Needle Detection and Localization in Ultrasound Images based on Deep Learning

MSc Thesis

Haitao Jiang

Delft University of Technology

TU Delft

Needle Detection and Localization in Ultrasound Images based on Deep Learning

by

Haitao Jiang

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday, October 31, 2023, at 15:30.

Thesis Committee:	Prof. dr. Jenny Dankelman, supervisor, TU Delft
	Dr. Nick van de Berg, supervisor, TU Delft & Erasmus MC
	Prof. dr. Nassir Navab, supervisor, TU Munich
	Dr. Zhongliang Jiang, supervisor, TU Munich
	Prof. dr. Theo van Walsum, Erasmus MC
Institution:	Delft University of Technology
Place:	Delft, the Netherlands
Project Duration:	Feb., 2023 - Oct., 2023

Cover Image: A picture of a phantom experiment that collects ultrasound images of needle covered by a part of python codes

Acknowledgments

Time is fleeting. After two years of study journey in Delft and in Munich, I met many challenges but have gained valuable knowledge and experience during this period. I would like to express my heartfelt gratitude to those who have provided assistance and guidance to me. First, I'm grateful for the supervision from my daily supervisors, Dr. Zhongliang and Dr. Nick van de Berg. Their supervision played a crucial role in helping me find the right path, and they were always patient in addressing my questions and helping me overcome any challenges encountered during my thesis. Second, I want to appreciate for my main supervisors, Prof. dr. Jenny Dankelman and Prof. dr. Nassir Navab. They offered unique insights and highlighted key issues that greatly contributed to the success of my thesis. Third, I want to thank my close friends who relieved my anxiety and provided emotional support when I was in a dilemma. Lastly, I have to express my deepest gratitude to my parents. Their unwavering support allowed me to pursue my studies in Europe without any worries. Without them, I would never have had the opportunity to come this far today."

Haitao Jiang
Oct. 24, 2023
Delft, Netherlands

Abstract

The technique of ultrasound-guided needle insertion is commonly employed in various clinical fields, including biopsy, anesthesia, brachytherapy, and ablation. However, the visibility of the needle in ultrasound (US) images remains a persistent challenge. To improve the guidance accuracy of needle insertion during interventions, it is crucial to develop a reliable technique that enhances the visibility of ultrasound needles and accurately detects their position and orientation. Recently, Deep Learning (DL) based segmentation methods have drawn attention because of their high efficiency and accurate results. In order to improve the model performance on challenging datasets, previous researchers modified the segmentation models by introducing spatial attention mechanism and temporal information. However, whether the approaches are effective for ultrasound images, especially for the needle insertion tasks remains unclear.

This thesis aims to investigate whether deep learning models are able to increase segmentation accuracy as well as localization accuracy in 2D ultrasound images, specifically focusing on introducing spatial attention and optical flow information into U-Net backbone. Spatial Mask Attention U-Net (SMA-UNet) and Optical Flow Attention U-Net (OFA-UNet) were therefore proposed. The hierarchical experiments were designed to evaluate the effects of training loss, mask width and optical flow methods, and then select an optimal configuration for the segmentation models. Furthermore, U-Net, Attention U-Net and two proposed models were validated on datasets collected from pork and beef phantoms, as well as patients. The evaluation results indicate that OFA-UNet has significant improvement in terms of segmentation metrics and geometrical errors compared to the U-Net baseline and the U-Net only considering the mask attention. Specifically, the model achieved Dice of 86.7%, IoU of 88.2%, Precision of 88.6%, tip error of 2.7 mm and angular error of 0.002 radians on the pork dataset. Furthermore, the OFA-UNet shows robustness and consistency in evaluation metrics across three different datasets, indicating its ability to adapt to varying complexities of US datasets.

Key words: Ultrasound, needle insertion, Deep Learning, spatial attention mechanism, temporal information

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
US	ultrasound
DL	Deep Learning
SMA-UNet	Spatial Mask Attention U-Net
OFA-UNet	Optical Flow Attention U-Net
2D	two dimensional
CNNs	Convolutional Neural Networks
ROI	region of interest
RANSAC	Random Sample Consensus
LK	Lucas Kanade
Dice	Dice Loss
BCE	Binary Cross Entropy
DiceBCE	Dice Binary Cross Entropy
DSC	Dice coefficient
IoU	Intersection over Union
TE	tip error
AE	angular error

List of Figures

List of Tables

1 Introduction

1.1 Needle Insertion and Ultrasound-guided interventions

Minimally invasive intervention is a kind of medical procedure that involves accessing and treating the affected area through small incisions or punctures in the skin, rather than open surgery. Incisions made during open surgery usually leave large wounds and tissue damage that may cause post-operation pain, longer healing time or even hemorrhage. With the advancement of medical technologies, a trend exists toward minimally invasive approaches, simplifying the intervention procedure, shortening the operation time, which results in a higher success rate of the treatment or inspection, and fewer repeats for the intervention or surgery.

Minimally invasive interventions involve the insertion of particular devices, for example, needles, through the skin into soft, inhomogeneous body tissue to reach a target (Abolhassani et al., 2007). In order to facilitate the clinician or surgeon to better operate and direct the needles during the intervention, image guidance technique that uses ultrasound (US) are widely applied to display the interventional activities inside the patient's body without any open incisions. It does not only visualize the anatomical structures but also provides the size, shape or even accurate localization information for the interventional guidance.

US imaging is a key imaging modality used in image-guided minimally invasive surgery. It uses high-frequency sound waves to produce images of the inside human body. In recent years, it has received significant attention due to its advantages of being widely available, non-ionizing, cost-effective and having real-time performance (Douglas et al., 2001). US offers a unique benefit of having a wide range of transducers that allow for flexibility in different healthcare settings, such as clinics, emergency rooms, op-

erating rooms, or even remote areas where access to larger imaging equipment may be limited (Scanlan et al., 2001). It can provide valuable information about the structure, size, shape, and movement of organs, as well as detect abnormalities, such as tumors, cysts, or fluid collections (Morgan et al., 2018). As a result, US-guided interventional procedures have been investigated and utilized in various clinical fields, such as biopsy (Hatada et al., 2000), regional anesthesia (Barrington and Kluger, 2013), ablation therapy (Sheafor et al., 1998), prenatal diagnosis and therapy (Oepkes et al., 2007), and cardiac interventions for structural and congenital heart disease (Jan et al., 2020).

Two-dimensional (2D) ultrasound is the most commonly used modality in echocardiography. The two dimensions presented are width (x-axis) and depth (y-axis) (Cootney, 2001). The standard ultrasound transducer for 2D echocardiography is the phased array transducer, which creates a sector-shaped ultrasound field. In case of 2D ultrasound images, data can usually be accessed by grabbing the frames at the display output of the ultrasound system (von Haxthausen et al., 2021). 2D US guidance is a widely used technique due to several advantages. Compared to other imaging modalities, 2D US has a shorter processing time and requires less data to be processed, enabling better real-time performance. In recent decades, multiple algorithms have been developed to enhance the image quality of 2D US. Additionally, 2D US is widely accessible and more affordable than other imaging modalities, which is crucial for clinical institutions and hospitals.

1.2 Problem Definition

One of the critical aspects of US-guided minimally invasive procedures is the percutaneous insertion of medical devices, particularly needles, into soft and heterogeneous body tissues to reach specific targets. The precision of needle placement directly influences the efficacy of treatments and the accuracy of diagnoses, varying based on the target organ and medical application. For instance, procedures such as prostate, kidney, and breast biopsies demand millimeter-level accuracy, while interventions in the brain, fetus, and eyes require even finer submillimeter precision (Abolhassani et al., 2007). However, achieving and maintaining such levels of accuracy is a complex challenge influenced by multiple factors, including the needle type, insertion depth, tissue deformation, and needle deflection (Rampersaud et al., 1999).

The problem of this thesis is the detection and localization of needles within the context of 2D ultrasound imaging. While ultrasound offers numerous advantages for real-time visualization and guidance during minimally invasive procedures, the precise identification and tracking of needles remain significant challenges. The most challenging aspect of needle-based intervention is for clinicians to align the 2D US plane and the instrument correctly. As illustrated in Figure 1.1, positioning the US plane parallel to

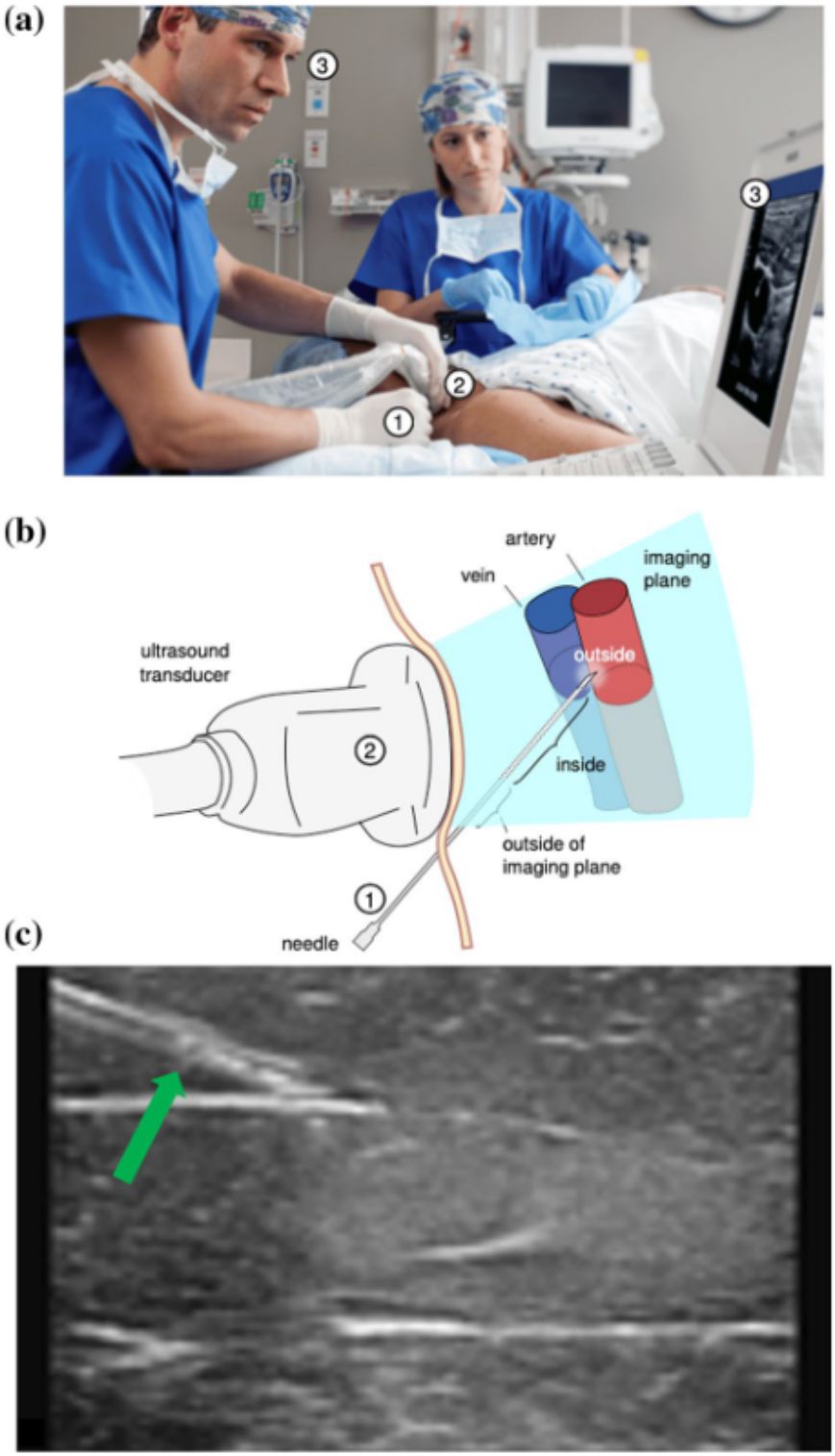


Figure 1.1: US-guided needle therapy (Pourtaherian et al., 2018): a clinical staff has to manage the multifold coordination of 1 needle, 2 US transducer, while 3 looking at the US screen (Courtesy of Philips Ultrasound). b Schematic representation of guiding a needle using US imaging, depicting an example situation for regional anesthesia, where the needle tip is outside the imaging plane and is approaching an erroneous target area. c B-mode US slice contains the needle, pointed by the green arrow

the needle is essential to visualize the needle in the US image accurately for in-plane imaging. During the procedure, clinical staff need to pay careful attention to aligning the needle and the US plane while monitoring the screen. A misaligned needle can result in the display of only the intersection of the needle and the US plane. Moreover, distinguishing the needle from the surrounding tissues in the B-mode images remains challenging for clinicians, as shown in Fig. 1.1 (c). Clinical staff require extra training to overcome this obstacle and interpret the instruments correctly. There is a pressing need for robust and accurate methods and algorithms to detect, locate, and monitor the positions of needles in 2D ultrasound images. Solving this problem will enhance the overall safety, efficiency, and effectiveness of minimally invasive interventions across a range of medical specialties.

1.3 Previous Research on Needle Detection

There are two common techniques for visualizing needles during ultrasound-guided interventions: hardware-based and image-based approaches. Although hardware-based methods have shown promising results by using some sensors, they come with a relatively high equipment cost, an additional calibration time, and require considerable specialized skills and training to operate, which hampers their broad acceptance in clinical practice. In contrast, image-based approaches have drawn much attention in the past decades. Owing to the fast advancement of artificial intelligence technology in recent years, the application of deep learning (DL) methods has revolutionized the way US images are acquired, processed and analyzed. DL has been proven that it has the potential to increase needle detection accuracy, improve detection process efficiency, and make the algorithm more robust for various equipment conditions and clinical applications.

Image segmentation is a kind of computer vision technique that assigns several labels to pixels of the image. The region of the image with the same label shares certain characteristics or identical semantic meanings for a specific task. Instance segmentation and semantic segmentation are two common types of image segmentation. The former one aims to detect each separate and distant object belonging to the same class (Yi et al., 2019) while the another classifies all pixels into corresponding categories (Guo et al., 2020). In medical segmentation fields, semantic segmentation draws more attention because it is able to detect and localize some specific structures, such as tumors, organs and vessels. It provides more information for doctors and helps them to better understand and analyze the image. Traditionally, medical image segmentation relies on manual annotation by clinical experts (Hesamian et al., 2019). However, this method is time-consuming, expensive, and prone to errors. Different experts may produce varying segmentation results due to their knowledge and experience, leading to inconsistent outcomes. Recently, the rapid development of DL has contributed to the advancement

of automatic image segmentation techniques, which largely improve the processing efficiency and provide convincing segmentation results. A literature review has been conducted to gain insight into the segmentation methods based on DL to detect and localize the needle in 2D US images. Some of them will be introduced as follows:

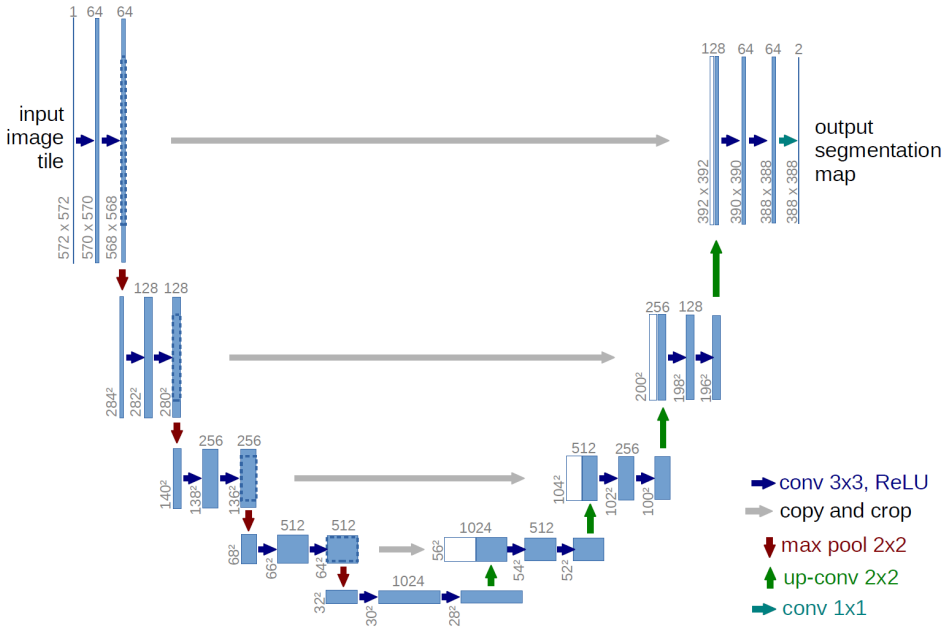


Figure 1.2: The model structure of U-Net model(Ronneberger et al., 2015)

1.3.1 2D CNNS

Convolutional neural networks (CNNs) are one branch of deep neural networks that are commonly used for image processing (Hesamian et al., 2019). They are typically composed of multiple convolutional layers, pooling layers, and fully connected layers, which are connected in different ways to construct a deep neural network. These models are trained using the back-propagation algorithm to adjust their weights and minimize the loss function. 2D CNNs have achieved great success in the area of 2D image processing. U-Net is one of the most famous CNNs algorithms for medical image segmentation proposed by Ronneberger et al. (2015). The network architecture is a U-shape structure consisting of a down-sampling path and an up-sampling path as shown in Fig. 1.2, which is also recognized as an encoder-decoder structure in DL. The down-sampling path is a typical CNN that compresses the input image and extracts textual features from spatial information. While the up-sampling path, commonly known as the expansive path, is composed of several up-sampling or de-convolutional layers to recover the image size and the semantic information. The most important feature of U-Net is the skip connections between the layers from the two paths with the same resolution and channel depth. These skip connections help in transmitting detailed information and assist in

recovering fine segmentation edges. The architecture of UNet can be expanded and modified according to the requirements of the task, such as adding or removing encoder and decoder layers, to accommodate different input image sizes and complexities.

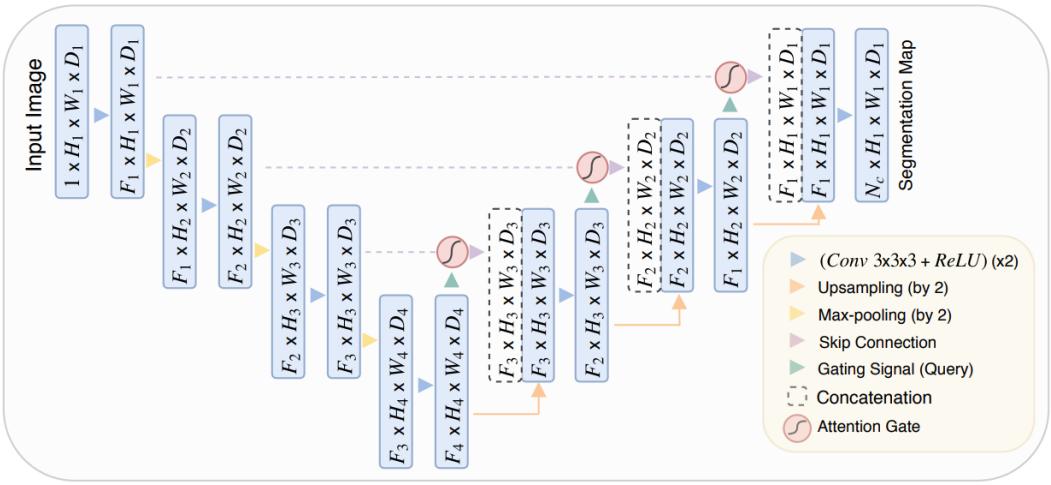


Figure 1.3: The structure of Attention U-Net (Oktay et al., 2018)

1.3.2 Attention Mechanism

The attention mechanism is a technique used in DL that enhances models by selectively focusing on important input elements, improving prediction accuracy and computational efficiency (Hore and Chatterjee, 2019). The idea of the attention mechanism is to mimic human vision attention that has a predetermined purpose and relies on specific tasks (Niu et al., 2021). It enables humans to focus attention on a certain object consciously and actively. These mechanisms in DL are specifically designed for different tasks and are often referred to as focused mechanisms. By incorporating attention mechanisms into DL models, researchers aim to enhance the performance and accuracy of various tasks by emphasizing relevant information and suppressing irrelevant features. These mechanisms allow the model to dynamically allocate its attention to different parts of the input, improving its ability to process complex data and make more informed decisions. Overall, attention mechanisms play a crucial role in DL by enabling the model to selectively attend to important aspects of the data and improve the overall performance.

One of the classical segmentation frameworks with the attention mechanism is 'Attention U-Net', which is proposed by Oktay et al., 2018. They came up with a novel attention gate model for medical imaging that automatically learns to focus on target structures of varying shapes and sizes. Compared to the original U-Net, three attention gates are plugged into the skip connection between the encoder and decoder (as shown in Fig. 1.3), which enables the model to suppress irrelevant regions in an input image

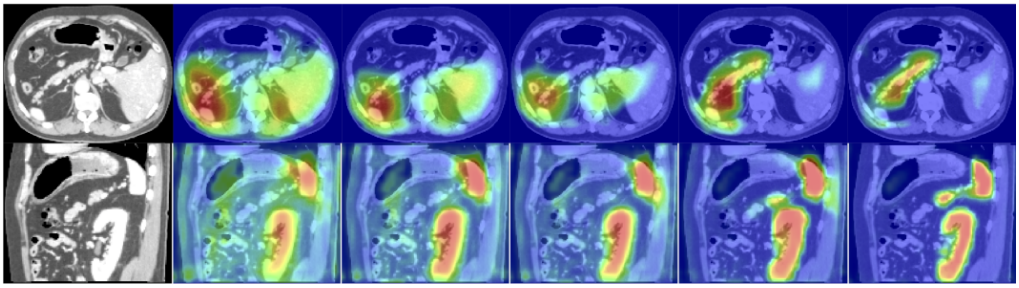


Figure 1.4: The model gradually learns to focus on the pancreas, kidney, and spleen (Oktay et al., 2018)

while highlighting salient features useful for a specific task. Fig. 1.4 visualize the process that the model gradually concentrates on the interested organs during training. Zhang et al. (2020) validated attention U-Net on the 2D US dataset with needles and the results proved the mechanism was able to improve the segmentation accuracy without increasing too many model parameters.

Another type of attention mechanism was employed by Amiri Tehrani Zade et al. (2023). They applied the bounding box around the ground truth of the needle as the attention mask or filtered kernel. Then They calculated the point-wise multiplication of the attention mask and the input image. It aims to solve the class imbalance problem in which the pixel ratio of the needle and background is too low. Therefore, it is helpful to train the network in important areas by attending more to foreground regions.

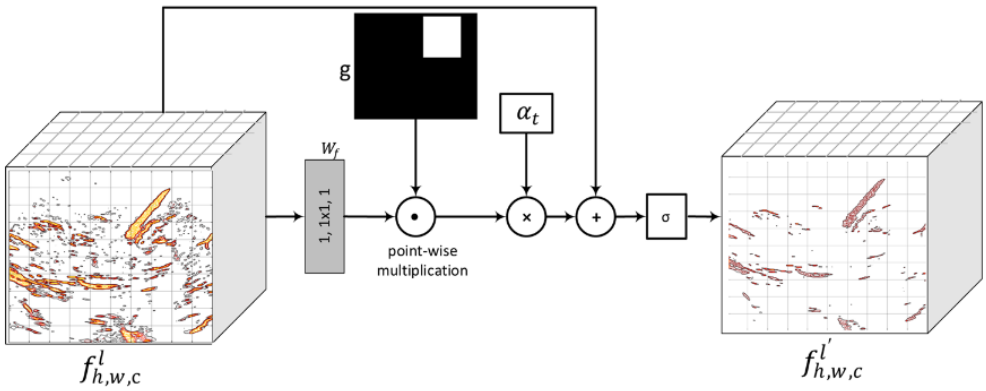


Figure 1.5: Attention mechanism with bounding box mask (Amiri Tehrani Zade et al., 2023)

1.3.3 Temporal Information

In the context of DL, temporal information refers to the consideration and utilization of time-related patterns and dependencies in data. It involves analyzing and understanding the sequential nature of data, where the order and timing of data points are crucial for accurate predictions or decisions (Wang et al., 2020). In the field of computer vision, temporal information is critical for tasks such as action recognition or video

analysis. By incorporating temporal information, DL models can capture the dynamics and movements within a sequence of frames, enabling more accurate recognition and understanding of actions or events.

In US images, particularly in the US-guided needle insertion task, the needle is continuously inserted into the body tissue. As a result, the variation between consecutive frames is minimal, making it possible for the DL model to learn and predict the needle location accurately.Mwikirize et al. (2019) first proposed a time-difference-based regression and classification CNNs to localize the needle in 2D US. By using pixel-wise logical operation, which captures the fine motion of the needle tip and feeds it into CNNs for detection, it is possible to determine the differences between two consecutive frames (Fig. 1.7). However, the methods processed the temporal information in a separate preprocessing algorithm, which slowed down the inference speed of CNNs. Based on the previous work, Mwikirize et al. (2021) combined the long short-term memory (LSTM) block with CNN to further employ the temporal information in US image sequence (Fig. 1.6). The method increased 30% of detection accuracy than the previous work and meanwhile achieved a detection rate of 15 frames per second. Amiri Tehrani Zade et al. (2023) utilized the GunnarFarneback algorithm as a traditional motion field estimation method. It estimates the motion for all pixels in the image and calculates the motion vectors by analyzing the changes in pixel intensity between two frames of a video. They enhanced the training input with temporal features extracted from the stack of consecutive frames and achieved a significant improvement.

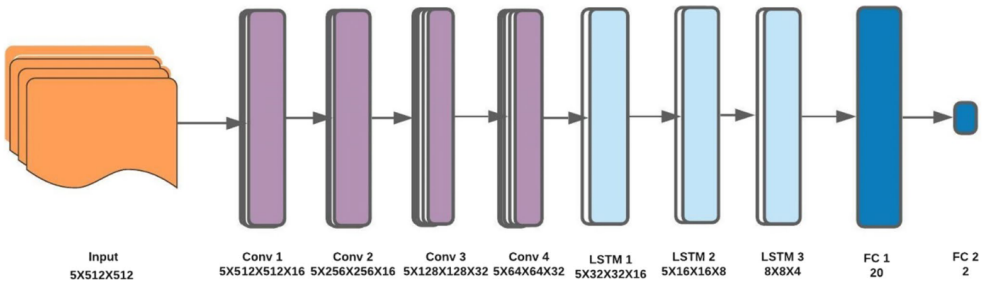


Figure 1.6: The model structure of CNN-LSTM model (Mwikirize et al., 2021)

1.4 Research Questions

The main research objective for this project is: **To better visualize and accurately locate the needle position during US-guided intervention**. To reach the research objective, research question has been proposed:

- Can the spatial attention mechanism and temporal information help improve the detection accuracy of the needle in US images?

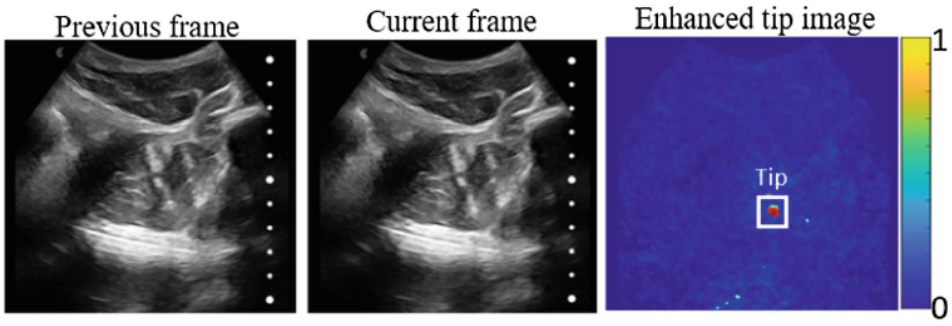


Figure 1.7: The needle tip enhancement with temporal information (Mwikirize et al., 2021)

1.5 Thesis Overview

The thesis is composed of four chapters. Chapter 1 presents the historical work done in the medical image segmentation field, which introduces the spatial attention mechanism and temporal information in previous work. Chapter 2 is concerned with the methodology and experiment setup used for this study. In particular, the mask attention UNet and its variants combined with optical flow will be illustrated. The experiment conditions are introduced in detail as well. In Chapter 3, results derived from multiple experiments are provided and the analysis of the capabilities and performance of the proposed models is conducted. Finally, the findings, conclusions, and outlook will be discussed in Chapter 4.

2 Methodology

The previous chapter outlined the research gaps in previous studies and the research objectives of the current thesis. The focus of this chapter is to discuss the methodology of the thesis, including the architecture of segmentation models, the experiment setup for the data collection, and the details for the model training and validation. § 2.1 will formally introduce the proposed Spatial Mask Attention U-Net (SMA-UNet) and Optical-Flow Attention U-Net (OFA-UNet). § 2.2 introduces the experiment setup and phantom construction for the US image collection. It also describes the process of labeling the ground truth of the needle in US image. Finally, § 2.3 discusses the implementation details for the model training and the evaluation metrics for the model validation.

2.1 Segmentation Model Architectrue

This section gives a detailed description of the two proposed segmentation models for needle segmentation in US images. The first SMA-UNet is inspired by the spatial attention mechanism which aims to suppress the background information and concentrate on the needle. The OFA-UNet utilizes the optical flow of the consecutive frames of US images to estimate the needle position for t+1 time and make it as the next frame's attention mask.

2.1.1 Spatial Mask Attention U-Net

As mentioned in § 1.3.2, the attention mechanism plays a key role in segmentation DL models because of its capability to concentrate on the region of interest (ROI) of input images. The attention U-Net was validated on the Panceras organs dataset and

has proven that it is able to automatically capture the relevant textural features and semantic meanings in encoder and decoder, respectively. However, there is a problem the needle is quite thin and the the needle only has a few pixels width in US images. It results in a significant imbalance between the foreground containing the needle and the background. While the organs occupy a considerable number of pixels throughout the entire ultrasound image. Consequently, the attention U-Net may not perform well on the needle detection task. Amiri Tehrani Zade et al. (2023) employed a bounding box around the ground truth of the needle as an attention mask to calculate point-wise multiplication with the input US image, which largely reduced the segmentation image size and made the model concentrate on the rectangular area where the needle may exist. However, the imbalance problem still exists in the ROI area. Therefore, the SMA-UNet is proposed by using an attention mask aligned with the needle shaft to address the category imbalance issue.

Model Architecture

The structure of the SMA-UNet is shown in Fig. 2.1. The baseline of the model is a classical U-Net model to segment the needle in the input US image. It also contains a contracting path to extract the textual features and an expansive path to recover the image. The contracting path indicated by the pink color is the encoder for the needle mask. Each block consists of convolutional, max-pooling, and activation layers to compress the input mask and make its image size and channel depth same as the corresponding layer. The red circle blocks represent the attention modules after each skip connection. They calculate the attention excitation between the skip connection and the encoded needle mask, the output of the attention is computed as equation 2.1.

$$M_o = M_c \oplus [I_e \otimes \mathrm{Sig}\,(M_e)] \tag{2.1}$$

where M_o, I_e and M_e stand for the output of attention mask, the extracted US image features and the encoded needle mask, respectively. $\oplus$ and $\otimes$ denote element-wise and channel-wise multiplication, respectively. Then the outputs are concatenated with the de-sampling layers to concentrate on the needle ROI. During the training stage, the needle masks are created by the ground truth of the needle in frame t-1. While in the inference stage, the mask will be created based on the output of the last frame, which enables the model to predict the needle in sequential US images.

Mask Creation

Since the segmentation output of the model probably not perfectly reflect the real needle position, it's possible to accumulate the error through the inference in a time-sequential US image. The potential solution is to enlarge the mask width to cover the region where the needle may exist. One of the morphological operations named dilation is employed to

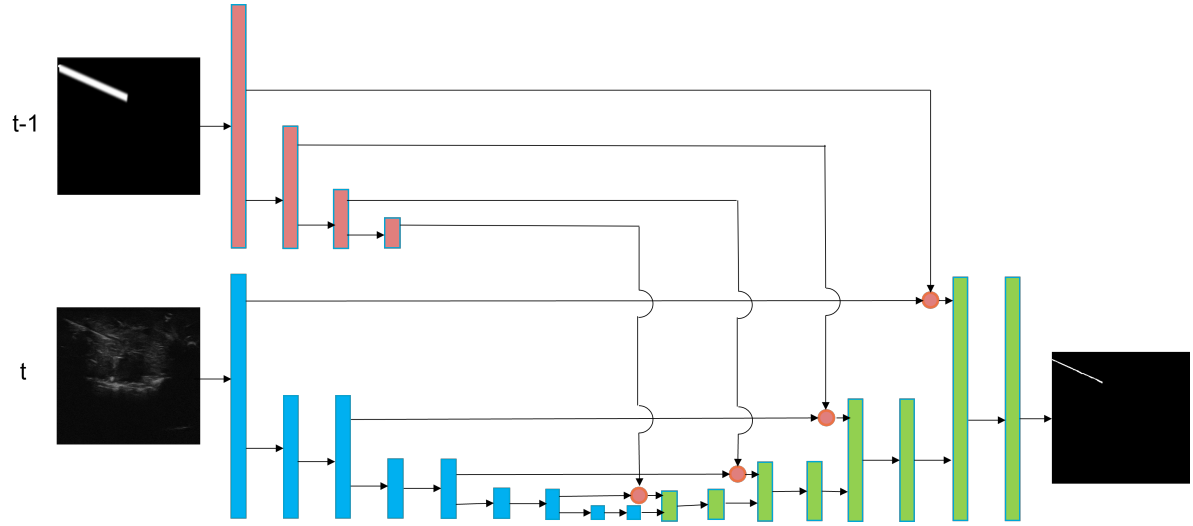


Figure 2.1: The model architecture of SMA-UNet. Blue and Green blocks represent the baseline U-Net structure. Red rectangular blocks represent the feature extraction path for ROI mask. Red dots represent attention blocks.

expand the needle mask width on the pixel level. Dilation is typically used to expand or grow the white regions (foreground) of a binary image while preserving its overall shape and connectivity (Serra, 2022). In dilation, a structuring element or a kernel is defined, typically a small binary image, and this element is overlaid on the input image. For each pixel in the input image, the dilation operation checks if there is any overlap between the structuring element and the corresponding neighborhood of the pixel. If any part of the structuring element overlaps with the foreground pixel(s) in the neighborhood, the corresponding pixel in the output image is set to white (foreground). This process is repeated for all pixels in the input image.

In mathematical description, If A and B are the image sets in N-space (E^N) contains elements a and b, repectively. The dilation of A by B is denoted by $A \oplus B$ and is defined by 2.2

$$A \oplus B = \{c \in E^N \mid c = a + b\}. \tag{2.2}$$

In order to choose the most suitable width for the attention mask, three sizes of structuring elements are chosen to construct the needle mask for segmentation. Fig 2.2 shows the mask created by dilation.

Post-processing

Post-processing is a common technique to process the output of the DL model aims to improve the accuracy and quality of medical image segmentation. Because the needle insertion process is slow to avoid misoperation and injury and the needle stiffness is high enough to endure the bending. It's reasonable to assume that the appearance of the needle in the US image is a straight line. However, the segmentation model sometimes categorizes the parts of the image which is not the needle into the needle

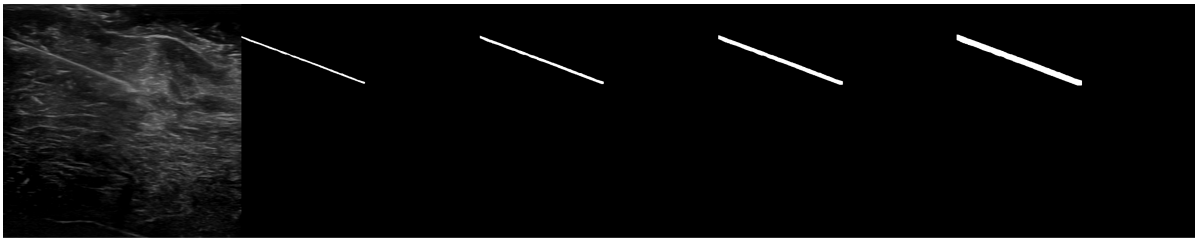


Figure 2.2: Different dilation results by three sizes of structuring elements. (a) the raw US image, (b) the ground truth of the needle, (c) dilation output by (3, 3), (d) (6, 6), (e) (10, 10) kernel, respectively

class. Moreover, the needle shaft probably is not continuous and has some breakpoints as shown in Fig. 2.3 These false outputs are so-called outliers with respect to the ground truth line of the needle. Intuitively, some line-fitting algorithms can be applied to remove the outliers. Random Sample Consensus (RANSAC) is an iterative method to estimate model parameters from a given dataset that contains outliers (Fishler, 1981). The RANSAC algorithm works by randomly selecting a subset of the data set, fitting a model to the selected subset, and determining the number of outliers. This process is repeated for a prescribed number of iterations. Fig. 2.3 (b) shows the output of the RANSACN algorithm and it successfully removes the outliers and fits the needle very well. While the linear fitting algorithm is affected by the outliers.

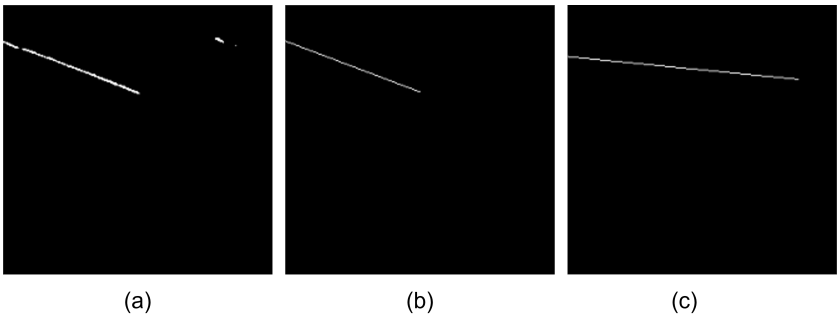


Figure 2.3: (a) the output of U-Net model, the red circle indicates the outlier and the yellow one indicates the breakpoint in needle line, (b) the fitting line of RANSAC algorithm, (c) the fitting line of Least Squares method

2.1.2 Optical Flow Attention U-Net

Although the SMA-UNet utilized the spatial attention mask generated from the last frame output to focus on the ROI of the needle and mitigate background interference, a problem remains that the needle tip movement fails to be captured by the SMA-UNet. Therefore, inspired by the (Amiri Tehrani Zade et al., 2023), an optical flow method is employed to estimate the motion filed and enhance the model input with temporal information.

Optical Flow

Optical flow is a fundamental concept in computer vision and image processing. It refers to the apparent motion of objects, edges, and surfaces in a visual scene caused by the relative motion between the observer (camera) and the scene itself. In other words, optical flow describes how pixels in an image appear to move from one frame to the next. There are two branches of approaches to calculating the optical flow, one is the traditional algorithm and another one is DL method.

Lucas and Kanade (1981) proposed the Lucas Kanade (LK) algorithm to calculate dense optical flow. It is a two-frame differential optical flow estimation algorithm, whose basic idea is based on the following three assumptions:

- Brightness constancy: The pixels of the target image in the scene appear to move from frame to frame without changing. For grayscale images (the same applies to color images), this means that the gray values of the pixels do not change with the tracking of the frame.
- Temporal continuity (small movement): The camera movement on the image changes slowly over time. In fact, this means that changes in time do not cause drastic changes in the position of the pixels, so that the gray values of the pixels can take the corresponding partial derivatives of the position.
- Spatial consistency: Adjacent points on the same surface in the scene have similar motion, and their projection onto the image plane is also close.

The LK algorithm for optical flow is based on the optical flow constraint $I_x u + I_y v + I_t = 0$. Here I are partial derivatives and u and v are the directional x and y components of the flow vectors. Theoretically, solving the optical flow equation gives us the optical flow.

However, the assumptions limit the application of LK algorithm. Learning-Based Optical Flow methods have drawn researchers' attention in recent years. Dosovitskiy et al. (2015) first proposed a DL-based architecture for optical flow estimation. Further, Ilg et al. (2017) proposed the 'Flownet 2.0' based on the previous work. They proposed an encoder-decoder network architecture, which consists of two components: an encoder and a decoder. The encoder takes in two consecutive frames and encodes them into feature maps, which are then used by the decoder to estimate the optical flow between the two frames. FlowNet2 also introduces a cost volume to further improve the accuracy of the optical flow estimation.

Model Architecture

The idea behind the OFA-UNet is inspired by the concept of analog signals, which are continuous physical properties such as force and temperature. In ultrasound scans of anatomy, the two consecutive B-mode images are continuous when the US recording frequency is high enough. This means that the position and geometry of the needle

remain similar between two consecutive images. This knowledge allows us to use the mask segmented from the previous frame to generate an attention map for more accurate segmentation of the current frame. Furthermore, to account for the tissue variation caused by the probe movement in the successive US images, the optical flow estimation method is utilized. The flow images F are estimated by feeding the current image I_t and the previous image I_{t-1} to the pre-trained FlowNet2. The potential mask of the current frame M_t is estimated based on the current flow image F_t and the mask of the previous frame M_{t-1} using equation 2.3.

$$
\begin{aligned}
P_{t-1}^{\text{mask}} &= \{p = (i,j) \mid m(i,j) == 1\} \quad m(i,j) \in M_{t-1} \\
P_{t}^{\text{mask}} &= \{p + F_t(p)\} \quad \forall p \in P_{t-1}^{\text{mask}} \\
M_t'(i,j) &= \begin{cases} 1 & \text{if } (i,j) \in P_t^{\text{mask}} \\ 0 & \text{others} \end{cases}
\end{aligned} \tag{2.3}
$$

where $p = (i, j)$ is the pixel position, P_{t-1}^{mask} is the set of the pixel positions, where the image value is one on M_{t-1}. P_t^{mask} is the prediction of the mask of the current frame based on the flow image. $M_t^{'}(i, j)$ is the predicted binary mask of the current image. The mask attention multiplication is the same as 2.1.

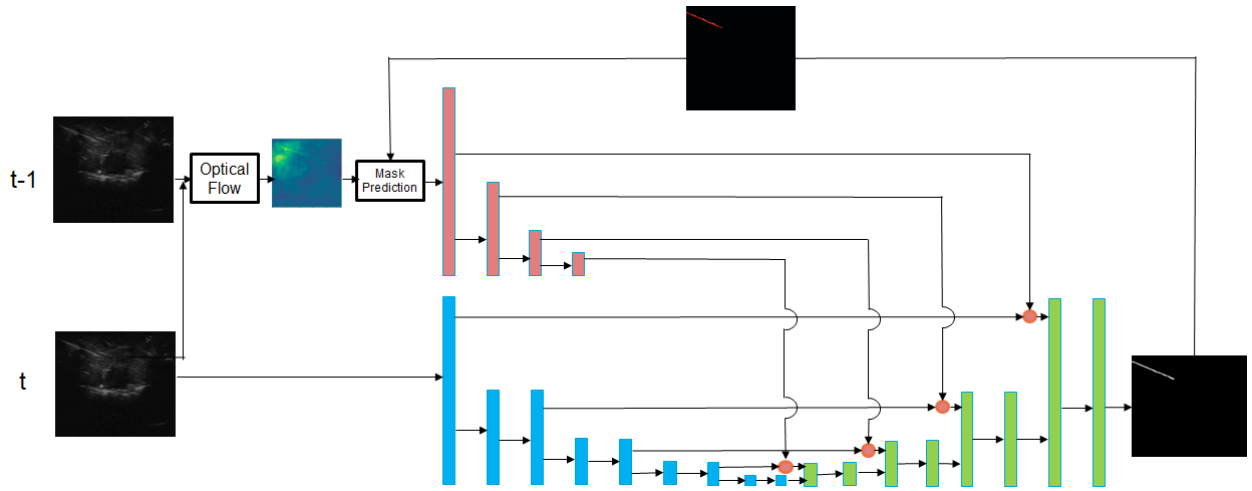


Figure 2.4: Illustration of the proposed optical flow-based UNet architecture

2.2 Data Preparation

In this section, the workflow of US data collection will be celebrated including the phantom construction, experiment setup, data pre-processing, data augmentation, and annotations.

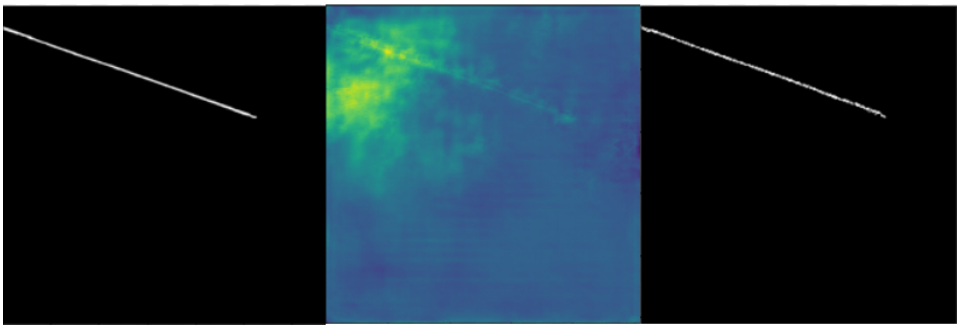


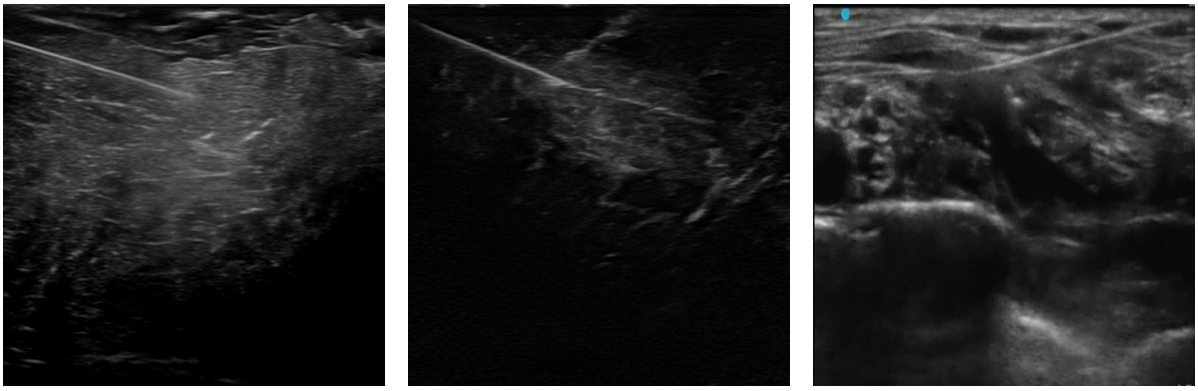
Figure 2.5: The prediction of the needle mask for frame t based on the optical flow and the ground truth of the previous frame

2.2.1 Dateset Description

In this work, both the *ex-vivo* dataset and the *in-vivo* dataset are utilized for model training and validation. *Ex-vivo* dataset is a widely used dataset type, that refers to a dataset composed of data collected outside the living organism or body. Normally the data is acquired from tissue samples, animal organs or other biological materials removed from a living organism. It enables the researchers to perform experiments in an isolated and controlled laboratory setting. *In-vivo* dataset is a kind of offline dataset from a live animal experiment or from patients. Because the dataset is recorded in a real clinical application, it usually reflects the complicated anatomical structures and clinical environment, which is the best approach to validate the developed algorithms (Yang et al., 2022).

Ex-vivo phantoms were constructed to mimic the complicated and clinical environment for needle insertion during minimally invasive interventions. Two phantoms were made from porcine and bovine tissue, respectively. The porcine tissue contains more intramuscular fat, which results in stronger echo reflections from the fat layers. Therefore, in ultrasound images, porcine tissue displays pronounced reflection patterns in US image (Fig. 2.6b). The bovine tissue has comparatively less intramuscular fat compared to the porcine tissue, and the muscle fibers are thicker. In ultrasound images, it exhibits a clearer layered texture (Fig. 2.6a).

In-vivo dataset is collected from patients and it commonly represents complicated anatomical structures such as vessels, bones, or organs (Fig. 2.6c). The collected US images possibly show the periodic deformation or pulse in a video because of the human breath or heart beating. Moreover, the individual differences of patients may result in the various appearances of the same target area. In this thesis, An *In-vivo* dataset which is open-source online was used (Tyagi, 2023). It was collected from patients for US-guided anesthesia. However, detailed information such as the needle type, target area and US configuration are limited.



(a) pork **(b)** beef **(c)** patient

Figure 2.6: Three sample US images selected from the pork, beef and patients datasets, respectively

2.2.2 Experiment Setup

In order to collect the *ex-vivo* dataset from porcine and bovine phantoms, a system was established. The overall experimental setup is shown in Fig. 2.7. A linear probe was installed on the end-effector of a KUKA robotic manipulator using a 3D-printed mount. A Siemens Ultrasound system was connected to the linear probe and visualized the real-time US image on the screen. The US images from the Siemens system were captured by a frame grabber and transferred to a personal computer. During the experiment, the robotic arm controlled the probe to descend until it contact with the phantom and a clear US image appeared on the screen. Then the needle was manually inserted into the phantom with an insertion angle around 20 degrees. The phantom's position was adjusted to ensure the in-plane appearance of the needle was displayed on the screen. The ImFusion software installed on the computer recorded the US video of the needle insertion process and stored data in a local disk. The collected US video frame rate is 30 frames per second.

Table 2.1: Hardware Setup

Equipment	Type
Robotic arm	KUKA LBR iiwa 7 R800, KUKA Roboter GmbH, Augsburg, Germany
Linear probe	12L3,Siemens AG, Erlangen, Germany
Ultrasound system	ACUSON Juniper Ultrasound System, Siemens AG, Erlangen, Germany
Frame grabber	DVI2USB 3.0, Epiphan Video, Ottawa, Canada

To ensure that the produced phantom can mimic a biological environment effectively, fresh samples of pork and beef were collected before the experiment. These samples were tightly wrapped in plastic wrap and placed in plastic containers to maintain the muscle tissue's density. A certain amount of water was added to the container to expel any air bubbles between the plastic wrap and the tissue sample. Additionally, ultrasonic

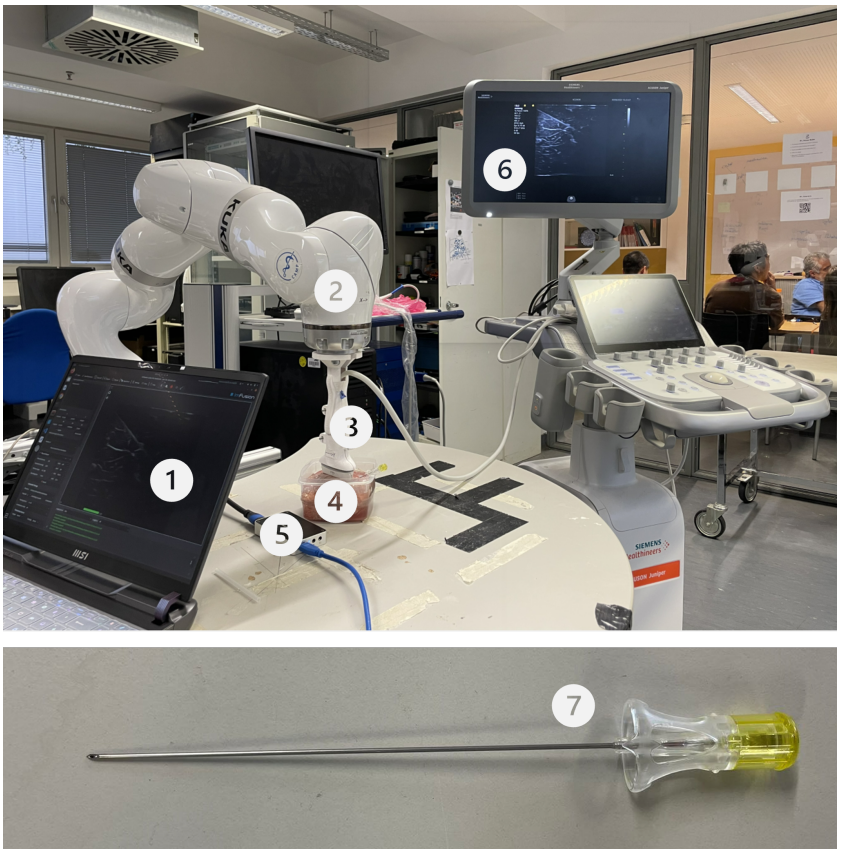


Figure 2.7: Experiment setup for US data collection. 1. PC, 2. KUKA robotic arm, 3. Linear US probe, 4. Phantom, 5. Frame grabber, 6. US system, 7. Needle (18 G, 1.27 mm)

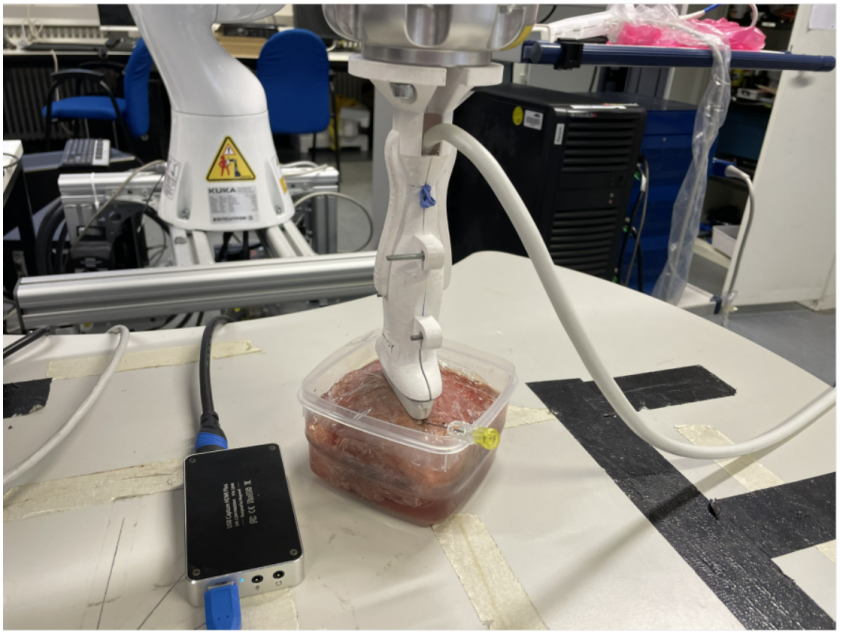


Figure 2.8: Coordinating the needle and US probe to ensure a clear in-plane needle is visualized in US image

couplant was applied to the plastic wrap to prevent noisy signals caused by air bubbles in the US image.

2.2.3 Data Processing

Pre-processing

The quality of the image data will significantly affect the results of DL models, especially in medical segmentation tasks. Therefore, several techniques were applied to the collected US data before they were used as the input for segmentation models. The pre-processing workflow in this study involved the following key steps:

- Conversion: Converting the 'IMF' files acquired in ImFusion software into the 'PNG' image format.
- Resizing: Adjusting the dimensions of all US images to a uniform size of (256, 256) to meet the input specifications of the segmentation model.
- Normalization: Ensuring that all US images are rescaled to fall within the range of [0, 1].

Expect the pre-processing process, data augmentation was also employed. Because the size of datasets acquired is limited, data augmentation is able to increase the size of datasets by using various transformations or modifications to the existing data. This significantly enhances the generalization and the performance of DL models. The data augmentation process has the following steps in this work:

- Rotation: Randomly rotating images by certain degrees (90°, 180°, 270°).
- Flip: Randomly flipping images either horizontally or vertically.
- Contrast Enhancement: Applying histogram equalization to adjust the contrast and brightness of images.

All of the above augmentation steps were randomly applied with a probability of 30%. Both the pre-processing and data augmentation were performed in OpenCV and Python 3.6.

Ground Truth labeling

Ground truth is crucial for training a DL model. During the training process of a supervised learning model, the model compares the output value with the ground truth value and adjusts the model parameters through back-propagation methods. Therefore, accurately labeling needles in ultrasound images is essential for training a precise and robust model. According to the literature review conducted earlier, there is still no standardized benchmark dataset for needle detection and localization in ultrasound images. As a result, the ground truth of the needle in collected datasets has to be manually

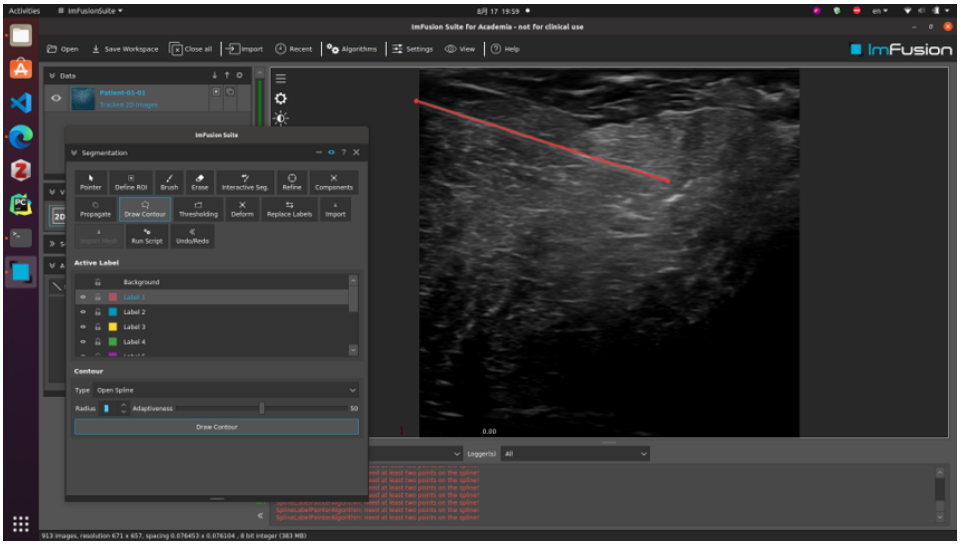


Figure 2.9: ImFusionSuite for labeling

annotated. In this work, the annotation process was performed using ImFusion software (Fig. 2.9), which provides convenient annotation tools. Since the bending of the needle during insertion is negligible, the needle can be assumed to appear as a straight bright line in the ultrasound image. As a result, the ground truth of the needle was drawn to cover the area of the needle, with a width of 5 pixels in ImFusion software. The needle annotations were assigned a value of 1, while the remaining pixels of the ultrasound images were recognized as the background and assigned a value of 0. The annotation process was guided by an expert sonographer at University Hospital rechts der Isar.

2.3 Implementation Details

This section will introduce all the DL model training details such as the framework of DL, optimizer, learning rate, batch size, training epochs, and early stopping strategy. Moreover, the experiments to compare the effects of different training loss, mask width and optical flow algorithms will be introduced. Lastly the metrics for model performance evaluation are discussed.

2.3.1 Experiments Design

Training loss is a crucial and fundamental factor in the field of deep learning. It plays a vital role in guiding the optimization process of neural networks and is a key metric used to assess the performance of models. Different loss functions are suitable for different types of tasks and datasets. For example, mean squared error (MSE) is typically used for regression tasks, while cross-entropy loss is commonly applied to classification tasks or segmentation tasks. In this work, three experiments were conducted to select the most suitable loss function for our needle segmentation task. As mentioned in § 2.1, the width

of the attention mask may affect the result of SMA-UNet. Therefore, three experiments were performed to train the SMA-UNet using three different width masks generated. Furthermore, the accuracy of the computed optical flow has a significant impact on generating the subsequent frame's attention mask, which affects the final model output and model stability. Therefore, two experiments were conducted to generate optical flow using the LK algorithm and the FlowNet2 algorithm, respectively. Lastly, with the selected configuration for model training, four models were validated on pork, beef and patients datasets for the model performance comparison and to evaluate the consistency of the segmentation model across different kinds of datasets.

2.3.2 Training Configuration

All the models were implemented in Pytorch 1.10.12 framework (Paszke et al., 2017) and trained on a computer with NVIDIA GeForce GTX TITAN X 12GB GPU at TUM IFL lab. The Adam optimizer with a learning rate of 0.001 was adopted for each experiment. The model parameters were initialized using the Kaiming initialization method (He et al., 2015). The datasets were divided into training dataset, testing dataset, and validation dataset, the details of dataset information are summarised in Table 2.2. Leave-one-out strategy for training was applied to test the trained model in each epoch and iterate the model parameter according to the testing results. To save the computational resources and avoid the overfitting problem, an early stopping strategy was taken. The maximum epoch was set as 100. The training process would automatically halt if the values of evaluation metrics have not improved for 20 epochs. The model parameters would be saved for every 10 epochs. The batch size was assigned as 16, which means 16 US images were utilized for training in one iteration. A larger batch size is able to best utilize parallel computing and accelerate the training and meanwhile, it will not cause GPU memory overflow. Therefore, the input dimension for models is hence $N \times C \times H \times W$, where N refers to batch size, C is the number of channels, in this case the value is 1 because of the grayscale image. H and W are the image height and width of the input data, respectively. Their values are all 256.

Table 2.2: Dataset partition

Dataset	Training	Testing	Validation
Bovine phantom	12 videos 4140 frames	460 frames	2 videos 617 frames
Porcine phantom	13 videos 2459 frames	273 frames	3 vidoes 598 frames
Patients	5 videos 843 frames	94 frames	1 video 123 frames

2.3.3 Loss Functions

As aforementioned, a suitable loss function plays a vital role in the optimization process of DL models and may yield a better performance of the model. In this work, three loss functions were selected, which are Dice Loss (Dice), Binary Cross Entropy (BCE) and Dice Binary Cross Entropy (DiceBCE).

Dice Loss

Dice loss is one of the commonly used functions in image segmentation tasks. Sorensen (1948) proposed a Dice similarity coefficient (DSC) to measure the similarity between two samples. In image segmentation tasks, it is used to assess the overlap or similarity between the predicted image and the ground truth. The original form of DSC is defined as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.4}$$

where $|X|$, $|Y|$ are the numbers of elements each sets. In the segmentation task, the mathematical form of Dice loss is modified from the original DSC as follows:

$$\mathcal{L}_{\text{Dice}}(y, \tilde{y}) = 1 - DSC = 1 - \frac{2\sum_{i=1}^{N} y_i \tilde{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \tilde{y}_i + smooth} \tag{2.5}$$

where, y_i refers to the output of the model that predicts the probability of the ground truth in pixel level, $\tilde{y}_i$ is the ground truth label of the pixel. *smooth* is a factor that prevents the denominator from being 0. It was set as 1 in this work.

Binary Cross Entropy

Cross entropy was first proposed in information theory. Then It is used to define a loss function in machine learning and optimization. Binary cross entropy, also known as log loss or logistic loss, is typically used for binary classification problems, where each element belongs to one of two classes. Therefore, it can also be applied in the needle segmentation task because the pixels of the needle were labeled as 1 and the pixels of the background were labeled as 0. BCE loss is defined as:

$$\mathcal{L}_{\text{BCE}}(y, \tilde{y}) = \frac{1}{N} \sum_{i=1}^{N} [y_i \log \tilde{y}_i + (1 - y_i) \log (1 - \tilde{y}_i)] \tag{2.6}$$

Dice Binary Cross Entropy

Dice Binary Cross Entropy is another commonly used loss function in segmentation tasks. It combines the Dice loss and the BCE loss to improve the performance of segmentation models. DiceBCE loss is defined as:

$$\mathcal{L}_{\text{DiceBCE}} = \alpha \mathcal{L}_{\text{Dice}} + (1 - \alpha) \mathcal{L}_{\text{BCE}} \tag{2.7}$$

		Actual	
		True (needle)	False (background)
Prediction	True (needle)	True Positives(TP)	False Positives(TP)
	False (background)	False Negative(FN)	True Negatives(TN)

Figure 2.10: Confusion Matrix

where α is a factor that controls the weights of two losses. In this work, the α was set as 0.5.

2.3.4 Evaluation Metrics

As mentioned in § 1.4, the research aims to determine the feasibility of using attention mechanisms and temporal information to improve segmentation accuracy and reduce localization errors compared to traditional CNN methods. Therefore, it is essential to define appropriate metrics for performance evaluation. In this project, two categories of metrics have been selected. The first category evaluates the performance of the model at the pixel level, focusing on segmentation. The second category measures the error between the predicted needle tip and the ground truth tip, as well as the angular error, from a geometrical perspective.

Segmentation Metrics

Accuracy is the most commonly used metric in classification and segmentation tasks. Accuracy is defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

where TP, TN, FP, FN represent True Positives, False Positives, False Negative, True Negatives, respectively. The confusion matrix elaborates the four concepts more clearly as Fig. 2.10 shows. Each column in a confusion matrix represents an actual class, while each row stands for the predicted class. The four elements in the matrix count the number of times instances of class A are classified as class B for all four pairs.

Actually, image segmentation can be assumed as a classification task. However, instead of assigning one label for a given image, each pixel in the input image is assigned a label. As a result, the model needs to produce output with the same dimensions as the input, providing the probability of each class. , the model needs to produce output with the same dimensions as the input, providing the probability of each class. Even if there is an imbalance between classes, it is possible to manually remove some of the images

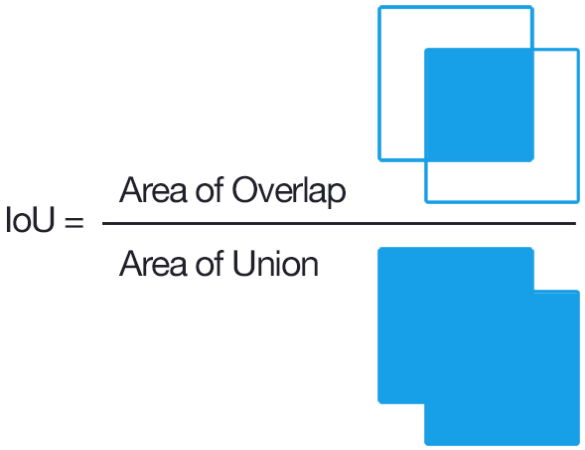


Figure 2.11: intersection over union

to ensure the balance of the dataset. However, this is not possible in the segmentation task, as the distribution of pixels in images is fixed and unchangeable. In the needle segmentation task, it is easy to find that there is a significant issue of sample imbalance. The needle is too narrow and only occupies hundreds of pixels while the background has tens of thousands of pixels. Assume that the input image actually has a needle but the model predicts that there is no needle in the image. In this case, the model's prediction is totally wrong. However, the accuracy of the model is almost 95% or even higher, which seems that the model has a correct prediction. Therefore, it is not possible to use accuracy as the evaluation metric for this project.

Considering the number of needle pixels is significantly low compared to the background, precision can be a valuable metric to deal with such imbalanced datasets. Precision highlights correct predictions, focuses on true positive predictions and minimizes the impact of false positives. Therefore, it provides a clear measurement of how well the model identifies that needle. Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

Regional-based metrics are usually used to evaluate complex tasks with imbalanced samples. These metrics provide a more fine-grained analysis compared to pixel-wise metrics. DSC is one of the regional-based metrics and has been introduced in § 2.3.3. Another way to calculate the DSC can be expressed as:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{2.10}$$

Intersection over Union (IoU) is another popular regional-based metric. It quantifies the accuracy of a segmentation model by measuring the intersection over union. IoU is typically used to assess the quality of individual segmented regions or objects. In this

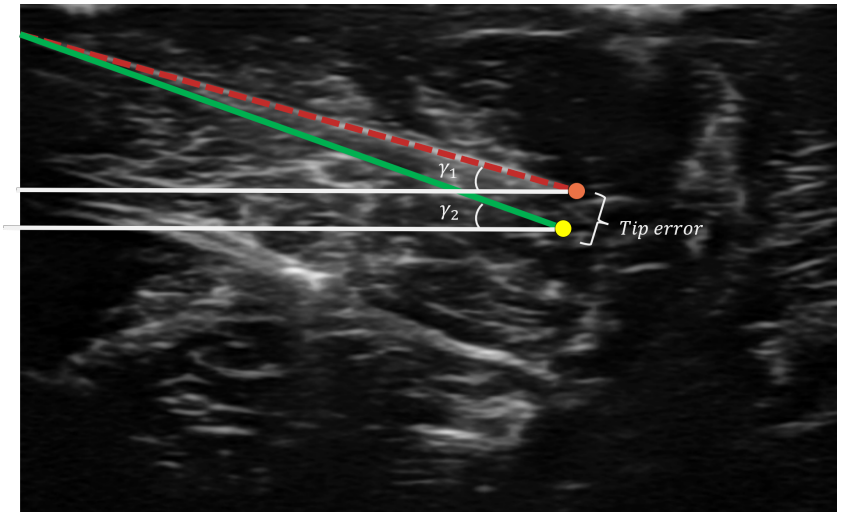


Figure 2.12: Illustration of tip error and angular error from geometric prospective

work, the needle segmentation is more important so IoU fits well for the evaluation. Fig. 2.11 illustrates the IoU. The mathematic form of IoU is defined as:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \tag{2.11}$$

where X, Y represent the predicted area and the ground truth, respectively. The IoU can also be calculated in terms of the confusion matrix:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2.12}$$

Geometric Metrics

The segmentation metrics evaluate the performance of the model at the pixel level and present numerical results, while geometric metrics are able to assess the accuracy of detecting needles in US images from a geometric perspective. According to the literature review done before, tip localization error and orientation angular error are commonly used in needle detection and localization tasks. Fig 2.12 shows the ground truth and predicted needle and factors used in calculating geometric errors.

Tip error (TE) is the Euclidean distance between the predicted needle tip and the ground truth tip. It can be expressed as:

$$TE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2} \tag{2.13}$$

where N represents the total number of images for validation, (x_i, x_i) and $(\tilde{y}_i, \tilde{y}_i)$ represents the coordinates of the predicted needle tip and the coordinates of the ground truth needle tip, respectively.

Angular error (AE) is the difference between the angle subtended by the predicted needle on the horizontal axis and the angle subtended by the ground truth on the horizontal axis. AE can be expressed as:

$$AE = \frac{1}{N}\sum_{i=1}^{N}|\gamma_i^1 - \gamma_i^2| \tag{2.14}$$

where γ_i^1, γ_i^2 represent the orientation angle of the predicted needle and the ground truth, respectively.

Statistical Analysis

Since several metrics have been selected to evaluate the performance of the proposed models, it is necessary to analyze the results and draw conclusions based on statistical analysis. For example, to select the best-performing training loss, we first need to employ 'analysis of variance (ANOVA)' to check whether there is a statistically significant difference in the metric results of the three losses. Then, the 'student t-test' method will be used to pairwise compare the three sets of result data and determine if there are significant differences between them. Finally, the comparison conclusion will be drawn based on the statistical analysis. In this work, if the p-value (probability value) associated with a statistical test is less than 0.05, the results are considered statistically significant. If the distribution of the results does not follow the normal distribution, a non-parametric analysis method like Mann-Whitney U test is used to evaluate the results. This approach ensures that the obtained findings are robust and reliable, providing a solid foundation for further interpretation and decision-making.

3

Results and Discussions

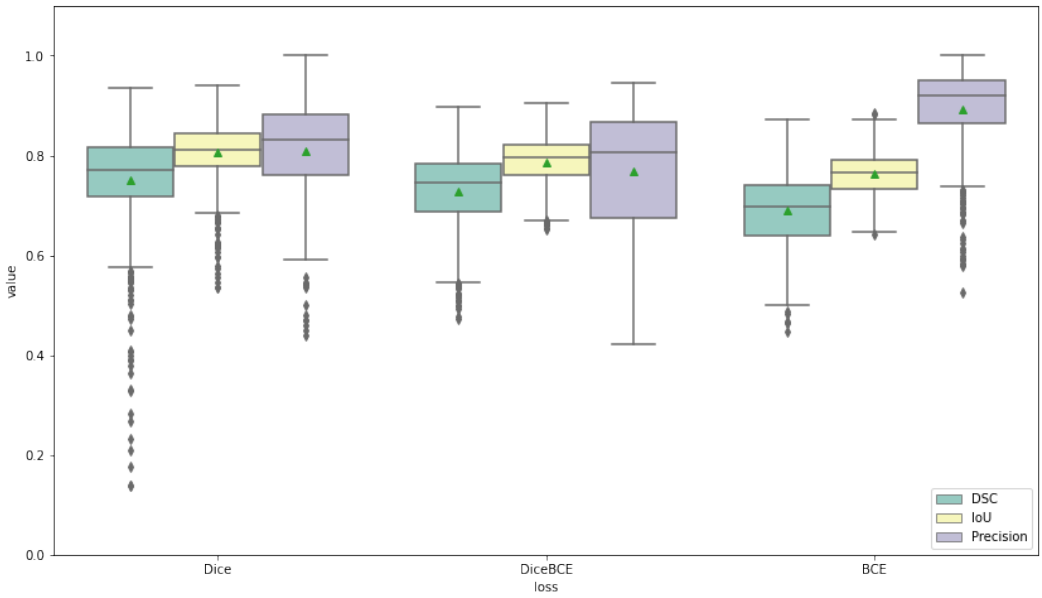
This chapter presents the experiment results evaluated on selected metrics. All models were trained on the training datasets and evaluated on separate validation datasets, which were not used during training. A comparison of the experiment results is made to assess the impact of training loss, mask width, and optical flow methods on segmentation models. The optimal model configurations are then selected for final validation on three datasets collected from pork, beef, and patients.

3.1 Training Loss Selection

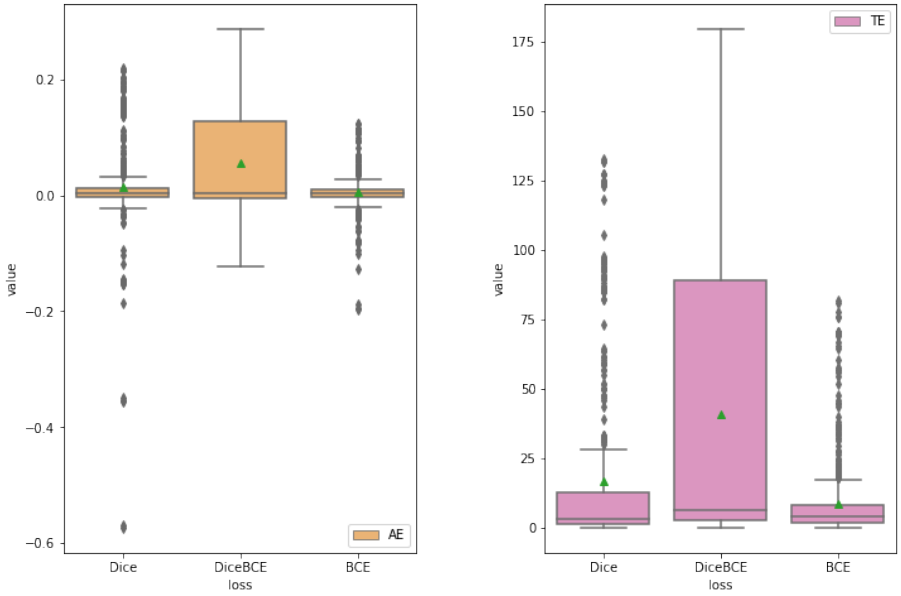
As mentioned in § 2.3.3, the loss function in training phase affects the convergence direction and the speed of a DL model. Different loss functions can guide the model to adjust its parameters in different ways during training, thus affecting the accuracy of model outputs. Therefore, three experiments were conducted by using Dice, DiceBCE and BCE to train three separate U-Net model on the pork dataset. Following are the quantitative results of three models and the discussion about the results.

3.1.1 Quantitative Results

The trained models were evaluated by five selected metrics and the results are presented in Fig. 3.1. It consists of three boxplots. Fig. 3.1a depicts the DSC, IoU, Precision based on segmentation metrics. Fig. 3.1b and Fig. 3.1c depict the angular error and tip error, respectively. All boxplots indicate the mean, median and standard deviation (SD). The results are recorded in Table 3.1.



(a) DSC, IoU and Precision



(b) AE

(c) TE

Figure 3.1: Bloxplots of evaluation results for U-Net models trained with three losses on pork validation dataset

Table 3.1: Performances of U-Net on pork dataset with different training losses

Metrics	Dice	DiceBCE	BCE
DSC (%)	**77.0 ± 4.8**	74.5 ± 4.4	69.8 ± 4.8
IoU (%)	**81.1 ± 3.1**	79.5 ± 2.9	76.6 ± 2.8
Precision(%)	**80.9 ± 5.7**	76.7 ± 7.9	89.3 ± 3.9
AE (10^{-3} rad)	**2.8 ± 7.2**	3.5 ± 1.5	3.3 ± 6.1
TE (mm)	**6.1 ± 4.2**	12.3 ± 10.4	7.9 ± 5.1

3.1.2 Discussion

In Fig. 3.1a, it can be observed that the mean values of DSC and IOU on models trained with Dice, DiceBCE, and BCE are sequentially decreasing while the distribution ranges of them are similar. Moreover, The mean of Presion on the model trained with BCE is higher than that of models with Dice and DiceBCE. The AE of DiceBCE has a large data distribution range and is not exactly normal distribution, therefore Mann-Whitney U test was performed. The outliers of Dice have a larger range than those of BCE. The trend is almost the same in TE as Fig. 3.1c shows. The DiceBCE has a largest distribution and the maximum outlier of Dice is larger than that of BCE. All the comparisons were proved to have statistically significant difference ($p<0.05$).

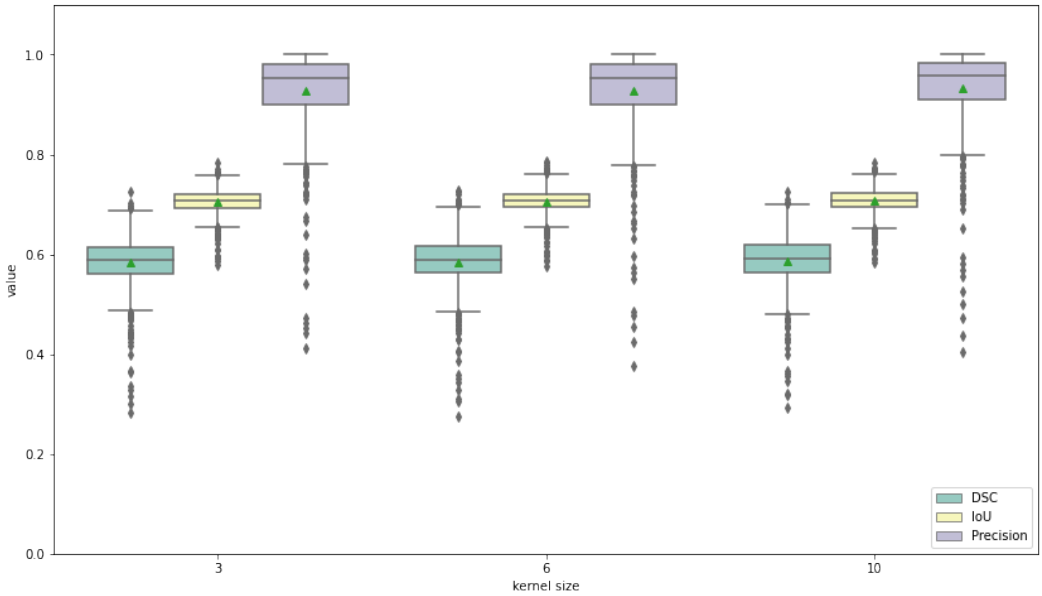
The overall performance of the model trained with Dice is the best among the three training losses, except for the metric Precision. One possible reason is the presence of a significant sample imbalance in the needle datasets. The Dice loss is able to focus on the prediction of positive samples, while the BCE loss is sensitive to imbalanced samples and tends to be affected by the dominant class. Both Dice and DiceBCE show similar performance on AE and TE, with means close to 0 and acceptable standard deviations. However, DiceBCE exhibits larger variance and higher values for both TE and AE compared to Dice alone. Therefore, it is not suitable for the needle detection task. According to Table 3.1, Dice achieves the best performance across all four metrics, making it the chosen training loss for the following experiments.

3.2 Mask Width Selection

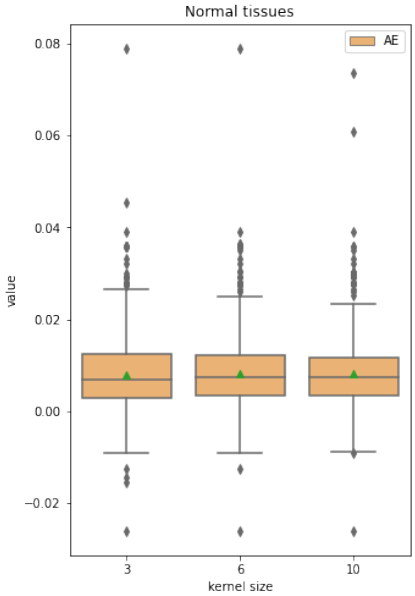
Both SMA-UNet and OFA-UNet incorporate the mask attention unit in their model structure to make the model concentrate on the ROI where the needle may exist. However, whether the mask width affect the model prediction is not clear. Therefore, three SMA-UNet models were trained on different masks which were generated by dilation operation. The following are the quantitative results of three models and a discussion about the results.

3.2.1 Quantitative Results

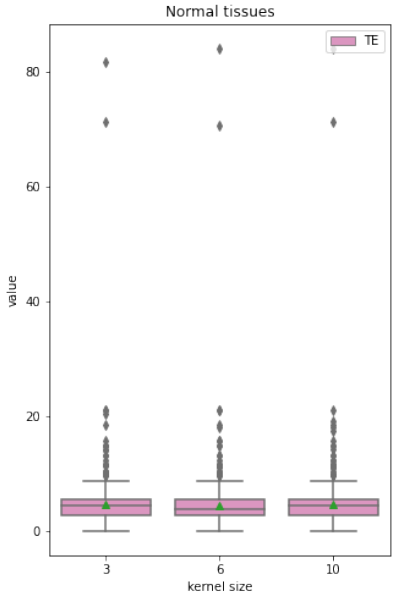
The trained SMA-UNet models were evaluated using five metrics, and the results are presented in Fig. 3.1. The evaluation includes three boxplots. Fig. 3.1a shows the DSC, IoU, and Precision based on segmentation metrics. Fig. 3.1b and Fig. 3.1c represent the angular error and tip error, respectively. All boxplots display the mean, median, and standard deviation (SD). The results are recorded in Table 3.1.



(a) DSC, IoU and Precision



(b) AE



(c) TE

Figure 3.2: Bloxplots of evaluation results for SMA-UNet models trained with three sizes of masks

Table 3.2: Performances of the SMA-UNet on pork dataset with different sizes of masks

Metrics	Kerner size: 3	Kerner size:6	Kerner size: 10
DSC (%)	**58.3 ± 5.4**	58.4± 5.6	58.7± 5.4
IoU (%)	**70.4 ± 2.5**	70.5 ± 2.7	70.7 ± 2.6
Precision(%)	**92.6 ± 8.4**	92.7 ± 8.2	93.3 ± 7.9
AE (10^{-3} rad)	**6.7 ± 8.4**	7.4 ± 8.4	7.4 ±8.3
TE (mm)	**4.5 ± 1.1**	4.7 ± 1.6	4.8 ± 1.6

3.2.2 Discussion

Since all results follow a normal distribution, a pairwise t-test was conducted between the results of three models on five metrics. T-test results indicated that all pairwise comparisons have no statistically significant difference (p-value > 0.05). Therefore, the mask width has no impact or little impact on the model performance. Fig. 3.2a illustrates that the three SMA-UNet metrics share a common feature: the Precision is relatively high and close to 1, while the DSC is lower than the IoU and Precision. The potential reason for this is that SMA-UNet uses the output of the last frame as the attention mask for the current frame prediction. Due to the insertion of the needle, it is difficult for the model to learn the needle feature near the needle tip, which leads to a decrease in true positives and an increase in false negatives. As a result, there is a high Precision and relatively low DSC. In terms of AE and TE, three models with different size mask also have similar distribution and few far outliers. Although different mask widths have no significant differences, according to the numerical results in Table 3.2, the model with mask generated by kernel size 3 has lower means on all metrics. It was chosen to be the mask size for the following experiments.

3.3 Optical Flow Algorithm Selection

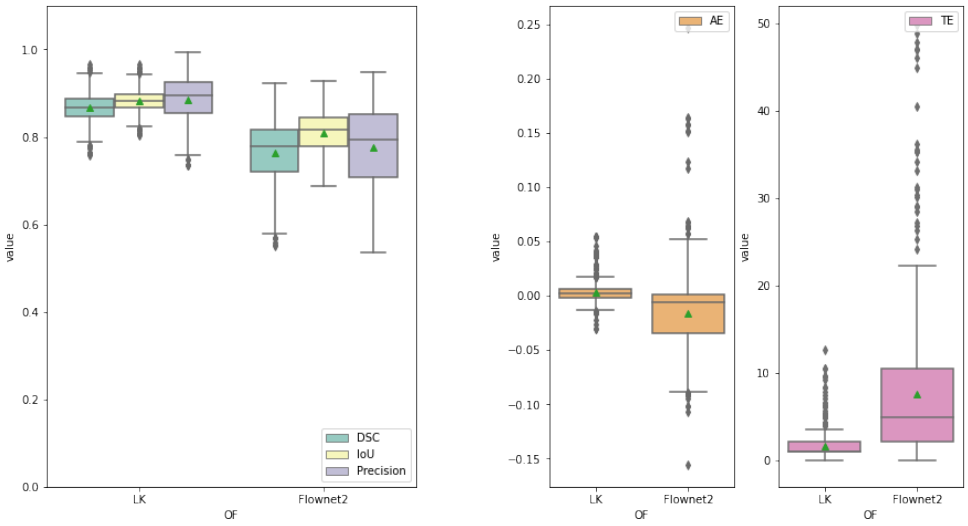
Optical flow calculates the motion fields between two adjacent US frames and predicts the current frame's attention mask based on the motion fields and the last frame's attention mask. The accuracy of calculated optical flow is vital for needle detection and localization. Two OFA-UNet models using LK algorithm and Flownet2 were trained to compare the performance of different optical flow methods. The following are the quantitative results of three models and a discussion about the results.

3.3.1 Quantitative Results

Three OFA-UNet models were assessed using five metrics, and the outcomes are presented in Fig. 3.3a. The evaluation includes three boxplots. Fig. 3.3a illustrates the DSC, IoU, and Precision based on segmentation metrics. Fig. 3.3b and Fig. 3.3c represent the angular error and tip error, respectively. All boxplots display the mean, median, and standard deviation (SD). The results are recorded in Table 3.3.

3.3.2 Discussion

Student t-tests were pairwise conducted between the metric results of LK and FlowNet2. The results showed statistically significant differences (p-value < 0.05). Fig. 3.3a illustrates that the means of DSC, IoU, and Precision using the LK algorithm are higher than those of the model using FlowNet2. Additionally, the variance of the three metric



(a) DSC, IoU and Precision **(b)** AE **(c)** TE

Figure 3.3: Bloxplots of evaluation results for OFA-UNet models trained with two optical flow methods on pork validation dataset

Table 3.3: Performances of the OFA-UNet on pork dataset with different optical flow algorithms

Metrics	LK	FlowNet2
DSC (%)	**86.7 ± 3.3**	76.2 ± 7.2
IoU (%)	**88.2 ± 2.5**	80.9 ± 4.6
Precision(%)	**88.6± 4.6**	77.6 ± 9.3
AE (10^{-3} rad)	**2.4 ± 3.8**	6.5 ± 1.3
TE (mm)	**2.7 ± 1.1**	9.9 ± 7.1

values using LK is smaller than that of FlowNet2. Fig. 3.3b and Fig. 3.3c demonstrate that the means of AE and TE in the model using LK are lower compared to the model using FlowNet2, and these means are close to 0. Conversely, FlowNet2 exhibits a larger variance in AE and TE, with the TE even reaching 50 mm as an outlier.

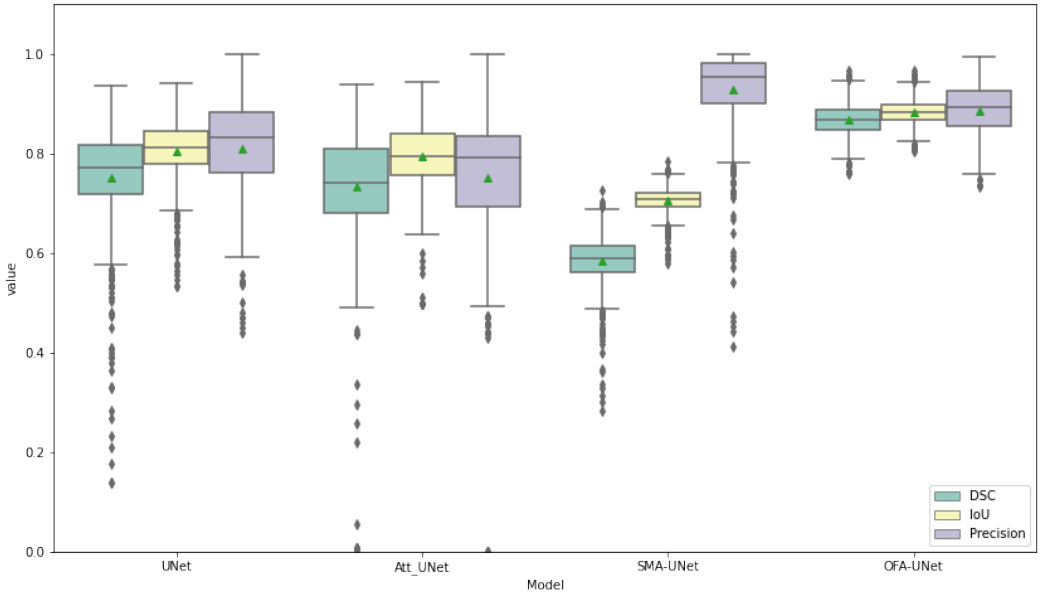
These metrics collectively indicate that the SMA-UNet with the LK algorithm performs significantly better than the model with FlowNet2. There are several reasons that can explain this phenomenon. Firstly, FlowNet2 was pre-trained on a dataset containing real-world object images, but it was not trained on a US dataset. This may result in poor performance of optical flow on US datasets. Secondly, FlowNet2 may have advantages in complex and dynamic scenes. However, in needle insertion scenarios, the texture features of the needle and tissue are not overly complex, and the LK algorithm tends to perform well based on mathematical modeling. Lastly, the Lucas-Kanade method typically performs well in scenes with small motion, while FlowNet2 is more suitable for handling large-scale and complex motions. In summary, the LK algorithm for optical flow exhibits better performance and is suitable for US images. Therefore, it was selected for the upcoming experiments.

3.4 Performances of Four Models on Three Datasets

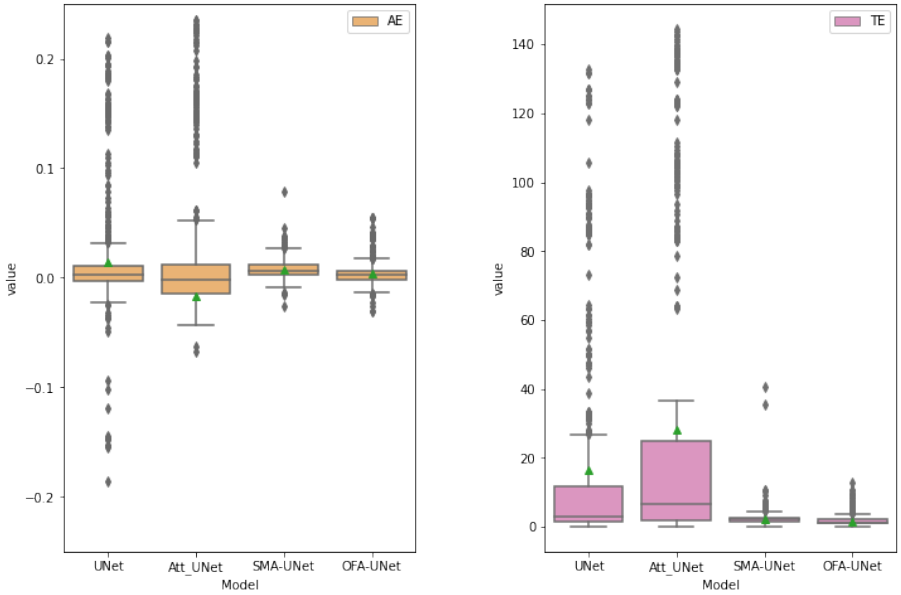
After the three selection sections above, the effects of training loss, mask width and optical flow methods on segmentation models are analyzed and the optimal training configuration is selected. In this section, UNet, Attention UNet, SMA-UNet and OFA-UNet were trained on pork, beef and patients datasets separately. Their validation results are also presented.

3.4.1 Quantitative Results

Four models trained on three datasets were evaluated using five metrics, and the outcomes are presented in Fig. 3.4, Fig. 3.5 and Fig. 3.6, respectively. Each figure includes three boxplots. Fig. 3.3a, Fig. 3.4a and Fig. 3.5a illustrate the DSC, IoU, and Precision based on segmentation metrics. Fig. 3.4b, Fig. 3.5b and Fig. 3.6b represent the angular error of models trained on three datasets, and Fig. 3.4c, Fig. 3.5c and Fig. 3.6c represent the angular error. All boxplots display the mean, median, and standard deviation (SD). The results of model performance on three datasets are recorded in Table 3.4, Table 3.4 and Table 3.4, respectively.



(a) DSC, IoU and Precision



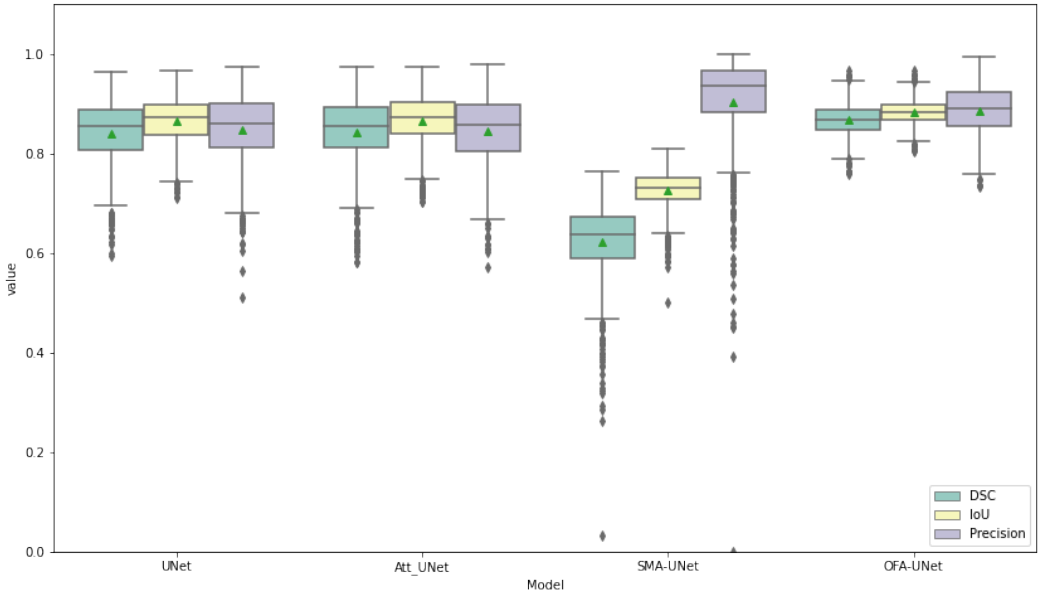
(b) AE

(c) TE

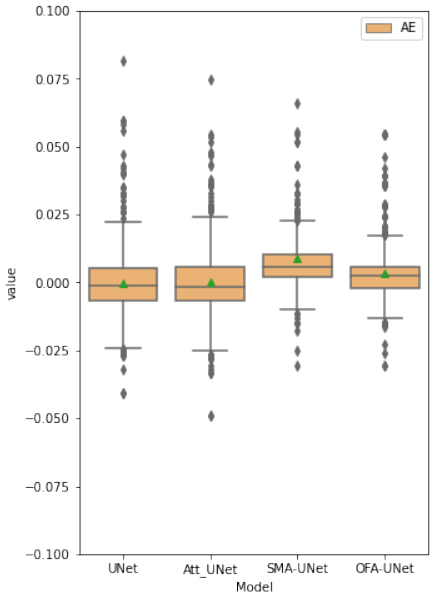
Figure 3.4: Bloxplots of evaluation results for four models on pork validation dataset

Table 3.4: Performances of four models on the pork dataset

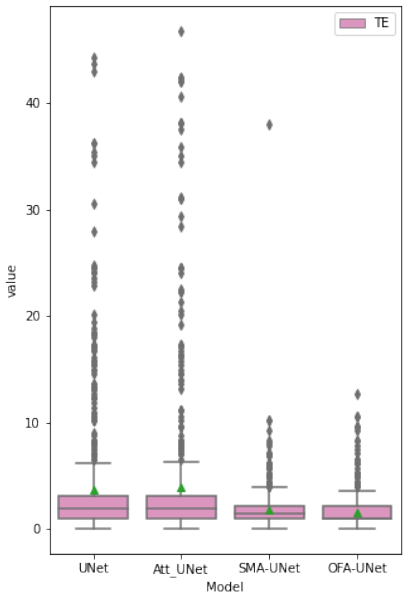
Metrics	U-Net	Attention U-Net	SMA-UNet	OFA-UNet
DSC (%)	77.0 ± 4.8	74.1 ± 6.4	58.3 ± 5.4	**86.7 ± 3.3**
IoU (%)	81.1 ± 3.1	79.3 ± 4.0	70.4 ± 2.5	**88.2 ± 2.5**
Precision(%)	80.9 ± 5.7	79.2 ± 5.0	92.6 ± 8.4	**88.6± 4.6**
AE (10^{-3} rad)	2.8 ± 7.2	1.9 ± 1.3	6.7 ± 8.4	**2.4 ± 3.8**
TE (mm)	6.1 ± 4.2	5.6 ± 3.5	4.5 ± 1.1	**2.7 ± 1.1**



(a) DSC, IoU, Precision



(b) AE

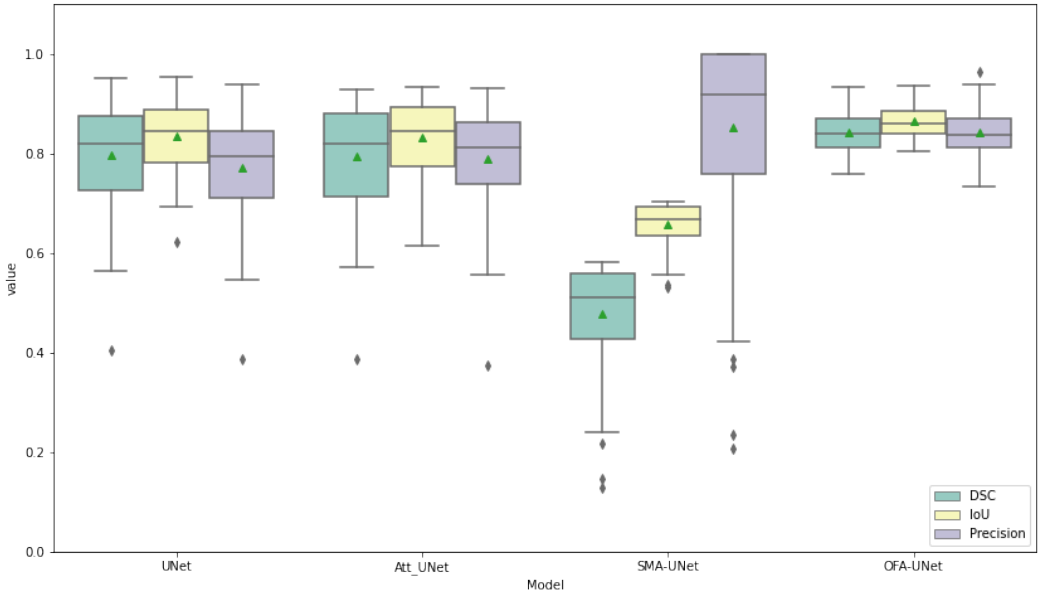


(c) TE

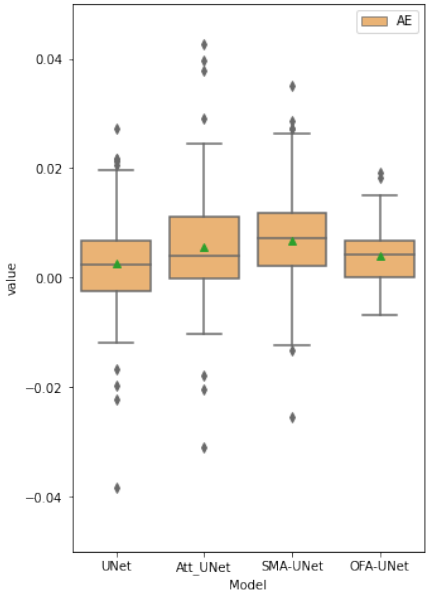
Figure 3.5: Bloxplots of evaluation results for four models on beef validation dataset

Table 3.5: Performances of four models on the beef dataset

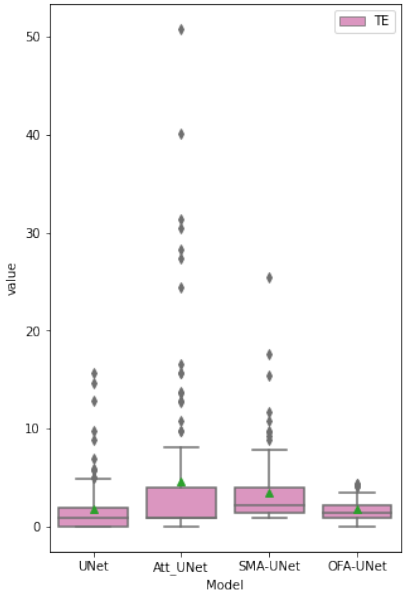
Metrics	U-Net	Attention U-Net	SMA-UNet	OFA-UNet
DSC (%)	80.2 ± 4.1	80.3 ± 4.4	62.1 ±8.0	**87.2 ± 4.1**
IoU (%)	84.7 ± 8.9	84.8 ± 9.1	72.5 ± 3.9	**88.3 ± 2.8**
Precision(%)	80.8 ± 4.7	80.5 ± 4.8	93.1 ± 10.2	**88.5 ± 4.5**
AE (10^{-3} rad)	51.7 ± 23.7	52.2 ± 23.8	8.8 ± 4.5	**3.4 ± 4.2**
TE (mm)	12.1 ± 11.4	12.8 ± 11.7	3.4 ± 2.1	**3.0 ± 1.8**



(a) DSC, IoU, Precision



(b) ae



(c) te

Figure 3.6: Bloxplots of evaluation results for four models on patients validation dataset

Table 3.6: Performances of four models on the beef dataset

Metrics	U-Net	Attention U-Net	SMA-UNet	OFA-UNet
DSC (%)	79.6 ± 10.2	79.3 ± 3.391	47.9 ± 9.8	**84.2 ± 4.0**
IoU (%)	83.5 ± 6.8	83.3 ± 6.9	65.7 ± 4.0	**86.4 ± 3.0**
Precision(%)	77.1 ± 9.9	79.0 ± 9.8	85.3 ± 18.1	**84.2 ± 4.8**
AE (10^{-3} rad)	2.5 ± 9.0	5.5 ± 10.3	6.7 ± 8.8	**4.2 ± 3.8**
TE (mm)	3.7 ± 3.2	9.1 ± 7.4	7.1 ± 6.1	**3.5 ± 2.0**

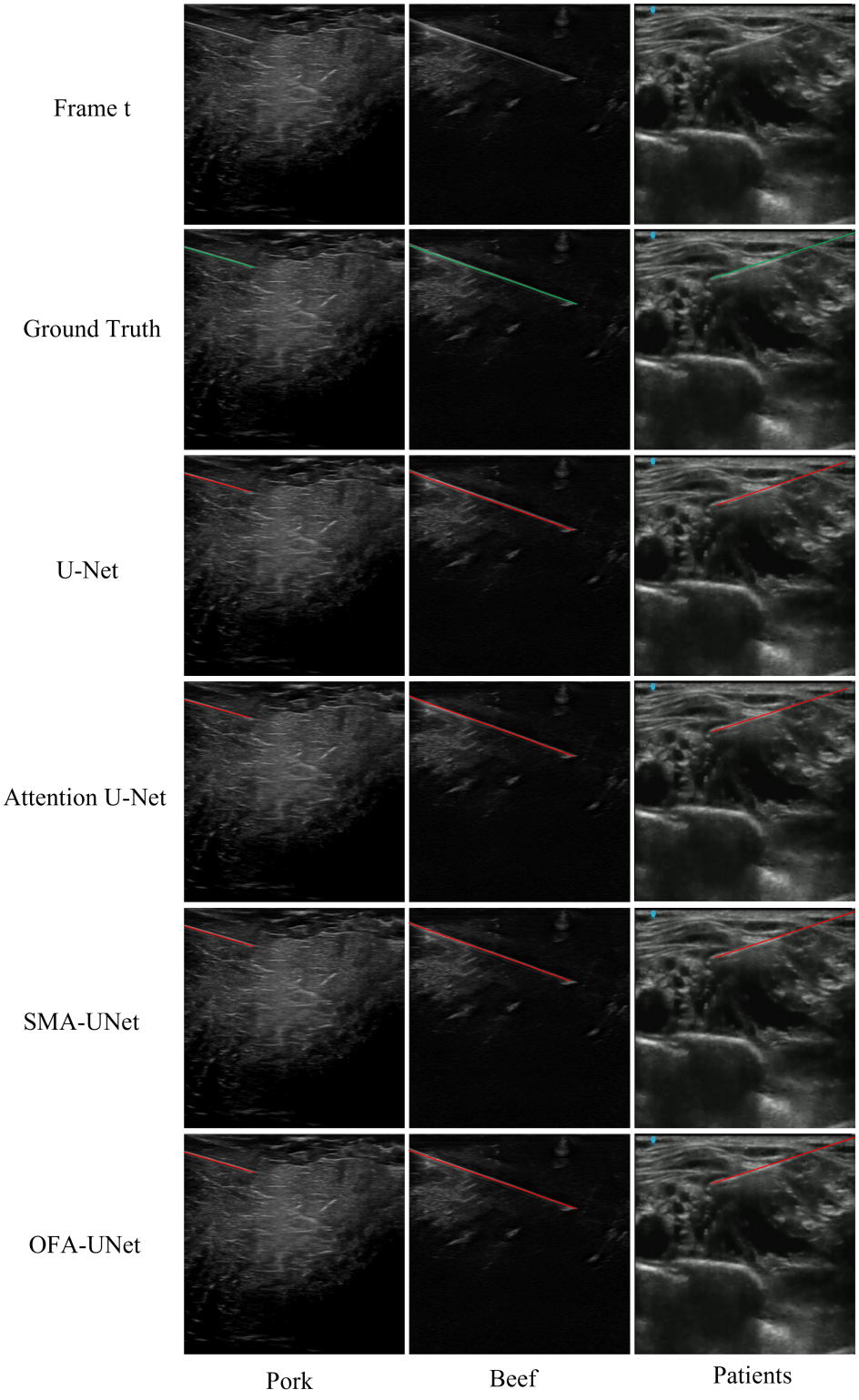


Figure 3.7: Qualitative results of needle detection by different models on pork, beef and patients US image. The green lines and red lines represent the ground truth of the needle and the predicted needle, respectively.

3.4.2 Qualitative Results

Sample qualitative results are shown in Fig. 3.7. Three sample US images were selected from pork, beef and patients datasets and the needle detection results of four models are visualized in US images. All four models successfully detected the needles. However, the U-Net and Attention U-Net models perform poorly when the needle visibility is not clear, particularly at the top right corner of the patient's US image. The SMA-UNet model performs better than U-Net and Attention U-Net, but it fails to capture the needle tip accurately compared to the ground truth. On the other hand, the OFA-UNet model exhibits the best performance and overcomes the interference of artifacts and poor visibility.

3.4.3 Discussion

Model Performance on Pork Dataset

In Fig. 3.4a, the results of UNet, SMA-UNet, and OFA-UNet have been analyzed previously. The results of UNet and Attention UNet do not have statistically significant differences in terms of five metrics (p-value > 0.05). SMA-UNet achieved the highest precision score compared to the other three models, but its DSC and IoU scores are the lowest. The variances of its metric values are smaller than those of UNet and Attention UNet. The performance of OFA-UNet is the best among all four models. Its DSC, IoU, and Precision means are the highest, and the variances are the smallest.

Regarding AE and TE, as shown in Fig. 3.4b and Fig. 3.4c, both SMA-UNet and OFA-UNet have a small distribution range close to 0. However, the range of outliers for SMA-UNet is larger than that of OFA-UNet. The AE and TE of UNet and Attention UNet are significantly larger than those of SMA-UNet and OFA-UNet, and they also have a larger number of biased outliers. The maximum TE for UNet and Attention UNet is around 140mm.

The performance of Attention UNet does not show a superiority compared to the UNet baseline according to the statistical analysis. However, the performance of SMA-UNet on geometrical metrics is much better than that of UNet and Attention UNet. This is because the needle object is too small compared to the whole image, and the self-adaptive attention in Attention UNet might learn artifacts that appear similar to the needle shape, thereby biasing the needle's prediction and causing a large variance. Nevertheless, the attention mask generated based on the last frame's prediction restricts the ROI of the US image, forcing the model to concentrate on the ROI of the needle during training iterations. OFA-UNet further improves the model's performance on segmentation metrics, and the model is more robust than the other models due to its smaller variances and fewer outliers. This validates the effectiveness of introducing temporal information into the model and demonstrates that such a mechanism can learn

the movement features of the needle in a video sequence.

Model Performance on Beef Dataset

Four models that were trained and validated on the beef dataset have similar performance in DSC, IoU, and Precision (Fig. 3.4a). The results of UNet and Attention UNet also do not have statistically significant differences in terms of five metrics (p-value $>$ 0.05). The outlier numbers of the three metrics of UNet and Attention UNet are smaller compared to these two models trained on the pork dataset. OFA-UNet still performs best on all three metrics and the variance ranges are the smallest.

In terms of AE, all four models achieve good results and are distributed within this range from -0.025 radians to 0.025 radians, the variance range is significantly smaller compared with models trained on the pork dataset. The mean of AE in SMA-UNet results is slightly larger than others. Regarding TE, UNet and Attention UNet have similar distributions and have many biased outliers. The maximum of them is almost 50 mm. However, the SMA-UNet and OFA-UNet have a smaller distribution range and most of the outliers are smaller than 10 mm.

As the statistical analysis indicates, UNet and Attention UNet have almost the same performance on the beef dataset, as well as on the pork dataset. The reason is also the same. With regard to the better performance of TE and AE on the beef dataset than on the pork dataset, the possible reason is that the US images of beef phantom have clearer texture structures and fewer artifacts than those of pork phantom.

Model Performance on Patients Dataset

The evaluation results of UNet and Attention UNet on the patients dataset, in terms of DSC, IoU, and Precision, show no statistically significant differences (p-value $>$ 0.05). However, SMA-UNet still exhibits a low DSC and a high Precision on the patient datasets. The distributions of the three metrics tend to approach 0. Among the models evaluated on the patients dataset, OFA-UNet achieves the best segmentation performance, similar to the performance of models on the pork and beef datasets.

In terms of AE and TE, the results of UNet are lower than those of Attention UNet and SMA-UNet (p-value $<$ 0.05). OFA-UNet has the smallest variance in AE and TE, with the fewest outliers near the upper quartile point.

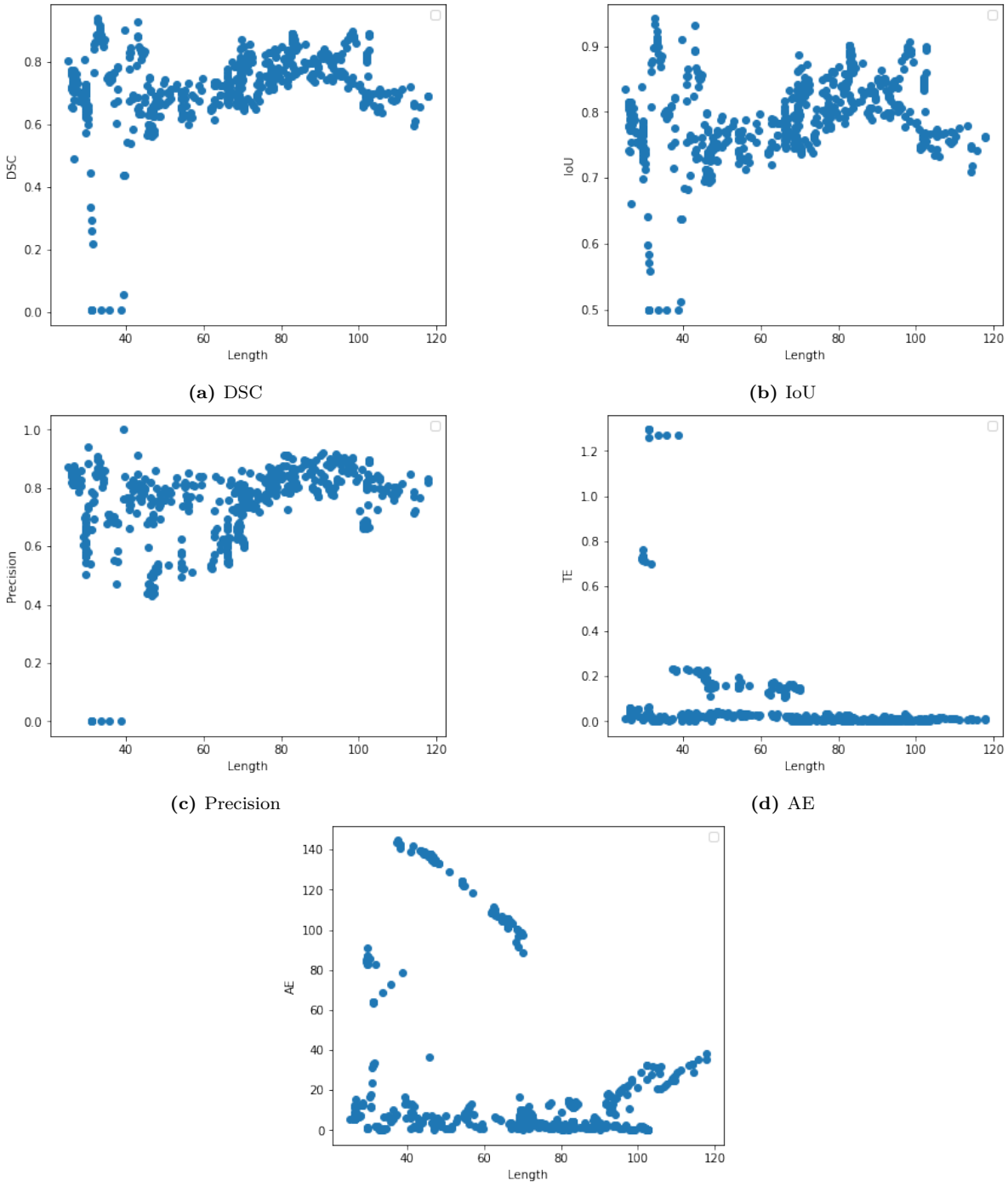
Unlike the metric results on the pork and beef datasets, UNet and Attention UNet do not exhibit statistically significant differences in segmentation metrics, but they differ in terms of AE and TE. Additionally, UNet performs the best in terms of AE and TE. This difference could be attributed to the small size of the validation set for patients, which consists of only 123 US images. The small validation set may introduce bias in the statistical results.

3.4.4 Outliers analysis

Most of the metric results show outliers in boxplots. For example, the boxplot of UNet's DSC in Fig. 3.4a has a long tail of outliers ranging from 0.1 to 0.6, and TE of UNet in Fig. 3.4c has outliers ranging from 30 to 130. However, it is difficult to determine the conditions under which these outliers appear and whether the model has poor performance based on the boxplot. Therefore, it is important to analyze the appearance of outliers in order to identify potential issues related to poor detection and localization of the needle.

During the needle insertion process under US guidance, all configurations were fixed, and the insertion angle of the needle was almost constant across all experiments to ensure optimal visualization of the needle. The only variable is the length of needle insertion. Therefore, five scatter plots are created to show the metric values across different insertion lengths (Figure 3.8).

Fig. 3.8a, 3.8b, and 3.8c reveal that most outliers occur at the beginning of the insertion (when the insertion length is less than 50 mm), and there is a tendency for the metric values to increase with the insertion length. However, the three metric results appear to drop after the insertion length reaches 90 mm. Fig. 3.8d and 3.8e show that outliers also exist at the beginning of the insertion when the insertion length is less than 70 mm. Both errors tend to decrease with the insertion length, while TE seems to increase after the insertion length reaches 90 mm. Five plots reveal that the segmentation model is sensitive to the length of the inserted text and is easily affected at the beginning and end of the insertion. This results in poor visibility at the edges of the US images.



(a) DSC

(b) IoU

(c) Precision

(d) AE

(e) TE

Figure 3.8: Relations between the segmentation model performances and the needle insertion depth. For each figure, the scatter plot shows the relation between metrics scores or errors of U-Net.

4

Conclusion and Outlook

4.1 Conclusion

In this project, SMA-UNet and OFA-UNet were proposed to capture the spatiotemporal feature of the inserting needle based on the spatial attention mechanism and temporal information. Two US datasets were established based on self-made pork and beef phantoms. To evaluate the influence of training loss, attention mask size and optical flow generated by different algorithms on needle detection and localization from segmentation and geometrical perspective by a hierarchical design flow. First, three UNet models were trained with Dice, DiceBCE and BCE losses. After comparing their performance, Dice loss was selected to be the optimal training loss for the model training as it is able to focus on the needle region regardless of the sample imbalance problem. Second, three sets of attention masks were created by dilation with three sizes of kernel. Then three SMA-UNet models were trained with the corresponding masks and validated on the pork dataset. The results show there are no statistically significant differences between the three models' performance on five metrics, which indicates that mask width has little impact on needle detection. The kernel size of 3 was chosen to generate the mask to avoid the model overfitting and increase the model robustness. Third, two OFA-UNet models were trained with LK and FlowNet2. Results indicate that the model using optical flow generated by LK algorithm achieved better performance than FlowNet2. Because the FlowNet2 was not fine-tuned on US datasets. Furthermore, UNet, Attention UNet, SMA-UNet and OFA-UNet were trained and evaluated on the pork, beef, and patients datasets with the selected training configuration. The results indicate that Attention UNet has no superiority to UNet across three kinds of datasets. SMA-UNet achieved better geometrical performance than UNet and Attention UNet, which proves

the effectiveness of the spatial attention mechanism. OFA-UNet performs best within five metrics across three datasets, which proves the introducing temporal information to the model is able to further capture the movement features of the needle during insertion. The performances of OFA-UNet are robust and consistent on three datasets, which demonstrate its adaptiveness to varying complexities of US datasets. Outliers analysis evaluates the needle detection errors across the needle insertion length. The results show the model outputs are less accurate at the beginning of insertion and limited by the visibility of the US image.

4.2 Outlook

Although the OFA-UNet has shown promising results in needle detection and localization, the performance of the model can be further improved through hyperparameter optimization. Since many hyperparameters, such as the optimizer and learning rates, were chosen experimentally, more combinations of hyperparameters need to be explored in order to achieve better performance.

In terms of US datasets, the insertion angle of the needle was almost fixed at around 20 degrees. It is unclear whether the model would achieve the same performance across various insertion angle ranges. Arif et al. (2018) and Van De Berg et al. (2019) investigated the needle visibility in relation to the insertion angle, which could provide guidance for collecting larger datasets with a wider range of angles. Moreover, most of the collected US images are in-plane, but this may not always be the case in clinical operations. Further exploration is needed to determine if the model can effectively handle out-of-plane scenarios.

Regarding the temporal information used in the OFA-UNet, the model was trained based on two consecutive frames, allowing it to capture short-term temporal information. However, it does not consider long-term sequential information. Mwikirize et al. (2021) combined long short-term memory (LSTM) recurrent neural networks with CNNs to learn spatiotemporal features from five consecutive frames, resulting in promising results. Therefore, it is possible to modify the OFA-UNet structure by introducing LSTM blocks for further exploration.

References

Abolhassani, N., Patel, R., & Moallem, M. (2007). Needle insertion into soft tissue: A survey. *Medical Engineering & Physics*, *29*(4), 413–431. https://doi.org/10.1016/j.medengphy.2006.07.003

Amiri Tehrani Zade, A., Jalili Aziz, M., Majedi, H., Mirbagheri, A., & Ahmadian, A. (2023). Spatiotemporal analysis of speckle dynamics to track invisible needle in ultrasound sequences using convolutional neural networks: A phantom study. *International Journal of Computer Assisted Radiology and Surgery*, 1–10.

Arif, M., Moelker, A., & van Walsum, T. (2018). Needle tip visibility in 3d ultrasound images. *Cardiovascular and Interventional Radiology*, *41*, 145–152.

Barrington, M. J., & Kluger, R. (2013). Ultrasound guidance reduces the risk of local anesthetic systemic toxicity following peripheral nerve blockade. *Regional anesthesia & pain medicine*, *38*(4), 289–299.

Cootney, R. W. (2001). Ultrasound imaging: Principles and applications in rodent research. *Ilar Journal*, *42*(3), 233–247.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. *Proceedings of the IEEE international conference on computer vision*, 2758–2766.

Douglas, B. R., Charboneau, J. W., & Reading, C. C. (2001). Ultrasound-guided intervention: Expanding horizons. *Radiologic Clinics of North America*, *39*(3), 415–428.

Fishler, M. A. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun, ACM*, *24*(6), 381–395.

Guo, D., Pei, Y., Zheng, K., Yu, H., Lu, Y., & Wang, S. (2020). Degraded image semantic segmentation with dense-gram networks. *IEEE Transactions on Image Processing*, *29*, 782–795. https://doi.org/10.1109/TIP.2019.2936111

Hatada, T., Ishii, H., Ichii, S., Okada, K., Fujiwara, Y., & Yamamura, T. (2000). Diagnostic value of ultrasound-guided fine-needle aspiration biopsy, core-needle biopsy, and evaluation of combined use in the diagnosis of breast lesions. *Journal of the American College of Surgeons*, *190*(3), 299–303.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, *32*, 582–596.

Hore, P., & Chatterjee, S. (2019). A comprehensive guide to attention mechanism in deep learning for everyone. *American Express*.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2462–2470.

Jan, M., Kalinšek, T. P., Štublar, J., Jelenc, M., Pernat, A., Žižek, D., & Lakič, N. (2020). Intra-cardiac ultrasound guided approach for catheter ablation of typical right free wall accessory pathways. *BMC Cardiovascular Disorders*, *20*, 1–8.

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *IJCAI'81: 7th international joint conference on Artificial intelligence*, *2*, 674–679.

Morgan, T. A., Jha, P., Poder, L., & Weinstein, S. (2018). Advanced ultrasound applications in the assessment of renal transplants: Contrast-enhanced ultrasound, elastography, and b-flow. *Abdominal radiology*, *43*, 2604–2614.

Mwikirize, C., Kimbowa, A. B., Imanirakiza, S., Katumba, A., Nosher, J. L., & Hacihaliloglu, I. (2021). Time-aware deep neural networks for needle tip localization in 2D ultrasound. *International Journal of Computer Assisted Radiology and Surgery*, *16*(5), 819–827. https://doi.org/10.1007/s11548-021-02361-w

Mwikirize, C., Nosher, J. L., & Hacihaliloglu, I. (2019). Learning needle tip localization from digital subtraction in 2D ultrasound. *International Journal of Computer Assisted Radiology and Surgery*, *14*(6), 1017–1026. https://doi.org/10.1007/s11548-019-01951-z

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. https://doi.org/https://doi.org/10.1016/j.neucom.2021.03.091

Oepkes, D., Devlieger, R., Lopriore, E., & Klumper, F. (2007). Successful ultrasound-guided laser treatment of fetal hydrops caused by pulmonary sequestration. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *29*(4), 457–459.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. *NIPS-W*.

Pourtaherian, A., Ghazvinian Zanjani, F., Zinger, S., Mihajlovic, N., Ng, G. C., Korsten, H. H. M., & de With, P. H. N. (2018). Robust and semantic needle detection in 3D ultrasound using orthogonal-plane convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, *13*(9), 1321–1333. https://doi.org/10.1007/s11548-018-1798-3

Rampersaud, Y., Simon, D., Foley, K., et al. (1999). Application-specific accuracy requirements for image guided spinal surgery. *Proceedings of the symposium of fourth computer assisted orthopaedic surgery*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241.

Scanlan, K. A., Propeck, P. A., & Lee Jr, F. T. (2001). Invasive procedures in the female pelvis: Value of transabdominal, endovaginal, and endorectal us guidance. *Radiographics*, *21*(2), 491–506.

Serra, J. (2022). Mathematical morphology. In *Encyclopedia of mathematical geosciences* (pp. 1–16). Springer.

Sheafor, D. H., Paulson, E. K., Simmons, C. M., DeLong, D. M., & Nelson, R. C. (1998). Abdominal percutaneous interventional procedures: Comparison of ct and us guidance. *Radiology*, *207*(3), 705–710.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, *5*, 1–34.

Tyagi, A. (2023). *Brachial plexus nerve segmentation*. Retrieved October 25, 2023, from https://github.com/Regional-US/brachial_plexus

Van De Berg, N. J., Sánchez-Margallo, J. A., van Dijke, A. P., Langø, T., & Van Den Dobbelsteen, J. J. (2019). A methodical quantification of needle visibility and echogenicity in ultrasound images. *Ultrasound in Medicine & Biology*, *45*(4), 998–1009.

von Haxthausen, F., Böttger, S., Wulff, D., Hagenah, J., García-Vázquez, V., & Ipsen, S. (2021). Medical robotics for ultrasound imaging: Current systems and future trends. *Current robotics reports*, *2*, 55–71.

Wang, S., Cao, J., & Philip, S. Y. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, *34*(8), 3681–3700.

Yang, H. X., Shan, C. F., Kolen, A. F., & de With, P. H. N. (2022). Medical instrument detection in ultrasound: A review. *ARTIFICIAL INTELLIGENCE REVIEW*. https://doi.org/10.1007/s10462-022-10287-1

Yi, J., Wu, P., Jiang, M., Huang, Q., Hoeppner, D. J., & Metaxas, D. N. (2019). Attentive neural cell instance segmentation. *Medical Image Analysis*, *55*, 228–240. https://doi.org/https://doi.org/10.1016/j.media.2019.05.004

Zhang, Y., Lei, Y., Qiu, R. L., Wang, T., Wang, H., Jani, A. B., Curran, W. J., Patel, P., Liu, T., & Yang, X. (2020). Multi-needle localization with attention u-net in us-guided hdr prostate brachytherapy. *Medical physics*, *47*(7), 2735–2745.