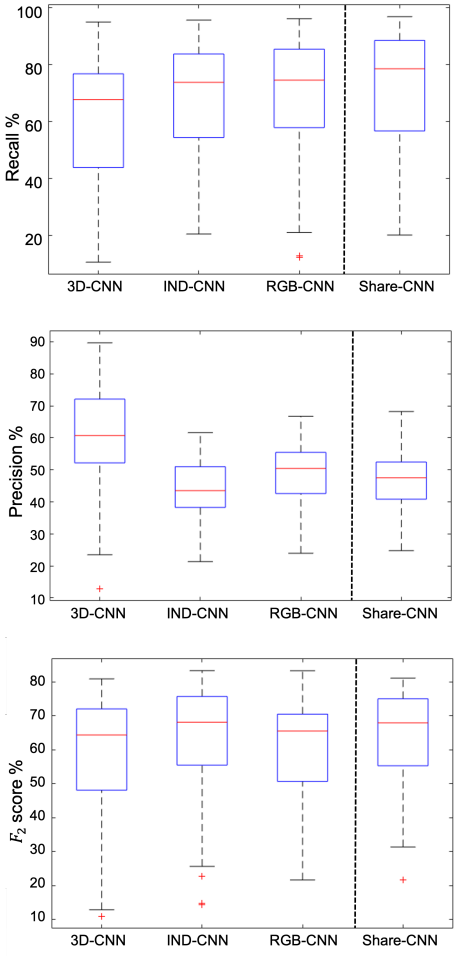


Figure 4.9 Boxplots of the performance comparison of the Share-CNN against 3D-CNN, IND-CNN, and RGB-CNN for three different metrics. The known CNNs boxplots are shown at the left side of the dashed line, while the boxplot of Share-CNN is depicted at the right side of the dashed line.

Table 4.1 Average performance of voxel-based classification (mean±std.)

Method	Recall (%)	Precision (%)	F_2 score (%)
GF-SVM [28]	29.9 ± 25.4	9.2 ± 8.8	19.0 ± 15.5
MF-AdaB [32]	61.2 ± 17.6	28.4 ± 16.6	45.5 ± 15.6
3D UNet [61]	30.3 ± 26.3	11.9 ± 12.7	21.4 ± 19.5
Share-CNN	72.3 ± 19.6	46.4 ± 8.5	63.8 ± 14.3

B. Comparison with different CNN methods

We further compare the Share-CNN with 3D-CNN, IND-CNN, and RGB-CNN. The training strategy of these CNNs is the same as the performed strategy for Share-CNN. The performance comparison is shown in Fig. 4.9, which results in the following findings.

- *Comparison to 3D-CNN:* When compared to 3D-CNN, the proposed Share-CNN has better Recall and higher F_2 score, while 3D-CNN achieves better precision. However, taking 3D data cubes as input, 3D-CNN has too many parameters in the network, requiring a large amount of training data. In contrast, the Share-CNN is much simpler. In terms of efficiency, 3D-CNN executes in about 10 mins. per volume on average, which is almost 5 times longer than the orthogonal slice approaches.
- *Comparison to IND-CNN:* The IND-CNN, which is designed to have multiple branches, delivers comparable performance as the proposed Share-CNN. This is because both networks fuse the high-level information in a similar fashion. However, the IND-CNN trains an individual CNN for each slice, which is computationally complex and leads to redundancy.
- *Comparison to RGB-CNN:* Compared to RGB-CNN, it can be observed that the Share-CNN achieves consistently higher performance. This can be explained by the fact that the RGB-CNN only exploits spatial correlation among different slices in the lower feature space.

C. Paired t-test between methods

In a further experiment, we have performed a paired t-test to identify clear differences of the proposed Share-CNN and other considered networks, i.e., MS-AdaB, RGB-CNN, IND-CNN, and 3D-CNN. We have adopted the F_2 score of each image as a measure for performing the t-test. In the paired t-tests, the significance level is set to 0.05. The detailed p-values for the paired t-tests are shown in Table 4.2. All obtained p-values are smaller than 0.05, except for IND-CNN. These results show that the Share-CNN performs significantly better than MF-AdaB, RGB-CNN and 3D-CNN methods (combined with the previous results).

Although IND-CNN shows little difference with Share-CNN, it has parameter redundancy that leads to overfitting and computational inefficiency.

Table 4.2 Paired t-test (p-value) between different methods.

Method	MS-AdaB	RGB-CNN	IND-CNN	3D-CNN
Share-CNN	$3.2{\cdot}10^{-14}$	$3.2{\cdot}10^{-6}$	$2.6{\cdot}10^{-3}$	$4.4{\cdot}10^{-1}$

D. Ablation study of CNNs

The Share-CNN includes two-stage training and a weighted loss function in the network. To better understand their influence on the classification performance, we have performed ablation studies for three different cases: 1) CNN without two-stage training (denoted as NoBoost), i.e., only trained on re-sampled images, 2) CNN with two-stage training but without weighted loss function (denoted as NoWeight), 3) the proposed CNN (denoted as Combine). The results of these ablation studies are listed in Table 4.3. For Share-CNN-NoBoost, although it receives relatively high Recall performance, the simple sampling strategy leads to the lowest Precision results, which makes the model-fitting more challenging. Furthermore, the weighted loss function can re-balance the information distribution during second-stage training and can maintain a higher recall, while omitting the non-catheter voxels. When compared with the no-weighted case, the weighted network versions provide lower variance in Precision and F_2 scores.

Table 4.3 Performance with ablation studies on proposed Share-CNN (mean±std.)

Method	Recall (%)	Precision (%)	F_2 score (%)
Share-CNN-NoBoost	92.4±8.6	12.0±8.5	35.2±17.2
Share-CNN-NoWeight	45.5±20.9	71.3±13.7	47.6±20.4
Share-CNN-Combine	72.3±19.6	46.4±8.5	63.8±14.3

Table 4.4 Performance comparison of CNNs with/without VOI (mean±std.)

Method	Recall (%)	Precision (%)	F_2 score (%)	Time (sec.)
VOI-90k-IND-CNN	53.3±17.7	58.8±11.7	53.4±15.3	6.9±0.4
VOI-190k-IND-CNN	62.6±19.2	52.6±10.7	59.2±15.9	15.1±1.3
IND-CNN	69.8 ± 20.1	47.7 ± 11.0	62.8 ± 16.1	110.5±59.0
VOI-90k-Share-CNN	53.7±16.4	59.1±11.0	53.9±13.9	6.5±0.4
VOI-190k-Share-CNN	63.1±17.8	53.0±10.0	59.8±14.1	14.1±1.2
Share-CNN	72.3 ± 19.6	46.4 ± 8.5	63.8 ± 14.3	103.4±55.7

E. Share-CNN Combined with VOI selection

Table 4.4 compares the performance of CNN with or without VOI, where different N values (adaptive thresholding to control the voxel cardinality) are considered. When sacrificing the voxel cardinality size (fewer voxels), the benefit is a reduced computational complexity, e.g. going from $\sim$100 secs. processing time to $\sim$10 secs. per volume, while the VOI selection is still able to reduce the number of false positives at the cost of a slight drop in F_2 score (for larger N). Although the VOI selection degrades the system performance, it dramatically decreases the number of voxels to be classified by the CNN. For comparison, IND-CNN is also included in the table, which shows a small performance degradation in efficiency and accuracy (with/without VOI selection). Moreover, IND-CNN has also more parameters in the model and is therefore more complex than Share-CNN. The execution time is measured using a Titan 1080Ti GPU and Python 3.7 on a standard PC with 32-GB RAM and 2.4-GHz CPU.

4.4.4 Catheter Localization

Based on voxel classification, the model-fitting is applied to the binary segmentation mask as discussed in Section 4.4.1, to localize the catheter (its skeleton and end-points) and remove the outliers. We employ the following metrics to measure the model-fitting performance: skeleton-based metrics, Volumetric Similarity (VS), and Average Hausdorff Distance (AHD) (the VS and AHD metrics are defined in Chapter 2). More specifically, the skeleton-based metrics include two specific types: (1) End-points error (EE), which is characterized by the average distance between corresponding end-points on the detected catheter and the end-points of the annotation; (2) The skeleton error (SE): the average distance between 5 equally-sampled points on the detected skeleton and the ground-truth skeleton. This error is defined as taking the shortest distance from each of the five points on the detected skeleton to the ground-truth skeleton. Those five shortest distances are averaged. The skeleton error has a more robust performance than the EE. This performance difference is explained by analyzing the difficult cases. For example, sometimes the catheter tip is attaching to the tissue, so that it is hard to distinguish the tip from the tissue in B-mode imaging, as shown in Fig. 4.8 (a). In such case, the EE metric will give a higher error than the SE. In any case, the SE metric has an inherently better accuracy because its definition is more generic. However, the EE could be more informative than the SE, because correctly localizing the tip of the catheter can facilitate the success of the intervention.

Here, we compare the catheter localization performance based on MF-AdaB, Share-CNN, VOI-90k-Share-CNN, and VOI-190k-Share-CNN. The localization performances are shown in Table 4.5, which are averages of a threefold cross-validation with five times model-fitting in each volume. The table shows that the proposed Share-CNN method achieves a better performance with a lower position error, which is smaller than the diameter of the catheter. Furthermore,

Table 4.5 Performance comparison on catheter localization. EE: end-point error, SE: skeleton-point error, VS: Volumetric Similarity, AHD: Average Hausdorff Distance. The numbers indicate mean±std.

Method	EE (mm)	SE (mm)	VS (%)	AHD (voxel)
MF-AdaB	3.33±2.76	2.91±2.55	67.3±20.7	6.71±7.72
Share-CNN	2.25±1.91	1.83±1.28	76.7±13.5	1.72±1.85
VOI-90k-Share-CNN	2.07±1.22	1.71±1.00	77.3±11.6	1.56±2.32
VOI-190k-Share-CNN	2.08±1.22	1.73±0.99	77.8±11.6	1.64±1.82

the results show that the VOI-based CNN can boost the localization precision in terms of the lowest error. When comparing the results in Table 4.5 and Table 4.4, the VOI pre-selection provides a lower F_2 score, but better localization accuracy. This is because VOI pre-selection provides a higher Precision performance, so that a better sparse volume can be achieved. The model-fitting relies on the SPD model-fitting, where fewer outliers would make randomly chosen control points more stable. Moreover, at the expense of 4% lower F_2 score through VOI-selection, we achieve 5-7 times faster voxel-based classification in the whole volume, which poses a clear trade-off between classification accuracy and computation efficiency. The whole processing chain based on VOI-190k-Share CNN takes about 18 secs. execution time (Frangi filtering: 1.5 secs., VOI selection: 0.3 secs., CNN: 14 secs. and SPD-RANSAC: 1.9 secs.). As a conclusion, at the expense of an acceptable degradation in segmentation performance by using coarse VOI selection, the overall execution efficiency is drastically improved, while preserving sufficient detection accuracy.

4.5 Conclusions

In this chapter, we have developed a coarse-to-fine-based catheter detection method in challenging 3D US data. The automated detection method is aiming to localize and segment the instrument from complex anatomical tissues in the received data. The proposed algorithm is characterized by three stages of processing with increasing accuracy: VOI pre-selection, CNN-based voxel-level classification and mode-fitting optimization. In the proposed framework, we have introduced a novel coarse-to-fine segmentation method by combing the *a-priori* model matching with state-of-the-art CNN classification, which can robustly and efficiently extract the catheter voxels from anatomical structures. Despite the relatively low segmentation score of about 60%, the method can be successfully applied because this detection rate is sufficient for readily initializing the first stages of processing (VOI pre-selection and CNN-based classification), so that the detection method can be used smoothly in the experiments.

The framework solves two key challenges for catheter detection in 3D US volumes: (1) efficient coarse candidate selection, and (2) robust and accurate fine

voxel classification. For the first challenge, VOI pre-selection is applied by combining Frangi vesselness filtering with adaptive thresholding, which efficiently and drastically discards the non-catheter voxels. As for the second challenge, a computation-efficient CNN is proposed to accurately and efficiently classify the resulting voxels from the VOI pre-selection. Based on the proposed novel framework, the overall performance of our method is much higher than the conventional handcrafted method with 10 times faster detection efficiency. With the classified voxels, the proposed method can localize the catheter with an average end-point error of about 2.1 mm, while it overall needs an execution time of 18 seconds per volume. Therefore, a higher efficiency can be achieved such that the detection algorithm is more acceptable for clinical application. The proposed algorithm already offers an order of magnitude faster execution, while the last speed up can come from algorithm optimization and clever mapping on the CPU.

The main contributions of this chapter are summarized in two aspects: (1) a novel design of VOI pre-selection, which drastically removes the background voxels without high computation cost, and (2) a novel design of a tri-planar CNN for voxel-level classification, which significantly reduces the computation cost and achieves better performance than the current literature. Further improvements are possible for improving the efficiency and accuracy for real-time performance. The remaining challenges are: (1) the VOI pre-selection still includes numerous irrelevant voxels, which is due to limited exploitation of discriminative information, and (2) the CNN-based classification cannot fully exploit the full semantic information due to the limited scope of the local 3D patch and the applied tri-planar strategy.

These remaining issues are explored in the next chapter (Chapter 5), which will investigate an advanced method that still focuses on a coarse-to-fine strategy, but with a much better efficiency and accuracy. Specifically, a slice-based neural network is employed to coarsely localize the instrument in the 3D US volume, which is then accurately segmented by a novel proposed 3D neural network. These networks will jointly improve the detection to a semi-real-time performance with a state-of-art accuracy comparable to the data from the latest literature.

CHAPTER 5

Instrument Segmentation by Patch-of-interest-based FCN

5.1 Introduction

The previous chapter has introduced an advanced framework to create an automated instrument segmentation system, which distinguishes the target in 3D US by a coarse-to-fine strategy. To delineate the instrument efficiently, a dedicated coarse pre-selection of voxels of interest and a fine voxel classification have been designed, which are evaluated to detect and segment the instrument in US images. It has been shown that the dedicated voxel-of-interest pre-selection can drastically reduce the computation task by a CNN for fine voxel classification. However, this proposal has still limitations for the clinical real-time applications, when considering the accuracy of coarse pre-selection and the overall efficiency. In addition, the tri-planar strategy limits the collection of semantic information by slicing, which constrains the fine segmentation performance based on voxel-level classification. This chapter aims at addressing these limitations.

5.1.1 Objective and Brief System Outline

The objective of this chapter is to model the automated medical instrument segmentation in 3D images by a region-of-interest-based semantic segmentation method. This method involves a fast but coarse region-of-interest selection and a fine semantic segmentation. As a result, the instrument in the 3D US data can be segmented more robustly and with higher efficiency (compared to Chapter 4). Following the general framework of *Segmentation-Modeling*, which has been proposed in the previous chapters, the proposed method consists of three key steps, as is depicted in Fig. 5.1. These steps are (1) efficient region-of-interest selection,

(2) fine semantic segmentation for medical instrument, and (3) catheter localization by model-fitting or other supplementary analysis.

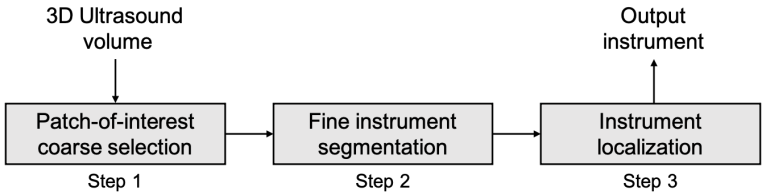


Figure 5.1 Diagram of an improved coarse-to-fine instrument detection and segmentation system. This chapter focuses on the first two steps of the diagram.

Based on the above approach, the key challenge is to design an efficient and robust coarse region selection, and then accurately perform (fine) semantic segmentation on the pre-selected space. More detailed challenges and corresponding solution directions are elaborated in the following subsections.

5.1.2 Specific Challenges for Interested Region Selection

In order to coarsely select the interested region or voxels with high efficiency from complex input US data, two essential aspects are considered for the coarse selection. (1) Efficiently selecting the region of interest, and (2) discriminate the instrument voxels and omit the non-instrument voxels as much as possible. Although Chapter 4 has addressed these challenges by a specifically designed Frangi filtering stage, this method has limitations because first, it is computationally expensive due to iterative filtering at the voxel-level, and second, a pre-defined vesselness filter limits the finding of discriminative information. We now briefly address these issues and objectives associated with region-of-interest selection and discriminative information exploration.

1. *Discriminative information exploration:* To better exploit discriminative information of the instrument, a machine learning method for voxel classification is commonly applied, such as voxel-level CNN classification. However, this method is computationally expensive and time-consuming. Another choice is to employ a state-of-the-art Fully Convolutional Network (FCN) on the image, which can better exploit the semantic information by a powerful GPU (such as widely adopted UNet [42]). Nevertheless, applying a 3D FCN on the whole 3D image for segmentation is challenging due to the complex network design, which requires a large amount of training images. Therefore, this network is difficult to be trained in our case.

2. *Efficient region selection:* By employing 2D FCN for semantic segmentation, the candidate instrument region can be localized. Nevertheless, the itera-

tive slice-based segmentation is still time-consuming and infeasible for efficient coarse selection, since it includes redundant computations for the slice-by-slice processing.

To address these challenges, two directions are jointly considered. First, as a compromise and to segment the instrument in 3D US, a 2D FCN is considered, which is applied on the decomposed 3D US data. By doing so, the 2D slices are extracted from the 3D volume, which thereby reduce the computation cost and yet still partly exploit the discriminative information. Second, to further improve the coarse selection efficiency, we propose to apply a spatial downsampling strategy in both input slices and 2D FCN design. By doing so, the interested regions containing the instrument can be obtained efficiently.

5.1.3 Specific Challenges for Fine Semantic Segmentation

With obtained coarse segmentation results, a fine segmentation is required on the selected region. A CNN-based fine voxel classification has been proposed in Chapter 4. However, this method is only classifying the remaining voxels with voxel-based classification, which is computationally expensive and less accurate than a 3D FCN. The involved challenges are now briefly listed below.

3. *Patch-based segmentation:* CNN-based classification only considers the selected candidate voxels for fine segmentation, which omits the misclassified voxels from the coarse selection steps and therefore obtains worse results than direct classification on the whole image. In addition, it would be computationally expensive when applying the CNN on all candidate regions, such as a larger cube containing the segmented instrument from the first step. Moreover, as discussed in the above, the 3D FCN is challenging of its own when applied to the whole US images.

4. *Exploration of FCN information:* With a 3D FCN method, one key challenge is the trade-off between network complexity and the amount of available training images. In most cases, the amount of training images are limited, which leads to a compact and simplified 3D FCN to reduce overfitting. Another choice is to consider transfer learning by a pre-trained network resulting from other tasks. Nevertheless, these pre-trained networks are commonly obtained in 2D format, which is not feasible for 3D US data.

With the above considerations, a patch-based 3D FCN is considered for the selected region, which can efficiently re-segment the candidate regions with better results. Here, a patch is a selection of a voxel volume, thereby being a three-dimensional information blob. To segment such 3D patches by a 2D pre-trained network, additional information has to be fused into the network. This extra information is supplied in the form of the direction of principal axis of the volume, which is then aggregated with a 3D compact FCN at the feature-map

level. In this way, the fine segmentation performance can be improved.

The sequel of this chapter is organized in the following way. Section 5.2 summarizes the related work in this field, detailing instrument detections by non-learning approaches and learning-based methods. Section 5.3 describes the proposed method, including every step of the coarse-to-fine segmentation. Sections 5.4 and 5.5 demonstrate and present the considered dataset, implementation details and experimental results. Finally, Section 5.6 concludes the chapter and presents some discussions on possible refinements.

5.2 Related Work

Image-based instrument detection or segmentation in 3D US has been studied during the past years, but the amount of studies in this area are still limited. From the viewpoint of methodology, these works can be classified into two categories: non-learning-based methods and learning-based methods.

5.2.1 Non-learning-based Methods

Before the popularity of machine learning in computer vision, conventional technologies were applied to 3D US data to detect medical instruments, by analyzing geometry and intensity properties of the instruments, such as shapes, intensity distribution, etc. In [62], the authors proposed to apply Principal Component Analysis (PCA) on thresholded 3D US volumetric data, which was derived by applying cluster analysis, to select the most likely region as the detected instrument. In [16], the Radon Transformation was applied on a 3D US volume to accumulate intensity values, which was used to localize a straight electrode in 3D US. Similar to the Radon Transformation, the authors of [63] proposed to detect the needle by a line-based description in 3D space using 3D Random Hough Transformation. With a more advanced spatial and instrument model description in 3D space, the authors of [20] proposed to apply a line filter in 3D US, which can roughly filter out needle-like structures in 3D images. Then the 3D RANdom SAmple Consensus (RANSAC) algorithm is applied to select the most likely region as the target instrument. In the same year, another publication [46] applied template matching on 3D US volume to detect the catheter with complex post-processing, which achieved successful detection results with a strong assumption of catheter direction in the images.

In the above approaches, RANSAC with line filtering achieved a more promising performance, because of a better thresholding for 3D US and efficient model description by RANSAC. However, the above approaches do still have limitations. (1) With limited discriminating information by the thresholding method, it is generally hard to extract accurate 3D regions for instrument detection. (2) Most of the above methods have been validated on simulated 3D

images or phantom data, which are significantly different than real clinical applications. (3) Strong assumptions on the instrument shape and direction are leading to an unsatisfactory generalization of the proposed methods. Therefore, the above methods do not fully exploit the information of the instruments.

5.2.2 Learning-based Methods

These techniques have been studied in recent years, which classify voxels into the binary instrument/non-instrument categorization. Handcrafted features were proposed by considering Frangi vesselness filter [48], Gabor filterbank [28], time-domain statistical feature [64] or multi-definition features [65], which achieved reasonable instrument segmentation results albeit with complex post-processing. However, these methods are less robust or partly inefficient when US images are recorded from a complex anatomical environment, due to the voxel-based processing.

Recently, deep learning, such as convolutional neural networks (CNNs) or fully convolutional neural networks (FCNs), have been intensively studied and applied for medical imaging-related areas [9]. CNNs are applied as a classifier to distinguish the category of the voxels in the 3D US, which are used to segment medical instruments in the 3D US image. A voxel-of-interest-based CNN pipeline [19] has been proposed to segment the instrument for cardiac intervention. The Frangi vesselness filter [15] is firstly applied to select the the possible voxels belonging to the instrument globally, and then a CNN is subsequently applied to classify the remaining voxels. This method avoids iterative voxel prediction on the full volume and achieves an inference time of 10 seconds on the average per volume. However, this efficiency is still far from real-time clinical application. More recently, slice-based semantic segmentation is applied to 3D US images to segment the instrument efficiently [66, 67]. However, this 2D approach has limited performance due to the slice-based strategy, which hampers the 3D information usage. Alternatively, patch-based 2.5D [68] and 3D [69] semantic segmentation methods have been proposed to segment the instrument in 3D US. Nevertheless, similar to voxel-based methods, straightforward iterative patch-based prediction on a full volume requires considerable computation time, which is not attractive for real-time applications (typically requires more than 10 seconds per volume). Furthermore, the segmentation performances [68, 69] are not optimized because of their limited information usage by a single network designing with limited training samples.

5.2.3 Direction of Proposed Method

To accurately and efficiently segment the instrument in 3D US by a semantic segmentation approach, a coarse-to-fine strategy is adopted for our method, which contains three levels of processing for the US image.

- *Coarse slice-based segmentation*: The 3D volume is decomposed into 2D slices along principal directions of the volumetric data (i.e. two principal directions parallel to US cone). These slices are efficiently segmented by a 2D segmentation network, yielding an initial coarse segmentation.
- *Patch-of-interest selection*: Based on the segmented slices, a 3D coarse segmentation result can be obtained by combining the segmentations of the slices into a coarse segmentation result. This 3D coarse result is divided into 3D patches from the original volumetric data, where patches are selected that contain (parts of) the initial coarse result. Then, corresponding patches from the original 3D volume constitute the coarse segmentation result, which is further processed for fine segmentation.
- *Fine patch-based segmentation*: The patches from the 3D US image are processed by a 3D network for fine segmentation, which exploits 3D contextual information within the selected patches for a better segmentation result. Therefore, the instrument can be segmented with accurate results, while avoiding expensive computations on irrelevant regions.

Compared to recent methods from literature, the proposed method efficiently exploits the discriminative information in coarse region selection, which provides a better estimation for regions of interest. In addition, a 3D patch-based network is considered for fine instrument segmentation. Therefore, the overall efficiency is improved by avoiding the computation cost of using a 3D network on non-instrument regions.

5.3 Methods

The proposed instrument segmentation method in 3D US images consists of three steps for a coarse-to-fine strategy, which is depicted in Fig. 5.2. More details are presented in the paragraphs below.

- *Step 1 – Slice-based UNet for coarse segmentation:* The input volume is decomposed to slices along principal directions of the volumetric coordinates, which are segmented by a slice-based 2D UNet. In addition, to further improve the detection efficiency, a spatial downsampling approach is applied to the 3D volume. This reduces the number of slices in the volumetric direction to one half of the amount of slice data. The spatial downscaling of the individually selected slices is performed in the slice-based UNet inference step. This downscaling is also implemented in the output layers of the network. As a consequence, the 3D volume is coarsely segmented with high efficiency.
- *Step 2 – Patch-of-interest (POI) selection:* Based on the initial coarse segmentation, the 3D coarse segmentation is obtained by combining the segmented

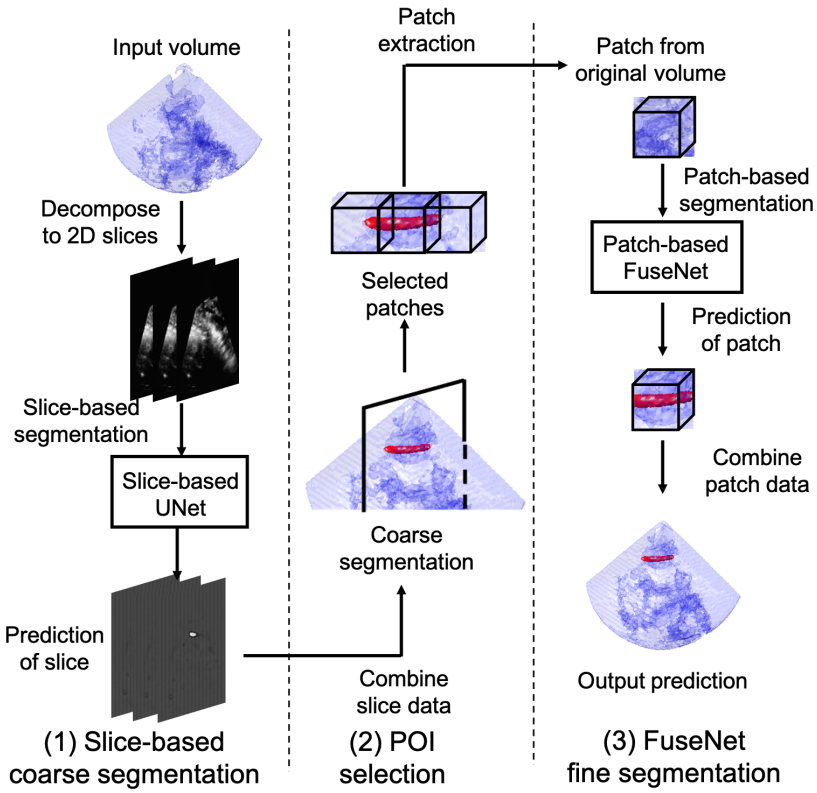


Figure 5.2 Overview of the proposed method. First, the input volumetric image is decomposed to 2D slices for a slice-based UNet segmentation. Second, the slice-based predictions are combined as an initial 3D coarse segmentation. In the second column, the coarse segmentation volume is divided into 3D patches, and the patches containing parts of the coarse segmentation of the instrument are selected. Third and in the right column, the selected patches are segmented by a patch-based FuseNet, which leads to fine segmentation results. The output volume is finally obtained based on the results of selected patches with the fine segmentation inside.

slices into a 3D volume. The 3D image is then divided into small patches, and the patches containing parts of the coarse segmentation results are selected with the coordinate information. Using that information, the corresponding patches are selected from the original 3D US volume. This collection of selected 3D patches forms the coarse segmentation result.

- *Step 3 – FuseNet fine segmentation:* Based on the selected patches from 3D US, a FuseNet is proposed to perform 3D semantic segmentation. These patches' voxels are re-segmented by the FuseNet, which thereby improves the final performance into a fine segmentation by exploiting more 3D semantic information. To better supervise the FuseNet, a hybrid loss function is introduced to enable the network to simultaneously learn the pixel-level and image-level discriminating information.

In the following subsections, the these steps are elaborated in detail. The subsections follow the order of the processing step from Fig. 5.2.

5.3.1 Slice-based UNet for Coarse Segmentation

When applying 3D UNet to the whole image for ROI-based feature extraction and segmentation [70], the key challenge is the limited GPU memory for complex 3D operations. Besides this, the instrument has a large variance in length and location inside the 3D space, which is typically ranging from 9 to 100 voxels. As a consequence, it is challenging to apply the feature-map-based ROI segmentation, which is designed for colorectal tumor segmentation [70]. Alternatively, when applying a patch-based segmentation, a two-stage coarse-to-fine strategy [20] is commonly applied to avoid exhaustive segmentation in 3D space. The 2D slice-based UNet was originally proposed to segment the instrument [66], which however processes limited 3D spatial information and obtains worse performance than 3D UNets. Moreover, iterative slice-by-slice prediction leads to redundant computations and reduces time efficiency, especially when considering this approach as a pre-processing method to extract detailed regions of interest. Based on these considerations, we propose a slice-based UNet with downsampled prediction, using spatial skipping of slices in the slice-taking direction, to improve the prediction calculation efficiency and provide an initial coarse segmentation result.

The framework for the coarse segmentation is shown in Fig. 5.3. Considering a simple example, the input volume with a size of M^3 voxels is decomposed into 2D slices along the lateral direction, which is denoted as slice set $\mathcal{M}$. This set is iteratively downsampled on slice basis to a ratio of K, and called $\mathcal{M}_K$. For each 2D plane in $\mathcal{M}_K$, we further extract its two adjacent slices in original slice set $\mathcal{M}$, however with a spatial gap d_1 (in voxels) between two adjacent slices. Based on these slices, a three-channel image is constituted to mimic RGB imaging and used

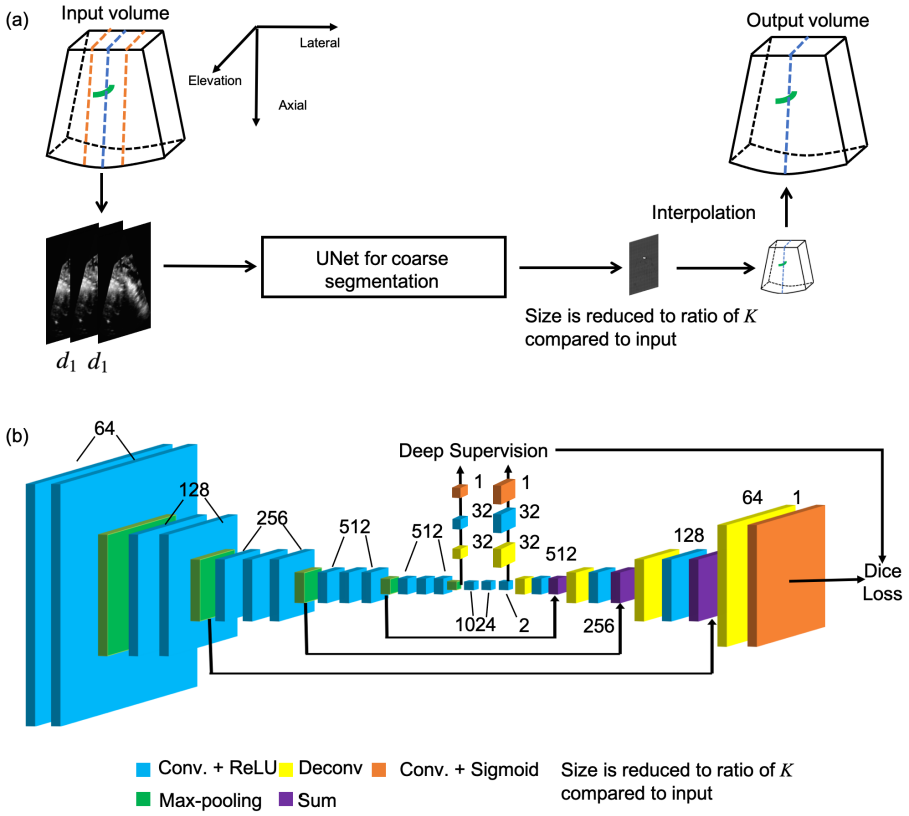


Figure 5.3 Coarse instrument segmentation framework and the corresponding slice-based UNet. (a) Slicing of the 3D volume and initial segmentation per slice (slices are spanned along axial and elevation directions in the drawing, and sliced in the lateral direction). (b) Detailed structure of Slice-based UNet architecture showing the dimensions of the convolutions.

as the input for the slice-based UNet. The three channels are originated from the spatial slice dimensions and the direction in which the slicing takes place.

The slice-based UNet is based on the VGG16 encoder [17], since it was proven to be a successful backbone for the instrument segmentation task in US images [66, 67]. As is shown in Fig. 5.3, the slice-based UNet has convolutional layers with 5-level Max-pooling (green layers at the left side of Fig. 5.3 (b)). After the last Max-pooling layer, the subsequent convolutional layers have kernel numbers 1024, 1024, and 2. After those layers, 4 deconvolutional layers are following, which all have equal kernel sizes of 2×2. Moreover, for each deconvolutional layer, an additional convolution operation is added to improve stability. To exploit more discriminating information at different scales, skipping connections are considered to construct the UNet structure. To further improve the perfor-

mance of UNet, deep supervision [71] is employed at different feature scales at the decoder side. With the proposed spatial downsampling slicing strategy, the output volume is obtained with a faster prediction for the initial coarse segmentation. To address the challenging case that the instrument is crossing the slices with small footprint, an orthogonal slicing strategy along the elevation and lateral directions is adopted in the coarse segmentation [66, 67] as the complementary to the spatial stride d_1. Because more spatial information of the instrument can be observed, this is better than the case when only applying single direction slicing during the training.

It is worth to mention that due to the nature of the ultrasound imaging, if the instrument orientation is parallel with the axial direction and thus perpendicular to the phased array planes of sound waves, the instrument wave reflection of the sound wave cannot be captured by the US transceiver. This occurs because there is only a small circle to reflect. Therefore, when slicing the 3D US volume, the axial direction is omitted, since the instrument is rarely positioned in parallel with the axial direction (the physician has learned to position the instrument in the orientations such that it can be successfully imaged). In addition, the processing of the US image can be done at a detailed resolution for finding the instrument. An example case that the instrument is crossing the slice with small footprint is depicted in Fig. 5.4 (a). The 2D UNet is pre-defined to segment the small footprint because a small spatial stride d_1 is applied in both available slicing directions for training to capture sufficient discriminative information.

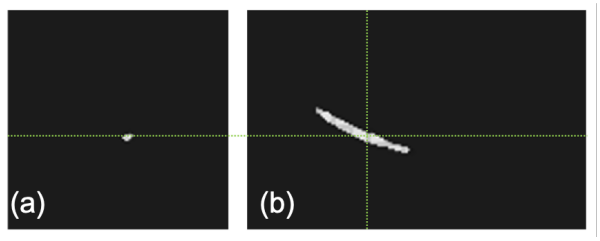


Figure 5.4 Example segmentation result of an extreme case where the instrument is crossing the slice with a small footprint. (a) Slice view from the plan perpendicular to the dominant instrument orientation. (b) Slice view from one alternative perpendicular direction in the same case.

5.3.2 Patch-of-interest (POI) Selection

This subsection applies the second stage of the processing US volume, in order to find the patches containing the instrument. As a consequence of the deconvolution operation and input set $\mathcal{M}_K$, the output prediction is downsampled with a ratio of K compared to the original input image in each dimension. The downsampled 3D prediction is later upsampled to its original size by interpolation. Thresholding and connectivity analysis are applied to select the two largest

connected components as the POI volume for patch extraction. Given the input image, the 3D volumetric data is divided into small non-overlapping patches with size D^3 voxels. By comparing the coarse segmentation and pre-allocated patches, patches containing coarse instrument predictions are extracted as the input for the FuseNet (the second CNN). It is worth to mention that because of the zero-padding in convolution operations for FuseNet, a $(D+S)^3$-voxels patch is actually extracted based on the patch above, where the S voxels serve as the compensation for information leakage at the patch boundary, which is depicted in Fig. 5.5. In this chapter, we select $D = 32$ and $S = 16$ based on empirical results, which yields a balance between efficiency and accuracy.

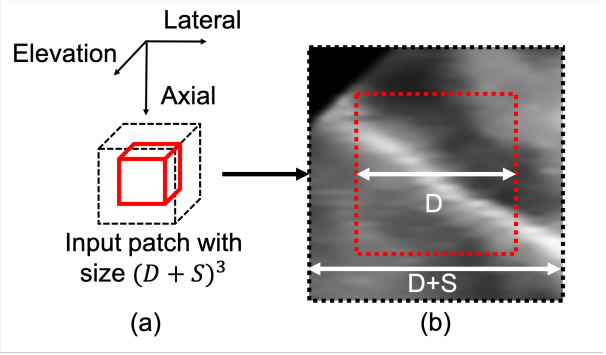


Figure 5.5 Input patch of FuseNet, which has a size of $(D+S)^3$ voxels. Parameter D is the non-overlapping patch size while S is the padding parameter to compensate for the boundary. (a) 3D patch visualization, where the red patch has a size of D^3 voxels and a dashed patch has a size of $(D+S)^3$ voxels. (b) 2D slice example from the 3D patch.

5.3.3 Patch-based FuseNet for Fine Segmentation

In this section, a novel FuseNet for 3D segmentation is proposed. The overview of the proposed FuseNet is shown in Fig. 5.6, which consists of two individual UNet variants with different spatial operations for 3D patch segmentation: a semi-3D Direction-Fused UNet (DF-UNet) and a full-3D Pyramid-UNet, which are depicted in Fig. 5.7 and Fig. 5.8, respectively. Intuitively, the DF-UNet exploits the intra-slice information by using a 2D feature extractor, while it is utilizing the inter-slice information by high-level tensor operations. This is denoted as 2.5D feature map, at the top right of Fig. 5.6. Nevertheless, this network cannot correctly analyze 3D contextual information, due to its 2D feature extraction. In contrast, the 3D Pyramid-UNet exploits the 3D information in a more straightforward way. However, this 3D UNet may not be properly trained with limited datasets. To make use of these two successful networks and compensate their limitations, FuseNet is proposed by creating a high-level feature fusion of these sub-networks. Moreover, it also addresses limitations for an individual network:

(1) an individual network may not properly exploit the spatial information, especially DF-UNet, and (2) a single network would lead to knowledge bias. As a solution, making an ensemble of several networks for prediction can typically improve the overall performance. From the nature of FuseNet, it exploits the semantic information from different dimensions and fuses them for instrument segmentation. The details of the Direction-Fused UNet and 3D Pyramid UNet are discussed in the sequel.

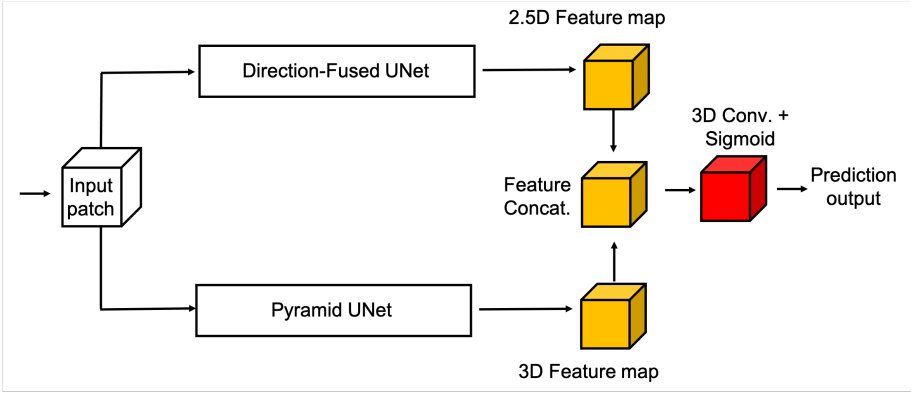


Figure 5.6 Overview of the patch-based FuseNet, which consists of Direction-Fused UNet, Pyramid UNet and a feature-fusion part. The feature maps from two different branches are concatenated and processed by external convolutional operations to generate the output prediction.

A. Direction-Fused UNet

An input patch with size $(D+S)^3$ voxels is decomposed into $D+S$ planes along each axis. For each plane, a 3-channel image is formed based on the adjacent images of the actual plane with a spatial gap of d_2 voxels along the axis. This is visualized in Fig. 5.7 and denoted as a 2.5D image at the bottom. As a result, the patch leads to $D+S$ different 3-channel images in each direction (padding is applied at the boundary plane). Then, each image is processed by the 2D UNet, which is based on a VGG16 encoder and a customized decoder. The encoder is based on the VGG16 network, from which the dense connections are removed. Then, the output of the encoder is filtered by three convolutional layers with filter numbers 1024, 1024, and 2 (the middle of Fig. 5.7 (b)). The decoder stage includes four deconvolution layers with masks of 2×2, 2×2, 2×2 and 4×4 pixels, in sequential order. Furthermore, after each deconvolution layer, an extra convolution operation is included to smooth the features. To limit the cost of the GPU memory, we perform summation instead of concatenation for skipping connections. As shown in Fig. 5.7 (a), the images are processed by the 2D UNet and its output features are stacked based on the plane's original positions, to construct feature maps along three axes together with high-dimensional

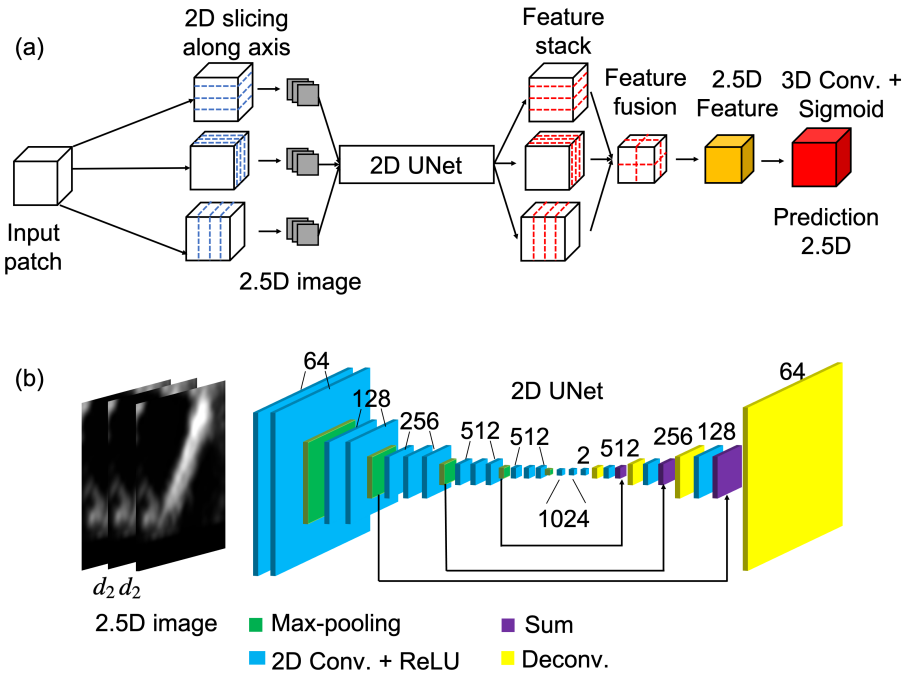


Figure 5.7 Overview of the patch-based Direction-Fused UNet, which is constructed based on a 2D UNet and feature-space operations. (a) Input 3D patch is decomposed into adjacent slices in three different directions simultaneously, which are processed by the 2D UNet to generate slice-based feature maps. The slice-based feature maps are then permuted to form feature maps in 3D space. (b) Input 2.5D image with stride d_2 as the input for 2D UNet.

transposition. Furthermore, feature maps from different axes are accumulated to form a fused feature map (top right of Fig. 5.7 (a)). Finally, the final prediction of the DF-UNet is obtained by applying a 3D convolution and a sigmoid layer (at the right most of Fig. 5.7 (a)). A further detail of the implementation is to accelerate the training and inference efficiency. This is achieved by extracting 3-channel images per direction to form a mini-batch for the 2D UNet, rather than a slice-by-slice processing of the feature extraction.

B. Pyramid-UNet

The proposed Pyramid-UNet is based on a customized 3D UNet, which has a simpler network architecture and avoids overfitting [61]. For a feature pyramid-based network that goes deeper in layers, the discriminating information at the low-level feature map vanishes (pixel-level, bottom branch of Fig. 5.8) and degrades the segmentation performance. Even though the UNet employs skipping connections to preserve the low-level information between encoder

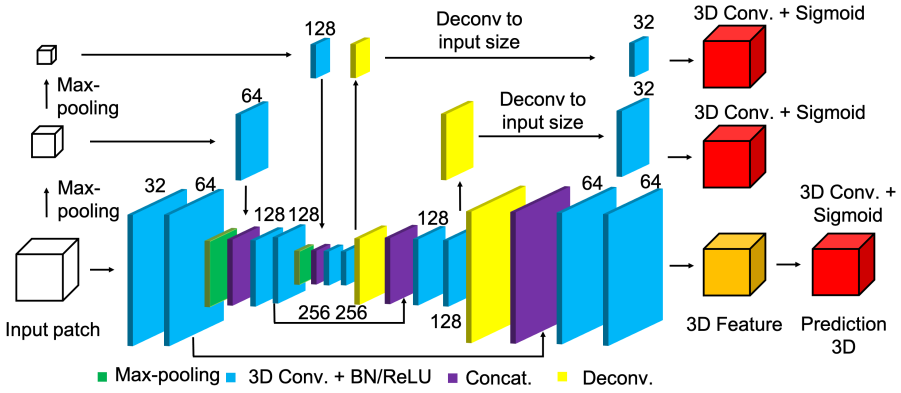


Figure 5.8 Overview of the patch-based Pyramid UNet, which is a 3D UNet with pyramidal input and output. The input 3D patch is reduced by max-pooling operations to generate low-level feature maps, which are concatenated with high-level features, so that the degradation of information is compensated. The pyramidal outputs are based on deep supervision, which therefore allows the network to preserve the discriminating information at different image scales.

and decoder, it still cannot adequately exploit the information at different image scales [43, 72]. To preserve more detailed information at different image scales, we consider to introduce pyramid input branches. The pyramidal inputs at different scales preserve low-level information within the encoding stage, which potentially compensates the vanishing of information during the feature extraction of the UNet. Furthermore, to better supervise and synchronize the features at different decoder scales, we also use deep supervision [71] and introduce an extra convolutional block for better stability. By introducing the pyramidal inputs and outputs to the UNet, the proposed network potentially preserves more information at different feature scales than the standard UNet for US images. As depicted in Fig. 5.8, the network has 32 kernels at the very beginning, which is gradually doubled when the information is passed to deeper layers in the right direction.

C. Feature Fusion

Based on the Direction-Fused UNet and the Pyramid UNet, feature fusion is performed to combine the features from different feature extractors. As shown in Fig. 5.6, feature maps extracted from two networks, e.g. denoted by dark-yellow cubes, are concatenated prior to convolution operations, which is followed by two convolutional layers with the filter numbers of 24 and 12 in the final 3D convolution layer (red cube). The final prediction of FuseNet results from the feature fuse layer and the sigmoid layer, which is indicated as a red cube in Fig. 5.6.

5.3.4 Hybrid Loss for Patch-based Fine Segmentation

To better supervise the overall patch-based FuseNet and to enforce it to learn more contextual information, we propose a hybrid loss function. This is in contrast with the conventional voxel-based difference, which can be trained with cross-entropy or Dice loss. The hybrid loss function includes a class-balanced focal loss (FL) and a contextual loss (CL). Using a predicted patch and corresponding ground truth, which are denoted as $\hat{Y}$ and Y, the hybrid loss function is defined as

$$\text{Loss}_{\text{Hybrid}}(\hat{Y}, Y) = \text{Loss}_{\text{FL}}(\hat{Y}, Y) + \text{Loss}_{\text{CL}}(\hat{Y}, Y), \tag{5.1}$$

where Loss_{FL} denotes the class-balanced focal loss and Loss_{CL} is the contextual loss. For each predicted output of the FuseNet, Eqn. (5.1) is applied with unity weight for each individual loss, except the outputs from deep supervision in the Pyramid UNet, which are assigned as 0.6 and 0.4 for the middle and the top branches in Fig. 5.8, respectively.

The loss function, such as Dice or cross-entropy, is typically applied for segmentation tasks in medical imaging. However, it is not optimal when segmented objects have large size variations and imbalanced class distributions in the ground truth [73]. Moreover, when the instrument has a small size in 3D space, then the boundary voxels, which are difficult to classify, are more critical than easily-classified voxels at the center part of the instrument. The commonly used loss functions may not be optimal, especially for the POI-based task, which requires a more accurate segmentation. Therefore, we have adopted the class-balanced focal loss function, which is based on the binary cross-entropy and the F-score loss [74, 67]. The latter loss term helps in steering the segmentation. The proposed focal loss is then defined by

$$\begin{aligned}\text{Loss}_{\text{FL}}(\hat{Y}, Y) = &-\Big(\sum_{i=1}^{N} \omega_{ci}(1-\hat{y}_{ci})^{\sigma}\log(\hat{y}_{ci}) + \sum_{i=1}^{N} \omega_{ni}(1-\hat{y}_{ni})^{\sigma}\log(\hat{y}_{ni})\Big) \\ &+ \Big(1 - \frac{(1+\beta^2)\sum_{i=1}^{N} y_{ci}\hat{y}_{ci}}{(1+\beta^2)\sum_{i=1}^{N} y_{ci}\hat{y}_{ci} + \beta^2\sum_{i=1}^{N} y_{ci}\hat{y}_{ni} + \sum_{i=1}^{N} y_{ni}\hat{y}_{ci}}\Big)^{\gamma},\end{aligned} \tag{5.2}$$

where y_{ci} denotes an instrument voxel from the ground truth, $\hat{y}_{ci}$ represents the voxel's prediction probability for the instrument class, while y_{ni} and $\hat{y}_{ni}$ are non-instrument voxels and their corresponding prediction probability, respectively. Parameters β and ω are controlling the weight between different classes, which are calculated as the square root of the inverse of the class ratio. Power parameters γ and σ are controlling the slope of the loss curve, which are empirically set to $\gamma = 0.3$ and $\sigma = 2$.

Conventionally, networks are learned by employing a voxel-based loss function, such as cross-entropy or Dice loss, which ignores the high-level differences between the prediction and ground truth at a global level. To allow the network

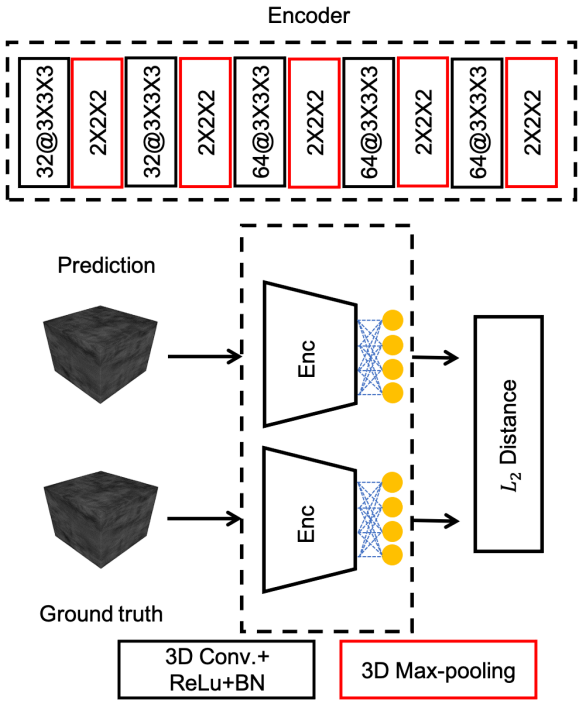


Figure 5.9 Overview of the patch-based contextual loss, which is constructed based on a shared encoder for both prediction and ground-truth patch. The feature-space differences are measured by L_2 distance. BN indicates the batch normalization.

to learn a better contextual representation or so-called high-level feature representation, we propose a contextual loss, which formulates the contextual difference in a high-level feature space. The prediction and ground truth are processed by a contextual encoder, which is depicted in Fig. 5.9, to generate high-level feature vectors in latent space, which are denoted as $S_{\hat{Y}}$ and S_Y, respectively. As a consequence, the contextual loss Loss_{CL} is characterized by

$$\text{Loss}_{\text{CL}}(\hat{Y}, Y) = ||\text{CE}(\hat{Y}) - \text{CE}(Y)||_2 = ||S_{\hat{Y}} - S_Y||_2, \tag{5.3}$$

where $||\cdot||_2$ denotes the L_2 distance and $\text{CE}(\cdot)$ represents the context encoder in Fig. 5.9.

5.4 Experiments and Implementation

5.4.1 Datasets Description

***Ex-vivo* RF-ablation catheter dataset**: The *ex-vivo* dataset consists of ninety-two 3D cardiac US images from eight porcine hearts. During the recording, the hearts

were placed in a water tank with an RF-ablation catheter (diameter≈2.3-3.3 mm) inside the heart chambers. The US probes were placed next to the heart, to capture the images containing the instrument. Our data collection setup is visualized in Fig. 5.10 (a) and (b). From the dataset, example 2D B-mode image is shown in Fig. 5.10 (c). The dataset includes volumes with a size ranging from $120 \times 69 \times 92$ to $294 \times 283 \times 202$ voxels, in which the voxel size was isotropically resampled to the range of 0.4–0.7 mm. The datasets were manually annotated by technicians with guidance and approval of clinical experts, to generate the binary segmentation mask as the ground truth.

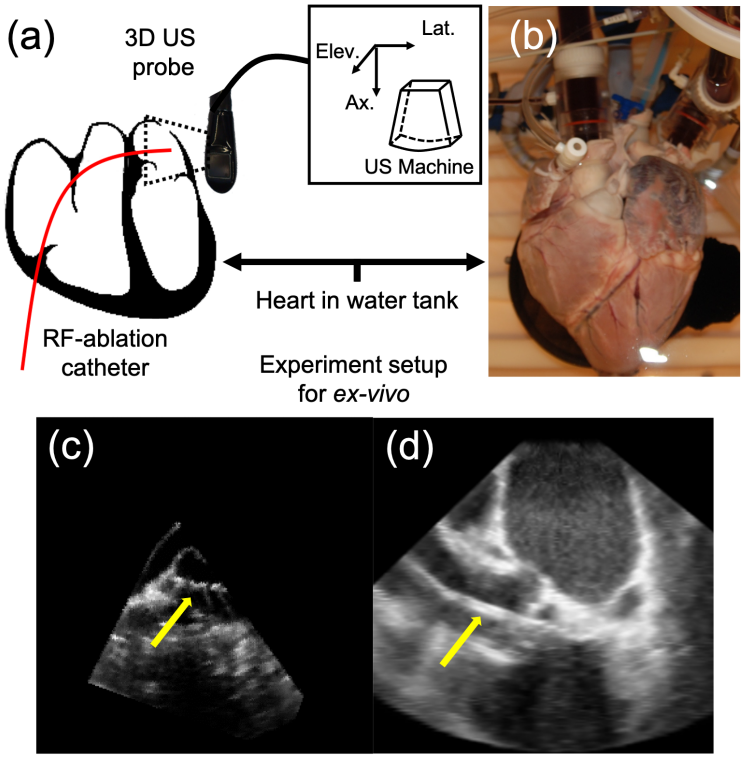


Figure 5.10 (a) Setup for *ex-vivo* 3D US dataset collection with an RF-ablation catheter. (b) Porcine heart placed in the water tank, and the US probe is placed under the heart while the catheter is going through the vein. (c) Example slice of *ex-vivo* recordings with the RF-ablation catheter. (d) Example slice of *in-vivo* recordings with the guidewire.

***In-vivo* TAVI guidewire dataset**: The collected *in-vivo* dataset includes eighteen volumes from two TAVI operations on humans. During the recording, the sonographer recorded images from different locations of the heart chamber without

any influence on the procedure. The volumes were recorded with a mean volume size of $201 \times 202 \times 302$, where the volume voxel size was resampled to 0.6 mm. The applied instrument in the *in-vivo* dataset is a guidewire (0.889 mm), having roughly one-third of the diameter of a catheter. The images were manually annotated in the same way as the *ex-vivo* dataset. An example image is shown in Fig. 5.10 (d).

5.4.2 Training Procedures

The proposed method has separate networks for coarse and fine segmentation tasks, thus the training procedures are individually described in the sequel. For the *ex-vivo* dataset, the images are randomly divided into 62/30 volumes for training/testing. Considering the limited data in the *in-vivo* dataset, a threefold cross-validation is applied with fine-tuning, which is based on the pre-trained *ex-vivo* model for the RF-ablation catheter.

A. Training for coarse segmentation: To train the slice-based UNet, each annotated instrument voxel in the ground truth is used as the center of the input sliced image, which introduces a translation invariance in a natural way to facilitate instrument segmentation. Non-instrument slices, i.e. slices using non-instrument voxels as the center point, are downsampled to the same size as the instrument voxels, to generate some images without instrument inside. The network is initialized based on a pre-trained VGG16 encoder, which is trained by the Adam optimizer with learning rate of 0.00001 using the Dice loss. The *ex-vivo* dataset is trained based on the above description, while the *in-vivo* dataset is trained with learning rate as 0.00001 for 2,000 iterations, based on the pre-trained *ex-vivo* model. During the training, rotation, mirroring and contrast transformations are applied on-the-fly to augment the amount of training images. Meanwhile, to learn the case that the instrument is crossing the slices, the slice sampling is randomly applied along elevation or lateral directions, following an orthogonal strategy. It is should be noticed that the downsampling strategy is only applied in the testing stage, to accelerate the inference efficiency and to avoid possible degradation of the information usage in the training phase.

B. Training for fine segmentation: The training patches are selected from instrument voxels [69], where an instrument voxel is used as the patch center. The Direction-fused UNet (DF-UNet) and the Pyramid-UNet are initially, separately trained by the input patches using the Adam optimizer with a mini-batch size of 4 and 8. More specifically, the learning rate for the DF-UNet is 0.0001 for transfer learning, while it is set to be 0.001 for the Pyramid-UNet to train from scratch. Each individual training is based on three epochs. Based on the pre-trained networks, i.e., DF-UNet and Pyramid-UNet, the feature-fusion part is then jointly trained with a learning rate of 0.00001 for one epoch, which finally generates the feature-fuse output. The network is trained first by a standard Dice loss function

to converge. Then, the parameters are fixed to globally learn the contextual encoder for 3,000 iterations, after which the whole networks are jointly trained by the proposed hybrid loss function until they jointly converge. The *ex-vivo* dataset is trained based on the above description, while the *in-vivo* dataset is trained with a learning rate of 0.00001 for 2,000 iterations based on the pre-trained *ex-vivo* model. This difference in training is caused by the limited training images for the *in-vivo* dataset. During the training, rotation, mirroring and contrast transformations are applied on-the-fly to augment the amount of training images.

5.4.3 Evaluation Metrics

To evaluate the performance of the proposed method, we consider the Dice score (DSC) and Hausdorff Distance (HD) as the evaluation metrics for different scenarios. As for coarse segmentation, DSC is used to evaluate the capabilities of the slice-based UNet, which means that for a higher DSC value a better POI selection can be achieved with fewer outliers. As for patch-based segmentation, DSC and HD are used to measure the network performances under different settings for the instrument segmentation task. Moreover, the average execution time for prediction is also considered for the framework comparison.

5.5 Experimental Results

In this section, we thoroughly validate the proposed POI-FuseNet with respect to accuracy and efficiency. Meanwhile, several ablation studies are also considered to validate the proposed components.

5.5.1 Ablation Studies

A. Coarse segmentation performance of the slice-based UNet: Several performance comparisons of the slice-based UNet are conducted in this section. First, the variations of the spatial gap d_1 between adjacent slices are validated from 0 to 5. Second, the variations of downsampling ratio K are tested, which is assigned to be 0.25, 0.5, and 1.0. The networks of the *ex-vivo* dataset are initialized based on the pre-trained VGG16 network for ImageNet, while the networks of the *in-vivo* dataset are initialized based on their corresponding *ex-vivo* models. The results are summarized by barplots in Fig. 5.11. Meanwhile, the inference efficiency for the downsampling ratio $K = 0.25$, 0.5 and 1.0 are about 0.2 sec., 0.6 sec., and 2.6 sec., respectively. These values are obtained by performing hybrid computations with both CPU and GPU, where the most time-consuming part is CPU-based slicing. It can be observed that a larger spatial gap d_1 provides a higher performance, which is because more spatial correlations are captured by the stride of slices. However, a too large stride d_1 may degrade the performance due to spatial decorrelation. For the downsampling ratio, $K = 0.5$ provides the best trade-off between efficiency and performance. Although $K = 0.25$ provides

a higher segmentation efficiency for 3D US, detailed spatial information is missing during the slicing, which therefore leads to unacceptable performance. As a consequence, we have experimentally selected hyperparameters $d_1 = 2$ and $K = 0.5$ for the coarse segmentation method, to select the patch-of-interest for further experiments. Because the length of the instrument inside 3D US volumes varies, the number of selected patches can range from 2 to 8, which is counted by automatically matching the coarse prediction to the pre-allocated patches in 3D US images. Based on statistical analysis of the results, the average number of selected patches by the coarse segmentation is about 5.

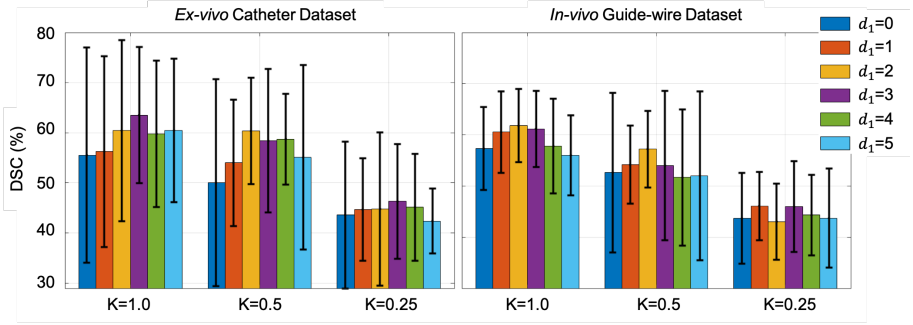


Figure 5.11 Barplots of DSC performances using the slice-based UNet for different settings of spatial stride d_1, downsampling ratio K, for *ex-vivo* and *in-vivo* datasets.

B. Ablation studies of proposed the DF-UNet: The ablation studies of the Direction-fused UNet (DF-UNet) on variations of the spatial stride d_2 between adjacent slices are validated from 0 to 5. Moreover, we also validate whether the DF-UNet achieves a higher performance than the UNet without DF, i.e. only a single branch of the three in Fig. 5.7 (a). The results are depicted in Fig. 5.12, which are trained with a standard Dice loss. The networks of the *ex-vivo* data are trained with an initialized pre-trained VGG16, since it provides a higher performance than when trained from scratch (w/o TL). Similar to the slice-based approach, the networks of the *in-vivo* dataset are initialized based on their corresponding *ex-vivo* models. From the results, the DF-UNet achieves a better performance than training from scratch or only considering a single direction by using the same d_2, which shows a more powerful capability of transferring the knowledge of the pre-trained *ex-vivo* dataset. Moreover, with a larger spatial gap, an improved 3D space information description can be achieved. As for the *in-vivo* dataset, spatial gap d_2 has less influence and variance on the performance than the results in the *ex-vivo* dataset, which is because of the lower image variation within the TAVI images. Furthermore, the proposed DF-UNet is consistently better than training from scratch and single-direction fusion in both datasets. Based on the results, the spatial stride is experimentally selected as

$d_2 = 3$ for further feature fusion.

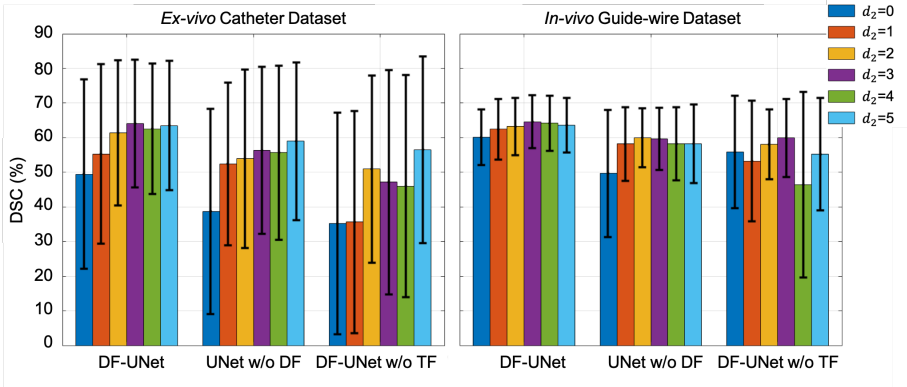


Figure 5.12 Barplots of DSC performances for different values of the spatial stride d_2 for the Direction-fused UNet (DF-UNet), UNet without DF (w/o DF) and DF-UNet without transfer learning (w/o TL) in *ex-vivo* dataset. Results of the DF-UNet, UNet w/o DF and DF-UNet w/o TL in the *in-vivo* dataset are also reported.

C. Ablation studies of the proposed Pyramid-UNet: Specifically, to validate the effectiveness of the components of the Pyramid-UNet, the following configurations are listed.

(1) Compact-UNet trained by the Dice loss (3D Pyramid-UNet without multiple input/output).
(2) Atrous Spatial Pyramid Pooling (ASPPv1) with dilation rate {1,2,4,8} based on the encoder of the proposed Compact-UNet under guidance of Deeplab v3+ [75], which is also trained with the Dice loss.
(3) Using two 3D convolution layers from the Compact-UNet, the ASPP operation is directly applied on the feature maps at original image resolution, which means that the ASPP is applied at low-level features directly (input layers). The adopted dilation rate is the same as in ASPPv1 and is trained with the Dice loss. The obtained model is denoted as ASPPv2.
(4) The proposed 3D Pyramid-UNet trained with Dice loss.

The results are summarized in Table 5.1. The basic backbone Compact-UNet provides the baseline performance for the ablation studies. Based on this architecture, a pyramid input-output structure is introduced to exploit multi-scale features, which improves the segmentation performance with better discriminating information extraction. However, this multiple input and output branches introduce more computation costs. As shown in Fig. 5.13, Compact-UNet generates a more noisy feature map where more outliers and blurry boundaries are

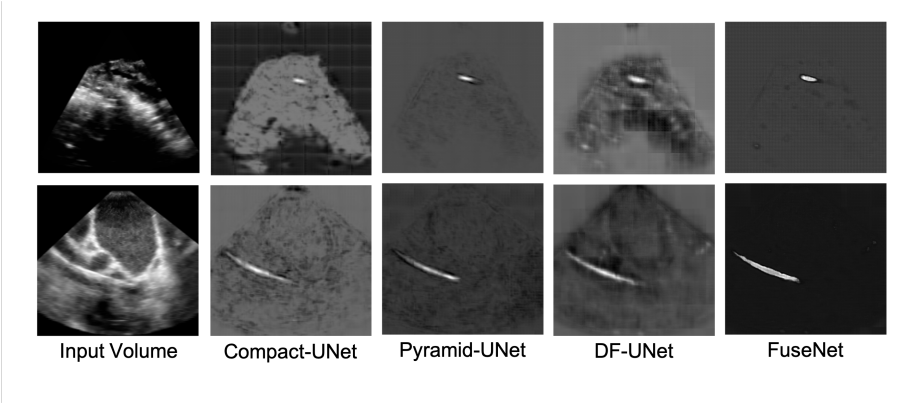


Figure 5.13 Example slices of selected feature maps from the Compact-UNet, Pyramid-UNet, DF-UNet and FuseNet. Images are re-scaled into the same intensity range for visualization purposes. The top row contains feature maps from the *ex-vivo* dataset, while the bottom row shows feature maps from the *in-vivo* dataset.

obtained than the Pyramid-UNet. Compared to ASPP networks, the proposed Pyramid-UNet has a better performance than both ASPP structures from our case. This performance difference occurs because the Pyramid-UNet structure exploits richer complex feature relationships at different image scales in both input and output branches. Initially, ASPP has been proposed after a complex and proper encoder for images, such as a VGG or ResNet encoder [43]. However, in our approach, the compact network encoder leads to less complex and discriminating feature maps for ASPP, which cannot represent sufficient information for further steps.

Table 5.1 Ablation studies of the proposed Pyramid-UNet, measured by the Dice score (DSC), Hausdorff Distance (HD) (expressed as mean±std), and average prediction times.

Method	RF-ablation Catheter *ex-vivo*		
	DSC (%)	HD (voxels)	Time (sec.)
Compact-UNet w. Dice	62.2±20.0	13.3±15.6	∼10
ASPPv1 w. Dice	63.8±16.7	11.8±18.6	∼11
ASPPv2 w. Dice	57.8±21.7	14.8±19.3	∼17
Pyramid-UNet w. Dice	65.8±18.9	11.3±13.8	∼11
Method	TAVI Guidewire *in-vivo*		
Compact-UNet w. Dice	63.8±9.2	9.8±5.5	∼12
ASPPv1 w. Dice	63.6±9.0	8.9±4.6	∼13
ASPPv2 w. Dice	62.6±9.5	8.5±3.4	∼19
Pyramid-UNet w. Dice	64.5±8.3	8.8±3.2	∼12

D. Ablation studies of the proposed FuseNet: Moreover, ablation studies on POI-FuseNet are also performed to validate the effectiveness of its components, which are gradually introduced to discuss the design of the proposed method. Besides these, also the training steps will be explained. The main steps are as follows.

(1) The Direction-fused UNet trained by the Dice loss (DF-UNet) is based on the configurations in the above sections. (2) The Pyramid-UNet is also trained with the Dice loss. (3) An EnsembleNet is trained with the Dice loss, which is our proposed FuseNet without feature fusion layer, which is instead replaced by the averaged output from two individually trained networks (DF-UNet and Pyramid-UNet) without joint training. (4) The FuseNet is pre-trained with the Dice loss. (5) Then, the FuseNet is trained with the Contextual loss (CL) under the guidance of the Dice loss. Specifically, we fail to obtain the result solely using CL without Dice loss, which shows the CL is a kind of compensation of pixel-level loss in high-level space. (6) The FuseNet is trained with the Focal loss (FL). (7) The FuseNet is trained with the Hybrid loss (HL), which combines the FL and CL.

The results are shown in Table 5.2. From the results, the DF-UNet and Pyramid-UNet obtain similar performances, but with different architecture and information-extraction steps. More specifically, the Pyramid-UNet directly extracts the 3D information from 3D space, which exploits semantic information in a straightforward manner. However, this network may not fully exploit information with limited training samples and complex 3D space. In contrast, the DF-UNet decomposes the 3D information into 2D slices with tensor operations, which process the 3D semantic information in a different way than the Pyramid-UNet. With this intra-slice feature extraction strategy, the DF-UNet could better exploit semantic information within the slices at the high-level feature space and combine them by the semi-3D operation. Nevertheless, the DF-UNet is rather time-consuming, due to the complex 2D-3D transformations in the network design. Furthermore, the DF-UNet cannot exploit 3D information due to its design nature. By integrating these two networks with the feature-level fusion, the proposed FuseNet achieves better performance than each individual network. From the demonstration in Fig. 5.13, fused features from two subnetworks compress backgrounds, such as tissue and chambers, while improving the confidence of the instrument-related voxels. More crucially, it is also better than a naive ensemble without feature fusion, which directly averages the predictions from two individual networks. Nevertheless, the overall prediction time is drastically increased, due to the dual network-based integration on a single GPU. Compared to a FuseNet that is trained by a standard Dice loss, the proposed hybrid loss can further improve the segmentation performance in both datasets. More specifically, when compared to FuseNet with Dice loss, both contextual loss (CL) and focal loss (FL) can improve the segmentation results in different aspects. As for CL, it encourages the contextual-level consistency between the prediction and

Table 5.2 Results of ablation studies of the proposed FuseNet, measured by the Dice Score (DSC), Hausdorff Distance (HD), which are expressed by their mean±std. The average execution time is also measured in seconds (w. means 'with').

Method	RF-ablation Catheter *ex-vivo*		
	DSC (%)	HD (voxels)	Time (sec.)
DF-UNet w. Dice	64.2±18.4	11.2±13.8	∼28
Pyramid-UNet w. Dice	65.8±18.9	11.3±13.8	∼11
EnsembleNet w. Dice	64.1±18.4	11.2±13.8	∼39
FuseNet w. Dice	67.7±15.9	10.1±13.0	∼41
FuseNet w. CL	68.9±14.7	8.8±10.2	∼41
FuseNet w. FL	69.1±11.1	9.0±9.1	∼41
FuseNet w. HL	**70.5±9.2**	**7.3±3.9**	∼41
Method	TAVI Guidewire *in-vivo*		
DF-UNet w. Dice	64.1±7.9	8.2±2.9	∼32
Pyramid-UNet w. Dice	64.5±8.3	8.8±3.2	∼12
EnsembleNet w. Dice	63.2±8.3	8.7±3.3	∼44
FuseNet w. Dice	65.0±8.3	8.3±3.2	∼47
FuseNet w. CL	65.8±8.2	8.1±3.0	∼47
FuseNet w. FL	65.9±8.0	8.0±3.0	∼47
FuseNet w. HL	**66.5±7.5**	**8.2±2.9**	∼47

annotation in high-level latent space. In contrast, the FL addresses the extremely imbalanced class distributions (instrument voxels are about 1% of the patch voxels) and focuses on the hard-classified voxels in 3D space, which leads to higher performance. Based on these two losses, the proposed hybrid loss (HL) achieves a higher performance than the individual CL and FL, and it also provides a lower standard deviation. It is worth to mention that the HL with contextual encoder is only considered in the training stage, such that the extra prediction complexity is not introduced during the testing phase. By comparing *ex-vivo* and *in-vivo* datasets, the hybrid loss has more influence on the *ex-vivo* data, which is explained by a larger image variance within the dataset. Since the performance of the network is improved by feature fusion, the prediction time is also increased, because of the augmented complexity of the proposed network architecture. All the GPU-based computations are measured on a GTX 1080Ti GPU using Python 3.6 with TensorFlow 1.10.

E. Ablation studies of the proposed POI-FuseNet: The results of the ablation studies of the proposed patch-of-interest (POI) strategy are presented in Table 5.3. Specifically, three different K values are validated.

The first validation involves the proposed POI-FuseNet trained with HL with downsampling ratio 0.25 (POI-FuseNet w. HL, K=0.25). Second, the proposed POI-FuseNet is validated with the setting but with downsampling ratio 0.5 (POI-

Table 5.3 Results of ablation studies of the proposed POI-FuseNet, measured by the Dice Score (DSC), which are expressed by their mean±std. The average execution time is also measured in seconds. The best settings are indicated in bold symbols (w. means 'with').

Method	RF-ablation Catheter *ex-vivo*		
	DSC (%)	HD (voxels)	Time (sec.)
FuseNet w. HL	70.5±9.2	7.3±3.9	∼41
POI-FuseNet w. HL, K=0.25	68.8±11.1	9.1±8.3	∼0.8
POI-FuseNet w. HL, K=0.5	**70.5±9.2**	**7.3±4.1**	**∼1.3**
POI-FuseNet w. HL, K=1.0	70.5±9.2	7.5±4.1	∼3.3
Method	TAVI Guidewire *in-vivo*		
FuseNet w. HL	66.5±7.5	8.2±2.9	∼47
POI-FuseNet w. HL, K=0.25	65.8±7.9	8.1±3.1	∼0.9
POI-FuseNet w. HL, K=0.5	**66.0±8.3**	**8.2±2.9**	**∼1.4**
POI-FuseNet w. HL, K=1.0	66.0±7.9	8.0±3.0	∼3.4

FuseNet w. HL, K=0.5). The third validation is same as the above, but without downsampling (POI-FuseNet w. HL, K=1.0).

With the introduction of the POI selection by coarse segmentation, the total prediction time per volume is drastically accelerated because the iteratively patch-based prediction is avoided. More specifically, when the downsampling ratio of slice-based UNet is increasing, the performance of POI-FuseNet is improved at the expense of time efficiency. There is only a small difference in performance between $K = 0.5$ and $K = 1.0$, while the performance is degraded using $K = 0.25$ in the *ex-vivo* dataset. This is because some catheters are partly missing from the slices due to the larger downsampling value, and therefore only a part of the catheter can be segmented. As a consequence, a higher K value provides a better generalization for the POI-FuseNet. Although the coarse segmentations have different performances due to the different K values, the final prediction performances of POI-FuseNet show not much differences in the end. Finally, the proposed FuseNet is trained on the *in-vivo* dataset by fine-tuning the parameters of the *ex-vivo* model, which achieves a 3-5% higher Dice score than training from scratch. Nevertheless, the overall framework of the proposed POI-FuseNet cannot be trained in an end-to-end style, which is limited by the coarse-to-fine strategy. This limitation is due to the large memory capacity requirement for complex 3D US images and limited training datasets, when compared to the state-of-the-art object segmentation methods in the current computer vision field. As a result, the feature maps at the full-image level cannot be used for POI purpose, such as in Mask R-CNN [76].

F. Ablation studies of different patch processing: Besides the above ablation studies in network configurations, we have also validated the proposed patch processing strategy, which considers zero-padding during the convolutional operations. This means that with a fixed patch size, i.e., $D + S = 48$, the parame-

ter D indicates a non-overlapping patch size and the parameter S is an extending parameter to compensate for the information leakage. The total patch size is empirically chosen to trade-off the network complexity, size of the instrument in 3D US and prediction efficiency [69]. As shown in Table 5.4, we have experimentally compared different combinations of D and S in terms of DSC and HD to validate the significant influence of these parameters. Based on the observations on the results in the table, a patch with $S = 0$ generates the worst segmentation result w.r.t full-volume cases, which is because the patch boundaries are affected by padding operations during the convolution operation. Even the skipping connections are applied to compensate the information leakage. In contrast, for a setting with $S = 32$, it provides slightly better performance than the case of $S = 16$, while significantly degrading the prediction efficiency (even compared to the POI-based condition, which is at least two times longer). As a conclusion, a proper combination strategy, i.e., (D, S), of patch-based segmentation should be considered, to provide stable results for semantic segmentation ($D = 32$, $S = 16$ in our case).

Table 5.4 Segmentation performances for different combinations of parameters D and S, measured by the Dice score (DSC) and Hausdorff Distance (HD), which are expressed by their mean±std. The (D, S) pair indicates the combination of different values of $D + S = 48$. The corresponding average execution times are also measured in seconds. Bold values denote the setting reported in this chapter, providing a good trade-off.

(D, S)	RF-ablation Catheter *ex-vivo*		
	DSC (%)	HD (voxel)	Time (sec.)
(16,32)	70.7±9.0	7.2±4.2	∼288
(32,16)	**70.5±9.3**	**7.3±3.9**	**∼41**
(48,0)	65.3±13.7	10.8±11.1	∼13
(D, S)	TAVI Guidewire *in-vivo*		
(16,32)	67.0±7.7	8.1±3.2	∼443
(32,16)	**66.5±7.5**	**8.2±2.9**	**∼47**
(48,0)	65.2±8.4	8.8±3.5	∼15

5.5.2 Performance Comparison with Learning-based Methods

We have compared the proposed method with the learning-based state-of-the-art instrument segmentation methods on our challenging datasets, such as Gabor feature extraction with the SVM classifier (GF-SVM) [28], multi-definition and multi-scale feature fusion with the AdaBoost classifier (MF-AdaB) [65], orthogonal slice-based ShareFCN [66], Voxel-of-interest-based CNN (VOI-CNN) for voxel-based classification [19]. Furthermore, other US-based segmentation methods, like Prenatal-UNet [61] and Compact-UNet with Anatomically Constrained neural network [77], i.e. ACNN, are also considered. The results are shown in Table 5.5, where DSC is the Dice score and HD is the Hausdorff Dis-

Table 5.5 Segmentation performances for different methods, measured by the Dice score (DSC), Hausdorff Distance (HD), which are expressed by their mean±std. The average execution time is measured in seconds. All the methods are validated on our datasets (symbol '-' means could not be calculated due to memory limit).

Method	RF-ablation Catheter *ex-vivo*		
	DSC (%)	HD (voxels)	Time (sec.)
GF-SVM [28]	3.3±8.5	-	>100
MF-AdaB [65]	36.5±19.0	19.1± 8.5	>100
ShareFCN [66]	52.8±21.0	15.6±16.7	∼3
VOI-CNN [19]	58.5±10.7	11.5± 7.7	∼10
Prenatal-UNet [61]	24.6±24.9	38.3±22.3	∼2
ACNN [77]	61.0±18.3	14.6±13.9	∼10
FuseNet	70.5± 9.3	7.3 ± 3.9	∼41
POI-FuseNet	**70.5**± 9.2	**7.3** ± 4.1	∼**1.3**
Method	TAVI Guidewire *in-vivo*		
	DSC (%)	HD (voxels)	Time (sec.)
GF-SVM [28]	1.0±1.7	-	>100
MF-AdaB [65]	37.6±23.3	23.9±18.2	>100
ShareFCN [66]	55.9±12.1	11.6± 7.8	∼4
VOI-CNN [19]	58.6± 7.9	11.0± 5.1	∼11
Prenatal-UNet [61]	53.2±14.7	18.8±11.1	∼2
ACNN [77]	63.9± 9.7	8.6 ± 4.7	∼12
FuseNet	66.5± 7.5	8.2 ± 2.9	∼47
POI-FuseNet	**66.0**± 8.3	**8.2** ± 2.9	∼**1.4**

tance. Meanwhile, some example results are visualized in Fig. 5.14, which qualitatively demonstrates the results of voxel-based classification, slice-based segmentation and patch-based segmentation on our datasets.

From the table, several observations and conclusions can be made, which we have clustered into five topics.

(1) *Comparison with handcrafted features*: As for handcrafted features with voxel-based classification, the performances are worse than with deep learning methods, which are due to the limited 3D information representation after feature extraction, especially for the Gabor filterbank method. First, it is a single-scale feature, which means it focuses on a specific spatial resolution. However, this extraction cannot handle our dataset with different spatial resolutions and instrument diameters. Second, the Gabor feature mainly focuses on boundary contrast at the edge in a homogeneous or semi-homogeneous background, which works properly in needle segmentation for anesthesia by 3D US. Nevertheless, in our case, the boundary of a catheter can be blurred with a lower contrast

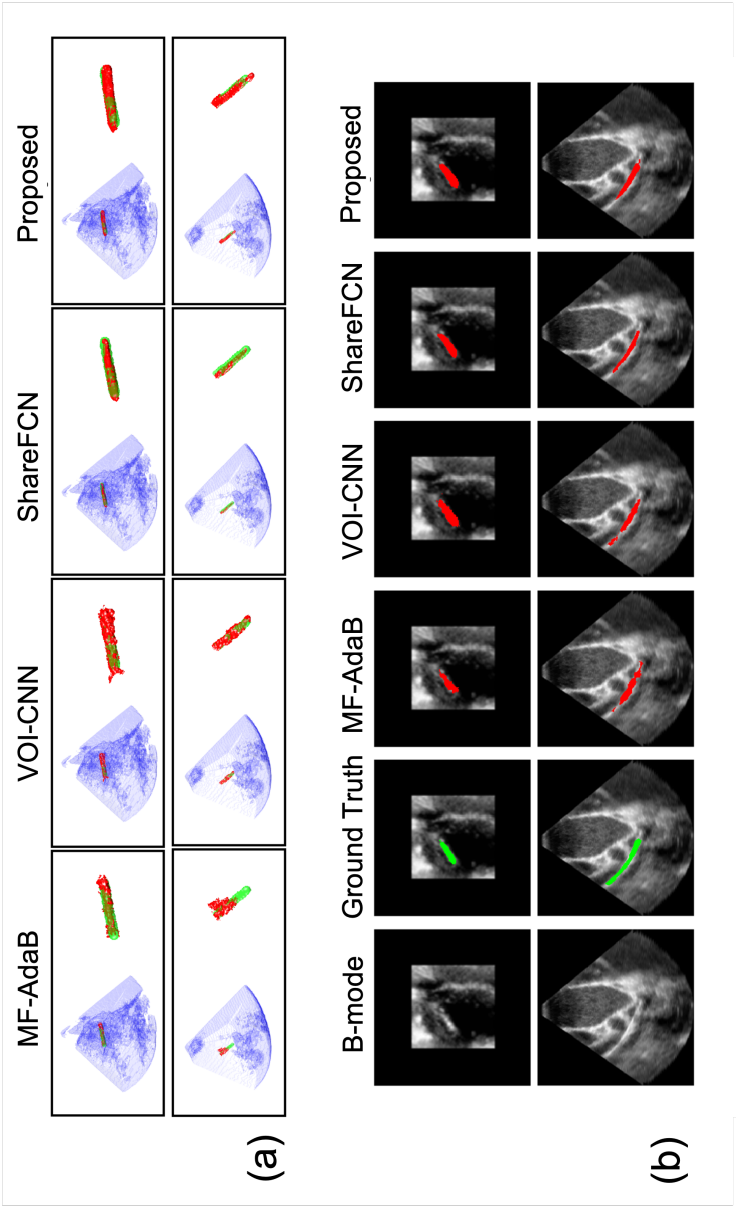


Figure 5.14 Instrument segmentation results of different methods on example US volumes. Top row: *ex-vivo*, bottom row: *in-vivo*. (a) Segmentation results displayed in 3D US volumes, where the 3D US volume is blue, the ground truth of the instrument is green, and the instrument segmentation is red. (b) Segmentation results displayed in 2D slices with corresponding original image, ground truth and segmentations. Green: ground truth. Red: segmentation. All the 2D images are corresponding to the above 3D images.

with anatomical background, which leads to significant difficulties for using a single-scale Gabor feature with a linear support vector machine. Third, the experimental settings are clearly different. The original paper applied a leave-one-out cross-validation method with an image-specific threshold to obtain the highest performance. However, in our case, we applied a dataset-level threshold, which leads to a much lower performance than the originally reported value. In contrast, the multi-scale approach with multi-definition features achieves much higher performance to about 36% DSC. Because of the multi-scale and definition features, more instrument-related information can be described from different viewpoints. Meanwhile, the non-linear Adaptive Boosting classifier provides a non-linear decision boundary which is more advanced than a simple LSVM, so that a higher performance is obtained. However, handcrafted feature design is strongly relying on experience and instrument model estimation, which therefore limits the information usage from training images. As a comparison, the proposed POI-FuseNet can extract much more discriminating information by considering a data-driven method with task-specific network design.

(2) *Comparison to VOI-CNN:* Compared to VOI-CNN, which is based on a pre-filtered voxel selection, the proposed patch-based semantic segmentation method portrays an improved performance. This is because the VOI-CNN method can degrade the true positive rate by introducing an imperfect voxel selection, whereas the POI-based method can overcome this degradation. More specifically for VOI-selection, the interested voxels are obtained from the remaining voxels after Frangi filtering, which cannot fully preserve the instrument voxels by a simple filtering, even when a CNN follows afterwards to classify the voxels' category [19]. In contrast, the POI-based approach selects the possible regions from the 3D volume using patches, which contains full-instrument-related voxels for a subsequent segmentation by the FuseNet. As a consequence, the instrument voxels are re-calculated by a more accurate CNN with a higher accuracy, yielding also a more accurate instrument segmentation.

(3) *UNet structure:* Compared to a more complex and generalized UNet [61], the proposed Compact-UNet achieves better performance, because of a smaller input size, simpler architecture, and task-specific design. Although the Prenatal-UNet achieves a fast prediction time as a result of larger patch size, i.e. 64^3 voxels, it is much more difficult to be trained for our instrument segmentation task, since the instrument occupies a small volume in a large 3D patch space. From Table 5.5 it follows that the proposed POI-FuseNet obtains a higher accuracy and efficiency than the Prenatal-UNet.

(4) *Contextual description:* When compared to the ACNN employing anatomical constrained knowledge to formulate the contextual information, the proposed method achieves a clearly increased performance. This is explained by the design of the ACNN, which includes a fixed pre-trained shape description

encoder for ground truth. This is not suitable for prediction with large variations in location and intensity values. Actually, considering the design of the ACNN, it is preferably used for anatomical structure segmentation with a fixed global location and size, which is easily learned and described by an Auto-Encoder. However, in our case, the instrument is located at any position of the input patches and within the patches. Moreover, the prediction of the input patch is ranging from zero to unity, instead of a fixed integer value. This variation leads to an encoder in ACNN that cannot perfectly represent the contextual information difference between the ground truth and the prediction. As a consequence, the ACNN cannot improve performance when compared with the proposed jointly trained approach. The jointly training procedure enables to adaptively learn the contextual information with varying instrument locations and intensity values.

(5) *Information exploitation in 2D and 3D:* Patch-based semantic segmentation approaches, i.e., Pyramid-UNet and DF-UNet, achieve a higher performance than SOTAs because of the richer spatial information usage. Moreover, the DF-UNet obtains better performance than the Compact-UNet, since the parameters are initialized from the pre-trained model, which is shown in ablation studies. With extracted feature maps from two independent networks, the FuseNet obtains more accurate segmentation results. This accuracy is explained by better hierarchical exploitation of contextual information among voxels.

From qualitative illustrations in Fig. 5.14, voxel-based classification methods, i.e. MF-AdB and VOI-CNN, generate a non-smooth surface, which results from voxel-by-voxel classification with a limited field-of-view. As for slice-based segmentation, i.e. ShareFCN, it generates a smoother surface and boundary, caused by the semantic information usage. However, when compared to the proposed patch-based segmentation, i.e. POI-FuseNet, the slice-based method has lower performance due to the limited and degraded spatial information compared to real 3D space.

In terms of the prediction efficiency, the proposed POI-FuseNet achieves a prediction efficiency of $\sim$1.3 seconds per volume, which includes $\sim$0.6 sec. POI pre-processing and 0.7 secs. patch-based refining segmentation. Because the patch-based segmentation re-calculates the voxels' category, the proposed method does not hamper the segmentation accuracy when compared to the VOI-based CNN [19]. From our experiments, voxel-based methods, such as GF-SVM, MF-AdaB, or LateCNN, consume more than 100 seconds on our computer platform, which is due to the iterative voxel-by-voxel calculations. As for patch-based methods on a full volume, they consume about 10-50 seconds to obtain the final prediction (based on the architecture of the networks), which is still far from real-time implementation. In contrast, the proposed POI-FuseNet preserves the segmentation accuracy by the re-calculation of voxels, but also improves the seg-

mentation efficiency. All the GPU-based methods are measured on a GTX-1080Ti GPU using Python 3.6 with TensorFlow 1.10.

It is worth to mention that a recent publication [78] has proposed to segment the needle from full-volume data. However, with our challenging dataset, we have failed to obtain a successful segmentation result, which may be explained by the much more simple network architecture and the more challenging task for cardiac US imaging. Moreover, when compared to our preliminary study [69], which applies morphology operations to connect the closest component and is therefore time-consuming for post-processing, the proposed POI-FuseNet omits this complicated post-processing, so that it is more efficient and robust. Otherwise, the morphology operations in 3D space would take more than 10 seconds of processing time for each data volume.

5.5.3 Performance Comparison with Non-learning-based Methods

Besides the comparisons to learning-based methods, we also compare the proposed method with non-learning-based methods, in particular Principal Component Analysis on thresholded 3D US (PCA) [62], Parallel Integral Projection (PIP) [16], Random Hough Transformation (RHT) [63] and line-filter-based RANSAC (Line-RANSAC) algorithm [20]. Since the above methods are using different approaches than direct segmentation, we adopt other metrics for comparison. Instead of testing volume using DSC and HD, we use success volume detection and the detection error as the metrics. These metrics represent success rate and the endpoint error of the instrument detection in terms of voxels. More specifically, the endpoint error is defined as the average distance of two endpoints of the instrument skeleton from the ground truth, i.e. tip and tail point, to the instrument axis obtained from the detection.

The experimental results are shown in Table 5.6. From the table, the proposed deep learning-based method shows 100% success rate with the lowest axis error, while traditional computer vision techniques have less detection rate with higher errors. The reasons are explained as follows. First, non-learning methods segment the images with simple and straightforward thresholding approaches (PIP does not apply thresholding). These approaches cannot extract accurate instrument-related voxels and omit background voxels, i.e. voxels from tissue and heart chambers. Second, non-learning-based methods are focusing on post-processing to localize the instrument, which heavily relies on the assumption that the background is not complex, while the instrument has a higher intensity distribution than the background. This also explains why these methods achieve promising results on simulated or phantom images. However, real tissue-based images for cardiac US are much more challenging for non-learning-based methods, which therefore obtain a much lower success rate. Third, it is worthwhile to discuss the PIP method, which relies on parallel intensity projection for thin instrument detection. As can be observed, it can be considered as a failure. This

occurs because the cardiac instrument has a similar intensity distribution as heart tissue, which is visible in Fig. 5.10. More crucially, the heart tissue occupies much more space than the instrument, such that the PIP method automatically converges to the direction with tissue passing through the estimated instrument axes, such as the heart wall of Fig. 5.10 (d). The experiments with non-learning-based methods further demonstrate the importance of the segmentation stage, which promises a robust and accurate instrument detection. It is important to mention that with a thicker instrument, it is much easier to detect the catheter than a guide-wire from 3D US data. Moreover, for catheter with a different diameter, a thicker tube would be easier to detect by a non-learning-based method, such as Line-RANSAC or RHT. However, the complex deep learning method can provide a more generalized result with sufficient training images.

Table 5.6 Detection performances for different non-learning methods, measured by success rate and average endpoint error (EE) in voxels (std. is excluded since we counted based on successful detection). All methods are validated on the earlier applied datasets.

Method	RF-ablation Catheter *ex-vivo*	
	Success rate (%)	EE (voxels)
PCA [62]	23.3%	3.7
RHT [63]	80%	9.6
PIP [16]	3.3%	13.2
Line-RANSAC [20]	76.7%	4.0
Proposed	100%	1.8
Method	TAVI Guidewire *in-vivo*	
PCA [62]	27.8%	4.3
RHT [63]	44.4%	12.6
PIP [16]	0%	-
Line-RANSAC [20]	38.9%	3.7
Proposed	100%	2.9

5.6 Discussion and Conclusions

This chapter has proposed an efficient and accurate coarse-to-fine instrument semantic segmentation method for 3D cardiac US images with high efficiency and accuracy. The proposed method is characterized by two key neural networks to follow a coarse-to-fine strategy. The proposed method solves two key challenges for medical instrument segmentation in 3D US: (1) efficient and accurate coarse region-of-interest selection, and (2) robust and accurate fine semantic segmentation. For the first challenge, a slice-based UNet is proposed by combining it with spatial downsampling, which efficiently extracts the regions containing instrument voxels. As for the second challenge, a robust and accurate semantic segmentation network is proposed, which exploits the contextual

voxel information for high segmentation performance. Based on the proposed method, the overall performance is much higher than conventional handcrafted techniques and a CNN-based classification method. In addition, the overall computational efficiency is drastically improved to about 1 second per volume. Therefore, a near real-time performance is achieved, so that the algorithm is more acceptable for clinical applications.

Discussion: Some aspects of our method still need further discussion and argumentation.
(1) To train the network, a voxel-level annotation is required, which is extremely challenging and laborious for low-quality 3D ultrasound imaging. Therefore, it is difficult to design a system for clinical usage that is based on large-scale image-based supervised learning using annotations.
(2) For each 3D US volume containing the instrument, the instrument should be visible in the B-mode US image for achieving a successful segmentation. However, this visibility is not always ensured during real clinical usage. Since the relative pose between the US probe and the instrument changes, the instrument may become invisible because of acoustic reflection. This aspect is quite fundamental and seriously hampers the robustness of the proposed method.
(3) The employed patch-based method in the second stage cannot fully exploit the semantic information within the whole image context, which is due to the limited GPU processing power. Therefore, a carefully designed FCN can lead to a higher performance with better exploration of contextual information.
(4) With the validation on an *in-vivo* dataset, the proposed method presents a significant value for clinical applications. Nevertheless, further study on extended *in-vivo* data is required to support further evaluation, because of the limited amount of covered volumes and patients. Also, the proposed method is only validated for two applications, i.e. a limited dataset for RF-ablation procedure simulation and static images from TAVI operations. Therefore, its stability and generalization, such as whether it can be used in other cardiac operations or US-guided interventions, still needs to be validated in future work.

Conclusion: The proposed method contains the following notable contributions. (1) A patch-based framework is applied for instrument segmentation in 3D US, which reduces the computation complexity and maintains the segmentation performance for the challenging segmentation task. The proposed framework is based on a patch-of-interest selector, which can efficiently select the most interesting regions in 3D US, thereby improving the segmentation speed for real-time applications. (2) The proposed FuseNet combines multi-defined features from 2.5D and 3D feature extraction, which improves the segmentation in complex 3D US volumetric data. With the proposed feature extraction networks, the FuseNet can extract direction-fused features and full 3D spatial features, which leads to better information usage than solely considering a 3D UNet.

(3) A hybrid loss function is proposed to guide the networks to simultaneously learn discriminative information at the pixel level and image level. This approach therefore improves the segmentation performances. The results of extensive validations experiments performed with the proposed method achieve a segmentation performance of about 70% Dice score and approximately 1-sec. execution time per volume.

In the next chapter, a novel coarse detection method and fine segmentation network will be investigated. This method can reduce the annotation effort while preserving the segmentation accuracy. Specifically, a reinforcement learning method is considered to coarsely detect the instrument in 3D US images, which is then segmented by a semi-supervised learning trained CNN network. This framework preserves the performance and efficiency, while drastically reducing the annotation effort for the deep learning training.

CHAPTER 6

Annotation Efficient Instrument Semantic Segmentation

6.1 Introduction

The previous chapter has introduced a semantic segmentation framework to detect medical instruments in 3D US, which is based on a patch-based coarse-to-fine strategy. In the first stage, the instrument region is coarsely localized based on the global segmentation results, after which the fine segmentation is performed in the instrument region. These approaches are all trained by using fully supervised learning. Therefore, the overall segmentation performance heavily relies on the annotation accuracy. Nevertheless, it is challenging to train CNN networks on a large-scale image dataset with carefully annotated ground truth. This challenge lies mainly in the effort for annotation, which is laborious and time-consuming. As a solution for this challenge, this chapter aims at addressing the training of deep learning networks with a more efficient approach for annotation.

6.1.1 Objective and Brief System Outline

The objective of this chapter is to develop an automated medical instrument segmentation method with only small annotation effort. This method involves a fast but coarse region-of-interest selection and a fine semantic segmentation, which largely follows the strategy of Chapter 5. Nevertheless, in contrast with the previous chapter, both stages of coarse detection and fine segmentation networks are trained with much less annotation effort than conventional fully supervised approaches. This is possible by adopting a method that consists of the same two key steps, as is depicted in Fig. 6.1, but now the implementation is completely different in the way that the networks are learned.

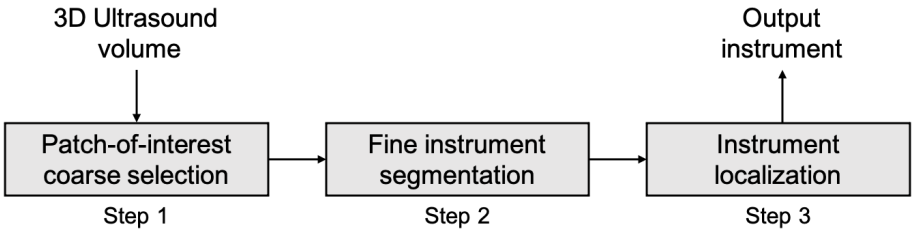


Figure 6.1 Diagram of a coarse-to-fine instrument localization and segmentation system. The diagram has a similar structure as in the previous chapter, but the training methods are different, and also the localization in the first stage is modified.

Based on the above approach, the key challenge is to design a coarse localization and accurate segmentation methods without large annotation effort, like following e.g. a semi-supervised approach. More detailed challenges and corresponding solution directions are elaborated below.

6.1.2 Challenges for Annotation Efficient Coarse-to-fine Segmentation

In order to coarsely select the interested region with high efficiency from complex input US data, Chapter 5 has proposed a slice-UNet for coarse segmentation. In addition, based on the coarsely selected region, a fine segmentation network has been proposed, which segments the selected regions by exploiting the fusion of multi-defined features from 2.5D and 3D feature extraction. However, this method requires careful voxel-level annotation to train the CNNs, which is expensive and laborious to obtain.

To address this limitation, network design in both coarse and fine stages are re-considered to work with a reduced annotation effort.

- *Coarse localization:* To coarsely localize the instrument and to train coarse segmentation, careful voxel-level annotation is required. With little annotation effort and localizing the instrument efficiently, a region-based annotation can be used, e.g. by indicating an instrument's center point. In this way, the network still can efficiently search the target location in the complex 3D US data, so that the laborious voxel-level annotation for all the training images can be avoided. For instance, reinforcement learning (RL) is a solution for learning the surroundings of the center point to contribute the coarse instrument localization.

- *Modified learning for fine segmentation:* With limited annotation available for fine segmentation, we aim at using all the US data available from the experiments, and using the coarse localization information. This can be achieved by employing a training method that learns from the selected surroundings of the coarse instrument localization, yet only with limited annotation

available. For example, semi-supervised learning (SSL) is a method that adopts unlabeled data for information exploration.

Based on the above challenges and directions, the followed approach of this chapter is to exploit the RL and SSL training techniques for the primary two steps for coarse and refined segmentation. The RL is most suited for coarse instrument localization, because it can work on regions where the instrument is considered to be present. Accordingly, the SSL training technique is attractive for the refinement segmentation, which is carried out as the second step.

The sequel of this chapter is organized in the following way. Section 6.2 summarizes the related work in this field, detailing coarse detections and SSL-based methods. Section 6.3 describes the proposed method, including every step of the coarse-to-fine segmentation and the integration of the proposed learning techniques. Sections 6.4 and 6.5 demonstrate and present the considered dataset, implementation details and experimental results. Finally, Section 6.6 concludes the chapter and presents some discussions on possible refinements.

6.2 Related Work

6.2.1 Coarse Detection

A coarse-to-fine strategy normally considers to first locate the instrument region and then perform fine segmentation. Detection networks can be employed to detect and localize the instrument region in the image. Typical examples for such networks are the Single Shot Multibox Detector (SSD) [79] and the Faster R-CNN [80], which are trained with available bounding-box annotations. These CNNs need to be trained on sufficient data, since they include regression networks with a complex architecture. In our task, because only limited 3D US images are available for training, the detection network cannot offer the desired performance.

Another approach to object detection in US is the reinforcement learning (RL) method for landmark detection [81], which iteratively finds the targets (called landmarks) in 3D images by a sliding window. Compared to the detection networks, which learn the discriminative information of the target by supervised learning, the RL method models agent-environment interaction to reach the task solution. Using a Deep Q-network (DQN), the RL method has been applied to several medical imaging applications with promising results [81, 82, 83]. The DQN models the movement prediction into a discrete value rather than continuous bounding boxes, which is easier to be trained with limited training images. In addition, the DQN employs a simplified decision CNN, which is less senstive to overfitting than detection networks.

6.2.2 Semi-supervised Learning

To reduce the annotation efforts for CNN training and leverage abundant unlabeled images, semi-supervised learning (SSL) methods [84, 85, 86, 87, 88, 89] have been studied for medical image segmentation. The most popular SSL methods follow a consistency-enforcing strategy [90, 91], which leverages the unlabeled data by constraining the network predictions to be consistent under perturbations of input or network parameters. A typical example is the student-teacher model, which is an implementation of a knowledge-distillation strategy [92]. Specifically, the teacher-student model has been proposed to distill the prediction distribution knowledge from a complex model (so-called teacher), which is then used to train a simplified and faster model (commonly denoted as student)[93]. The recent SSL methods exploit the teacher-student approach [94], which train a teacher model based on labeled images, and then the labeled and unlabeled image predictions from the teacher model are used as supervision for training the student model. However, for a standard teacher-student model, unlabeled images cannot be learned by the teacher model, which may lead to unstable predictions for student supervision. Alternatively, the mean teacher (MT) [90] model exploits the unlabeled information in both teacher and student models simultaneously, which achieves state-of-the-art performance in a variety of applications. Nevertheless, several limitations exist for a standard MT model in segmentation tasks. First, a typical MT model expects to minimize the distance between the predictions from two models [90]. However, a direct distance measure without prediction selection would lead to performance degradation, which is caused by too many less confident sample points. As a result, it is challenging for image segmentation tasks with many unreliable prediction points. Meanwhile, the soft information components of predicted results are not properly exploited because of the simple distance measurement. Second, the updating of temporal parameters in the MT model leads to information correlation, which unfortunately introduces a knowledge bias [95].

To address the above issues, several solutions have been proposed recently. An uncertainty-aware self-ensembling model is proposed in [88, 89] to make use of certainty estimations for the segmentation of unlabeled images, which enhances the segmentation performance with limited annotations. Although uncertainty-aware methods [88, 89] achieve superior performances, they are all based on the mean teacher approach with exponential moving averaging (EMA) for parameter updating, which still incurs a parameter-correlation problem between teacher and student models. To overcome the network weight bias from EMA, a Dual-Student model has been proposed to perform interactive prediction refinement [95]. Although the Dual-Student model achieves a better performance than a the MT method, it only exploits the pixel-level information without considering the contextual information, which may not be sufficient for the semantic segmentation task. To deal with this problem, a novel hybrid constraint on predictions is proposed in this chapter, which better exploits

the voxel-level and contextual-level information simultaneously in unlabeled images.

Based on the above literature analysis, we first adopt the DQN method as the localization approach to minimize the annotation effort for coarse region selection, while it also yields an efficient inference procedure. Second, to finely segment the selected regions with an annotation-efficient solution, we propose a novel SSL training scheme based on the uncertainty estimation, imposing both a voxel-level and image-level constraint. In this way, an annotation-efficient solution can be obtained for medical instrument segmentation in 3D US.

6.3 Proposed Method

6.3.1 Direction of Proposed Method

To accurately and efficiently segment the instrument in 3D US, which is trained by an annotation-efficient approach, the proposed method contains two levels of processing for the US image.

- *Coarse region detection*: The 3D volume is processed by a regional patch-based RL framework (e.g. DQN), which learns to coarsely localize the instrument region in 3D US with an environment-action policy. The method is trained with a simplified annotation, which reduces the annotation effort compared to the existing instrument localization methods.
- *Fine patch-based segmentation*: The patches from 3D US images are processed by a 3D network for fine segmentation, which is trained by a novel SSL framework. The proposal is able to exploit unlabeled information at both the voxel and contextual level, which leverages abundant unlabeled images for instrument segmentation.

Compared to recent methods from literature, the proposed method efficiently exploits the contextual information in coarse region selection, which offers a robust estimation for regions of interest with the annotation-efficient solution. In addition, a 3D patch-based network is considered for fine instrument segmentation, which is trained on a large number of unlabeled images while using a few labeled images only. Therefore, the overall annotation effort is significantly reduced to train a successful segmentation framework.

As shown in Fig. 6.2, the proposed coarse-to-fine instrument segmentation framework includes two stages. First, the instrument's center point is localized by the Deep Q-Network (DQN). Second, the Dual-UNet, trained by the SSL framework, is applied on local patches around the estimated location for fine instrument segmentation.

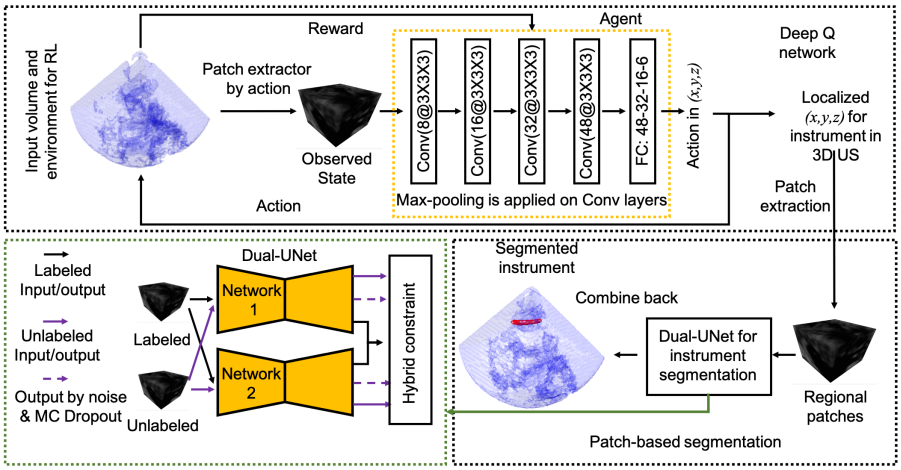


Figure 6.2 Schematic diagram of the proposed framework. (1) Top box: The input 3D volumetric data is processed by a coarse localization algorithm based on the deep Q-network, which localizes the instrument center point in 3D space. (2) Bottom box: Local patches around the detected points are extracted and segmented by the Dual-UNet, which is trained by the proposed SSL scheme (bottom-left green box). The output of the Dual-UNet is the average result of two predictions. The predicted patches are combined back to generate the final prediction output.

6.3.2 Deep Q Learning as a Coarse Selection

The detection of the instrument's center point in 3D US is modeled under the reinforcement learning (RL) framework, which is inspired by behavioral psychology. The system performs an interaction between the agent, i.e. human or a learning agent, with the 3D US image as an environment. Specifically, RL is defined as a computational method to learn the interaction with the environment so as to maximize the cumulative reward signals [81]: The learning agent with its current observation state s interacts with the environment $\mathcal{E}$, by performing successive actions $a \in \mathcal{A}$ to maximize the expected reward r.

As shown in Fig. 6.2 (upper part), we define the observation state in Cartesian coordinates as a 3D observation patch with a size of d^3 voxels w.r.t. the patch center point (x, y, z). To interact with the input image (enviroment) $\mathcal{E}$, the action space $\mathcal{A}$ has six elements, which are defined as $\{\pm a_x, \pm a_y, \pm a_z\}_{\text{scale}}$ w.r.t. three different axes with a resolution of step size 'scale', based on a coarse-to-fine multi-scale strategy [82, 83]. Based on the setting of scale, e.g. in our case scale $\in \{9,3,1\}$, a multi-scale strategy is performed from large to small scale values in the action space. For instance with scale=9, parameter $+a_x$ indicates that the observation region is moved forward along the x-axis by 9 voxels. When the action output falls into local oscillation, i.e. making the observation moving around the region, the scale is reduced to 3 and the procedure continues in a refined way. Based on a

multi-scale strategy, the observation-agent interactions can efficiently converge to the target location [82, 83]. With the observed state s, the agent makes a decision about an action from $\mathcal{A}$, to iteratively update the location of the 3D patch. After each action, a reward r of the RL system is specified by

$$r = \text{sign}\left(D(Pt_g, Pt_{t-1})/\text{scale} - D(Pt_g, Pt_t)/\text{scale}\right), \tag{6.1}$$

where D denotes the Euclidean distance between two points, Pt_g being the ground-truth point, Pt_t is the current state, Pt_{t-1} is the previous state, while scale represents the step size scale in the environment $\mathcal{E}$, see [81]. As a result, the reward $r \in \{-1, 0, +1\}$ indicates whether the agent invokes the patch to move forward or leave it to the instrument center point. With the obtained reward, the optimized action policy can be implemented by learning a state-action value function $Q(s, a)$, which can be approximated by the Deep Q-Network (DQN) [82].

The state-action value function $Q(s, a)$ is commonly called Q-function for Q-learning, which maps input states to corresponding actions. Nevertheless, the commonly employed Q-table [81] leads to the curse of dimensionality when using 3D images, and is therefore impossible to be implemented in practice. Alternatively, the observation-action strategy serves as a dimension-reduction projection, such that it can be approximated by a CNN for image-based observation. As a result, the Q-function is approximated by the CNN, as proposed in the yellow-dotted box in Fig. 6.2 (within top box). To train the DQN, the corresponding loss function is defined as:

$$\mathcal{L}_{\text{DQN}} = E_{s,r,a,\hat{a}\sim M}[\ \left(r + \gamma_{\text{DQN}} \cdot \max_{\hat{a}} Q(\hat{a}, \hat{s}; \widetilde{\omega}) - Q(a, s; \omega)\right)^2\ ], \tag{6.2}$$

where the future reward discount parameter γ_{DQN} is set to 0.9, $\hat{a}$ and $\hat{s}$ are action and observed state in the next step, respectively. Parameter M is the experience replay to de-correlate the random samples. Parameters ω and $\widetilde{\omega}$ are trainable parameters of the Q-networks for the current and target network, respectively. The architecture of the adopted Q-network is depicted in Fig. 6.2, where four recent patches are used as the input [82, 83]. The search is based on historical prediction and is terminated after local oscillation.

6.3.3 Semi-supervised Dual-UNet for Segmentation

With the coarse localization of the instrument in 3D US, the instrument is then segmented by the proposed patch-based Dual-UNet, which is trained by a hybrid constrained SSL framework. Given the training patches containing N labeled patches $\{(x_i, y_i)\}_{i=1}^{N}$ and M unlabeled patches $\{x_j\}_{j=1}^{M}$, where x is the 3D input patch and y is the corresponding 3D annotation, the task is to minimize the following hybrid loss function in Eqn. (6.3):

$$\mathcal{L}_{\text{hybrid}} = L_{\text{sup}} + L_{\text{semi}}, \tag{6.3}$$

where the L_{sup} means the standard supervised loss and L_{semi} represents the proposed constraints for semi-supervised learning. They are introduced in the paragraphs below.

A. Supervised loss function L_{sup}

In this chapter, we consider the standard cross-entropy and Dice hybrid loss function as the supervised loss. Given the label y and its corresponding prediction $\hat{y}$, the supervised loss L_{sup} is defined as

$$L_{\text{sup}} = \text{BCE}(y, \hat{y}) + \text{DICE}(y, \hat{y}), \tag{6.4}$$

where the BCE and DICE are abbreviations for binary cross-entropy and binary Dice loss as defined in [96].

B. Semi-supervised loss function L_{semi}

To exploit the unlabeled image under the supervised signal from labeled data, we propose an SSL training scheme, based on a novel hybrid constraint, which employs a Dual-UNet as the segmentation network. The proposed Dual-UNet structure is motivated by the mean-teacher architecture, which learns the network parameters by updating a student network from a teacher network [88, 89]. Intuitively, this method introduces two networks whose parameters are highly correlated due to the strategy of performing an exponential moving average (EMA) on the updating process. As a result, the obtained knowledge is biased and may be not discriminative enough [95]. Alternatively, we propose to use two *independent* networks, to learn the discriminating information by knowledge interaction through uncertainty constraints. Specifically, the hybrid constraint consists of two types of constraints: a voxel-level constraint and a contextual-level constraint. As for the voxel-level constraint, an intra-network uncertainty constraint (L_{intra}) and an inter-network uncertainty constraint (L_{inter}) are defined to exploit the voxel-level discriminating information for the predictions of the unlabeled images. These two constraints are based on the uncertainty estimation of the predictions, which select the most confident predictions as the supervised signal. Therefore, the unlabeled images' predictions are communicated between two individual networks, which enforce the networks to generate similar predictions with different parameter values. In addition, context-level prediction constraints, i.e. label-wise constraint L_{LCont} and network-wise constraint L_{NCont}, are introduced to exploit the semantic information between unlabeled and labeled predictions, and contextual similarities between networks, respectively. As a result, the context-level constraints can leverage complementary information for voxel-level uncertainty estimations. In our work, we consider the Compact-UNet as a backbone architecture [69], which has proven to be successful in segmenting instruments in 3D volumetric data, as shown in Fig. 6.3. Details of the hybrid constraint components are provided in the sequel.

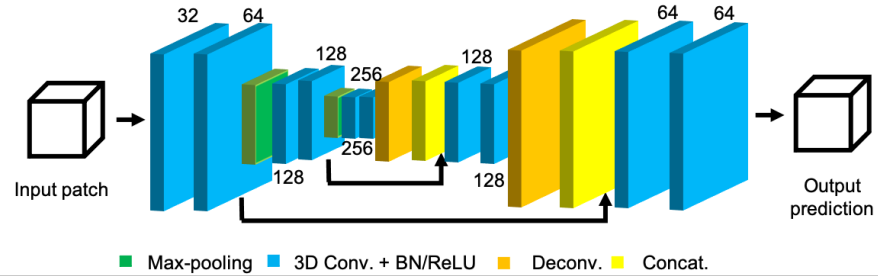


Figure 6.3 Overview of the backbone Compact-UNet. The architecture is simplified for the patch-based binary segmentation in 3D US images. The numbers indicate the filter aperture sizes of $3 \times 3 \times 3 \times N$, where N indicates the number in the figure.

B.1. Intra-network Uncertainty Constraint L_{intra}: Although there is some literature directly using the prediction from a network to guide the unsupervised learning [86, 92], the direct usage of the predictions may include noisy and misclassified voxels, which leads to unsatisfactory results. To generate reliable predictions from history and use them to guide the network to gradually learn discriminating information, we design an uncertainty constraint for each individual network. Given an input patch, T predictions are generated by T times forward passes, based on a Monte Carlo Dropout (MCD) approach and patch input with Gaussian noise (GN) [97]. Therefore, the estimated probability map for a class is obtained by the average of T times predictions for an input patch, resulting in the averaged prediction $\hat{P}_q$ for network $q \in \{1, 2\}$. Based on the above probability maps, the uncertainty $\hat{U}_q$ of this map is measured by $\hat{U}_q = -\sum_c \hat{P}_q \log(\hat{P}_q)$ for c different classes and the related loss constraint for network q is formulated by:

$$L^q_{\text{intra}} = \frac{\sum(\mathcal{I}(\hat{U}_q < \tau_1) \odot ||\hat{y}_q - \hat{P}_q||)}{\sum \mathcal{I}(\hat{U}_q < \tau_1)}, \tag{6.5}$$

where $\sum$ is the sum of all voxels of the considered patch. Symbol $\mathcal{I}$ is a binary indicator function, τ_1 denotes a threshold to measure the uncertainty [89], which selects the most reliable voxels by binary voxel-level multiplication $\odot$. Parameter $\hat{y}_q$ is the prediction for network q with $q \in \{1, 2\}$. By following this approach, the proposed strategy is approximately equal to the mean-teacher method with the history step equal to unity in the methods [88, 89]. Intuitively, this constraint selects the reliable voxels from Bayesian predictions, where only the most confident points are selected to guide the network.

B.2. Inter-network Uncertainty Constraint L_{inter}: Besides the above uncertainty constraint for each network, we also propose an uncertainty constraint to measure the prediction consistency between two individual networks with the purpose to constrain the knowledge and avoid bias [95]. The proposed inter-network uncer-

tainty constraint enables the networks to learn the discriminating information, by comparing the predictions between two networks with stable voxel selection. With the above definitions of normal prediction ($\hat{y}_q$) and the averaged Bayesian prediction ($\hat{P}_q$), their corresponding binary predictions are obtained as C_q and $\hat{C}_q$, respectively, which are thresholded by 0.5 for a fair class distribution. Based on these, more stable voxels for each network are defined as

$$\mathcal{S}_q = \mathcal{I}(C_q \odot \hat{C}_q) \odot (\mathcal{I}(U_q < \tau_2) \oplus \mathcal{I}(\hat{U}_q < \tau_2)), \tag{6.6}$$

where U_q is the uncertainty based on the normal output and $\hat{U}_q$ represents the uncertainty based on Bayesian output. Parameter τ_2 is a stronger threshold to select the more stable voxels for the Network q than in case of using Eqn. (6.5). By using a voxel-based logical operator OR ($\oplus$), stable instrument voxels are loosely selected to find the matched prediction voxels from the same-class prediction. Furthermore, we also define the voxel-level probability distance $D_q = ||\hat{y}_q - \hat{P}_q||$, which indicates the predictions' consistency. With definitions of stable voxels and probability distances, the less stable voxels in the stable samples are optimized to enhance the overall voxel confidence between two networks. Specifically, the inter-network uncertainty constraint L_{inter} for Network 1 is formulated by:

$$L^1_{\text{inter}} = \frac{\sum(((\mathcal{S}_1 \odot \mathcal{S}_2 \odot \mathcal{I}(D_1 > D_2)) \oplus (\overline{\mathcal{S}_1 \odot \mathcal{S}_2} \odot \mathcal{S}_2)) \odot ||\hat{y}_1 - \hat{y}_2||)}{\sum((\mathcal{S}_1 \odot \mathcal{S}_2 \odot \mathcal{I}(D_1 > D_2)) \oplus (\overline{\mathcal{S}_1 \odot \mathcal{S}_2} \odot \mathcal{S}_2))}, \tag{6.7}$$

where $||\cdot||$ is the probability distance at the voxel level and $\overline{(\cdot)}$ stands for a binary NOT operation. Intuitively, the operation $\mathcal{S}_1 \odot \mathcal{S}_2 \odot \mathcal{I}(D_1 > D_2)$ selects the less stable voxels from Network 1 by comparing the probability distance from the two networks' stable voxels. As for function $\overline{\mathcal{S}_1 \odot \mathcal{S}_2} \odot \mathcal{S}_2$, if the voxels are not stable for both networks but are stable for the Network 2, then these voxels' information are used to guide the Network 1 to generate a similar prediction. This uncertainty constraint enables the unsupervised signal communication between two individual networks, and train the Network 1. A similar expression with mirrored indexes is applied to Network 2.

B.3. Label-wise Contextual Constraint L_{LCont}: The above loss constraints on the intra-/inter-network consider voxel-level consistency of paired predictions, i.e., the predictions from two networks for the same input, while ignoring the differences between labeled and unlabeled predictions at the contextual level. To learn the prediction consistency at the contextual level, we also introduce a contextual constraint, based on the implementation of adversarial learning. Specifically, the labeled and unlabeled predictions are analyzed by a classifier, as shown in Fig. 6.4, to generate the image class: labeled or unlabeled, which are used to generate the binary cross-entropy (BCE). L_{LCont} is defined as

$$L^q_{\text{LCont}} = -\text{BCE}(\hat{Cls}_q, Cls_q), \tag{6.8}$$

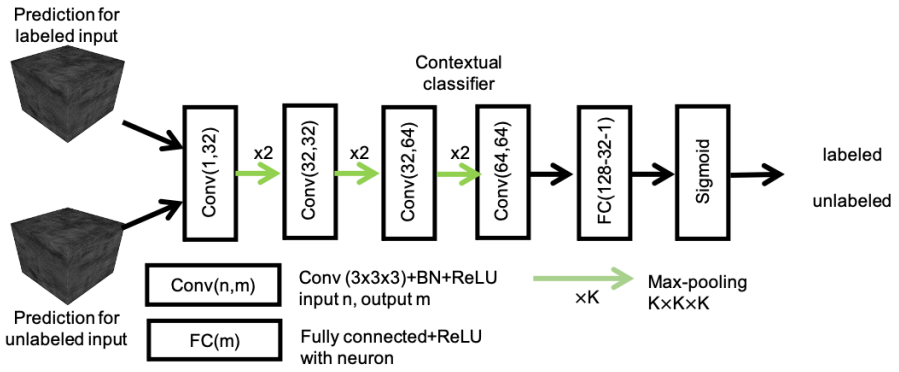


Figure 6.4 Architecture of the proposed classifier for L_{LCont}. The network distinguishes the input is labeled or not.

where the $\hat{Cls}_q$ is the predicted class whether the input prediction having a corresponding annotation or not, while Cls_q is the prior knowledge of the prediction having an annotation or not. The negative sign is considered to maximize the similarity between the labeled predictions and unlabeled predictions, while the BCE is minimized to distinguish them.

B.4. Network-wise Contextual Constraint L_{NCont}: The label-wise contextual constraint focuses on the contextual difference between labeled predictions and unlabeled predictions, which ignores the contextual information consistency between two individual networks. To fully exploit this contextual information at network-level, a network-wise contextual constraint is introduced as well. Specifically, it has two different processing steps for labeled and unlabeled predictions. (1) The labeled images' predictions and corresponding ground truth are processed by an encoder to generate contextual vectors, which are used to measure the latent space similarity between prediction and ground truth. (2) As for the unlabeled predictions from the two networks, their contextual vectors are measured to enforce themselves to be as similar as possible. The contextual encoder (CE) has a similar structure with that of Fig. 6.4, but excludes the FC layers and adds one extra Conv layer (kernel number of 64). The L_{NCont} is defined as the vector distance by norm-2, and consisting of several components:

$$\begin{aligned} L_{\text{NCont}} =& ||CE(\hat{y}_1^l) - CE(y)|| + ||CE(\hat{y}_2^l) - CE(y)|| \\ &+ ||CE(\hat{y}_1^u) - CE(\hat{y}_2^u)||, \end{aligned} \tag{6.9}$$

where $\hat{y}$ and y are predictions and corresponding annotations, respectively. Parameters l and u represent labeled and unlabeled patches. This network-wise constraint compensates the intra-network contextual information usage in L_{LCont} and enforces the information interaction in a similar way as L_{inter}. Based on the design, the contextual encoder CE is trained properly from the supervised signal,

which is simultaneously used to enforce the unlabeled predictions from different UNets to be the same.

C. Hybrid loss

Based on the above-defined constraints, the proposed hybrid loss function is defined by a weighted summation of the previous constraints, leading to

$$\mathcal{L}^q_{\text{hybrid}} = L^q_{\text{sup}} + \alpha(L^q_{\text{intra}} + L^q_{\text{inter}}) + \beta L^q_{\text{LCont}} + \gamma L^q_{\text{NCont}}. \tag{6.10}$$

Here, parameter $q \in \{1, 2\}$ is the network number in Fig. 6.2, L^q_{sup} denotes the supervised loss of network q, the term L^q_{intra} is the uncertainty constraint to measure the consistency between outputs for network q, term L^q_{inter} is an uncertainty constraint to measure the consistency between the two networks, which exploits the information from two independent networks to enhance the performance. Term L^q_{LCont} is a label-wise contextual constraint to compensate the information usage of the voxel-level predictions, based on knowledge between labeled and unlabeled predictions. Term L_{NCont} is the network-wise contextual constraint based on the contextual difference between predictions and labels from different networks. Finally, coefficients α, β and γ are parameters to balance the weight between different loss components.

Intuitively, L_{sup} makes use of labeled information to guide the networks to converge to correct predictions and optimize the direction in hyperparameter space. In contrast to supervised information, L_{intra} focuses on the information uncertainty for each individual network. Specifically, it considers MCD and GN to generate noisy and less confident predictions, which are binarized by threshold τ_1 to select the reliable predictions in the patch. With these selected voxels, the probability distances between normal predictions to these voxels are minimized to enhance the confidence of the network, which avoids the voxels with low confidence or noise from a common Π-model. However, in contrast to the uncertainty-aware network [88, 89], which employ two separate networks with historical parameter correlation, the proposed method ensembles two networks into one with a history step equal to unity. Moreover, instead of intra-network information usage, L_{inter} focuses on voxel-level uncertainty interactions, which omits the parameter correlations and generates more diverse network parameters from the random procedures, such as parameter initializations of networks, and applied MCD and GN techniques. In detail, the L_{inter} loss is designed to select more stable voxels based on predictions, which are used to reduce the probability distance between the predictions of these stable voxels from two networks. As described in the definitions, L_{LCont} is considered to maximize the prediction similarity between labeled and unlabeled outputs, while L_{NCont} is used to enforce a higher contextual similarity between the predictions of the two networks.

6.4 Experiments

6.4.1 Datasets and Preprocessing

Ex-vivo RF-ablation catheter dataset: To validate our instrument segmentation method, we collected an *ex-vivo* dataset on RF-ablation catheter for cardiac intervention, consisting of 88 3D cardiac US volumetric images from 8 porcine hearts. During the recording, the heart was placed in a water tank with the RF-ablation catheter (with a diameter of 2.3-3.3 mm) inside the heart chambers. The phased-array US probe (X7-2t with 2,500 elements by Philips Medical Systems, Best, the Netherlands) was placed next to the interested chambers to capture the images containing the catheter, which was monitored by a US console (EPIQ 7 by Philips). For each recording, we pulled the catheter out and re-inserted it into the heart chamber, and placed the probe with different locations and view angles, to minimize the overlap among images. The obtained volumetric images are re-sampled to the volume size of $160 \times 160 \times 160$ voxels (where padding is applied at the boundary to make the volume such that it has equal size in each direction), which leads to a voxel size range of 0.3-0.8 mm. All the volumes are manually annotated at voxel level. Moreover, the catheter centers are also marked as the target location for DQN. To validate the proposed method, 60 volumes are randomly selected as training set, 7 volumes are used as validation images and 21 volumes are used as testing images. To train the DQN, 60 volumes with target location are used to learn the action policy. To train the Dual-UNet, 6, 12 and 18 of 60 volumes are selected as the labeled images, while the remainder are the unlabeled images for SSL training.

In-vivo RF-ablation catheter dataset: To further validate the generalization of the proposed method, an *in-vivo* RF-ablation catheter dataset was collected from 2 live porcine hearts, which includes 13 images with the RF-ablation catheter in the heart chambers. The data collection was approved by an ethical committee, and recorded at Utrecht University, the Netherlands. The images were collected by a phased-array US probe (X7-2t with 2,500 elements by Philips Medical Systems, Best, the Netherlands). During the recording, a medical doctor manipulated the catheter to reach different regions of the heart chamber, where the RF-ablation procedures were performed. Similar to the *ex-vivo* dataset, the images are re-sampled to the volume size of $160 \times 160 \times 160$ voxels. The obtained images are manually annotated at voxel-level. All 13 volumes are used to validate the generalization of the model, which is trained on the above *ex-vivo* dataset.

In-vivo TAVI guide-wire dataset: We also collected an *in-vivo* TAVI guide-wire dataset including 18 volumes from 2 TAVI operations. The study was approved by the institutional review board of Philips (ICBE) and the Catharina Hospital Eindhoven (Medical Research Ethics Committees United, MEC-U; study ID: non-

WMO 2017-106). Patients approved the use of anonymous data for retrospective analysis. During the recording, the sonographer captured images at different locations in the chamber without interfering the procedure. The volumes were recorded with a mean volume size of $201 \times 202 \times 302$ voxels. Similar to the above *ex-vivo* dataset, volumes are re-sampled to have a volume size of $160 \times 160 \times 160$ voxels. The guide-wire (0.889 mm) has the thickness of about 3-5 voxels due to spatial distortion. The images are manually annotated by a technical expert to generate the binary segmentation mask as the ground truth. The guide-wire center point is marked for DQN training. We have randomly divided the dataset into three parts: 12 volumes for training, 2 volumes for validation and 4 volumes for testing. Specifically for the training images, 2, 4 and 6 volumes of 12 images are selected as the labeled images, while the rest are used as the unlabeled images for SSL training.

6.4.2 Implementation Details and Training Process

We have implemented the proposed framework in Python 3.7 with *TensorFlow* 1.10, using a standard PC with a Titan 1080Ti GPU. We have trained the DQN with the Adam [98] optimizer (learning rate 10^{-4}) for 40 and 20 epochs until converging on the validation datasets for *ex-vivo* and *in-vivo* datasets, respectively. Replay memory is set to be 100k with random sampling. Parameters of the target network are updated for every 2,500 steps. When considering the efficiency and accuracy of the DQN, we define the input state space to 55^3 voxels for resized images with the size of 96^3 voxels for both *ex-vivo* and *in-vivo* datasets, to ensure that the observations can contain sufficient contextual information of instrument. The total training times for the two datasets are 32 and 16 hours, respectively.

As for SSL training, the patches are generated by applying random translations based on the annotated instrument center points. Moreover, the data augmentations with rotation, mirror and intensity re-scales are applied. To adapt the UNet as a Bayesian network [99] and generate uncertainty prediction, dropout layers with rate 0.5 are inserted prior to the convolutional layers. Gaussian random noise is also considered during uncertainty estimation. For the uncertainty estimation suggested by [89], $T = 8$ is used to balance the efficiency and quality of the estimation. *Ex-vivo* dataset training is terminated after the loss has converged on the validation dataset, or after 10,000 steps with mini-batch size of 4 using the Adam optimizer (learning rate 10^{-4}) [98], which includes 2 labeled and 2 unlabeled patches. Meanwhile, the training on the *in-vivo* dataset is terminated after the loss converged on the validation dataset with mini-batch size of 4 (learning rate 10^{-4}). Hyperparameters α, β and γ are empirically chosen as 0.1, 0.002 and 0.1, respectively, to balance different loss components. Moreover, a ramp-up weighting coefficient strategy is considered for threshold parameters to balance the components confidence during the training. Thresholds τ_1 and τ_2 are experimentally selected, based on an uncertainty function with probability as 0.5

and 0.7 w.r.t. uncertainty estimation $U = -\sum_c p\log(p)$, respectively. The total training times for the two datasets are 14 and 7 hours, respectively.

6.4.3 Evaluation Metrics

As for the experiments using DQN as pre-selection, the metric is the Euclidean distance between the detected instrument center point and ground-truth center point in terms of voxels, which is denoted as 'Dis'. Moreover, the success rate of off-line localization is also considered as an evaluation metric, to evaluate the detection performance.

To evaluate the overall segmentation performance of the proposed method, we consider the Dice score (DSC), Volumetric Similarity (VS), and 95-% Hausdorff Distance (95HD) as evaluation metrics.

6.5 Results

6.5.1 Performance of DQN for Instrument Localization

In this section experimentally compares the following methods for instrument localization on our datasets: Single Shot MultiBox Detector (SSD) [79], Faster R-CNN [80], and the proposed DQN method. We consider different input volume sizes: 96^3, 128^3 and 160^3 voxels, which involves different amounts of contextual information within the fixed observation space, i.e. 55^3 voxels for a fixed network structure in DQN. The SSD and Faster R-CNN models are extended into 3D space with a modified backbone network, i.e. 3D ResNet, which reduces the GPU memory usage, increases the learning speed and keeps sufficient contextual information for handling small instruments. The object scale parameters are defined based on the size of the instruments for different feature map sizes (five object scales for SSD as $\{0.4, 0.4, 0.7, 0.7, 0.9\}$ and two sizes of $\{48, 72\}$ for Faster R-CNN). The bounding boxes are generated based on the center point of the instrument in 3D space, with the same size in each direction, which is equal to the length of the instrument. We have failed to train the network with bounding box following the shape of the instrument (e.g. a box of size [6, 6, 50] for a catheter), which is too small to map the boxes in the high-level feature maps. Because there is only one instrument in the image, the number of detected results is set to be unity after the non-maximum suppression. Data augmentations of shifting, rotation, and resizing are applied during the training. The results are summarized in Table 6.1, where the distance metric is expressed as the number of voxels in the resolution of 160^3 voxels.

Comparing the SSD and Faster R-CNN, the DQN method provides a stable and higher performance for several reasons. First, the DQN method includes a simpler CNN network with discrete prediction space compared to SSD and Faster R-CNN, thus it is easier to be trained with a limited number of images. In contrast, the complex network backbone structures of the detection networks are

Table 6.1 Detection accuracy of different coarse detection methods, which are ranked by detection distance (Dis) using mean±std for successful detections. The case is considered successful when Dis<10 voxels.

RF-ablation Catheter	**DQN**	
Volume Size	Dis (voxel)	Success rate
96^3 voxels	3.8 ± 1.8	$21/21$
128^3 voxels	4.0 ± 2.4	$21/21$
160^3 voxels	4.7 ± 3.3	$18/21$
RF-ablation Catheter	**SSD**	
Volume Size	Dis (voxel)	Success rate
96^3 voxels	-	$0/21$
128^3 voxels	5.8 ± 2.1	$8/21$
160^3 voxels	5.3 ± 2.6	$15/21$
RF-ablation Catheter	**Faster R-CNN**	
Volume Size	Dis (voxel)	Success rate
96^3 voxels	-	$0/21$
128^3 voxels	5.1 ± 0.6	$3/21$
160^3 voxels	5.7 ± 2.2	$7/21$
TAVI Guide-wire	**DQN**	
Volume Size	Dis (voxel)	Success rate
96^3 voxels	2.4 ± 1.0	$4/4$
128^3 voxels	3.2 ± 2.4	$4/4$
160^3 voxels	3.4 ± 2.3	$4/4$
TAVI Guide-wire	**SSD**	
Volume Size	Dis (voxel)	Success rate
96^3 voxels	-	$0/4$
128^3 voxels	-	$0/4$
160^3 voxels	-	$0/4$
TAVI Guide-wire	**Faster R-CNN**	
Volume Size	Dis (voxel)	Success rate
96^3 voxels	-	$0/4$
128^3 voxels	-	$0/4$
160^3 voxels	-	$0/4$

easier to overfit on the limited training images. Second, for both SSD and Faster R-CNN, the bounding boxes are analyzed based on high-level feature maps in the whole image, where it is hard for the networks to learn the discriminating information of very small objects with extremely imbalanced class distributions. Third, these methods have different objective functions. The DQN is trained to maximize the reward function based on the distance metric loss. In contrast, the SSD and Faster R-CNN are trained by regression loss and classification loss, which are difficult to be trained on the limited training data. As can be observed, when the input volume size decreases, the performances of SSD and Faster R-CNN are also degrading accordingly, which is because less discriminating information is extracted from high-level feature maps. As a conclusion, the bounding-box-based locators are not feasible for our challenging task with only limited training data.

From the results, the DQN method provides the best accuracy on both datasets in the environment of 96^3 voxels. A larger volume size leads to lower performance and a higher failure rate. For a fixed observation size (55^3 voxels) for the agent of DQN, the larger volume size means the observation may not capture the whole instrument. In contrast, a smaller volume size would ensure the observation covers the whole instrument. Because of the multi-scale spatial steps for interaction, the localization of DQN takes around 0.2, 0.3 and 0.7 seconds for the 96^3, 128^3 and 160^3 cases, respectively, while the SSD and Faster R-CNN are faster, e.g., about 0.1 seconds for 160^3 voxels. The *ex-vivo* dataset obtains inferior performances due to higher variation of the images when compared to TAVI operations, which has almost fixed anatomical structures.

6.5.2 Comparisons with Other Methods

With the detected instrument center point, patches with the size of 48^3 voxels are extracted around the point for semantic segmentation (i.e. 2 patches for each direction and 2^3 patches in total). Performance comparison with the state-of-the-art methods is presented below.

We compare the proposed method with the state-of-the-art SSL methods, including Bayesian UNet (B-UNet) [97], Π-model [86], Adversarial-based segmentation (AdSeg)[84], multi-task attention-based SSL (MA-SSL)[87], uncertainty-aware-based mean-teacher (UA-MT),[89] and teacher-student-based (TS) knowledge distillation [93] (all of them are based on the Compact-UNet as backbone [69]). Specifically, the TS model trains the teacher part with a more complex model by increasing the filter number by a factor of two. The teacher is first trained on labeled images, which is then used to generate the soft-prediction of unlabeled images for the student model. The soften parameter of the TS model is set to 5 for unlabeled images with loss weight 0.5. Results are shown in Table 6.2, which depicts that the proposed method clearly outperforms the state-of-the-art SSL approaches. Examples of segmentation results of different SSL methods are

Table 6.2 Segmentation performance for different methods, expressed in Dice Score (DSC), Volumetric Similarity (VS), and 95-% Hausdorff Distance (95HD), which are shown by mean±std. Combination (L, U) represents (Labeled, Unlabeled) images for SSL training. All methods are based on DQN pre-selection results. The results of the proposed method are denoted in bold.

Method	# *Images*	*RF-ablation Catheter*			# *Images*	*TAVI Guide-wire*		
	(L, U)	DSC %	VS %	95HD (voxels)	(L, U)	DSC %	VS %	95HD (voxels)
B-UNet[97]	(6, 0)	43.0±26.1	62.9±30.8	13.4±13.5	(2, 0)	34.5±25.0	46.2±28.3	9.5±11.9
AdSeg[84]	(6, 54)	44.7±25.1	66.1±28.6	13.1±11.6	(2, 10)	41.5±25.0	68.9±22.5	11.8±10.8
Π-model[86]	(6, 54)	29.2±19.8	53.0±26.0	31.5±7.1	(2, 10)	22.6±15.6	46.5±23.9	22.9±9.0
MA-SSL[87]	(6, 54)	41.3±24.1	64.5±28.5	18.3±13.2	(2, 10)	41.2±24.8	62.7±29.9	8.9±13.4
UA-MT[89]	(6, 54)	45.5±25.3	70.8±29.5	12.7±13.2	(2, 10)	36.0±27.8	75.0±14.4	13.2±12.4
KD-TS [93]	(6, 54)	40.2±26.9	53.4±35.4	14.7±16.3	(2, 10)	37.6±26.4	60.9±27.2	10.5±6.6
Proposed	(6, 54)	**54.3±15.0**	**82.7±16.8**	**8.2±7.6**	(2, 10)	**43.2±23.9**	**81.9±13.2**	**7.5±6.1**
B-UNet[97]	(12, 0)	60.5±14.2	84.9±18.3	9.5±8.6	(4, 0)	45.0±28.3	75.8±32.9	5.6±5.3
AdSeg[84]	(12, 48)	62.1±11.9	88.5±10.5	9.1±9.5	(4, 8)	50.0±28.7	85.0±24.2	5.3±4.9
Π-model[86]	(12, 48)	29.3±20.1	68.6±23.6	30.6±5.8	(4, 8)	21.4±11.9	45.3±12.6	22.6±10.2
MA-SSL[87]	(12, 48)	61.7±11.5	87.3±12.4	7.9±7.8	(4, 8)	46.9±25.2	75.0±38.2	9.3±13.5
UA-MT[89]	(12, 48)	61.2±12.9	89.2±8.9	9.3±7.9	(4, 8)	49.1±22.5	88.4±11.5	4.0±5.2
KD-TS [93]	(12, 48)	62.0±11.7	89.7±10.2	11.8±11.8	(4, 8)	48.4±28.7	84.0±17.5	4.8±6.1
Proposed	(12, 48)	**65.6±9.5**	**90.2±5.6**	**5.1±3.9**	(4, 8)	**53.2±21.7**	**89.7±10.7**	**3.8±4.7**
B-UNet[97]	(18, 0)	64.1±9.8	87.3±8.8	5.9±5.0	(6, 0)	61.3±9.4	88.4±5.9	4.2±5.5
AdSeg[84]	(18, 42)	66.2±8.7	89.1±5.8	10.3±11.1	(6, 6)	60.6±7.7	94.0±2.6	5.3±5.2
Π-model[86]	(18, 42)	28.8±21.9	55.1±23.9	27.7±9.1	(6, 6)	32.1±7.4	60.2±11.1	20.2±10.4
MA-SSL[87]	(18, 42)	62.3±13.0	83.0±9.7	7.9±8.0	(6, 6)	59.2±3.2	87.1±6.6	3.5±4.2
UA-MT[89]	(18, 42)	66.3±9.2	89.5±8.7	4.3±3.2	(6, 6)	64.7±7.3	92.0±3.8	1.8±0.6
KD-TS [93]	(18, 42)	66.3.±8.5	88.7±8.0	4.6±4.8	(6, 6)	64.7±8.1	89.6±5.0	2.2±1.2
Proposed	(18, 42)	**69.1±7.3**	**92.8±4.6**	**3.0±2.1**	(6, 6)	**68.6±7.9**	**96.5±3.3**	**1.7±0.6**
Share-CNN[19]	(60, 0)	58.4±12.6	81.3±16.2	8.0±6.4	(12, 0)	56.4±13.3	87.2±10.2	5.7±6.6
Compact-UNet[69]	(60, 0)	66.8±7.3	88.0±7.1	4.0±3.1	(12, 0)	63.2±6.6	94.9±2.1	1.9±0.5
Pyramid-UNet[69]	(60, 0)	**70.6±6.5**	**90.1±6.5**	**3.0±2.3**	(12, 0)	**67.4±6.4**	**96.1±1.8**	**1.5±0.5**

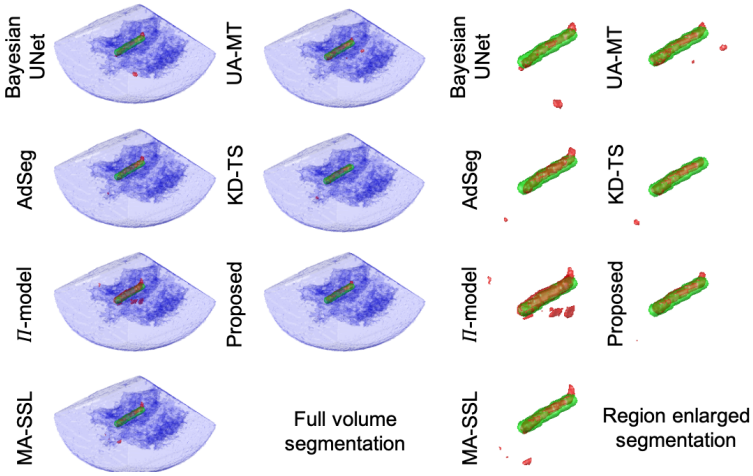


Figure 6.5 Examples results via different methods for (L, U)=(18, 42), which are corresponding to Table 6.2. Left: full volume. Right: the enlarged region including the catheter. Green: annotation, red: segmentation, and blue: heart tissue. All the results are obtained based on the coarse detection.

shown in Fig. 6.5, where 18 annotated images are used for training. It can be observed that the proposed method provides less outliers than other methods because of the effective coarse segmentation step and better uncertainty constraints.

Table 6.3 Paired t-tests (p-value) for different SSL methods compared to our method. The expression e-n stands for $\times 10^{-n}$.

Dataset	B-UNet	AdSeg	Π-model	MA-SSL	UA-MT	KD-TS
RF-ablation Catheter	8.7 e-4	1.6 e-3	3.4 e-8	6.1 e-4	4.0 e-3	4.6 e-4
TAVI Guidewire	6.9 e-3	3.5 e-2	3.3 e-3	2.2 e-2	2.3 e-3	5.9 e-2

From the table, it can be observed that when the number of labeled images increases, the segmentation performances are improved w.r.t available supervised information except for the Π-model, which is due to the unreliable information of the predicted unlabeled images. Because of the unreliable information in challenging 3D US images, the Π-model obtains a lower performance than a simple Bayesian UNet. Compared to AdSeg, MA-SSL and UA-MT, the proposed method achieves a better performance, since it exploits uncertainty and contextual information, which improves the information usage of unlabeled images. To further validate the performance differences between different SSL methods, we have performed a paired t-test with $\alpha = 0.05$ on two datasets using the DSC metric in a one-tailed test, which are summarized in Table 6.3. From the table, the proposed method gives larger statistical differences than the other methods on the *RF-ablation Catheter* dataset, i.e. p-value$<$0.01. In contrast, the proposed method

has less statistical difference on the *TAVI Guide-wire* data, especially for the KD-TS model, which is because of limited testing data with only 4 images.

We have also compared the proposed method with supervised learning methods at the bottom of Table 6.2. The proposed method obtains better results than voxel-wise Share-CNN for catheter segmentation [19]. The proposed method also outperforms the supervised learning method with Compact-UNet as backbone structure, while it achieves similar performance to a more complex Pyramid-UNet. This is explained by two reasons. (1) The proposed constraint provides more confident and accurate predictions on unlabeled images, which guides the network to exploit the most meaningful information. However, for the supervised learning methods with a standard loss, it has less capability of the discriminating information because noisy labels may exist. (2) The proposed method represents a semi-supervised learning by a multi-task learning approach, which exploits the information for different tasks, such that knowledge is better learned. It is worth to mention that the proposed method is statistically better than Share-CNN (p-value$<$0.01), while it has less difference to Compact-UNet (p-value$\sim$0.03). Compared to Pyramid-UNet, there is no statistical difference between the segmentation results. These results show that the proposed SSL method achieves state-of-the-art performance with much less annotation effort.

Although the Dual-UNet has a more complex architecture than standard UNet, it employs unlabeled images as the support and guidance for SSL training, which achieves comparable performances with fully supervised learning. Finally, from the experiments, the proposed two-stage scheme executes in approximately 1 second per volume (0.2-0.3 seconds for DQN pre-selection and 0.7 seconds for patch-based segmentation). As a comparison, exhaustive patch-based segmentation requires more than 10 seconds per volume [69], while a voxel-of-interest-based CNN method takes about 10 seconds [19]. Therefore, the proposed method is 10 times faster.

6.5.3 Ablation Study of Different Loss Components

The ablation studies on different constraint components are summarized in Table 6.4, where different numbers of labeled and unlabeled images are considered. More specifically, the UNet with Mont Carlo operation is denoted as the baseline and backbone structure for the proposed method, which is trained with L_{sup} only. For the proposed SSL, constraint components are added one by one to validate their effectiveness.

Several conclusions can be drawn from the table. (1) The simple backbone UNet with supervised loss can learn more discriminating information with the number of available annotations increasing, which however obtains worse performance than the Dual-UNet. This is because randomly initialized parameters and dropout operations in the Dual-Unet avoid the learning bias with higher network diversity, which results in more stable predictions. (2) Compared to

Table 6.4 Segmentation performance for loss components using the Dice Score (DSC), Volumetric Similarity (VS), and 95-% Hausdorff Distance (95HD), which are shown in mean±std. Combination (L, U) denotes (Labeled, Unlabeled) images for SSL training. DU stands for Dual-UNet. The highest performances are depicted in bold fonts.

Method	# *Images*	*RF-ablation Catheter*			# *Images*	*TAVI Guide-wire*		
	(L, U)	DSC %	VS %	95HD (voxels)	(L, U)	DSC %	VS %	95HD (voxels)
UNet-L_{sup}	(6, 0)	43.0±26.1	62.9±30.8	13.4±13.5	(2, 0)	34.5±25.0	46.2±28.3	9.5±11.9
DU-L_{sup}	(6, 0)	46.4±24.2	65.3±31.0	12.6±12.6	(2, 0)	38.4±26.8	59.1±22.0	9.4±10.5
DU-$L_{\text{sup+intra}}$	(6, 54)	47.0±22.0	72.3±21.2	12.1±11.5	(2, 10)	39.3±21.2	65.4±23.5	10.9±11.7
DU-$L_{\text{sup+inter}}$	(6, 54)	49.2±22.6	71.9±25.9	11.9±11.6	(2, 10)	39.9±25.8	69.1±18.8	8.2±6.5
DU-$L_{\text{sup+intra+inter}}$	(6, 54)	51.2±22.3	72.9±28.8	10.9±8.3	(2, 10)	40.7±25.3	67.6±27.6	9.7±5.6
DU-$L_{\text{sup+intra+inter+LCont}}$	(6, 54)	54.0±17.4	76.5±22.8	8.4±8.4	(2, 10)	41.2±22.2	76.6±18.5	**7.5±5.9**
DU-$L_{\text{sup+intra+inter+NCont}}$	(6, 54)	53.2±17.2	79.4±22.3	11.9±10.2	(2, 10)	41.3±22.5	75.7±15.4	8.2±10.6
DU-$L_{\text{sup+intra+inter+LCont+NCont}}$	(6, 54)	**54.3±15.0**	**82.7±16.8**	**8.2±7.6**	(2, 10)	**43.2±23.9**	**81.9±13.2**	7.5±6.1
UNet-L_{sup}	(12, 0)	60.5±14.2	84.9±18.3	9.5±8.6	(4, 0)	45.0±28.3	75.8±32.9	5.6±5.3
DU-L_{sup}	(12, 0)	61.2±10.0	86.4±8.5	9.7±6.7	(4, 0)	47.0±30.3	78.2±31.8	4.9±5.4
DU-$L_{\text{sup+intra}}$	(12, 48)	62.3±9.6	87.7±11.8	9.7±10.6	(4, 8)	48.4±23.7	81.2±18.1	4.7±6.3
DU-$L_{\text{sup+inter}}$	(12, 48)	61.6±9.4	87.7±8.6	8.4±7.2	(4, 8)	48.1±25.9	82.7±22.0	4.7±6.2
DU-$L_{\text{sup+intra+inter}}$	(12, 48)	62.6±10.4	88.1±9.0	9.1±10.3	(4, 8)	49.5±27.7	82.9±18.9	4.6±6.6
DU-$L_{\text{sup+intra+inter+LCont}}$	(12, 48)	64.6±8.9	88.5±8.5	7.0±8.0	(4, 8)	50.5±22.3	85.5±14.5	**4.5±5.5**
DU-$L_{\text{sup+intra+inter+NCont}}$	(12, 48)	63.2±10.5	85.9±7.2	7.8±7.2	(4, 8)	50.0±24.7	84.7±20.1	4.5±5.7
DU-$L_{\text{sup+intra+inter+LCont+NCont}}$	(12, 48)	**65.6±9.5**	**90.2±5.6**	**5.1±3.9**	(4, 8)	**53.2±21.7**	**89.7±10.7**	4.6±4.7
UNet-L_{sup}	(18, 0)	64.1±9.8	87.3±8.8	5.9±5.0	(6, 0)	61.3±9.4	88.4±5.9	4.2±5.5
DU-L_{sup}	(18, 0)	65.1±8.9	87.5±8.8	5.5±3.5	(6, 0)	62.6±6.6	90.4±5.4	2.2±1.1
DU-$L_{\text{sup+intra}}$	(18, 42)	66.9±7.7	88.6±7.6	4.9±5.0	(6, 6)	63.5±10.5	92.3±2.6	2.9±1.3
DU-$L_{\text{sup+inter}}$	(18, 42)	66.5±9.3	89.8±6.5	5.2±6.6	(6, 6)	65.0±9.4	92.1±4.5	2.4±0.8
DU-$L_{\text{sup+intra+inter}}$	(18, 42)	67.7±8.5	89.1±8.0	3.5±2.2	(6, 6)	65.2±8.0	93.1±2.7	1.9±1.0
DU-$L_{\text{sup+intra+inter+LCont}}$	(18, 42)	68.8±7.2	90.9±4.7	3.3±2.2	(6, 6)	66.3±5.2	93.7±4.0	1.8±0.4
DU-$L_{\text{sup+intra+inter+NCont}}$	(18, 42)	68.9±7.5	92.3±5.2	4.1±3.3	(6, 6)	66.2±9.0	92.8±2.0	**1.6±0.5**
DU-$L_{\text{sup+intra+inter+LCont+NCont}}$	(18, 42)	**69.1±7.3**	**92.8±4.6**	**3.0±2.1**	(6, 6)	**68.6±7.9**	**96.5±3.3**	1.7±0.6

the case with only supervised loss, adding voxel-level constraints, i.e. L_{intra} and L_{inter}, allows to select the stable voxels from uncertainty estimations, which thereby exploits the discriminating information from unlabeled images' predictions. More specifically, the L_{intra} constraint focuses on prediction uncertainty within the network, while the L_{inter} constraint exploits the uncertainties of the predictions between two individual networks. The results indicate that both constraints improve the performance and are complementary to each other. (3) The contextual-level constraint including label-wise and network-wise constraints, also contributes to further performance improvement. Specifically, the label-wise constraint exploits the contextual similarity between labeled and unlabeled images' predictions, while the network-wise constraint focuses on prediction similarity between different networks of the Dual-UNet. (4) The proposed hybrid loss gives more significant performance improvement when the amount of labeled images are small, which indicates the proposed method is able to exploit the discriminating information from unlabeled images.

It can be observed that as the number of annotated images increases, the variance of the segmentation performance is decreasing. This is because a more confident guidance is obtained from available annotations. In the following ablation studies, we have chosen the cases with the most annotated volumes for both datasets, i.e. the (18, 42) and (6, 6) combinations for labeled and unlabeled training images.

6.5.4 Ablation Study of Patch Size of Dual-UNet

To investigate the influence of the patch size, the input patch size of 32^3, 48^3 and 64^3 are examined. The results are shown in Table 6.5. From the results, patches with 32^3 voxels obtain a slightly worse performance than 48^3 voxels, which however requires about 3 seconds execution time because more patches are required for a fixed volume size after DQN pre-selection (64^3 voxels). In contrast, patches with 64^3 voxels require similar execution time to 48^3 voxels (0.7 second), but obtain a much worse performance with a higher GPU memory usage (we set batch=1 for this case). Although a larger contextual information can be captured, the data are easily overfitted compared to the case with a smaller patch size. As a result, the optimal patch size is considered to be 48^3 voxels.

6.5.5 Ablation Study of DQN Pre-selection

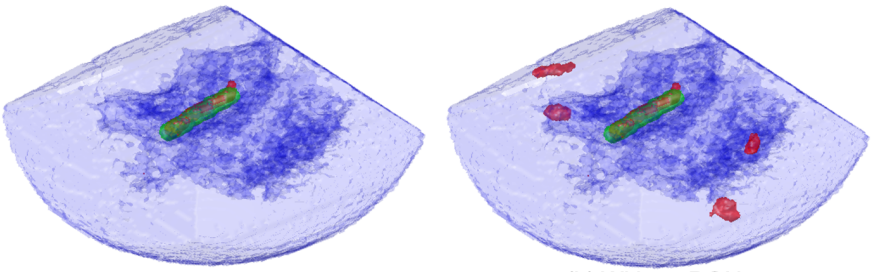
Experimental results with and without DQN pre-selection are summarized in Table 6.6. As can be observed, the DQN pre-selection improves the overall segmentation performance. Example images with and without DQN as the pre-selection are shown in Fig. 6.6, which demonstrates that the DQN pre-selection can avoid outliers outside the instrument region.

Table 6.5 Ablation studies of different patch size for the Dual-UNet. Performance are evaluated by the Dice Score (DSC), Volumetric Similarity (VS), and 95-% Hausdorff Distance (95HD), which are shown in mean±std. The best performances are printed in bold.

Patch Size	*RF-ablation Catheter*		
	DSC %	VS %	95HD (voxels)
32^3 voxels	68.5±7.9	91.2±6.4	3.3±1.9
48^3 voxels	**69.1±7.3**	**92.8±4.6**	**3.0±2.1**
64^3 voxels	66.3±9.6	85.2±11.4	3.7±2.4
Patch Size	*TAVI Guide-wire*		
	DSC %	VS %	95HD (voxels)
32^3 voxels	66.7±9.5	92.8±3.8	**1.6±0.3**
48^3 voxels	**68.6±7.9**	**96.5±3.3**	1.7±0.6
64^3 voxels	62.3±10.0	94.0±3.6	1.9±0.9

Table 6.6 Ablation studies of DQN pre-selection. Segmentation performances are evaluated by the Dice Score (DSC), Volumetric Similarity (VS), and 95-% Hausdorff Distance (95HD), which are shown in mean±std. The best performances are printed in bold.

Patch Size	*RF-ablation Catheter*		
	DSC %	VS %	95HD (voxels)
w/o DQN	44.9±21.3	80.0±13.7	50.0±22.2
w DQN	**69.1±7.3**	**92.8±4.6**	**3.0±2.1**
Patch Size	*TAVI Guide-wire*		
	DSC %	VS %	95HD (voxels)
w/o DQN	57.8±13.9	87.1±5.9	32.3±22.3
w DQN	**68.6±7.9**	**96.5±3.3**	**1.7±0.6**



(a) With DQN pre-selection

(b) Without DQN pre-selection

Figure 6.6 Example volumes of the segmentation results with and without DQN as the pre-selection. Green: ground truth, red: segmentation result.

6.5.6 Generalization Against Different Recording Settings

To further validate the generalization of the proposed method, the trained models of DQN and SSL segmentation steps are directly applied on the *in-vivo* RF-ablation catheter dataset (the models are trained on the *ex-vivo* RF-ablation catheter dataset with 18 labeled images). The proposed DQN successfully detects the catheter with an accuracy of 6.7 ± 2.4 voxels. Although this is slightly worse than the results on the *ex-vivo* RF-ablation catheter dataset, it still can localize the catheter with 100% successful rate, which highlights the generalization of the DQN method. Based on the pre-selected regions from DQN, the UNet segmentation networks are applied to segment the catheter, where the results are summarized in Table 6.7. As can be observed, although the performances of the proposed method are somewhat degraded, the overall performance is still acceptable, since the proposed method produces better results than other state-of-the-art methods.

Table 6.7 Segmentation performance for different methods on the *in-vivo* RF-ablation catheter dataset. Segmentation performances are evaluated by the Dice Score (DSC), Volumetric Similarity (VS), and 95-% Hausdorff Distance (95HD), which are shown in mean±std. The proposed method is printed in bold.

Patch Size	*in-vivo* RF-ablation catheter		
	DSC %	VS %	95HD (voxels)
B-UNet[97]	30.3±20.9	61.7±24.6	10.2±10.4
AdSeg[84]	58.9±5.4	82.3±8.9	7.4±6.5
Π-model[86]	38.5±12.8	60.4±21.2	16.1±12.0
MA-SSL[87]	47.3±14.9	82.2±22.5	8.5±3.5
UA-MT[89]	52.4±5.1	83.9±9.6	11.4±4.8
KD-TS [93]	52.8±7.4	80.1±18.8	10.3±4.0
Proposed	**63.8±10.1**	**86.3±12.9**	**6.2±5.1**

6.6 Discussions and Conclusions

This chapter has proposed a method to achieve annotation-efficient deep learning for instrument segmentation in 3D volumetric US data. A crucial aspect of the proposed method is that it requires less annotation effort yet obtains similar performance as the fully supervised learning methods. Our method avoids laborious and careful annotation work in the 3D dataset by employing the reinforcement learning and semi-supervised learning techniques. These subsequent learning techniques are exploited for training a deep q-network and Dual-UNet for coarse detection and fine segmentation, respectively. This approach significantly reduces the challenges for training an acceptable model with sufficient ground truth, and also compares favorably to a state-of-the-art coarse-to-fine segmentation strategy. Although the voxel-level annotations are

still required for a part of the training images, the amount of required labeled images is much smaller than the number of unlabeled training images, which clearly improves the annotation efficiency of the training framework.

Discussion

Several aspects of our method still need further discussion and argumentation.

(1) *Random noise*: The Monto Carlo method in the Bayesian network introduces random noise during the training, of which the uncertainty estimation requires more training iterations to converge and stabilize. The random noise in the training procedure complicates the convergence of the network and therefore makes it harder for training.

(2) *Complex model*: As stated in Dual-Student [95], the two individual networks can have different complexity and can even have more than three branches to learn the discriminative information. However, due to the size of 3D UNets and computation complexity, it is difficult to achieve these forms on a standard GPU with limited memory. Because of this GPU memory limitation, the proposed method still focuses on a patch-based processing method, which cannot exploit the full contextual information contained in the whole image.

(3) *Model generalization*: As can be observed from the results of the generalization analysis, the proposed method still has a performance degradation when applied to unseen datasets under different recording settings. A recent study [100] has shown that the pre-trained self-supervised feature learning can improve the generalized performance for the segmentation by better network initialization instead of training from scratch, which can be considered as a research direction for improving this generalization capability in future work.

(4) *Statistical confidence*: It is worth to mention that, although the statistical analysis shows the difference between different methods, the number of testing samples is limited and effectively less than 30 images. A larger testing dataset is required for further validation in the future, which indicates a more complex and larger size dataset should be constituted for further validation.

(5) *Clinical validation with in-vivo data*: Finally, artifacts and speckle noise are commonly existing in US imaging, which hampers the segmentation performance. Because of the difficulty of collecting realistic *in-vivo* data, only limited *in-vivo* data is used in our experiments. This *in-vivo* data limitation does not enable a thorough validation of the proposed method, as the noise components are commonly occurring in clinical practice. Further *in-vivo* data collection and validation should be performed in future to fully validate the effectiveness and robustness of the proposed method.

Conclusion

The proposed method contains the following contributions. First, a DQN-driven instrument localization scheme is proposed, which learns to coarsely localize the instrument region in 3D US with an environment-action policy. The application of the deep q-network (DQN) is quite novel for medical instrument localization. The DQN is trained with a simplified annotation, which reduces the annotation effort compared to existing instrument localization methods. Second, a Dual-UNet is proposed for subsequent fine segmentation, which is trained by a novel semi-supervised learning (SSL) framework. The proposed training strategy is novel for medical imaging and able to exploit unlabeled information at both the voxel and contextual level, which leverages abundant unlabeled images for instrument segmentation. Third, despite the data limitations mentioned in the preceding discussion, the proposed method is thoroughly evaluated on multiple challenging datasets compared to the existing methods, using an *ex-vivo* RF-ablation catheter dataset, an *in-vivo* TAVI guide-wire dataset, and an *in-vivo* validation dataset. These validations also include extensive ablation experiments on the proposed method. The proposed method achieves a segmentation performance of about 70% Dice score and approximately a one-second execution time per volume, which offers both high performance and computation efficiency.

In the next chapter (Chapter 7), a novel dimension reduction method will be investigated for medical instrument detection in a 3D US volume. This method can reduce the computation cost and yet improve the detection efficiency on the full-volume processing, which bypasses the compromise using patch-based processing with contextual limitations. Specifically, a dimension-reduction module for general CNNs is designed, which drastically reduces the complexity of the CNN network, while preserving the detection performance to the state-of-the-art methods.

CHAPTER 7

Multi-dimensional CNN for Instrument Detection in 3D US

7.1 Introduction

The previous chapters have introduced several semantic segmentation frameworks to localize and segment a medical instrument in 3D US, which are based on first performing careful voxel-level annotation and then carrying out a second subsequent dense voxel-level classification procedure. Based on the semantic segmentation in 3D US, the target instrument is identified and localized in the complex 3D US data. Although some trade-off techniques are applied to accelerate the overall efficiency, such as e.g. using 2D slice-UNet, tri-planar classification or applying a coarse-to-fine strategy, these results still require one order-of-magnitude improvement to achieve real-time performance. Moreover, regional 3D patch-based methods (in Chapters 5 and 6) cannot fully exploit the 3D contextual information during the information encoding, performed at the encoder side of the 3D UNet. These challenges are mainly due to the huge data size of the 3D images, complex 3D CNN design and limited computing platform capabilities. As a solution for these challenges, this chapter aims at addressing the design of a CNN that can handle the complex 3D operations at the whole image level with high computation efficiency.

7.1.1 Objective and Brief System Outline

The objective of this chapter is to design a CNN to automatically detect and localize the medical instrument in full 3D processing, albeit with reduced computation cost. This method is based on a dimension-reduction procedure for CNN feature maps, which largely reduces the feature-map size, thereby yielding a sim-

pler CNN decoder, going by this reduction from a 3D decoder to a 2D decoder. In addition and in contrast to the previous chapters using dense voxel-level classification, the prediction of the instrument is generated by a trained annotated skeleton model, which reduces the manual annotation efforts for CNN training. The proposed framework and its principal stages are depicted in Fig. 7.1.

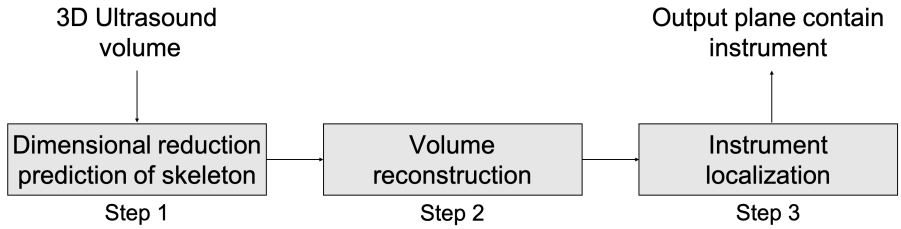


Figure 7.1 Diagram of a dimension-reduced medical instrument detection and localization system. The diagram has a dimension-reduction step to compress a 3D volume into 2D planes, which are used to predict the instrument skeleton in 2D images. Then, these 2D images are exploited to reconstruct the 3D data for instrument detection and localization.

Based on the above concept, the key challenge is to design 3D-2D CNN by a procedure for effective feature-map reduction and associated loss function for the network training. More detailed challenges and corresponding solution directions are elaborated below.

7.1.2 Challenges for Dimension-reduction CNN

In order to localize the instrument in 3D US by processing whole volumes with high efficiency and reduced annotation effort, the following considerations have to be addressed to achieve the correct detection and localization of the target.

- *Process the whole 3D US volume efficiently:* To process the whole 3D US data volume in a single GPU hardware with limited memory capacity, a common practice is to reduce the network design for 3D CNN, such as reducing the number of filters or the number of scales. However, these approaches are still focusing on the full-3D format, which is still complex for processing and requires high GPU cost (because a common UNet structure has a mirrored decoder with high computational cost). Instead of this, an alternative common choice in literature is to reduce the size of the 3D decoder, which is nevertheless still processing complex 3D information.
- *Reduce the annotation effort:* The common practice of dense classification for segmentation is to train a CNN with voxel-level data based on careful voxel-by-voxel annotations, which is rather laborious work. Moreover, the extremely imbalanced class distribution of the skeleton annotation in 3D images increases the challenges for training a successful network.

To reduce the computation load and improve the efficiency, one possible direction of the CNN design is to apply a feature-map-reduction procedure, which compresses the 3D features to a 2D format. As a result, the feature maps are processed in 2D format for faster processing, yielding to a lower computational cost than common 3D CNN processing. In addition, the annotation effort can be reduced by considering the simple skeleton annotation than dense voxel-level annotation, which can be obtained by finding the instrument tip and tail of the straight instrument and then connecting them together via a line piece. To make use of this type of annotation, the loss function should be carefully considered by exploiting voxel and contextual level information.

Based on the above challenges and directions, the followed approach of this chapter is to exploit a dimension-reduction module to 3D-2D CNN and its loss functions, in order to exploit the skeleton-based annotation. Based on these design steps, the medical instrument can be localized in complex 3D US with high efficiency and lower annotation effort.

The sequel of this chapter is organized in the following way. Section 7.2 summarizes the related work in this field, detailing efficient medical instrument detection methods based on machine learning. Section 7.3 describes the details of the proposed method. Sections 7.4 and 7.5 demonstrate and present the considered dataset, implementation details and experimental results. Finally, Section 7.6 concludes the chapter and presents some discussions on possible refinements in the future.

7.2 Related Work on Representative Literature by Machine Learning

The case study of finding an instrument in 3D US data is only scarcely explored in literature because it is difficult and the data is noisy. Therefore, this related work section is based on a selection of representative literature, which describes instrument detection by machine learning approaches.

Recently, Pourtaherian *et al.* [28] have studied instrument detection algorithms in 3D US. Their method distinguishes the candidate instrument-like voxels by analyzing these voxels with Gabor features. These features introduce more discriminating information on the distribution of local orientations. After the voxel-based classification, the instrument is localized with a pre-defined semantic model. Pourtaherian *et al.* performed an experiment on catheter detection in an *in-vitro* dataset, which implicitly shows the necessity for further validation on *ex-vivo* or *in-vivo* datasets. Yang *et al.* [65] employed more discriminating features for a supervised learning method with a multi-scale approach, to capture more contextual information. Although the authors achieved satisfying performance on different *ex-vivo* and *in-vivo* datasets, the limited capacity of handcrafted features leads to outliers after segmentation, which requests a complex model-fitting or post-processing to finally detect the catheter. Furthermore, handcrafted fea-

ture design requires experience and effort, which is then gradually replaced by deep learning methods, e.g. CNN methods.

CNNs have achieved a significant success in different recognition tasks in the medical imaging area [9]. Researchers have proposed medical instrument detection methods using this deep learning approach in many different applications and modalities. For example, tri-planar CNN methods for voxel-wise classification were already introduced for instrument segmentation [66, 19]. However, these approaches require the network to iteratively predict all the voxels in 3D US, leading to a high computation cost, which is therefore not suitable for real-time applications. Although Yang *et al.* [19] have proposed a pre-filtering-based acceleration method, the 10-sec. prediction time is still too long for a real-time application, which is typically around 5-10 frames per second in 3D US-guided operation for most clinical scenarios. Slice-based FCN [66] was proposed to segment an instrument in 3D US, by decomposing the volume into adjacent slices using a transfer-learned 2D FCN. Although the authors' method achieved impressive segmentation results for instrument segmentation, the capacity of 3D contextual information extraction is limited due to the slice-based strategy. To overcome the 3D information compromise, Yang *et al.* [69] proposed a patch-based 3D UNet to segment a cardiac catheter in 3D US, which achieved the satisfied performance. Nevertheless, iterative patch-based operations in a 3D volume hamper the interpretation of the contextual and semantic information of the whole image. Instead of patch-based approaches, Arif *et al.* [78] proposed a full-3D CNN method for instrument segmentation with 3D UNet as a backbone, which shows an impressive performance on a challenging dataset. However, the true 3D operations at both the encoder and decoder sides severely complicate the network structure and are typically constrained by the bounded GPU memory size.

Based on the above literature analysis and considering the network complexity, full-image information usage and time efficiency, this chapter aims at detecting the instrument by a multi-dimensional mixture CNN. The considered method can detect the medical instrument in 3D volumetric data at full-image level, while achieving high detection efficiency. In addition, a specifically designed loss function is employed to improve the annotation efficiency and preserves the detection accuracy.

7.3 Proposed Method

This section proposes a framework to detect a medical instrument in 3D volumetric B-mode ultrasound images using a Multi-dimensional Mixture Network (MixDNet). The proposed framework is depicted in Fig. 7.2, in which the input 3D volume is processed by MixDNet. The outputs of the MixDNet are estimated instrument skeletons, which are projected along the principal volumetric axes.

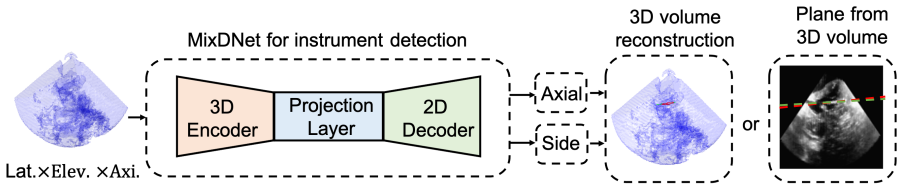


Figure 7.2 Framework of the proposed medical instrument detection method (for testing), where the green dashed line in the output image is the ground truth, while the red dashed line indicates the detected instrument (or its principal orientation axis). The 3D ultrasound volume is processed by MixDNet to generate the instrument prediction on the projected planes along two of three available principal axial, lateral and elevation directions. Based on the prediction on the projected planes, the detected instrument is reconstructed and visualized in a 3D volume, or visualized in a 2D plane by slice selection.

Because this concept unfolds minimally over two axes, multiple skeletons are used. Based on the estimated skeletons oriented in two of the three available axial, lateral and elevation directions, the instrument is detected in 3D space, and visualized in a 2D plane or 3D space. In addition, the voxel-level and image-level loss functions are proposed to exploit the skeleton-based annotations, which successfully train the network for the instrument localization.

7.3.1 Construction of MixDNet

The proposed MixDNet is depicted in Fig. 7.3. In contrast with a standard encoder-decoder architecture like 3D UNet [42] or Feature Pyramid Nets [72], we propose a hybrid 2D-3D dimension architecture, which consists of a 3D encoder, a 3D-to-2D information-reduction layer and a 2D decoder. For the input 3D US volume, a 3D encoder is applied for high-level feature extraction. The encoded information is processed by a projection layer to extract the most discriminating information along the principal axes, which extracts the relevant information through the dimensions and channels. Then, the compressed features are processed by a 2D decoder, which decodes the projected features to generate the instrument skeleton along the principal axes. More specifically, there are two individual branches of dimension reduction to extract the dimension information along the axial or side direction simultaneously, i.e. axial and lateral (or elevation) direction (following the nature of the US cone) in Fig. 7.3. To reduce the complexity of the network, reduction blocks at the same image size are shared. Moreover, the 2D decoder parameters are shared in different directions.

A. 3D Encoder:
Considering the limited GPU memory size of hardware and the complex 3D convolutional kernels, we have designed a compact 3D encoder to avoid GPU memory overflow and network overfitting. Specifically, the 3D encoder includes

Figure 7.3 Illustration of the proposed Multi-dimensional Mixture Network (MixDNet). The network consists of three main parts: (a) information Encoder in 3D space; (b) dimension reduction along the axial and lateral (side) axes, note that each image scale has its own reduction block; (c) information Decoder in 2D space, which generates the projected predictions of the instrument in 2D space alone the two directions. Based on the prediction from the 2D decoder, a multi-level loss is applied to supervise the network. (Conv is convolutional operation, IN is instance normalization, ReLU is rectangular activation function, Deconv is de-convolutional operation.)

Figure 7.4 Detailed flow of the dimension-reduction module in the middle of Fig. 7.3. The input feature map is processed by three different branches along one of three axes (along D axis in this figure, D is an example): Max-pooling, Avg-pooling and Convolution operations. The feature maps processed by Max-pooing and Avg-pooling are concatenated and processed by a convolution operation, which is then fused with the Convolution branch by summation. The block color is the operation indicated at the right side of the diagram and applied onto the feature maps, while the size of the result is depicted below the applied operation block.

a stack of 3D convolutional and Max-pooling layers. After each convolutional layer, ReLU and and Instance Normalization (IN) layers are inserted to accelerate the convergence. Specifically, the IN normalizes the input samples at single-image level, instead of conventional batch normalization working on the set of batch images, which is less sensitive to the image variations of the input batch [101].

B. 3D-to-2D reduction module:
As for the 3D-to-2D reduction module, it is a spatial channel-based attentional module, which can extract the most relevant information along a specific dimension, while reducing the size of feature mappings. More specifically, we design a dimension-reduction (or dimension-projection) block, which is based on three different operations along one of three principal axes, as is depicted in Fig. 7.4. The block extracts the first-order statistics along the interested axis, by maximizing and averaging all possible discriminative information. Then, the output tensors are concatenated and processed by a convolution operation to reduce the channel size. Meanwhile, a series of convolution operations are applied on the input feature maps. First, the channel information is compressed, which is then followed by two convolutions to obtain the dimension-reduced feature map. Finally, the obtained feature maps are accumulated to obtain the final result. This approach is similar to an attention operation, dealing with spatial and channel information, but consisting of different ways to summarize them. As for the Avg-pooling-based branch, it can summarize all information along one dimension, while the Max-pooling operation focuses only on the maximized signal responses and ignore the minor information. The convolution-based branch can summarize the channel-based information, which acts as a compensation for the above non-parametric approaches. As a consequence, the information can be summarized properly and spatial dimensions are reduced. As shown in Fig. 7.3, for each feature scale, a scale-specific dimension-reduction block is designed to fit the convolutional channels. In this paper, the side view is chosen to be the lateral direction, because of the fixed pose between instrument and tissue in the datasets.

C. 2D Decoder:
Based on the reduced feature maps from the dimension-reduction blocks, a 2D decoder is designed to formulate the output, which describes the instrument skeletons along the axes. The decoder consists of 2D convolutional layers, followed by a ReLU and Instance Normalization. De-convolutional layers are applied to upsample the feature maps. More details of the 2D decoder network are depicted in Fig. 7.3.

The motivation of the proposed structure is explained in follows. (1) The conventional 3D U-Net structure requests a complex encoder and decoder, which

unfortunately increases the memory usage for limited GPU hardware. Moreover, with the input volume size increased, GPU memory can easily overflow with a larger mini-batch size, which therefore increases the difficulties to train the network. (2) In the proposed structure, the decoder part is simplified from 3D space to 2D space, which is based on the prior knowledge of instrument shapes in 3D images and to reduce the decoder redundancy. More crucially, the dimension-reduced outputs drastically decrease the class-imbalance challenges, which makes it easier to train the network. To overcome these challenges and apply the detection on a whole data volume, we design the proposed structure to facilitate efficient detection of the instrument.

7.3.2 Multi-level Loss Function

The input of the proposed network is a full-3D B-mode volumetric image, while the output is in 2D planes, indicating the instrument skeleton projection along the axes, see Fig. 7.5 as an example to generate a prediction for one dimension. To guide the MixDNet to learn and generate the correct skeletons in 2D projected images, we design a multi-level loss function, which is generally formulated by

$$\mathcal{L}(\hat{Y}, Y) = \alpha \mathcal{L}_{\text{pixel}}(\hat{Y}, Y) + \beta \mathcal{L}_{\text{image}}(\hat{Y}, Y), \tag{7.1}$$

where the $\hat{Y}$ is the network prediction and Y is ground truth. Loss component $\mathcal{L}_{\text{pixel}}$ focuses on the prediction of the projected instrument skeleton in a 2D plane at the pixel level, while the loss component $\mathcal{L}_{\text{image}}$ concentrates on a high-level description of the skeleton at the 2D image level. Parameters α and β are weight parameters to balance the individual loss functions. More specifically, component $\mathcal{L}_{\text{pixel}}$ is defined as a weighted binary cross entropy (BCE), specified by

$$\mathcal{L}_{\text{pixel}}(\hat{Y}, Y) = -\sum_{j=1}^{N} w_j^i y_j^i \log(\hat{y_j}^i) - \sum_{j=1}^{N} w_j^n y_j^n \log(\hat{y_j}^n), \tag{7.2}$$

where N denotes the number of pixels for each 2D prediction or ground-truth image, superscript i represents the instrument skeleton pixels and superscript n denotes the group of the non-instrument pixels. The class weight parameter w is a hyperparameter to control the weight between two different classes, which is employed because of the extreme imbalance between classes in the ground-truth images. Moreover, deep supervisions [102] are employed at the 2D decoding blocks with weight set to 0.1, which are applied after dimension-reduction operations in Fig. 7.3.

Besides the pixel-level loss function, we also define an image-level loss function, which enforces the MixDNet to learn high-level information to properly match the predictions and ground truth in 2D planes. As described in Fig. 7.5, the constructed projection images, together with corresponding ground-truth images, are processed by a shared contextual encoder (CE) to generate the high-level descriptor [69], which can describe the input images in a latent space. For

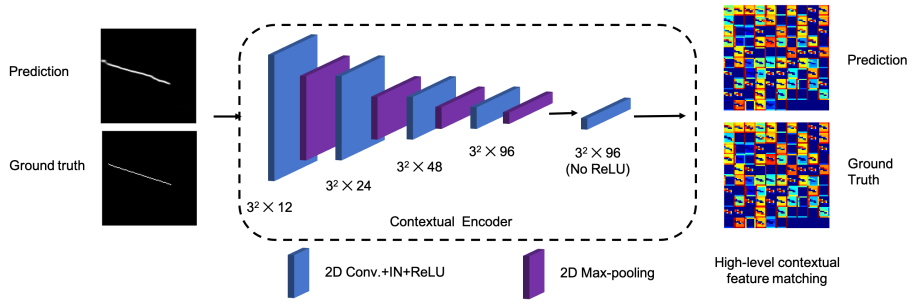


Figure 7.5 Overview of the contextual encoder for image-level loss function. The ground truth and prediction are processed by an encoder to generate high-level feature maps, which are matched to measure the high-level similarity in an encoded feature space. The high-level feature maps for ground truth and prediction are represented by heat maps as shown at the right side. (IN: instance normalization)

the description of each view, its corresponding loss function is defined as the distance between the descriptor of the prediction and its corresponding ground truth, leading to

$$\mathcal{L}_{\text{image}}(\hat{Y}, Y) = ||\text{CE}(\hat{Y}) - \text{CE}(Y)||_2, \tag{7.3}$$

where function CE$(\cdot)$ denotes the contextual encoding net for latent space projection, $||\cdot||_2$ stands for the norm-2 distance. As a consequence, Eqn. (7.3) also holds for any other predictions along different axes. It is worth to mention that the CE$(\cdot)$ is projection function, which projects complex information into a latent high-level space. The loss function $\mathcal{L}_{\text{image}}$ measures the similarity between two images in the latent space, and therefore can be sensitive to the shape and location differences at the contextual view. Based on the preceding definitions for $\mathcal{L}_{\text{pixel}}$ and $\mathcal{L}_{\text{image}}$, the overall loss function along the two individual axes can be formulated as a weighted summation, based on the predictions and ground-truth pairs, i.e. $\{\hat{Y}_{\text{axial}}, Y_{\text{axial}}\}$ and $\{\hat{Y}_{\text{side}}, Y_{\text{side}}\}$.

7.3.3 Instrument Detection based on 2D Projections

The MixDNet generates the estimated instrument skeletons along different axes as 2D predicted images, i.e. $I_{\text{2D-axial}}$ and $I_{\text{2D-side}}$, as shown in Fig. 7.3. Based on the 2D predictions, the instrument in the 3D volume is obtained by replicating the 2D images along the feature map reduction-directions. As a result, the imaged 3D volume I_{3D} with the detected instrument is obtained by

$$I_{\text{3D}} = \text{Rep}(I_{\text{2D-axial}}, \theta_{\text{axial}}) + \text{Rep}(I_{\text{2D-side}}, \theta_{\text{side}}), \tag{7.4}$$

where the Rep$(\cdot, \theta)$ represents the replication function of the 2D prediction along the specific direction θ, such as e.g. θ_{side} for indicating a side-view direction.

Based on the reconstructed volume, a simple threshold and RANSAC model-fitting are applied on the sparse volume to find the instrument. Another choice to detect the instrument from 2D planes is plane extraction, which is inspired by clinical usage. In practice, sonographers prefer to automatically visualize the plane containing the instrument, i.e. the instrument axis is in-plane, which can avoid a complex plane searching to find the correct instrument plane in 3D volumetric data. Exploiting the natural property of ultrasound imaging, i.e. the propagation of sound waves is always taking place along the axial direction of the ultrasound probe, the instrument detection can be formulated by two steps: (1) extract the plane containing the instrument along the axial direction of the probe, (2) based on the prediction alongside the side-view axis, the instrument can be extracted from the 2D plane that is spanned between the axial and side-view axes. These steps are illustrated in Fig. 7.6 with an example.

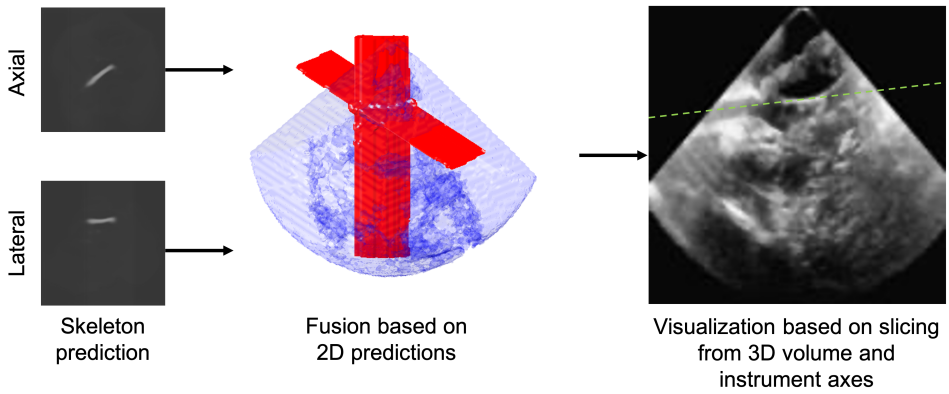


Figure 7.6 Based on the skeleton predictions in 2D planes, the 3D instrument skeleton can be obtained by replicating planes. Therefore, the 2D B-mode slice containing the instrument is obtained by model-fitting and axis estimation along the US cone (the green dot lines are the axis of the detected instrument in slice containing the instrument).

7.4 Experiments

7.4.1 Datasets and Preprocessing

To validate the proposed instrument detection method, we have collected two different datasets for different ultrasound-guided operation tasks: RF-ablation operation for cardiac intervention and needle-based interventions for regional anesthesia.

RF-ablation catheter dataset: We have collected 94 3D cardiac US data volumes from eight porcine hearts. During the recording, the heart was placed in a water tank with a RF-ablation catheter for Electrophysiology (with a diameter of 2.3–3.3 mm) inside the heart chambers. The phase-array US probe (X7-2t with 2,500 elements by Philips Medical Systems, Best, the Netherlands) was placed next to the intended chambers to capture the images containing the catheter, which was monitored by a US console (EPIQ 7 by Philips). The obtained volumetric images are re-sampled to create a volume size of $160 \times 160 \times 160$ voxels (where padding is applied at the boundary to make the volume such that it has equal sizes in each direction, called isotropic), which leads to a voxel size ranging from 0.3–0.8 mm. An example 2D slicing image is portrayed by Fig. 7.7.

Anesthesia needle dataset: A dataset based on needle-based intervention is collected by a motorized VL13-5 linear-array (VL13-5 with 192 elements by Philips Medical Systems, Best, the Netherlands) from chicken breast, which was monitored by a US console (iU22 xMATRIX by Philips). The dataset includes 20 volumetric images with two different types of needles: 17G (diameter of 1.47 mm) and 22G (diameter of 0.72 mm). For each type of needle, 10 images are collected. To ensure the image independence of tissue appearance, needles are inserted into different locations of the chicken breast. The images are isotropically re-sampled to obtain a voxel size of 0.3 mm, which leads to a volume size of $128 \times 128 \times 128$ voxels. An example of a 2D slicing image is shown in Fig. 7.7. In contrast with the catheter dataset, the needle has a clear contrast to the surrounding tissues because of the different material and medium.

Annotation: The ground-truth binary skeleton mask for both datasets are generated by connecting the annotated instrument endpoints, which are annotated by a technical expert. This annotation strategy can reduce the annotation difficulty and effort in 3D US volumetric data, compared to voxel-level annotation [28, 78]. Although the author of [78] proposed to use a dilation operation to instrument voxels, the deformation and blurry boundaries of the instrument lead to voxel-based category uncertainties in the automatically generated ground truth for network learning. However, with skeleton-based training giving a sparse annotation only, it is more challenging for the network to learn the semantic information when compared to the dense annotation-based instrument segmentation.

7.4.2 Implementation Details and Training Process

Considering the limited dataset and GPU memory (11 GB for a 1080Ti GPU), the proposed MixDNet has 12 convolutional kernels in the first layer, which are gradually doubled after each max-pooling operation, except for the deepest level where two convolutional layers are applied with kernel size $3^3 \times 96 \times 96$ voxels. As a consequence, the MixDNet has a number of hyper-parameters that is

approximately 1.1 Million, which is smaller than a standard 3D UNet or similar architectures (commonly about 5-10 Million parameters).

We have trained the proposed network using a stochastic gradient descent update with the Adam optimizer. As for the catheter dataset, the initial learning rate is set to 0.001 for a mini-batch size equal to 4. The learning rate is reduced for every 100 epochs by a factor of 0.1 and the training is terminated after 200 epochs. During the experiment, 64 volumes are randomly selected as training data and 30 volumes are used as testing images. With respect to the needle dataset, due to the limited amount of images for training, fivefold cross-validation is applied. This means that 20 images are randomly divided in to two parts: 16 images are used for training, while the rest is applied for testing. The procedure is repeated five times to obtain the overall performance on the average. The initial learning rate is set to be 0.001 for a mini-batch with size 8 and is terminated after 500 epochs. During the training, mirroring is applied on elevation/lateral directions and rotation along the axial direction to preserve the global shape information of the US data cone. Data-shifting operations within 16 pixels along the elevation or lateral directions are randomly applied on-the-fly during the training stage. Moreover, intensity jittering, and image resizing with a scale range of 0.8-1.2 are also randomly applied. During the training, the total number of observed images for the network are 12,800 and 8,000, for the catheter dataset and the needle dataset after data augmentation, respectively. Data augmentation facilitates the network to learn more invariant information of the dataset and avoid overfitting [103]. The parameters α and β are empirically selected as 1.0 and 0.01, respectively. The proposed method is implemented in Python 3.7 with TensorFlow 1.10.

7.4.3 Evaluation Metrics

Since the ground truth of our datasets is an instrument skeleton indicated by a line having a diameter of one voxel, standard evaluation metrics such as the Dice score, are not feasible to evaluate the proposed method. Since the ground truth of the instrument skeleton is a line-based prediction with one voxel diameter in the 3D space, the Average Hausdorff Distance (AHD) is considered as an evaluation metric ([104]), because it is more sensitive to voxel mismatch between annotated voxels and the prediction voxels. Moreover, the AHD is less sensitive to outliers than a standard Hausdorff Distance, since the detected skeleton of the instrument is sometimes thicker than the ground truth. Moreover, the axis localization error is defined with two different metrics: the endpoints error (EE) and the orientation error (OE), which are introduced in Chapter 2. The EE is defined as the maximum distance of two endpoints on the instrument skeleton from the ground truth, i.e. tip and tail point of the ground truth, to the instrument axis obtained from 3D reconstruction (conservative measure). Similarly, the OE is defined as the angle difference between the detected instrument axis to the ground-truth skeleton.

Finally, the execution time for computing the prediction per volume is evaluated in the experiments.

7.5 Results

This section is organized in two parts. First, ablation studies are performed of different loss types and effectiveness of going from 3D to the 2D using the dimension-reduction module. Second, a performance comparison with different state-of-the-art (SOTA) medical instrument detection methods in 3D US is evaluated. In the experiments, we have considered a RANSAC-based model-fitting as a post-processing step to detect the instrument, which introduces randomness in the results. We have conducted the detection session five times and have chosen the worst-case results of each method. The proposed method and the SOTA methods are compared based on these fivefold experiments.

7.5.1 Ablation Study

Two ablation studies have been performed to validate the proposed method. First, the proposed method with different types of loss functions is compared, i.e. only standard BCE loss, only with pixel-level loss (weighted BCE), only with image-level loss and the hybrid loss (see Eqn. (7.1)), as proposed with MixDNet. Second, with the proposed multi-level loss function, we have performed another ablation study on the effectiveness of different 2D dimension-reduction methods, which are discussed in Section 7.3 for each individual branch and their ensemble, i.e. Max-pooling, Avg-pooling, Convolution, Concatenation of Max-pooling and Avg-pooling (denoted as Max+Avg), and the proposed method. Meanwhile, we also consider the comparison between a true 3D network and the proposed MixDNet, i.e. excluding the dimension-reduction blocks and using full-3D convolutions in the decoder (due to the limited GPU memory for full-volume operations on a 1080Ti with 11 GB memory, we apply a mini-batch size equal to unity for the full-3D model). The results of all above ablation studies are listed in Table 7.1 and Table 7.2.

From the results in Table 7.1, it can be observed that the multi-level loss can provide a higher performance with both datasets. However, the network cannot learn meaningful semantic information with only considering image-level loss. Because the randomly initialized contextual encoder cannot generate a correct feature representation of both ground truth and prediction from the untrained network, this fails to guide and constrain the segmentation network to learn meaningful knowledge after training iterations. As a consequence, the image-level loss can be only considered as complementary for the pixel-level loss. It is worth to mention that the BCE loss can lead to a failure in detection, due to extremely imbalanced classes, which underestimate the detection and generate empty predictions. To further validate the proposed method, a paired t-test is

Table 7.1 Ablation study on different loss types for two datasets, using the proposed network, which are evaluated by the Average Hausdorff Distance (AHD), Endpoints Error (EE) and Orientation Error (OE) using mean±std. The term 'Failed' means that we have failed to obtain the results. The $*$ means a statistical difference under the metric with a significance level of 0.05.

Loss	Catheter		
	AHD (voxel)	EE (voxel)	OE (degree)
BCE	7.2 ± 8.6	7.5 ± 19.0	16.1 ± 19.2
Pixel-level loss	2.6 ± 1.1	2.6 ± 0.7	10.5 ± 6.5
Image-level loss	Failed	Failed	Failed
Proposed	2.4 ± 0.9	2.3 ± 0.5	$7.3 \pm 2.1^*$
Loss	Needle		
	AHD (voxel)	EE (voxel)	OE (degree)
BCE	16.4 ± 19.6	11.4 ± 19.8	23.3 ± 35.7
Pixel-level loss	5.2 ± 6.7	2.9 ± 1.0	5.5 ± 3.7
Image-level loss	Failed	Failed	Failed
Proposed	$3.2 \pm 2.2^*$	2.5 ± 0.6	5.0 ± 4.9

performed with a significance level of 0.05, based on the maximum value of 5 execution cycles rather than multiple comparisons (same as the rest t-tests). As shown in Table 7.1, the proposed multi-level loss does not offer a statistically significant performance improvement in most metrics. However, we have observed it performs much better than the pixel-level loss and the standard BCE loss with the OE metric on the catheter dataset. On the needle dataset, the proposed multi-level loss performs better than others with the AHD metric.

As shown in Table 7.2, for catheter detection, the proposed dimension reduction significantly outperforms other modules, except for Max+Avg. Compared to the Max+Avg approach, the proposed technique does not show statistically different results with the AHD and EE metrics, but performs better with the OE metric. For needle detection, most of the examined dimension-reduction modules do not show statistically significant differences, only 3D UNet performs much worse. Considering the needle dataset is with very limited data, more investigations need to be performed on a larger dataset in the near future.

Example feature maps after the proposed dimension-reduction blocks are shown in Fig. 7.7 for two different datasets. As can be observed, the feature maps represent discriminating information from local texture to high-level locations, ranging from Shallow scale to Deep scale. By comparing the instrument areas to the B-mode slice, the instrument can be found with a high contrast in feature maps. However, when it comes to black regions in B-mode, e.g. empty areas, the corresponding feature maps look rather noisy, which is because they are obtained by compressing the non-instrument information. This figure demonstrates that the proposed block can extract the discriminating information along the specific

Table 7.2 Ablation study on different dimension-reduction modules for two different datasets, which are evaluated by the Average Hausdorff Distance (AHD), Endpoints Error (EE) and Orientation Error (OE) using mean±std. The symbol $*$ stands for a statistical difference using the metric with a significance level of 0.05.

Method	Catheter		
	AHD (voxel)	EE (voxel)	OE (degree)
3D UNet	5.9 ± 6.7	5.3 ± 6.0	26.2 ± 26.2
Max-pooling	3.2 ± 2.1	2.8 ± 1.0	9.4 ± 4.5
Avg-pooling	3.5 ± 6.6	2.7 ± 2.7	8.8 ± 4.4
Convolution	3.1 ± 1.3	3.0 ± 1.0	11.0 ± 5.6
Max+Avg	2.7 ± 1.3	2.4 ± 0.6	9.3 ± 4.2
Proposed	2.4 ± 0.9	2.3 ± 0.5	$7.3 \pm 2.1^*$
Method	Needle		
	AHD (voxel)	EE (voxel)	OE (degree)
3D UNet	19.2 ± 20.2	8.9 ± 15.2	11.6 ± 11.3
Max-pooling	4.5 ± 7.0	2.5 ± 0.9	5.2 ± 3.9
Avg-pooling	4.3 ± 5.0	2.7 ± 0.7	5.1 ± 3.5
Convolution	4.3 ± 4.6	2.8 ± 0.7	5.4 ± 3.2
Max+Avg	3.6 ± 2.4	2.7 ± 1.0	5.3 ± 3.7
Proposed	3.2 ± 2.2	2.5 ± 0.6	5.0 ± 4.9

Table 7.3 Performance comparisons with SOTA methods for catheter detection (using the catheter dataset), which are evaluated by Average Hausdorff Distance (AHD), Endpoints Error (EE), Orientation Error (OE) and execution time of the inference stage.

Method	Catheter			
	AHD (voxel)	EE (voxel)	OE (degree)	Time (sec.)
Handcrafted [28, 65]	6.6 ± 10.4	6.5 ± 7.8	17.0 ± 17.9	> 600
VOI-PatchCNN [67]	2.8 ± 1.5	2.8 ± 1.2	8.2 ± 3.2	~ 10
SliceFCN [67]	4.1 ± 5.3	3.7 ± 4.9	9.2 ± 5.3	~ 1.0
Pyramid-UNet [69]	2.6 ± 1.9	2.4 ± 1.0	7.5 ± 3.3	~ 48.0
Proposed	2.4 ± 0.9	2.3 ± 0.5	7.3 ± 2.1	~ 0.12

feature map axis. With further operations, the instrument skeleton is predicted in 2D output images, of which examples are shown in Fig. 7.6.

7.5.2 Performance Comparison with SOTA

We have compared the proposed method to many different state-of-the-art (SOTA) medical instrument detection methods in 3D US with respect to the metrics of AHD, EE, OE and inference time. For fair comparison, we have implemented and evaluated all SOTA methods on our datasets with *2 voxels dilation* for

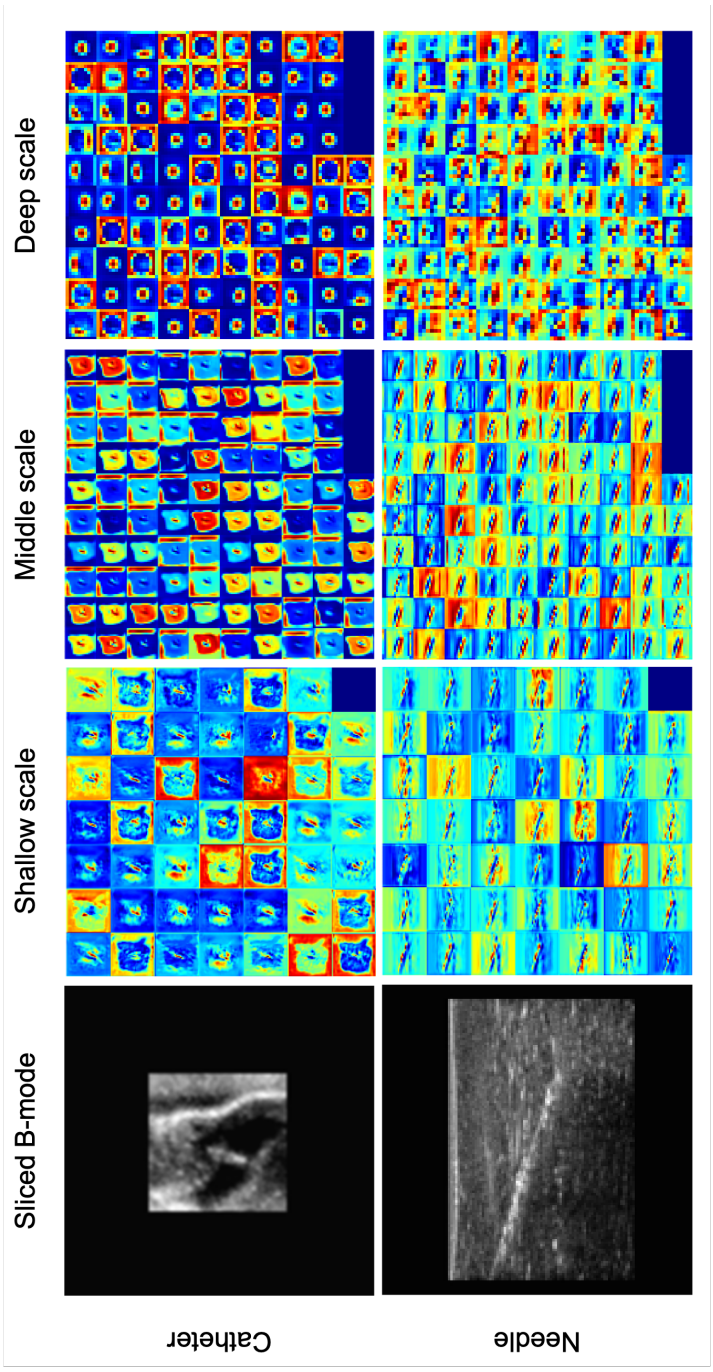


Figure 7.7 Heat maps of feature activation after the dimension-reduction blocks at shallow, middle and deep scales of the network, which are depicted in Fig. 7.3 as summation symbols. Feature maps are re-scaled for visualization. The results show that the dimension-reduction blocks extract the most instrument-relevant information from 3D tensors. The top-row images are from the catheter dataset and the bottom-row images are from the needle data.

Table 7.4 Performance comparisons with SOTA methods for needle detection (using the needle dataset), which are evaluated by Average Hausdorff Distance (AHD), Endpoints Error (EE), Orientation Error (OE) and execution time of inference stage.

Method	Needle			
	AHD (voxel)	EE (voxel)	OE (degree)	Time (sec.)
Handcrafted [28, 65]	7.8 ± 8.7	7.6 ± 10.3	8.3 ± 9.0	> 120
PatchCNN [66]	4.9 ± 7.3	2.8 ± 0.7	5.9 ± 4.7	> 240
ShareFCN [66]	5.9 ± 8.3	2.6 ± 0.7	5.1 ± 2.1	~ 1.0
3D UNet [78]	19.2 ± 20.2	8.9 ± 15.2	11.6 ± 11.3	~ 0.2
Proposed	3.2 ± 2.2	2.5 ± 0.6	5.0 ± 4.9	~ 0.06

skeleton annotation, instead of voxel-accurate annotation, as suggested by [78]. This approach is adopted, since we have failed to obtain successful predictions with our data, due to the extremely imbalanced skeleton annotations for CNN-based segmentation methods, or the occurrence of too much false positives after the handcrafted feature classification method. This method failure in obtaining predictions also indicates that the SOTA segmentation methods are not feasible for skeleton-based detection in 3D space, while the proposed method can handle this and is feasible. The detailed results are listed in Table 7.3 and Table 7.4.

For catheter detection in 3D US, the proposed method is compared in Table 7.3 with several catheter detection methods in 3D US volumetric data, based on multi-scale and definition features with an AdaBoost classifier (handcrafted from [28, 65]), voxel-of-interest-based patch-wise CNN (VOI-PatchCNN from [19]), slice-based 2D FCN for 3D US (SliceFCN from [67]) and Pyramid UNet for patch-based segmentation ([69]). Unfortunately, no other works are available for catheters by machine learning methods. From the results, the proposed method achieves a higher detection accuracy with better efficiency. It should be noticed that the proposed method is solely based on the annotated skeleton, instead of voxel-level annotation, which provides less information than the SOTA methods. However, experimental results show that our method achieves still a higher performance even with more challenging training conditions. The proposed approach is therefore more challenging than the reported SOTA references. Moreover, when considering the execution-time efficiency, the proposed MixD-Net architecture achieves more than 8 times faster inference. The obtained fast and accurate results present a promising performance for the requirement of real-time applications. It is worth to mention that although the patch-based UNet can be accelerated by patch-of-interest pre-selection, which achieves about 1-second execution time per volume [105], the proposed method is still faster than the coarse-to-fine segmentation approach.

For the needle detection in 3D US, the proposed method is compared in Table 7.4 with several needle detection methods in 3D US volumetric data, based on multi-scale and definition features with an AdaBoost classifier (handcrafted from

[28, 65]), Patch-wise CNN ([66]), ShareFCN ([66]) and 3D UNet ([78]). From the results, the proposed method achieves a higher performance than state-of-the-art methods with better efficiency. It is mentioned here that because of the different tasks and datasets, we could not obtain the results based on the structure of [78]. This is explained by the shallow network in the experiment with just 4 kernels in the first layer, which cannot handle our complex dataset. In contrast with this, a more complex and shallower 3D UNet is mentioned and referred to in Table 7.2 and Table 7.4, which obtains a much worse performance due to the extreme class imbalance and skeleton annotation.

Moreover, paired t-tests are performed based on the metric of the endpoints-error (EE) with a significance level of 0.05. For the catheter dataset, the proposed method is statistically better than handcrafted, VOI-PatchCNN and SliceFCN methods, while there is no statistical difference between the Pyramid-UNet and the proposed method. For the needle dataset, the proposed method performs statistically better than handcrafted, PatchCNN and 3D UNet methods, while there is no statistical difference between the ShareFCN and the proposed method. However, the proposed method achieves much faster inference time than these state-of-the-art methods.

In terms of efficiency in execution time, the inference time is about 0.12 sec. per volume for the catheter dataset and about 0.06 sec. per volume for the needle dataset on a GTX 1080Ti GPU. It is important to remark that our method achieves about 6 and 3 seconds inference time for catheter and needle datasets, respectively, on a standard CPU (2.4-GHz quadcore 8th-generation i5 processor). This efficiency improvement clearly indicates lower hardware requirements for the intended real-time application.

7.6 Discussion and Conclusion

This chapter has proposed a highly efficient and accurate medical instrument detection algorithm for usage in 3D volumetric US data by employing deep learning techniques. This efficiency and good performance is achieved by proposing a novel dimension-reduction module to reduce CNN complexity by reducing 3D feature maps to a 2D compressed format along principal axes of interest. This algorithm is trained by a multi-level loss function with skeleton-based annotation, which ensures an accurate detection result. A important aspect of the proposed method is that despite its high-speed inference time compared to state-of-the-art methods, it still achieves comparable detection accuracy. The proposed dimension-reduction technique for high-speed calculation avoids overfitting and reduces the computational cost. In addition, the skeleton-based annotation avoids accurate voxel-level ground truth, which clearly improves the annotation efficiency.

Discussion

To apply our method in real-time applications, there are still a few discussion points, as listed below.

(1) *Clinical validation with in-vivo data*: Further validation on *in-vivo* datasets is still needed to support the clinical value for the proposed method, which is considered as future work. More specifically, there are some limitations for both validated datasets. Since the US images during the procedures are different than those in our datasets, these differences can introduce and pose challenges. As for the catheter dataset, the isolated hearts were placed in a water tank. The heart chambers were filled with water to mimic the intervention procedure, although it should be blood for a real heart. This discrepancy can lead to different image noise and contrast levels between the instrument and the surrounding background. For the needle dataset, the real procedure would introduce more complex subcutaneous tissue and vessel structures, which complicate processing of the acquired images. As a consequence, to make the proposed method suitable for real clinical practice, more studies should be conducted on *in-vivo* data in the future.

(2) *Video data*: The proposed method is applied to static images instead of 3D US video, which is commonly used during interventions. Therefore, a further study in 4D US is necessary in the future, which exploits the temporal information and possibly increases the detection efficiency. However, this 4D data requires more advanced processing to handle the temporal domain with image-to-image fluctuations.

Conclusion

The proposed method contains the following contributions. First, a novel multi-dimensional hybrid structure for instrument detection in 3D US is proposed. With this approach, network complexity is reduced and overfitting can be better avoided when compared to the traditional full-3D networks. Second, the obtained structure is based on a specifically designed dimension-reduction block, which reduces the spatial information from 3D to 2D, while extracting the most relevant instrument information along specific directions in the data volume. Third, to train the CNN, the proposed multi-level loss function allows the network to learn the information at pixel-level and image-level simultaneously, so that more context is obtained. In this chapter, the proposed algorithm achieves a similar or higher performance than state-of-the-art methods with 3-8 times higher computation efficiency, thereby paving the way for real-time applications.

CHAPTER 8

Conclusions

This thesis has presented the techniques for medical instrument detection in 3D ultrasound volumetric data. In this chapter, the contributions and final conclusions presented in individual chapters to implement this system are summarized. The research questions of Chapter 1 are reconsidered and addressed with our findings.

8.1 Conclusions on Individual Chapters

Chapter 2: This chapter presents an extensive summary of the commonly considered techniques and methodologies for image-based instrument detection methods in 3D volumetric datasets. We have introduced the well-known feature analysis techniques for instrument-related information measurement. For classification and detection of the instrument, machine learning techniques, such as support vector machine, adaptive boosting and deep neural networks are introduced. Specifically, conventional machine learning algorithms and state-of-the-art deep learning methods are compared. We have found that support vector machine or adaptive boosting algorithms can address the medical instrument detection tasks with a proper quality. When a larger amount of data are available, deep learning approaches are also attractive as a technical solution, because of their superior performance and robustness at the expense of more computation power. A model-fitting algorithm is also introduced for instrument detection in segmented images, to extract the instrument location and orientation in 3D volumes. Finally, evaluation method and commonly applied metrics are discussed, such as the Dice score, endpoint-error measurement, the Hausdorff distance.

Chapter 3: This chapter presents novel feature representations for catheter processing in 3D ultrasound volumes. We have proposed a generic method of multi-scale and multi-definition feature analysis to detect and localize the catheter position and orientation in challenging 3D US volumes. With the specifically proposed multi-scale/definition feature descriptors for catheter voxel classification, the catheter voxels in the imbalanced 3D US volumes can be detected by a standard binary adaptive boosting classifier. The classification results of the proposed features achieve an F_1 score of 52-83% on different experimental datasets (*in-vitro* to *in-vivo*), which is superior to the state-of-the-art computer vision techniques. Subsequently, the model of a curved catheter is fitted to the detected voxels by the proposed Sparse-plus-dense RANSAC model-fitting algorithm. The proposed complete system achieves an average localization error (endpoint-error) of about 2 mm in the challenging datasets.

Chapter 4: For efficient instrument detection, an efficient novel voxel-of-interest-based catheter detection framework is proposed in this chapter. To reduce the overall computation complexity, we have introduced a voxel-of-interest pre-selection step, prior to the voxel-level CNN-based classification, which is achieved by employing an efficient pre-modeled Frangi vesselness filter. Subsequently, a novel tri-planar CNN is proposed to classify the selected voxels in the 3D volumes, which partitions the local 3D patches to orthogonal 2D slices to reduce the computational cost. The evaluation on the challenging *ex-vivo* dataset demonstrates that the proposed method drastically accelerates the overall detection speed with more than 10 times improvement (from several minutes to around 18 seconds per volume), while preserving a segmentation score of about 60%. In addition, based on the segmented volumes, the SPD-RANSAC is applied to localize the catheter with a 2.1-mm localization error giving state-of-the-art performance.

Chapter 5: This chapter presents a novel deep learning method for efficient and robust localization of the medical instrument in 3D volumes. This method is based on a patch-of-interest strategy, for which an interested patch-selection algorithm and an efficient 2D FCN are employed to coarsely segment the instrument in a slice-by-slice manner. Subsequently, based on this coarse segmentation, the patches of interest are further partitioned and processed by a novel FuseNet, which is constructed by two individual and parallel CNNs for 2.5D and 3D information processing and subsequent fusing their feature maps. Additionally, to train the overall network, a novel hybrid loss function is introduced, which simultaneously learns discriminative information at the pixel level and image level, to enhance the segmentation performance. Extensive validation has demonstrated that the proposed framework achieves a segmentation performance about 70% Dice score and processing speed of about 1 second per

volume, which is much faster and better than the state-of-the-art techniques.

Chapter 6: Accurate voxel-level annotation is expensive and laborious to obtain for CNN training. In this chapter, to address this challenge for medical instrument detection, a novel semi-supervised learning framework is proposed, which consists of a coarse patch selection and fine segmentation. To achieve a patch selection with minimum annotation effort, a deep reinforcement learning technique is employed, which is adaptive to the image content to localize the region-of-interest. Based on the selected patches, a fine segmentation network is applied based on semi-supervised training, which is achieved by training under the guidance and constraint of an uncertainty estimation from the proposed Dual-UNet predictions. With extensive validation on *ex-vivo* and *in-vivo* datasets, the proposed method achieves a segmentation performance of about 70% and approximately a one-second execution time per volume. Meanwhile, the required annotation images are only about 30% of the total set of training images, which indicates promising results for annotation-efficient solutions.

Chapter 7: Full-image based CNN processing is expensive, time-consuming and a challenge for 3D volumetric data. To address these issues, a novel multi-dimensional processing technique is presented in this chapter. The proposed multi-dimensional method consists of a 3D encoder, 3D-to-2D dimensional reduction module and a 2D decoder, which reduces the computational complexity from 3D to 2D operations by projecting the feature maps along the interested principal axes. In addition, a multi-level loss function that focuses on both pixel-level consistency and image-level consistency using a skeleton-based annotation is employed, to enable the network to learn the information at different image context levels. Extensive validation on catheter and needle datasets demonstrate that the proposed method is 3-8 times faster than the current techniques, while it maintains a comparable detection accuracy to the state-of-the-art methods.

8.2 Discussion on the Research Questions

This section elaborates on the proposed methods and solutions with respect to the research questions formulated in Section 1.6.

RQ1. Features and modeling of the instrument for an automated detection system

RQ1a. What are good discriminative shape features for a medical instrument?

In Chapter 3, several 3D discriminative features are proposed for instrument voxel classification, which include novel Hessian matrix-based features, Gabor features and statistical features. In that chapter, a thorough validation is performed for classifying the instrument using those features on different

challenging datasets (*in-vitro* to *in-vivo*) at voxel level, which has demonstrated that the proposed multi-scale and multi-definition features achieve an F_1 score of 52-83%. In addition, the later chapters, such as Chapter 5, with deep learning methods also exploit the multi-scale discriminative information of the instrument by using pyramidal down- and up-scaling structures, such as commonly considered in the UNet architecture. Throughout the thesis, multiple experiments have been conducted to show promising results on different datasets. In conclusion, with experimental validations, the deep learning approaches with multi-resolution scaling features are attractive for modeling instrument discriminative information.

RQ1b. Is it possible to model a curved instrument in 3D image data based on position information, from the initial classification of voxels?

Based on the classification results in Chapter 3, a Sparse-plus-dense RANSAC algorithm is proposed by spline model-fitting techniques, which successfully models the curved catheter in 3D volumetric data. The proposed model-fitting algorithm is indeed based on the initial voxel classification, which achieves an average localization error of about 2 mm in challenging datasets. However, in some of the succeeding chapters where the detection plays a role, and are based on the deep learning approaches, the curvatures of the instruments are measured or processed by the proposed Sparse-plus-dense RANSAC algorithm as a post-processing step, because it provides a stable solution. In that case, the curvature processing is based on the deep learning semantic segmentation.

RQ2. Pre-modeling and robustness of the detection system

RQ2a. For efficient removal of irrelevant voxels, is pre-modeling and selecting the voxel points in 3D volume data a feasible solution?

In Chapter 4, an efficient Frangi vesselness pre-modeling stage of the catheter is proposed that removes the background voxels and selects the catheter-like voxels, prior to subsequent CNN classification. The proposed pre-selection can efficiently select the voxels-of-interest while removing more than 99% of the irrelevant voxels with a speed in the order of a second. Similarly, a local patch-of-interest pre-selection is also proposed to efficiently select the voxel points from the 3D volume, as shown in Chapters 5 and 6. That method has proven that it can efficiently exclude the background voxels by pre-selection and can be combined successfully with subsequent deep learning networks. In conclusion, both of these approaches have demonstrated their feasibility.

RQ2b. Can we directly describe the instrument within a deep neural network using (partially) 3D US data, to potentially improve the detection accuracy or improve the detection efficiency?

To directly learn the discriminative instrument information in 3D US data within one network, an experiment is conducted in Chapter 4. This experiment shows that the F_2 score is comparable to more efficient methods, and only indicates a higher precision at the cost of a slightly lower recall. Hence it is possible, but there are more efficient ways to find the instruments. In the same chapter, a tri-planar CNN is proposed, which partitions the 3D local patches into orthogonal 2D slices to reduce the computational cost, while still preserving the information for instrument detection. In this way, the 3D information is partially used for classification. The evaluation on the challenging *ex-vivo* dataset demonstrates that the proposed CNN improves the segmentation performance to a score of about 60%, which is higher than the state-of-the-art methods.

RQ3. Exploitation of the 3D context and near-real-time instrument detection

RQ3a. Can we implement an efficient and robust region-of-interest (ROI) instrument detection by means of a deep learning method?

To efficiently and robustly select the ROI by a deep learning method, a patch-of-interest strategy is proposed in Chapter 5, which employs an efficient 2D FCN that coarsely segments the instrument following a slice-by-slice strategy. Extensive validations on *ex-vivo* and *in-vivo* datasets show that the proposed pre-selection can segment the instrument with a Dice score of about 60% within only 0.6 seconds. The results ensure a successful and efficient selection of the ROIs. In addition, the pre-selection using such ROI method enables an efficient subsequent segmentation, which can be obtained with an overall execution speed of about 1 second per volume.

RQ3b. Can the instrument be robustly segmented by deep learning after applying the ROI methods? Does this technique provide a more meaningful semantic model and improve the robustness against challenging volumes containing anatomical structures?

For robust and accurate segmentation of the instrument after the ROI method, a novel FuseNet is proposed, which is constructed by two individual and parallel CNNs for 2.5D and 3D information processing. Since both 2.5D and 3D information are processed by the FuseNet, a more meaningful segmentation model can be achieved for challenging 3D US images, so that anatomical structures in the background or surroundings can be better handled. Extensive validation has demonstrated that the proposed framework achieves a segmentation performance of about 70% Dice score, which clearly outperforms the state-of-the-art methods. However, these Dice scores are based on overlap in the segmentation, which is not same as being meaningful. We estimate that some of these improved segmentations are certainly more meaningful, while others do not improve the understanding.

RQ4. Annotation-efficient training of a deep learning method for instrument detection

RQ4a. Is it possible to train an efficient coarse localization method to find the sub-volume containing the instrument without requiring accurate voxel-level annotation?

In Chapter 6, to efficiently localize the instrument without accurate voxel-level annotation, a deep reinforcement learning method is adopted. This method can interact adaptively with the input image content to efficiently localize the region of interest. Compared to bounding-box-based methods, the adopted and proposed deep Q-network has a specific simple architecture, enabling easy training and achieving superior performances for our challenging datasets, resulting in an accuracy of about 4 voxels. Similarly, this is also shown in Chapter 7, where the instrument is coarsely localized by a multi-dimensional method without employing accurate voxel-level annotation, e.g. just using an instrument skeleton, which achieves a localization error of about 3-4 voxels on the average. These two proposed approaches form a good basis for efficient coarse localization of the instrument.

RQ4b. How should the region of interest be segmented with a deep learning method, when the network is trained by only a small amount of annotated data (or even without) at voxel-level and a large amount of unannotated data?

For training the segmentation network using both a small amount of annotated data at voxel-level and a large amount of unannotated data, an annotation-efficient semi-supervised learning method is proposed. In order to generate the constrained loss functions for using the unlabeled images during the training, an uncertainty estimation based on Bayesian predictions is employed, which is then exploited to train the segmentation network. With extensive validation on *ex-vivo* and *in-vivo* datasets, the proposed method achieves a segmentation performance of about 70% Dice score, while the required annotation images are reduced to only about 30% of the total amount of training images.

In fully training the network without any voxel-level annotations, a multi-level loss function is considered in Chapter 7 to facilitate the network to learn the information from different image contexts. Extensive validation experiments on catheter and needle datasets demonstrate the proposed approach achieves a comparable detection accuracy (end-points error: 2-3 voxels, orientation error: 5-7 degrees) to the state-of-the-art methods employing semantic segmentation CNNs.

RQ5. Real-time detection of medical instruments

RQ5. Is it possible to reduce the complexity of a 3D model by decreasing the number of dimensions in the neural network modeling, rather than reducing the image-scale/resolution, CNN filter sizes, etc.?

To reduce the complexity of a 3D model for medical instrument detection, a novel multi-dimensional CNN is proposed in Chapter 7. The proposed method avoids to reduce the image-scale, resolution and CNN filter sizes, but instead reduces the dimensionality of the feature maps from 3D to 2D. This approach drastically decreases the complexity of the decoder compared to commonly considered 3D CNN models. The proposed method achieves an inference time of about 0.1 seconds, being 10 times faster than the commonly considered methods, which require at least 1-second execution time.

8.3 Utilization and Outlook

The fast development of image processing and artificial intelligence enable to automate and perform more sophisticated and complicated tasks in many areas of the growing information society. In the medical domain, US imaging is one of the essential modalities for treatment and diagnostics, which has been substantially benefiting from this trend, and is already merging within fetal and cardiac monitoring. Image-guided minimally-invasive intervention is another important application direction of US imaging, as it can provide simpler and better outcomes with the aid of AI enabled systems. The research presented in this thesis has proposed several solutions for automated medical instrument detection and localization in 3D US volumetric data, which is acquired by standard cardio US transducers with low spatial resolution. The research experiments have shown that the developed algorithms are accurate and robust for the detection task by employing deep learning techniques, which are however computationally complex and expensive to be trained. Therefore, further optimization is required for a real-time and robust performance in mature future applications with high-resolution US images.

As for high-resolution US imaging, one of the limitations of the current work is that the processing-throughput rate of this work is, though approaching the speed of clinical support, not yet sufficient. For the support of live interventions, a real-time execution of about 10-20 Hz video frame rate is minimally required. Therefore, the computational complexity of the considered approaches and related systems have a direct influence on the applicability in the clinical domain. Acceleration of the execution efficiency of the proposed method should be further optimized in the future, such as a more compact yet accurate deep learning network design or the throughput optimization of total processing pipeline. In parallel to software optimization, the actual hardware development also enables a much faster execution time of a complex algorithm, because parallel GPU processing is still expanding over time. As a consequence, it is easier to achieve higher processing speeds in the upcoming years, which natually leads to a better applicability of this work.

In terms of other applications or modalities, the proposed concepts and/or methods can be easily adopted to different types of operations, such as needle-based biopsy taking under the guidance of US imaging. In addition, thanks for the learning capability of the CNNs, many other types of devices, such as stents, valves, pacemaker leads, etc., can be learned by an end-to-end training strategy. With the ongoing generalization of the proposed methodology, the work in this thesis can be easily transferred to other image-guided operations, within US imaging or outside the scope of acoustic imaging, such as X-Ray modalities for real-time guidance.

From the clinical perspective and considering the unique properties of ultrasound like real-time execution and radiation-free imaging, US imaging is increasingly becoming one of the important imaging modalities in the near future for treatments and diagnostics, when compared to expensive CT with radiation or MRI imaging. Additionally, US imaging is assumed to be widely applied in many different applications in regional or local hospitals to be supplementary to CT or X-Ray imaging, or even replace them for simple tasks, because it is less expensive and can be equipped with more advanced image processing software. With the trend of more affordable US machines and advanced US analysis in the future, a reduction of the hospital referring cost and a better early outcome for patients may be achieved. Future systems with integration of the proposed artificial intelligence algorithms should eventually improve the outcomes of the operation and treatment, and facilitate clinicians in training and operations. This enables a substantially broader usage of US-guided analysis, which indicates a bright future of US-based image processing.

Bibliography

[1] B. R. Douglas, J. W. Charboneau, and C. C. Reading, "Ultrasound-guided intervention: expanding horizons," *Radiologic Clinics of North America*, vol. 39, no. 3, pp. 415–428, 2001.

[2] I. M. Germano, *Advanced techniques in image-guided brain and spine surgery*. Thieme Medical Publishers, Incorporated, 2002.

[3] T. M. Peters, "Image-guidance for surgical procedures," *Physics in Medicine & Biology*, vol. 51, no. 14, p. R505, 2006.

[4] K. Cleary and T. M. Peters, "Image-guided interventions: technology review and clinical applications," *Annual review of biomedical engineering*, vol. 12, pp. 119–142, 2010.

[5] X. Wu, "Fast catheter segmentation and tracking based on x-ray fluoroscopic and echocardiographic modalities for catheter-based cardiac minimally invasive interventions," 2015.

[6] A. Pourtaherian, "Robust needle detection and visualization for 3d ultrasound image-guided interventions," Ph.D. dissertation, Department of Electrical Engineering, 9 2018, proefschrift.

[7] T. M. Peters, "Image-guided surgery: from x-rays to virtual reality," *Computer methods in biomechanics and biomedical engineering*, vol. 4, no. 1, pp. 27–57, 2001.

[8] M. Mischi, "Contrast echocardiography for cardiac quantifications." 2002.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[10] W. Xia *et al.*, "In-plane ultrasonic needle tracking using a fiber-optic hydrophone," *Medical Physics*, vol. 42, no. 10, pp. 5983–5991, 2015.

[11] J. Krücker, S. Xu, N. Glossop, A. Viswanathan, J. Borgert, H. Schulz, and B. J. Wood, "Electromagnetic tracking for thermal ablation and biopsy guidance: clinical evaluation of spatial accuracy," *Journal of Vascular and Interventional Radiology*, vol. 18, no. 9, pp. 1141–1150, 2007.

[12] C. Nadeau *et al.*, "Intensity-based visual servoing for instrument and tissue tracking in 3d ultrasound volumes," *IEEE TASE*, vol. 12, no. 1, pp. 367–371, 2014.

[13] K. J. Draper, C. C. Blake, L. Gowman, D. B. Downey, and A. Fenster, "An algorithm for automatic needle localization in ultrasound-guided breast biopsies," *Medical physics*, vol. 27, no. 8, pp. 1971–1979, 2000.

[14] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[15] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 1998, pp. 130–137.

[16] M. Barva, M. Uhercik, J.-M. Mari, J. Kybic, J.-R. Duhamel, H. Liebgott, V. Hlavác, and C. Cachard, "Parallel integral projection transform for straight electrode localization in 3-d ultrasound images," *ieee transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 55, no. 7, 2008.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah *et al.*, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[19] H. Yang, C. Shan, A. F. Kolen *et al.*, "Catheter localization in 3d ultrasound using voxel-of-interest-based convnets for cardiac intervention," *International journal of computer assisted radiology and surgery*, vol. 14, no. 6, pp. 1069–1077, 2019.

[20] Y. Zhao, C. Cachard, and H. Liebgott, "Automatic needle detection and tracking in 3d ultrasound using an roi-based ransac and kalman method," *Ultrasonic imaging*, vol. 35, no. 4, pp. 283–306, 2013.

[21] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[22] H. Yang, C. Shan, A. F. Kolen, and P. H. de With, "Deep q-network-driven catheter segmentation in 3d us by hybrid constrained semi-supervised learning and dual-unet," in *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2020.

[23] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[24] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[25] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 1, pp. 55–73, 1990.

[26] R. Mehrotra, K. R. Namuduri, and N. Ranganathan, "Gabor filter-based edge detection," *Pattern recognition*, vol. 25, no. 12, pp. 1479–1494, 1992.

[27] M. Kaya and O. Bebek, "Gabor filter based localization of needles in ultrasound guided robotic interventions," in *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*. IEEE, 2014, pp. 112–117.

[28] A. Pourtaherian, H. Scholten, L. Kusters, S. Zinger, N. Mihajlovic, A. Kolen, F. Zou, G. Ng *et al.*, "Medical instrument detection in 3-dimensional ultrasound data volumes," *IEEE Transactions on Medical Imaging*, 2017.

[29] I. Hacihaliloglu, R. Abugharbieh, A. Hodgson, and R. Rohling, "Bone segmentation and fracture detection in ultrasound using 3d local phase features," *Medical image computing and computer-assisted intervention–MICCAI 2008*, pp. 287–295, 2008.

[30] J.-K. Kamarainen, V. Kyrki, and H. Kalviainen, "Invariance properties of gabor filter-based features-overview and applications," *IEEE Transactions on image processing*, vol. 15, no. 5, pp. 1088–1099, 2006.

[31] L. Antiga, "Generalizing vesselness with respect to dimensionality and shape," *The Insight Journal*, vol. 3, pp. 1–14, 2007.

[32] H. Yang, A. Pourtaherian, C. Shan, A. F. Kolen *et al.*, "Feature study on catheter detection in three-dimensional ultrasound," in *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10576. International Society for Optics and Photonics, 2018, p. 105760V.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[35] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[38] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[39] M. Ranzato, C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *Advances in neural information processing systems*, 2007, pp. 1137–1144.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[41] R. Sun, "Optimization for deep learning: theory and algorithms," *arXiv preprint arXiv:1912.08957*, 2019.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] M. G. Linguraru, N. V. Vasilyev, P. J. Del Nido, and R. D. Howe, "Statistical segmentation of surgical instruments in 3-d ultrasound images," *Ultrasound in medicine & biology*, vol. 33, no. 9, pp. 1428–1437, 2007.

[45] M. Aboofazeli, P. Abolmaesumi, P. Mousavi, and G. Fichtinger, "A new scheme for curved needle segmentation in three-dimensional ultrasound images," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*. IEEE, 2009, pp. 1067–1070.

[46] K. Cao, D. Mills, and K. A. Patwardhan, "Automated catheter detection in volumetric ultrasound," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 37–40.

[47] M. Uherčík, J. Kybic, H. Liebgott, and C. Cachard, "Model fitting using ransac for surgical tool localization in 3-d ultrasound images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 8, pp. 1907–1916, 2010.

[48] M. Uherčík, J. Kybic, Y. Zhao, C. Cachard, and H. Liebgott, "Line filtering for surgical tool localization in 3d ultrasound images," *Computers in biology and medicine*, vol. 43, no. 12, pp. 2036–2045, 2013.

[49] A. Pourtaherian, S. Zinger, P. de With, H. H. Korsten, and N. Mihajlovic, "Gabor-based needle detection and tracking in three-dimensional ultrasound data volumes," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3602–3606.

[50] A. Pourtaherian, S. Zinger, N. Mihajlovic, J. Huang, G. C. Ng, H. H. Korsten *et al.*, "Multi-resolution gabor wavelet feature extraction for needle detection in 3d ultrasound," in *Eighth International Conference on Machine Vision*. International Society for Optics and Photonics, 2015, pp. 987 513–987 513.

[51] Y. Zhao, Y. Shen, A. Bernard, C. Cachard, and H. Liebgott, "Evaluation and comparison of current biopsy needle localization and tracking methods using 3d ultrasound," *Ultrasonics*, vol. 73, pp. 206–220, 2017.

[52] C. Papalazarou, P. H. de With, and P. Rongen, "Sparse-plus-dense-ransac for estimation of multiple complex curvilinear models in 2d and 3d," *Pattern Recognition*, vol. 46, no. 3, pp. 925–935, 2013.

[53] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.

[54] C. De Boor, *A practical guide to splines*. Springer-Verlag New York, 1978, vol. 27.

[55] P. Ambrosini *et al.*, "Fully automatic and real-time catheter segmentation in x-ray fluoroscopy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 577–585.

[56] T. Tan, J.-J. Mordang, J. Zelst, A. Grivegnée, A. Gubern-Mérida, J. Melendez, R. M. Mann, W. Zhang *et al.*, "Computer-aided detection of breast cancers using haar-like features in automated 3d breast ultrasound," *Medical physics*, vol. 42, no. 4, pp. 1498–1504, 2015.

[57] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 212–220.

[58] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey *et al.*, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2014, pp. 520–527.

[59] A. Pourtaherian, F. G. Zanjani, S. Zinger, N. Mihajlovic, G. Ng, H. Korsten *et al.*, "Improving needle detection in 3d ultrasound using orthogonal-plane convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 610–618.

[60] H. Yang, C. Shan, A. F. Kolen, and P. H. de With, "Catheter detection in 3d ultrasound using triplanar-based convolutional neural networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 371–375.

[61] X. Yang, L. Yu, S. Li, X. Wang, N. Wang, J. Qin, D. Ni, and P.-A. Heng, "Towards automatic semantic segmentation in volumetric ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 711–719.

[62] P. M. Novotny, J. W. Cannon, and R. D. Howe, "Tool localization in 3d ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2003, pp. 969–970.

[63] H. Zhou, W. Qiu, M. Ding, and S. Zhang, "Automatic needle segmentation in 3d ultrasound images using 3d hough transform," in *MIPPR 2007: Medical Imaging, Parallel Processing of Images, and Optimization Techniques*, vol. 6789. International Society for Optics and Photonics, 2007, p. 67890R.

[64] P. Beigi, R. Rohling, T. Salcudean, V. A. Lessoway, and G. C. Ng, "Detection of an invisible needle in ultrasound using a probabilistic svm and time-domain features," *Ultrasonics*, vol. 78, pp. 18–22, 2017.

[65] H. Yang, C. Shan, A. Pourtaherian, A. F. Kolen *et al.*, "Catheter segmentation in three-dimensional ultrasound images by feature fusion and model fitting," *Journal of Medical Imaging*, vol. 6, no. 1, p. 015001, 2019.

[66] A. Pourtaherian, F. G. Zanjani, S. Zinger, N. Mihajlovic, G. C. Ng, H. H. Korsten *et al.*, "Robust and semantic needle detection in 3d ultrasound using orthogonal-plane convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 13, no. 9, pp. 1321–1333, 2018.

[67] H. Yang, C. Shan, A. F. Kolen, and P. H. 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)*. IEEE, 2019, pp. 260–264.

[68] H. Yang, C. Shan, A. F. Kolen, and H. de With Peter, "Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1122–1126.

[69] H. Yang, C. Shan, T. Tan, A. F. Kolen *et al.*, "Transferring from ex-vivo to in-vivo: Instrument localization in 3d cardiac ultrasound using pyramid-unet with hybrid loss," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 263–271.

[70] Y.-J. Huang, Q. Dou, Z.-X. Wang, L.-Z. Liu, Y. Jin, C.-F. Li, L. Wang, H. Chen *et al.*, "3d roi-aware u-net for accurate and efficient colorectal tumor segmentation," *arXiv preprint arXiv:1806.10342*, 2018.

[71] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 149–157.

[72] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[73] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[74] F. Isensee, J. Petersen, S. A. Kohl, P. F. Jäger, and K. H. Maier-Hein, "nnu-net: Breaking the spell on successful medical image segmentation," *arXiv preprint arXiv:1904.08128*, 2019.

[75] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[76] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[77] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[78] M. Arif, A. Moelker, and T. van Walsum, "Automatic needle detection and real-time bi-planar needle visualization during 3d ultrasound scanning of the liver," *Medical image analysis*, vol. 53, pp. 104–110, 2019.

[79] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.

[80] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[81] F.-C. Ghesu *et al.*, "Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans," *IEEE T-PAMI*, vol. 41, no. 1, pp. 176–189, 2017.

[82] A. Alansary *et al.*, "Evaluating reinforcement learning agents for anatomical landmark detection," *Medical image analysis*, vol. 53, pp. 156–164, 2019.

[83] ——, "Automatic view planning with multi-scale deep reinforcement learning agents," in *MICCAI*. Springer, 2018, pp. 277–285.

[84] Y. Zhang *et al.*, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *MICCAI*. Springer, 2017, pp. 408–416.

[85] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *MICCAI*. Springer, 2018, pp. 370–378.

[86] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," *arXiv preprint arXiv:1808.03887*, 2018.

[87] S. Chen *et al.*, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *MICCAI*. Springer, 2019, pp. 457–465.

[88] S. Sedai *et al.*, "Uncertainty guided semi-supervised segmentation of retinal layers in oct images," in *MICCAI*. Springer, 2019, pp. 282–290.

[89] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *MICCAI*. Springer, 2019, pp. 605–613.

[90] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[91] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE TMI*, 2020.

[92] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *arXiv preprint arXiv:2004.05937*, 2020.

[93] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[94] C. Gong, X. Chang, M. Fang, and J. Yang, "Teaching semi-supervised classifier via generalized distillation." in *IJCAI*, 2018, pp. 2156–2162.

[95] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proceedings of the IEEE ICCV*, 2019, pp. 6728–6736.

[96] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[97] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[98] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[99] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[100] X. Zhuang *et al.*, "Self-supervised feature learning for 3d medical images by playing a rubik's cube," in *MICCAI*. Springer, 2019, pp. 420–428.

[101] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[102] Q. Dou *et al.*, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.

[103] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[104] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, p. 29, 2015.

[105] H. Yang, C. Shan, A. Bouwman, A. F. Kolen, and P. H. N. de with, "Efficient and robust instrument segmentation in 3d ultrasound using patch-of-interest-fusenet with hybrid loss," *Medical Image Analysis*, 2020.

Acronyms

Adam Adaptive Moment Estimation

AdB Adaptive Boosting

AI Artificial Intelligence

ASPP Atrous Spatial Pyramid Pooling

B-mode Brightness mode

BN Batch Normalization

CE Cross Entropy

CL Contextual Loss

CNNs Convolutional Neural Networks

ConvNet Convolutional Neural Network

CRF Conditional Random Field

CT Computed Tomography

DF-UNet Direction-fused UNet

DL Deep Learning

DQN Deep Q-Network

DSC Dice Score

EE End-points Error

EMA Exponential Moving Averaging

FCN Fully Convolutional Neural Network

FL Focal Loss

FN False Negative

FP False Positive

GF Gabor-like Feature

GPU Graphics Processing Unit

HD Hausdorff Distance

HL Hybrid Loss

IN Instance Normalization

LDA Linear Discriminant Analysis

LOOCV Leave-One-Out Cross-Validation

LSVM Linear Support Vector Machine

MCD Monte Carlo Dropout

MF Multi-scale and Multi-definition Feature

MixDNet Multi-dimensional Mixture Network

MLP Multi-layer Perceptron

MRI Magnetic Resonance Imaging

MT Mean-Teacher

OE Orientation Error

PC Personal Computer

PCA Principal Component Analysis

PIP Parallel Integral Projection

PR Precision Recall

POI Patch-of-interest

PVA PolyVinyl Alcohol

RANSAC RANdom SAmple Consensus

RF-ablation Radiofrequency Ablation

ReLU Rectified Linear Unit

RHT Random Hough Transformation

RL Reinforcement Learning

ROI Region-of-interest

SE Skeleton Error

SGD Stochastic Gradient Descent

SNR Signal-to-Noise Ratio

SOTA State-of-the-art

SP Specificity

SPD-RANSAC Sparse-plus-dense RANdom SAmple Consensus

SSD Single Shot Detectoor

SSL Semi-supervised Learning

SVM Support Vector Machine

TAVI Transcatheter Aortic Valve Implantation

TE Tip Error

TEE TransEsophaegal Echocardiography

TL Transfer Learning

TN True Negative

TP True Positive

TTE Transthoracic Echocardiography

US Ultrasound

VOI Voxel-of-interest

VS Volumetric Similarity

WSL Weakly-supervised Learning

Acknowledgement

Everyone knows the Ph.D. is a title, but it is not just a title for an individual person working for it. For me, as part of my life, pursuing a Ph.D. means many different things within the research, and beyond it, with opportunities and challenges. But whatever I faced during the past years, finally I am here, finalized my research as a Ph.D. candidate and now continue to move on.

I would like to thank my supervisor Prof. Peter. H.N. de With for giving me the chance to start my Ph.D. journey and leading me to the area of medical imaging and computer vision in our VCA group. I still remember the afternoon that I received his call for my Ph.D. interview, where he explained project overview, but also asked critical questions. Thanks to him for this opportunity, and finally, I had the chance to join the group for research. During past years, I have learned a lot from him, not only about academic knowledge, but ambitions and attitude for science and technologies as a researcher. I still remember several evening meetings with him about the paper revision, no matter how busy he was, he always can manage the schedule and helped me to catch the due, even in Christmas and summer holidays. He also teaches me a lot outside the academic life: to be positive, optimistic, and have interests in different areas, like music, food, and beer. It is my great honor to be one of Peter's students, and I believe I will remember my time in VCA group forever.

I want to express my gratitude to my co-promotor Dr. Alexander F. Kolen, for his patient guidance, encouragement, and of course, the support for experimental setup throughout my Ph.D. time. Alex has supported me a lot and in Philips administration stuffs and helped me have good experience in my Philips time. He always encouraged me to try challenges and always to be optimistic. Many thanks to Alex for granting me a chance to have this Ph.D. position, as I still remember our first meeting, he explained to me the ambitions and applications of my topics in clinical applications, and believe we can have achievement in the end, and obviously, we did it. Thank you Alex, for many years support.

I would like to thank Prof. Caifeng Shan for his kind help and supports during my past years. Caifeng helped me a lot and spent so much precious time in my research direction, paper writing and paper revisions. Your great contribution has increased my outputs quality significantly. You also guided me to have the right timeline of my Ph.D. path, which is most challenging and important for a

fresh Ph.D. student. Without you, I believe I would lose my mind for a long time at my early stage of the research.

I am grateful to all my friends and colleagues in the VCA group. Of course, the first and the most important lady, Anja. You are so kindly to me, and helped me a lot in our daily administration things, also you are willing to have some small jokes to make us more closer to each other. Many times, your efforts make my life easier during my Ph.D. period. Also I really thanks Joy for her kind support and suggestions in my past years. At the first time I joined the group, she guided me so many detailed things as a fresh Ph.D., from common registration to order a new workstation. She also gave me many suggestions about how to be a good Ph.D. candidate in the university, and also many interesting talks within and outside the academic research. I also thank Chenyang for our interesting lunch talk and great suggestions in research framework and topic discussions. Same to Arash and Farhad, although you guys left the group for years, but I still remember our discussions about my topics and guidances for my initial deep learning study. Without you, I might face more challenges in my early stage research. Also many thanks to Fons and Joost, we had many interesting experiences in Houston for SPIE and Shenzhen for MICCAI, although the travels are busy and tired, we had great time and valuable discussions during those periods. Thanks to Marco Mamprim, we used to have interesting travel to Athens for ICIP, where we climbed mountain at the center of the city. Thanks to Cheng, for her kindly and in-time replies for my administration questions. Many thanks to other colleagues including but not limited Xikai, Xin, Francesca, Panos, Anweshan, Patrick, Liang, Amir, and Marco, etc. for our interesting discussions and talks during past years.

I would like to thank all my dear friends for their supports. First, and most important, thank my bro, Jianfei, without his supporting and time we shared, my Ph.D. time would be much more challenging and difficult than I can imagine. Also thank you Hao, who has many interesting discussions with me within and outside the academic topics. Although his is working and living in Sweden, our chats are always in time and meaningful. My special thanks are also send to my friend and colleague Tao, he has helped me so much in deep learning discussion and also job findings. I really would like to show my thanks to Dinglin, Lynae, Jingxi, Eric, Neal and Chris, who have kindly expressed their concerns when I faced problems during past years. I would like to share my thanks to my dear room mates Shengling and Fan, we shared many interesting days in topics of comic, video game and life during my last time in Eindhoven, which are precious to me in my whole life time. Thanks to Sushan and Jiahang, for their meaningful suggestions about the career plan in both academe and industry areas. I also would like to thanks the other friends including but not limited Xuming, Xue, Yan, Yulun, Tang for our time since I started to live outside China. It is hard to imagine how my life abroad would be without their company.

Finally, I would like to express my deepest gratitude to my families, who have been supporting me for pursuing a Ph.D. I really thanks my wife, Qian, for her endless support and love during past decade. I also would like to thank my parents, Bin and Miao, for their love, encouragement, and always being proud of me.

Publication list

Journal articles

[J-1] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Medical Instrument Detection in Ultrasound-Guided Interventions: A Review", Artificial Intelligence Review, *minor revision*.

[J-2] **H.Yang**, C. Shan, A. Bouwman, L. Dekker, A. F. Kolen, and P. H. N. de With, "Medical Instrument Segmentation in 3D US by Hybrid Constrained Semi-Supervised Learning", *IEEE Journal of Biomedical and Health Informatics (JBHI)*, vol. 26, no. 2, pp. 762-773. (2022).

[J-3] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Weakly-supervised Learning For Catheter Segmentation in 3D Frustum Ultrasound", *Computerized Medical Imaging and Graphics (CMIG)*, 96, 102037 (2022).

[J-4] **H.Yang**, C. Shan, A. Bouwan, A. F. Kolen, and P. H. N. de With, "Efficient and Robust Instrument Segmentation in 3D Ultrasound Using Patch-of-Interest-FuseNet with Hybrid Loss," *Medical Image Analysis (MedIA)*, vol 67, 101842 (2021).

[J-5] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, " Efficient Medical Instrument Detection in 3D Volumetric Ultrasound Data," *IEEE Transaction on Biomedical Engineering (TBME)* vol 68, no. 3, pages 1034-1043 (2021).

[J-6] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, " Catheter localization in 3D ultrasound using voxel-of-interest-based ConvNets for cardiac intervention," *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, vol 14, pages 1069–1077 (2019).

[J-7] **H.Yang**, C. Shan, A. Pourtaherian, A. F. Kolen, and P. H. N. de With, " Catheter segmentation in three-dimensional ultrasound images by feature fusion and model fitting," *Journal of Medical Imaging (JMI)*, 6(1), p.015001 (2019).

International conference proceedings

[C-1] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Deep Q-Network-Driven Catheter Segmentation in 3D US by Hybrid Constrained Semi-Supervised Learning and Dual-UNet," *Proceedings of the 23rd International Conference on Medical Image Computing & Computer Assisted Intervention. (MICCAI 2020)*, Lima, Peru, 2020.

[C-2] L. Min, **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Feasibility study of catheter segmentation in 3D Frustum Ultrasounds by DCNN," *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling. (SPIE MI 2020)*, Huston, USA, 2020.

[C-3] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Automated catheter localization in volumetric ultrasound using 3D patch-wise U-net with focal loss," *Proceedings of IEEE 26th International Conference on Image Processing. (IEEE ICIP 2019)*, Taipei, China, 2020.

[C-4] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Efficient catheter segmentation in 3D cardiac ultrasound using slice-based FCN with deep supervision and F-score loss," *Proceedings of IEEE 26th International Conference on Image Processing. (IEEE ICIP 2019)*, Taipei, China, 2020.

[C-5] **H.Yang**, C. Shan, T.Tan, A. F. Kolen, and P. H. N. de With, "Transferring from ex-vivo to in-vivo: instrument localization in 3D cardiac ultrasound using Pyramid-UNet with hybrid loss," *Proceedings of the 22nd International Conference on Medical Image Computing & Computer Assisted Intervention. (MICCAI 2019)*, Shenzhen, China, 2019.

[C-6] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Improving catheter segmentation & location in 3D cardiac ultrasound using direction-fused fcn," *Proceedings of IEEE 16th International Symposium on Biomedical Imaging. (IEEE ISBI 2019)*, Venice, Italy, 2019.

[C-7] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Catheter detection in 3D ultrasound using triplanar-based convolutional neural networks," *Proceedings of IEEE 25th International Conference on Image Processing. (IEEE ICIP 2018)*, Athens, Greece, 2018.

[C-8] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Feature study on catheter detection in three-dimensional ultrasound," *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling. (SPIE MI 2018)*, Huston, USA, 2018.

International patent applications

[P-1] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Image-processing method and apparatus for object detection or identification," Patent No. EP3815617A1, Published in May, 2021.

[P-2] **H.Yang**, C. Shan, A. F. Kolen, and P. H. N. de With, "Identifying an interventional device in medical images," Patent No. WO2020089416A1, Published in May, 2020.

Curriculum vitae

Hongxu Yang was born in Taiyuan, China, in 1990. He received his Bachelor degree (BSc.) in Electrical Engineering jointly from Tianjin University and Nankai University in 2014. Since then, he continued with a Master degree (MSc.) of Electrical Engineering at the Eindhoven University of Technology (TU/e). During his MSc. study, Hongxu completed his master thesis in 2016 at IMEC/Holst Centre, working on an EEG-based authentication system with an encryption algorithm, which was presented at the European Signal Processing Conference 2017. Directly after his master studies, Hongxu continued his research as a PhD candidate in the Video Coding and Architectures research group, as part of the Signal Processing Systems department at the Electrical Engineering Faculty of the TU/e. With his dissertation work, he has developed efficient and accurate algorithms for medical instrument detection in three-dimensional ultrasound imaging by employing deep learning approaches. This research has resulted in 2 filed patent applications and over 15 publications in international peer-reviewed scientific journals and top international conferences in the area of computer vision and medical imaging. Publications have been issued in the journal on Medical Image Analysis, IEEE Transactions on Biomedical Engineering and Medical Image Computing and Computer-Assisted Intervention. In his spare time, you can find Hongxu playing video games, reading novels, and cooking food for the family.