國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

時序模組輔助基於注意力機制特徵提取器用於超音波影像分割

Attention-based Feature Extractor with Temporal Module for Ultrasound Image Sequence Segmentation

王擇翔

Ze-Siang Wang

指導教授: 李明穗 博士

Advisor: Ming-Sui Lee, Ph.D.

中華民國 112 年 8 月

August, 2023

doi:10.6342/NTU202302561

國立臺灣大學碩士學位論文

口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

時序模組輔助基於注意力機制特徵提取器
用於超音波影像分割
Attention-based Feature Extractor with Temporal Module for Ultrasound Image Sequence Segmentation

本論文係王擇翔君（學號 R10922125）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 112 年 7 月 31 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Computer Science and Information Engineering on 31 July 2023 have examined a Master's thesis entitled above presented by ZE-SIANG WANG (student ID: R10922125) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李明穗 _____ _____

（指導教授 Advisor）

_____ _____ _____

系主任/所長 Director: _____

Acknowledgements

時光飛逝，碩班生活在忙碌又充實的步調中很快就來到了尾聲，這兩年不論是在研究上或是生活中，我都受到了太多人的幫助，這本論文絕不是我一個人能夠完成的。

首先，我要謝謝李明穗老師，在我當時還什麼都不會的情況下，讓我可以加入 IVLab 跟大家一起學習。在研究上給予建議和方向，不厭其煩的陪我討論問題。我學到的不只是專業領域的知識，還有做研究的精神和解決問題的能力。

此外，也要感謝曾文萱醫師在每次的討論中都給我許多建議及方向，幫助我能扎實地完成這篇論文。

接著我要謝謝柏維，跟我一起修課、吃飯、喝酒、運動，不論遇到什麼問題都能從他那邊得到解決方法，在研究方面也給了我很多建議。

也謝謝實驗室的學長們，世耘、偉綸、昱霖、柏燁，在我初來乍到時，給了我許多研究上的建議及指導，並且大家也都會在實驗室聊天、射飛鏢，營造出一個歡樂的實驗室氛圍。

謝謝瑋君一路以來的陪伴，在遇到困難時有你的陪伴，不斷地給我加油打氣，也陪我一起度過許多歡樂的時光。

最後也要謝謝我的父母家人，無條件的支持我完成學業，讓我沒有後顧之憂

doi:10.6342/NTU202302561

地能夠專注在學業上，對你們的感謝溢於言表。

doi:10.6342/NTU202302561

摘要

與他人溝通是我們日常生活中的一項基本能力。然而，患有聲帶萎縮的人在與他人溝通方面存在困難。值得慶幸的是，一種稱為注射增強的治療方法被創造來解決這種情況並在多年來被證明是有效的，且廣泛應用於許多聲帶疾病。在大多數情況下，醫生會將玻尿酸（Hyaluronic Acid）注射到患者的聲帶中，以改善聲門間隙並幫助聲帶正常閉合。過去，醫生必須從病人的發聲去判斷聲帶恢復情況以及是否需要補充玻尿酸。近來，使用超音波影像來分析玻尿酸在人體內殘留情況和作用位置是可行的。隨著電腦視覺領域的發展，可以使用電腦去幫助醫生追蹤玻尿酸在人體中降解作用以及估算出玻尿酸殘留體積。儘管基於 CNN 的模型在圖像分割任務中取得了優異的性能，但由於卷積運算的局部性，使得它們仍然無法學習全局和遠程信息。此外，當前大多數分割模型只關注分割任務中的空間特徵，忽略時間特徵。然而，時間特徵對於醫生推斷玻尿酸體積也很重要。因此，我們認為時間信息對於模型正確預測玻尿酸也是很重要的。在本研究中，我們提出了 AFTNet（注意力特徵時間網絡），其中包含基於注意力機制的特徵提取器和時間模組。借助基於注意力的特徵提取器和時間模組，我們的模型不僅可以更有效的學習全局和遠程信息，還可以更好地學習目標影片的時間特徵。我們將此模型應用於我們提出的患者喉嚨數據集，不僅能協助醫生解決難以判斷的鈣化以及雜訊案例，其性能優於基於 CNN 的模型和基於 Transformer 的模型。

關鍵字：超音波影像分割、Transformer、卷積神經網路、循環神經網絡、注射喉

doi:10.6342/NTU202302561

成形術

doi:10.6342/NTU202302561

Abstract

Communicating with other people is a basic ability in our daily life. However, those who suffering from vocal cord atrophy have trouble communicating with others. Thankfully, a treatment method called injection laryngoplasty is created to solve this situation, which being proved effective over the years and widely applied to many vocal cord disorders. In most cases, doctors inject hyaluronic acid (HA) into patients' vocal cord to improve the glottal gaps and help vocal cord close properly. Previously, doctors have to judge the patients' voice to check the recovery and determine whether to complement HA. Recently, to observe how HA remains and works at, it is feasible to analyze on ultrasound image sequences. With the development of computer vision, doctors can employ computer-assisting method to track degradation of HA and estimate HA volume in human body. Although CNN-based models have achieved excellent performance in image segmentation tasks, they still can not learn global and long-range information because of the locality of convolution operation. Besides, most current segmentation models only focus

doi:10.6342/NTU202302561

on spatial features, ignoring temporal features in segmentation task. However, temporal features are also important for doctors to inference HA position. Therefore, we believe temporal information is also critical for the models to predict HA position correctly. In this study, we proposed AFTNet(Attention Feature Temporal Network), which contains attention-based feature extractor and temporal module. With the benefit of attention-based feature extractor and temporal module, our model can not only better learn global and long-range dependencies, but temporal features of the target videos. We apply this model to our proposed Patient Throat Dataset, which not only assists doctors in difficult-to-diagnose calcified and noise cases, but outperforms both CNN-based and Transformer-based models.

Keywords: Ultrasound Image Segmentation, Transformer, Convolution Neural Network, Recurrent Neural Network, Injection Laryngoplasty

doi:10.6342/NTU202302561

Contents

doi:10.6342/NTU202302561

doi:10.6342/NTU202302561

List of Figures

doi:10.6342/NTU202302561

doi:10.6342/NTU202302561

List of Tables

doi:10.6342/NTU202302561

Chapter 1 Introduction

Figure 1.1: Ultrasound imaging process.

In current society, there are more and more industries that use voice as the main tool, and the dependence on voice is also increasing, such as teachers, salesmen and singers. Voice plays an critical role in daily communication, professional performance, and even artistic performance and has become an indispensable part of daily life. It is worth mentioning that the onset age of vocal cord atrophy, which is due to the overuse of vocal cord, tending to be younger. Vocal cord atrophy is the thinning of one or both vocal muscles. Since one vocal cord cannot meet another one, patients need to take extreme effort to force vocal cords to close well during voicing. Patients with vocal cord antrophy may have a hoarse, husky, or weak voice due to disrupted or obstructed vocal cord vibrations, resulting in changes in voice quality. Damaged vocal cords can lead to a sore or uncomfortable feeling in the throat. Speaking or exerting force to produce sound may further irritate the damaged vocal cords. The more hoarse the patient's voice becomes, the more

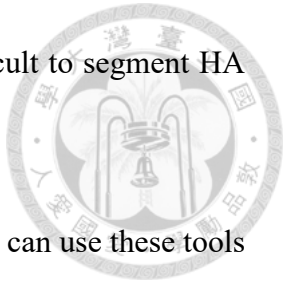
doi:10.6342/NTU202302561

they tend to exert force to produce sound. Exerting more force to produce sound can lead to throat pain and incomplete vocal cord closure. This process would turn into a vicious circle, making the condition even worse and causing problems in daily life. Thankfully, a treatment called injection augmentation has been developed into effective practice over the years and be widely applied to many vocal cord disorders. In most cases, in order to treat vocal cord atrophy or vocal cord paralysis, the vocal cord would be injected with Hyaluronic Acid (HA) to improve the glottal gaps and help vocal cord close properly. The function of HA is to serve as a volume filler, enabling the vocal cords to close properly. Doctors often judge if there is a need of complement HA by patients' voice. There was not effective and efficient method to estimate volume and track the degradation of HA. Recently, to observe how HA remains and works at vocal cord, doctors employ ultra-sound image sequences due to its convenience and rapidity. Although we could examine the change of HA in the throat with ultrasound image sequences, it is still not possible to directly estimate the volume through ultrasound imaging. The resolution of ultrasound images is low and the technical threshold of manipulators influence a lot, it is challenging to interpret HA volume in ultrasound images correctly.

In this task, we calculate the HA volume of ultrasonic image sequences, which can be used to track and estimate the trend of changes in the HA volume and help doctor make clinical judgements. Figure 1.1 illustrate the process of ultrasound imaging. The patients will undergo ultrasound imaging at two, eight and twenty-four weeks after the injection of HA and we can estimate the HA volume through their ultrasound images. The estimated HA volume can help the doctor judge the patients recovery and determine whether to complement HA into patients' throat. Figure 1.2 illustrate the ultrasound throat image sequences at two, eight and twenty-four weeks after the injection of HA. As time goes

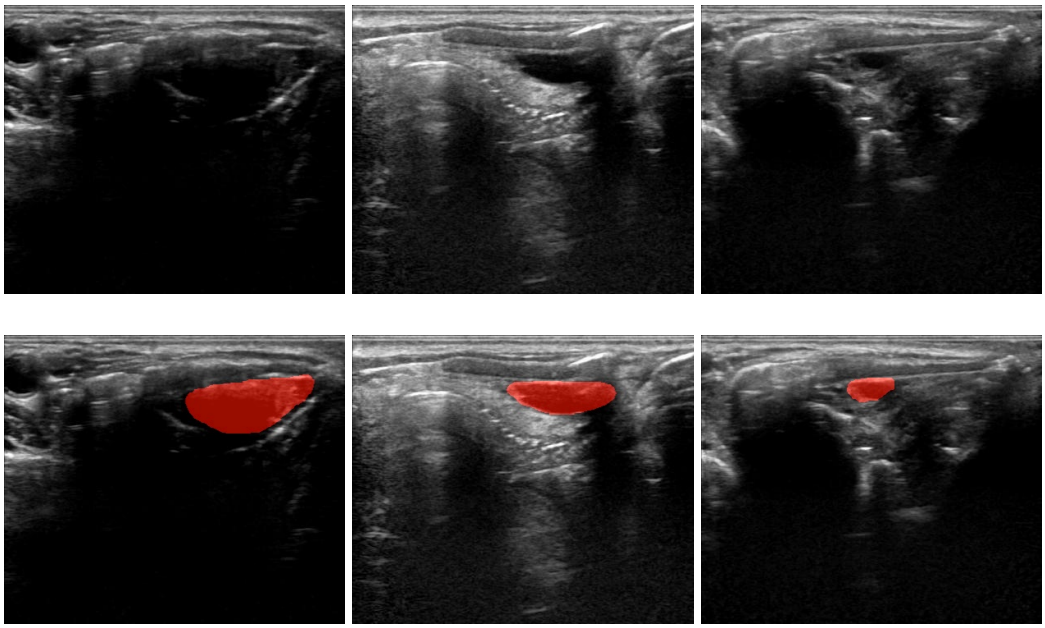
doi:10.6342/NTU202302561

by, the volume of HA will become smaller, and it will become difficult to segment HA volume.



Benefiting from advances in the field of computer vision, people can use these tools to perform medical image analysis. Among them, image segmentation plays an important role in medical image analysis. Image segmentation technology could help doctors not only segment target cells, but segment objects injected into the patients body. Convolutional neural networks (CNNs) has dominated in image segmentation domain for a long time. Starting from FCN [13], which segments images by classifying every pixels to corresponding labels. Then, U-shape architecture [6, 14, 15] made a significant progress in medical image segmentation, which contains an encoder to perform down-sampling, a decoder to perform up-sampling and skip connection to extract multi-scale feature during down-sampling steps. However, CNN-based models can not learn global and long-range dependencies because of the locality of convolution operation. Hence, with the success in natural language processing (NLP), transformer [8] can also be applied in computer vision field, which could learn global and long-range information. Nevertheless, the above-mentioned models make a prediction just with single image, the prediction lack of temporal information. Therefore, we proposed a model with attention-based feature extractor and temporal module, which can not only learn global and long-range information, but learn temporal information like the doctor who infers HA area in the current frame and through the image sequences. With temporal module, we can better segment HA area even if the edge of HA is not clear. The proposed AFTNet (Attention Feature Temporal Network) achieves strong performance on our proposed Patient Throat Dataset. It not only outperforms the CNN-based models and transformer-based models, but not cost too much computation resources.

doi:10.6342/NTU202302561



(a) 2 weeks (b) 8 weeks (c) 24 weeks

Figure 1.2: Patient ultrasound throat image sequences at different times after the injection of HA. The 1st row shows ultrasound images in different time while the 2nd row shows the HA area segmented by doctor. The main difference from other ultrasound image dataset is that the throat is surrounded by other tissues and the HA area to be segmented is not obvious.

doi:10.6342/NTU202302561

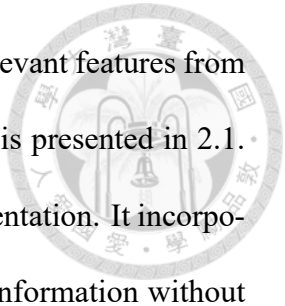
Chapter 2 Related Work

In this chapter, we first introduce CNN-based models in image segmentation tasks and the drawbacks of convolution operations. What follows is self attention/Transformer to complement CNNs. With aids of self attention and Transformer, it is effective to reduce the impact of convolution operation in CNN-based models. However, it still has some drawback to learn long range and global information. Transformer is employed in computer vision due to its success in Natural Language Processing (NLP). With self attention mechanism utilized in Transformer, the model can learn long range dependencies of the images. In the end is the brief introduction of Transformer based backbone in computer vision.

2.1 CNN-based Models

Convolutional neural networks (CNNs) have been widely used in image segmentation tasks and have achieved significant success [6, 10, 16]. U-Net [15] is a popular and widely used architecture for medical image segmentation. It consists of an encoder path and a decoder path. The encoder path performs downsampling operations to capture context and extract high-level features. The decoder path performs upsampling operations to recover spatial resolution and combine the low-level and high-level features for segmen-

doi:10.6342/NTU202302561



tation. U-Net has skip connections that allow the decoder to access relevant features from the encoder, aiding in precise localization, the architecture of U-Net is presented in 2.1. DeepLab [5] is a state-of-the-art convolutional model for image segmentation. It incorporates atrous (dilated) convolutions to capture multi-scale contextual information without significantly increasing the computational cost. DeepLab uses a combination of dilated convolutions, pooling, and skip connections to improve segmentation performance. It also includes a final up-sampling step to obtain dense pixel-level predictions. However, there are some limitation of using CNN-based models in image segmentation. It has difficulty in capturing fine details: CNNs, especially those with down-sampling operations like pooling or strides, can lead to a loss of fine-grained details during the encoding process. This down-sampling can make it challenging for CNNs to accurately capture small objects or intricate boundaries in segmentation tasks. Also, CNNs process images in a local and sequential manner, with limited access to global spatial context. While skip connections and encoder-decoder architectures partially address this issue, they may not capture long-range dependencies as effectively as other architectures like transformers.

2.2 Self-attention/Transformer to complement CNNs

Self-attention is a mechanism that can complement convolutional neural networks (CNNs). By incorporating self-attention mechanisms, models can capture relationships between distant spatial locations and leverage global context, allowing them to capture long-range dependencies [1, 4]. Recently, Transformers, which is composed of encoder and decoder, originally introduced for natural language processing, have also been employed to computer vision tasks and have achieved remarkable results [3, 12]. The transformer architecture relies heavily on self-attention mechanisms. Instead of using convolu-

doi:10.6342/NTU202302561

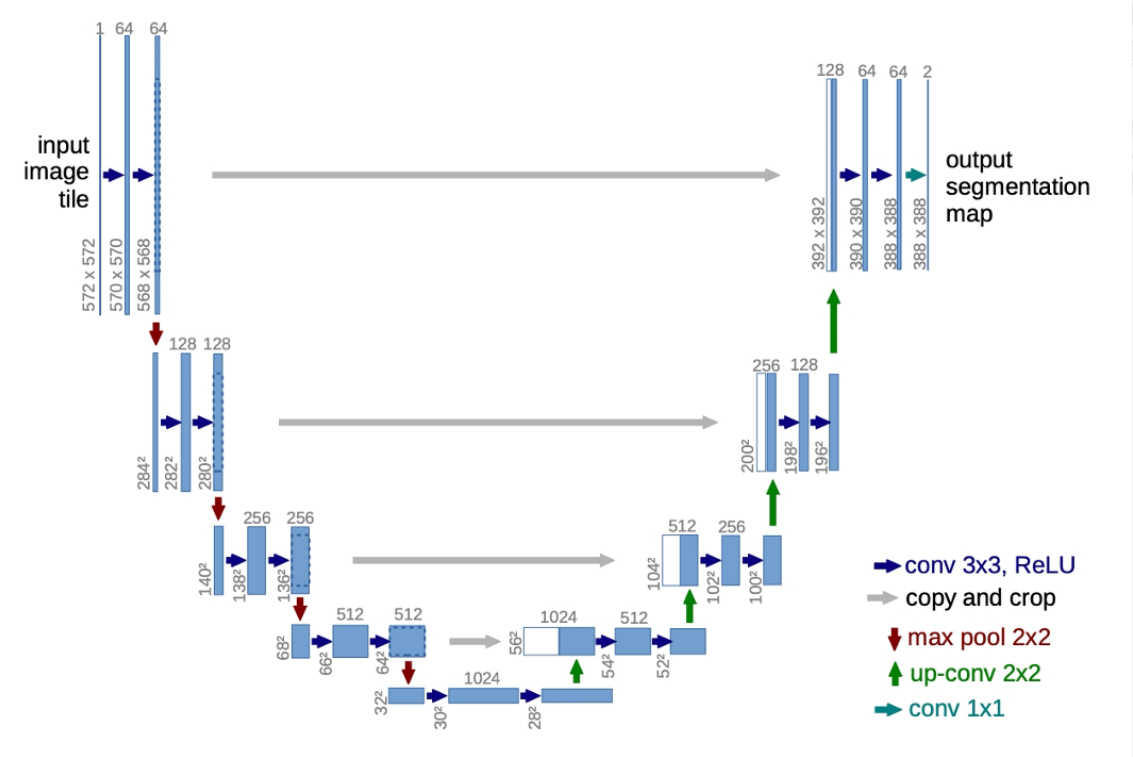


Figure 2.1: The architecture of U-Net [15].

tional operation, transformers utilize self-attention layers to capture relationships between spatial locations. By attending to all positions within an input sequence, transformers can effectively model long-range dependencies and capture global context. Transformer-based models have shown great potential in tasks such as image classification, object detection, and image segmentation. By incorporating self-attention mechanisms, these models can capture relationships between distant spatial locations and leverage global context, allowing them to handle tasks that require capturing long-range dependencies or modeling fine-grained details. The combination of CNNs and self-attention provides a powerful framework for addressing various computer vision challenges.

doi:10.6342/NTU202302561

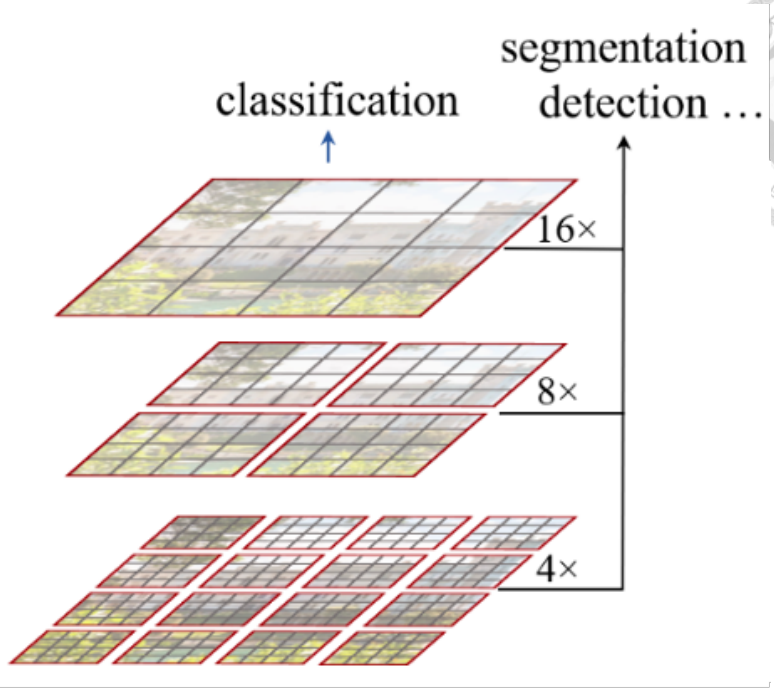


Figure 2.2: The hierarchical feature maps of Swin Transformer [12].

2.3 Transformer based vision backbones

Transformer-based vision backbones have gained significant attention in recent years for their effectiveness in various computer vision tasks. Unlike traditional convolutional neural networks (CNNs) [9, 11, 18], which rely on convolutions for feature extraction, transformer-based models utilize self-attention mechanisms to capture global dependencies and learn spatial relationships.Vision Transformer (ViT) [8] was one of the first transformer-based models to be introduced for vision tasks. It applies the transformer architecture to image classification by dividing the input image into patches and transforming them into sequence-like data. The model leverages self-attention to capture the relationships between patches and learns representations for classification tasks. DeiT [19] incorporates distillation techniques and data augmentations to improve generalization and achieve state-of-the-art performance in image classification tasks. Swin Transformer [12] introduces a hierarchical design that captures dependencies across different scales, as illustrated

doi:10.6342/NTU202302561

in 2.2, allowing it to model fine-grained details and global context effectively. The Swin Transformer has demonstrated strong performance in image classification, object detection, and semantic segmentation tasks. Its hierarchical design, shifted windows, and attention mechanisms contribute to its ability to capture fine-grained details and global context efficiently. Swin Transformer represents an exciting advancement in vision transformers and offers a promising alternative to traditional convolutional neural networks.

doi:10.6342/NTU202302561

Chapter 3 Method

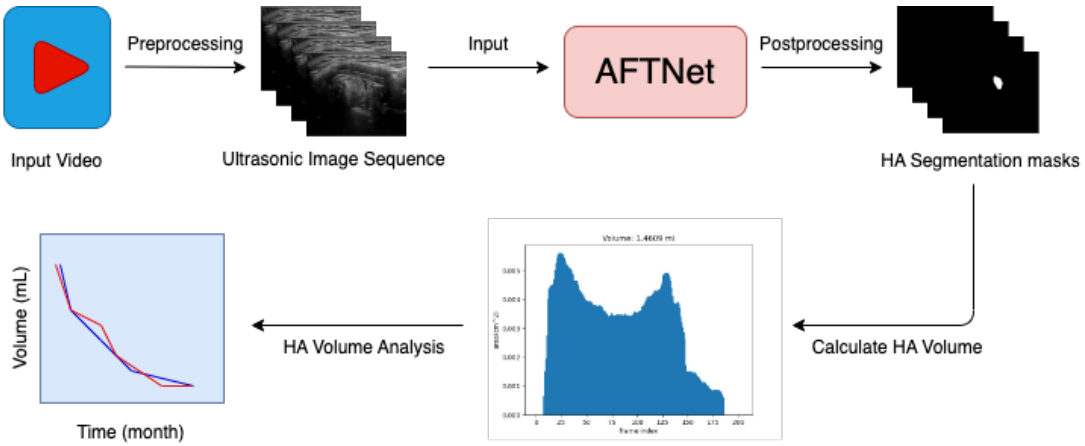


Figure 3.1: System overview.

3.1 System Overview

An overview of proposed system is presented in Figure 3.1, which contains five main procedures: Data Preprocessing, AFTNet, Data postprocessing, calculating HA volume and HA volume analysis.

Given an ultrasonic video, it first split the video into image frames, which then input to AFTNet. The model not only extract spatial features, but temporal features of the image sequences. With attention-based spatial feature extractor, the model will learn long range dependencies and global context, and temporal let the prediction more reliable. Afterward, by applying postprocessing method, the noise of prediction map will be erased and the

doi:10.6342/NTU202302561

unexpected results such as predicted area jump too far and suddenly appear or disappear will also be removed. Finally, we can estimate HA volume of the prediction map and track the degradation of HA for doctors to take corresponding treatments.

3.2 Data Preprocessing

Because ultrasound data are video format, we convert input data into image frames at first. For each frame, we concatenate its previous δ frames and last δ frames like inference method of doctors to better extract temporal features. The schematic diagram is illustrated in Figure 3.2. Afterward, we feed processed frames to the proposed model with basic data augmentation, such as rotation and changing contrast in order to enrich the diversity of training data.

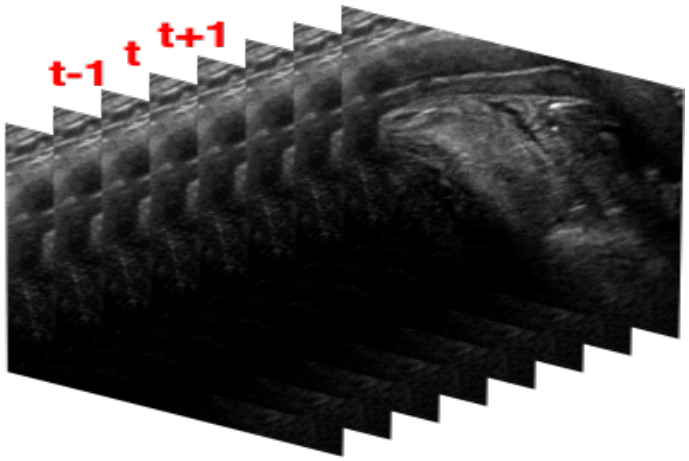


Figure 3.2: Concatenate first δ frames and last δ frames with the current frame.

3.3 Model

Inspired by [7, 21] model architecture, we proposed a network for ultrasound image sequence segmentation, which feature extractor and temporal module are based on Swin-UNet [2] and C-LSTM [17] respectively, called AFTNet (Attention Feature Temporal

doi:10.6342/NTU202302561

Network). AFTNet consists of feature extractor, temporal module and refining module, architecture is illustrated in Figure 3.3. Feature extractor is for spatial feature extraction, whose attention-based feature extractor block is composed of swin transformer block [12] can learn global and long-range semantic information to better deal with unclear edge or noise conditions of the ultrasound images. In order to segment HA area like doctor who infer from current frame through multiple frames, we employ temporal module to make our network extract temporal features from image frames, which is helpful to segment HA area within multiple frames. Refining module can further refine the prediction mask from temporal module to get a refined result.

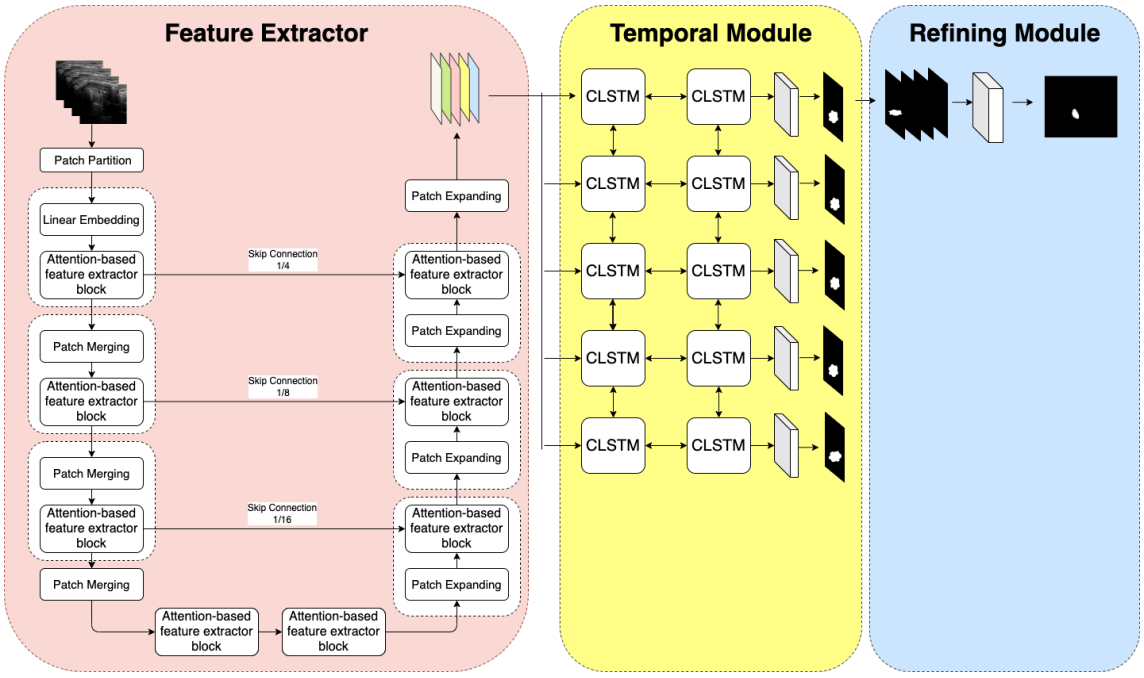


Figure 3.3: AFTNet contains three module : Feature Extractor, Temporal Module, Refining Module.

3.3.1 Feature Extractor

Our feature extractor architecture is based on Swin-Unet [2]. It consists encoder, bottleneck, decoder and skip connections. Different from the standard multi-head self attention (MSA) in conventional Transformer block, we construct attention-based fea-

doi:10.6342/NTU202302561

ture extractor block based on Swin Transformer block [12], which uses shifted window based multi-head self attention to learn representation instead of global self attention. The window-based multi-head self attention and the shifted window-based multi-head self attention are applied in the two successive blocks, a MLP with GELU activation function are deposed after each attention block. A layer normalization layer is employed before each window-based MSA and MLP, and residual connections are used after each module as illustrated in Figure 3.4. In order to perform down-sampling, patch merging layer is applied to reduce the feature map to half of the original resolution. Since Transformer is too deep to be converged [20], we construct bottleneck with two successive Swin Transformer blocks to learn the deep feature representation. In decoder, Swin Transformer blocks are used to learn representation and patch expanding layers are used to implement up-sampling. In patch expanding layers, the feature map is reshaped into a higher resolution with $2\times$ up-sampling and the feature dimension is reduced to half of the original feature map. Besides, proposed in UNet[15], skip connections are also applied here to combine multi-scale features from encoder and decoder, which enable the direct concatenate features of encoder and the corresponding decoder. Skip connection can let the model maintain features of different spatial resolution, which is beneficial to preserving different spatial resolution information and handling different size of objects segmentation.

3.3.2 Temporal Module

Based on [17], we implement our temporal module inspired by doctor who interprets HA area of the current ultrasound images through the ultrasound image sequences. First of all, instead of feeding feature map of the original ultrasound image resolution, the feature map being fed into temporal module are one-eighth of the original ultrasound images.

doi:10.6342/NTU202302561

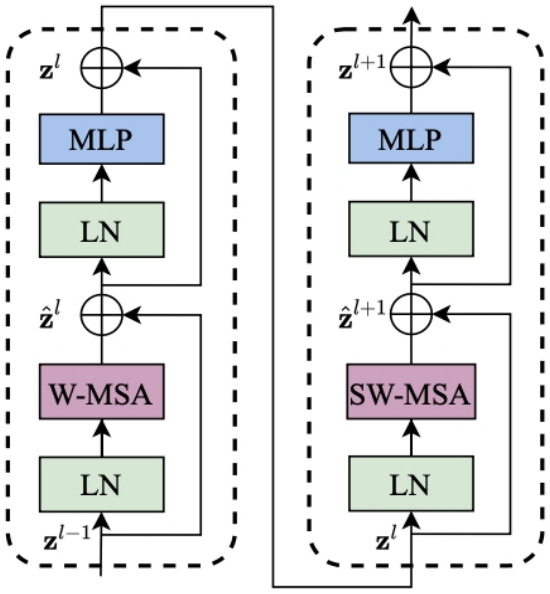


Figure 3.4: Two successive Swin Transformer Blocks [12].

Then, feature maps are fed into Bi-directional C-LSTM uses previous information and future information during prediction. The output of each time point t of the C-LSTM is a pixel-wise feature map with temporal information. Then, we concatenate the current frame with forward and backward C-LSTMs followed by convolution layers and ReLUs, obtaining the final temporal prediction map.

3.3.3 Refining Module

In refining module, we refine the final prediction map through compressing the prediction map from temporal module. We input the prediction map to 1×1 convolution followed by ReLU.

3.4 Data Postprocessing

We perform three main step to make our prediction more reliable. First of all, we set a threshold to check if the predicted mask area is larger than threshold and choose the

doi:10.6342/NTU202302561

largest connected component. Because HA tends to get together, this can help us remove very small predicted regions. Then, we calculate local mean and global mean, respectively. Local mean is the average coordinates of the mask within current few frames, while global mean is the average coordinates of the mask in the whole ultrasound image frames. With local mean and global mean, we can remove predicted mask which move too far from other frames, making the prediction more reasonable. Finally, by calculating sliding window distance, we can avoid mask appear or disappear suddenly.

3.5 HA Volume Analysis

In this step, we get the predicted volume of the input ultrasound images sequences after the injection of HA in two week, eight weeks and twenty-four weeks, respectively. We estimate HA volume through binary prediction masks of the ultrasound image sequences. We calculate HA volume based on mask areas of each image frame, which are formulated as follows:

$$V = \frac{((\sum_{i=1}^{n-1} A[i] + \sum_{i=2}^{n} A[i]) \times h)}{2n} \tag{3.1}$$

With the benefit of the predicted volume, doctors can better track how much HA remains in patients' throat and take corresponding treatments if needed. The experiment result of HA volume estimation are presented in Section 4.4.

doi:10.6342/NTU202302561

Chapter 4 Experiments

We conduct experiments on our proposed Patient Throat Dataset. The dataset is collected by the doctor, and consists of ultrasound images of patients' throat after two weeks, eight weeks and twenty-four weeks respectively. In the following, we first introduce our proposed Patient Throat Dataset, which can be divided into four different categories. Then, experiments are conducted on Patient Throat Dataset, comparing our model with the previous excellent segmentation models. Afterward, we compare our proposed model with other excellent segmentation models with different cases of Patient Throat Dataset. The ablation study of the important modules of our proposed model is presented in the end.

4.1 Datasets

4.1.1 Patient Throat Dataset

We proposed Patient Throat Dataset, which consists of ultrasonic image sequences of patients' throats after two weeks, eight weeks and twenty-four weeks of the injection of HA respectively. There are 82 videos in our dataset, and being annotated by the doctor with CVAT. The ground truth volumes are computed through the ground truth annotation of the dataset. The biggest difference from other ultrasound image dataset is that the HA

doi:10.6342/NTU202302561

in our dataset are blocked by other tissues and the edges of HA are not obvious. Besides, the dataset can be divided into four cases : normal cases, calcified cases, noisy cases and the final cases are calcified, contour reappear in the end of the video. The ultrasonic images and ground truth of different cases are presented in Figure 4.1. The following is an introduction to the four different types of the proposed Patient Throat Dataset :

Normal Cases The type is composed of clear ultrasound image sequences, and the edge of HA are obvious.

Calcified Cases Patients' throats are calcified so that their HA in ultrasound images tend to be white and the edge between tissue and HA become unclear. Therefore, this type of ultrasound images are challenging to segment correctly.

Noisy Cases During the ultrasound imaging process, there may be air between the probe and the skin. There are trivial things, such as tissue inflammation and blood clots in patients' throats. It will make the videos look noisy. Hence, this type of ultrasound images are the most challenging cases of our proposed dataset.

Calcified, contour reappear in the end In the beginning of the type of ultrasound image sequences are normal while in the end of the video, there are calcified condition in patients' throat tissue. The type of cases are a little harder than that of in normal cases.

4.1.2 Image Phantom Dataset

This dataset is proposed to simulate the ultrasound images of throat, consisting of two types : normal and calcified types, as illustrated in Figure 4.2. The main function of the

doi:10.6342/NTU202302561

dataset is to help calculate the HA volume because the practical HA volume injected into the patients may be affected by many factors such as moisture and the inflammation of the throat tissue of the patients. This condition may cause a discrepancy between the injected HA into the throat and the actual situation. Besides, with this simulation approach, it's possible to increase the quantity of training data. The method used to create this dataset is that the doctor can make a jelly of a specific volume to simulate various situation of the patients' throats, such as normal and calcified types. As a result, the dataset could be divided into two types : normal and calcified. The detail description of the two types are illustrated below :

Normal This type is proposed to simulate the normal condition of patient throat. It consists of 1(mL), 2(mL) and 5(mL) jelly volume. The edges between the HA and the other tissues are obvious.

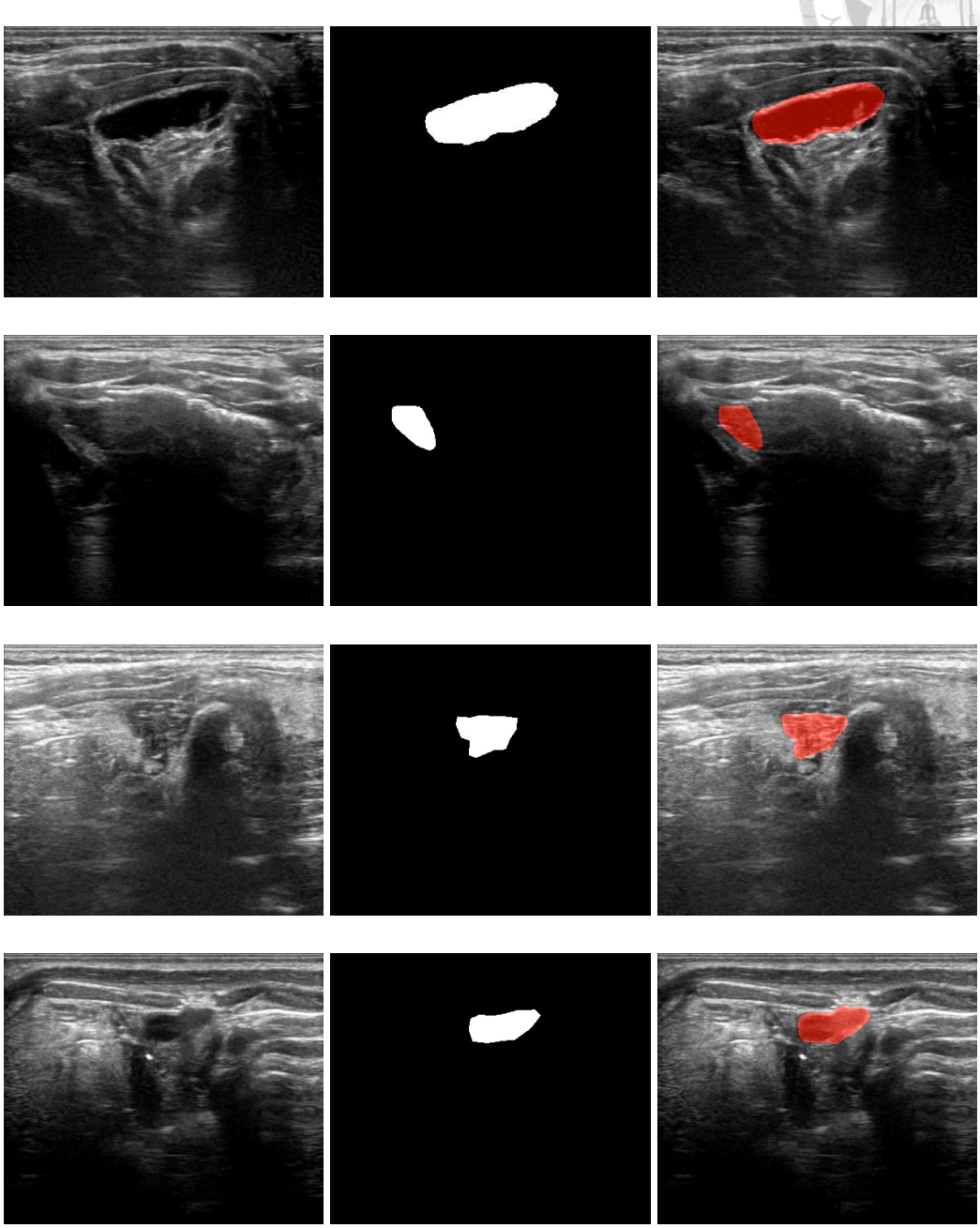
Calcified This type is proposed to simulate the calcified condition of patient throat. It consists of 1(mL), 2(mL) and 5(mL) jelly volume. The edges of this type cases are not obvious. Therefore, the cases of this type are more difficult that the normal ones.

4.2 Implementation Detail

Experiments were implemented on Python 3.8 and Pytorch 1.11. For all training cases, data augmentation such as random jitter and random flipping are used to increase training data diversity. We input eight ultrasound image frames to the model once at a time. Each of image frame and patch size are set as 352×416 and 4 respectively. The model was trained with AdamW optimizer. The learning rate is set as 0.0001 and is controlled

doi:10.6342/NTU202302561

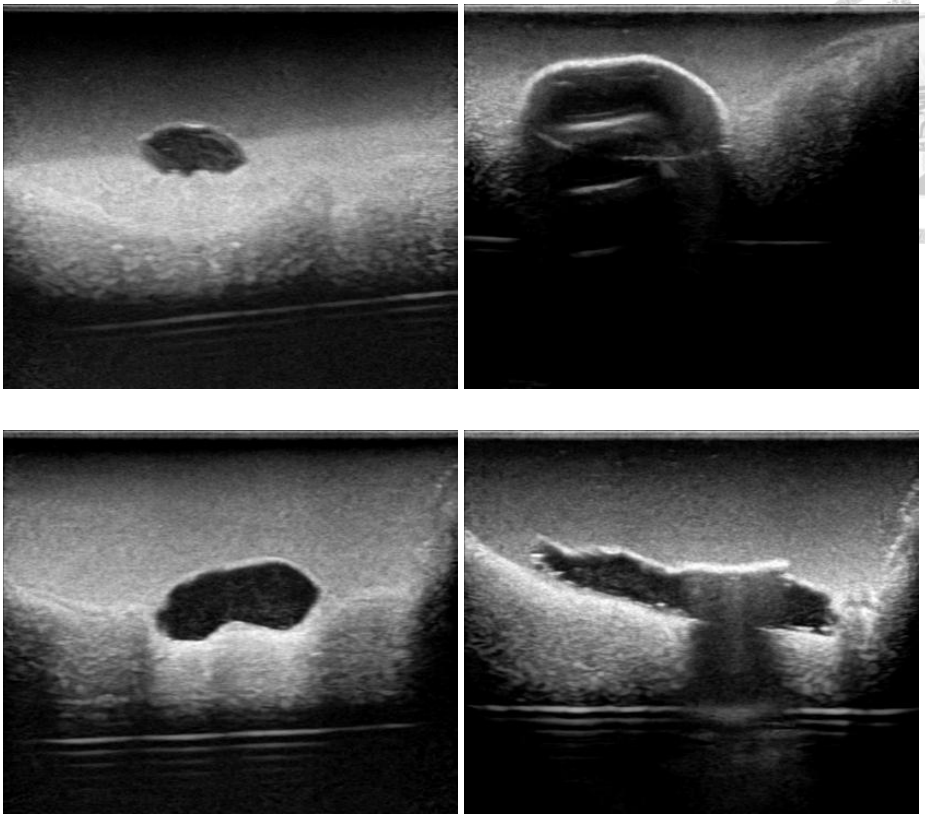
by multi-step learning rate (MultiStepLR) scheduler. We train and inference on a single NVIDIA RTX A6000 GPU.



(a) Ultrasound Image (b) Mask (c) Visualization

Figure 4.1: Four different cases of our proposed Patient Throat Dataset. From top to bottom is normal, calcified, noisy and calcified, contour reappear in the end.

doi:10.6342/NTU202302561



(a) Normal Image (b) Calcified Image

Figure 4.2: Image Phantom Dataset. (a) is the example images of normal type of Image Phantom Dataset. The the edge of the target object is clear. (b) is the example images of calcified type of Image Phantom Dataset. The edge of the target area in calcified cases are not obvious. Therefore, temporal information is critical for this condition.

4.3 Comparison with Other Segmentation Models

Table 4.1 presents comparison to other segmentation models, including both Transformer-based and CNN-based models. Different from other 2D segmentation models, our proposed model performs much better because it not only learn spatial information of the image sequences, but learn temporal information with temporal module. First of all, as Figure 4.5 presents, the edge of target HA is clear and the area is large. Although the case is easy to segment, the proposed model can segment more precisely than other models. Then, Figure 4.8 shows the result on calcified case of Patient Throat Dataset, our model can precisely segment HA even if the edge of HA is not obvious. Because HA would

doi:10.6342/NTU202302561

degrade over time, it will become difficult to track the residual state of HA in the human body. Fortunately, our model can segment HA very well even if the remaining volume is small, the result reveal in Figure 4.7. The reason of our model can segment HA more precisely is that the temporal module predict HA area like doctors who judge it through multiple ultrasonic image frames. The predicted area of Swin-UNet is too small while our model segment more precisely. Finally, the most challenging cases of our proposed Patient Throat Dataset, are noisy cases, our model can still segment HA accurately, see Figure 4.6. Because there are trivial things in ultrasound image sequences of noisy cases, it is challenging for every models to segment HA. With global feature extractor and temporal information, our model can outperform other models.

Table 4.2 reveals comparison with other segmentation models in the different cases of our proposed dataset. We can observe that our proposed model perform better in all of cases in Patient Throat Dataset.

Table 4.1: Results on Patient Throat Dataset.

Model	IoU(%)	Dice(%)	Parameters
V-Net [14]	48.51	63.61	132.05M
U-Net [15]	56.07	70.16	24.44M
DeepLabV3 [5]	56.96	70.33	22.43M
Swin-Unet [2]	55.64	69.05	27.16M
TCSNet [22]	58.39	72.04	24.45M
Ours	**62.66**	**75.98**	28.61M

Table 4.2: Results of different cases on Patient Throat Dataset.

	Normal		Calcified		Noisy		Calcified, contour reappear in the end	
	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)
V-Net	56.37	71.49	46.93	61.82	29.32	44.30	53.51	68.93
U-Net	58.35	71.83	57.01	72.24	43.07	58.76	65.94	79.05
DeepLabV3	60.27	73.76	55.65	68.46	39.20	54.86	67.15	79.95
Swin-Unet	57.67	70.80	54.23	67.95	37.87	52.03	61.14	75.76
TCSNet	64.78	78.16	56.57	70.76	42.75	58.36	63.22	76.68
Ours	**66.49**	**79.29**	**61.98**	**71.52**	**48.57**	**63.69**	**68.86**	**81.05**

doi:10.6342/NTU202302561

4.4 HA Volume Estimation

In Table 4.3 shows the estimated volume of our proposed model on Patient Throat Dataset at different period of the injection of HA. The error between the estimated volume and real volume is within 2%. Figure 4.3 presents the degradation of HA volume at different period after the injection of HA. As the figure illustrated, the predicted volumes of our proposed model are very close to the ground truth volumes.

Table 4.4 presents the estimated volume of our proposed model on Image Phantom Dataset with different type. From the results, it can be seen that the predicted volumes are very close to the actual volumes. However, as illustrated in Figure 4.4, case 6 of the calcified type, the right-hand portion was not predicted. Therefore, the error rate between the estimated volume and actual volume is a little higher.

Table 4.3: Estimated volume of HA on Patient Throat Dataset.

Time	Estimated Volume(mL)	Actual Volume(mL)	Error Rate(%)
2 weeks	**1.4197**	1.3801	2.87
8 weeks	**0.7004**	0.7131	1.79
24 weeks	**0.4713**	0.4674	0.84

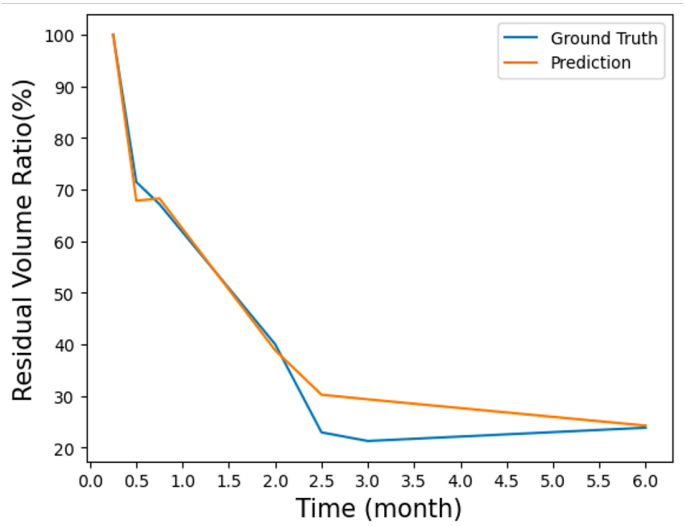
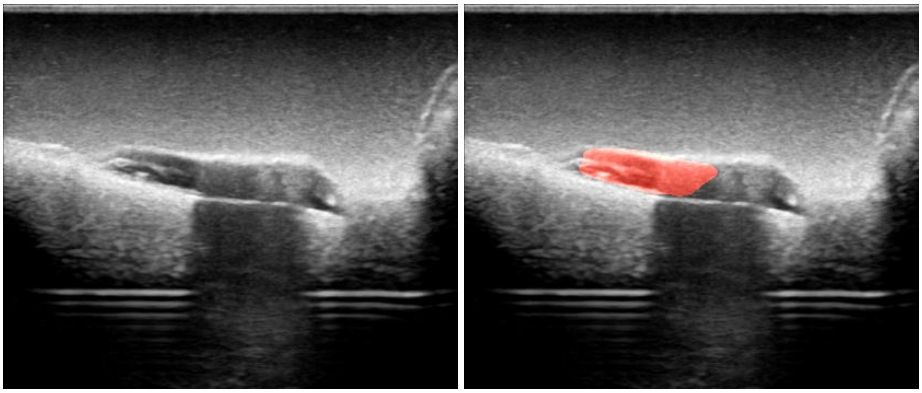


Figure 4.3: The degradation of HA volume trends in patient throat over time on Patient Throat Dataset.

doi:10.6342/NTU202302561

Table 4.4: Estimated volume of HA on Image Phantom Dataset.

Type	Case Number	Estimated Volume(mL)	Actual Volume(mL)	Error Rate(%)
Normal	1	**0.969**	1	3.1
Normal	2	**0.940**	1	6
Normal	3	**1.956**	2	2.2
Normal	4	**1.937**	2	3.3
Normal	5	**5.005**	5	0.1
Normal	6	**5.224**	5	4.4
Calcified	1	**0.987**	1	1.3
Calcified	2	**1.072**	1	7.2
Calcified	3	**0.957**	1	4.3
Calcified	4	**2.011**	2	0.5
Calcified	5	**2.062**	2	3.1
Calcified	6	**2.154**	2	7.5
Calcified	7	**5.308**	5	6.1
Calcified	8	**4.922**	5	1.5
Calcified	9	**5.213**	5	4.2



(a) Original image (b) Predict result

Figure 4.4: Image Phantom Dataset.

4.5 Ablation Study

In the ablation analysis, we will explore how the important components of our models influence performance. Here we conduct two experiments to verify that our temporal module and postprocessing can improve our performance.

doi:10.6342/NTU202302561

4.5.1 Temporal Module

Table 4.5 shows that with temporal module, our model can learn the detailed temporal prediction. The reason why temporal module improves performance is that our model can learn how to segment HA precisely through the past frames and the future frames like doctors interpret clinically. Because, sometimes, the doctor will interpret the ultrasound image from the back to the front frames. With such a module, we can segment ultrasound image sequences even if the edge of HA is not obvious or the HA area is tiny.

Table 4.5: The impact of temporal module.

Model	IoU(%)	Dice(%)
w/o Temporal Module	60.64	72.05
$w/$ Temporal Module	**62.66**	**75.98**

4.5.2 Postprocessing

Table 4.6 shows that with our postprocessing, the performance become better. The reason why postprocessing can improve performance is that it can avoid some cases misjudging HA volume by setting a threshold to remove misprediction parts. Also, by calculating local mean and global mean of masks, it can avoid HA disappear or appear suddenly.

Table 4.6: The impact of postprocessing.

Model	IoU(%)	Dice(%)
w/o Postprocess	61.76	74.69
$w/$ Postprocess	**62.66**	**75.98**

doi:10.6342/NTU202302561

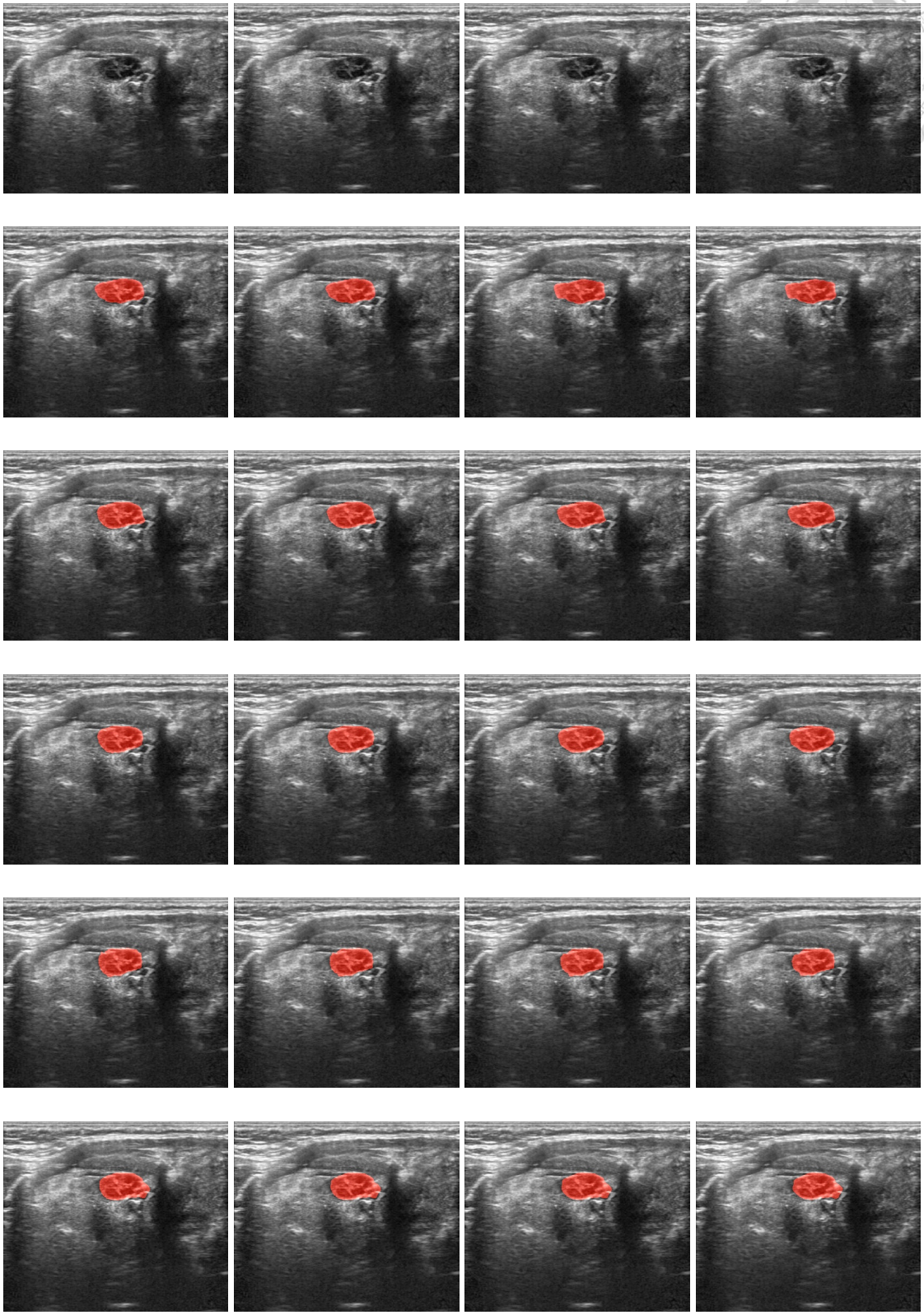


Figure 4.5: First to sixth columns in order are : original images, ground truth, UNet prediction, DeepLabV3 prediction, Swin-UNet prediction and our prediction. The experiment inference on normal cases of Patient Throat Dataset. In this case, although the area of HA is clear, our proposed model segment the HA more precisely than other models.

doi:10.6342/NTU202302561

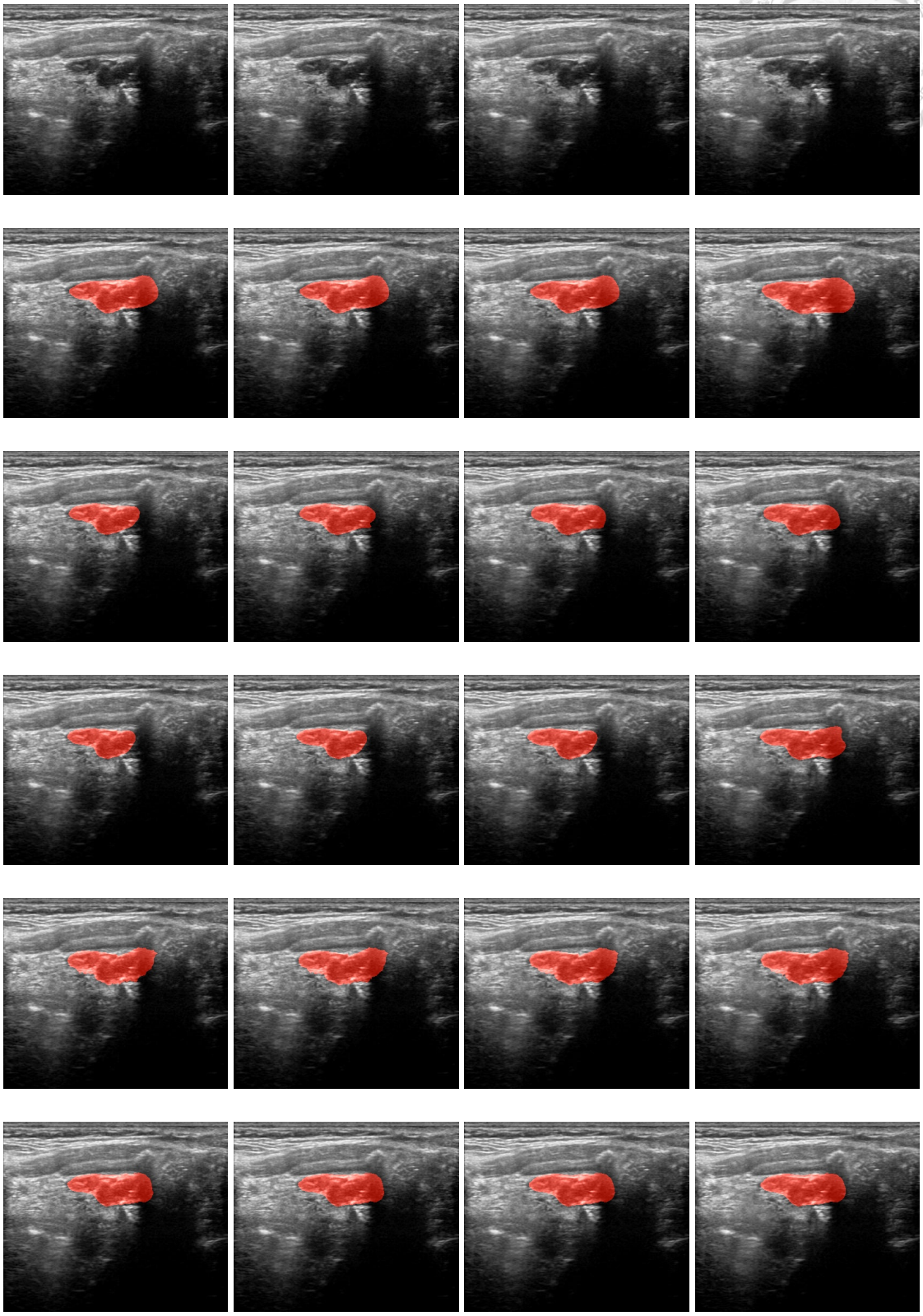


Figure 4.6: First to sixth columns in order are : original images, ground truth, UNet prediction, DeepLabV3 prediction, Swin-UNet prediction and our prediction. The experiment inference on noisy cases of Patient Throat Dataset.

doi:10.6342/NTU202302561

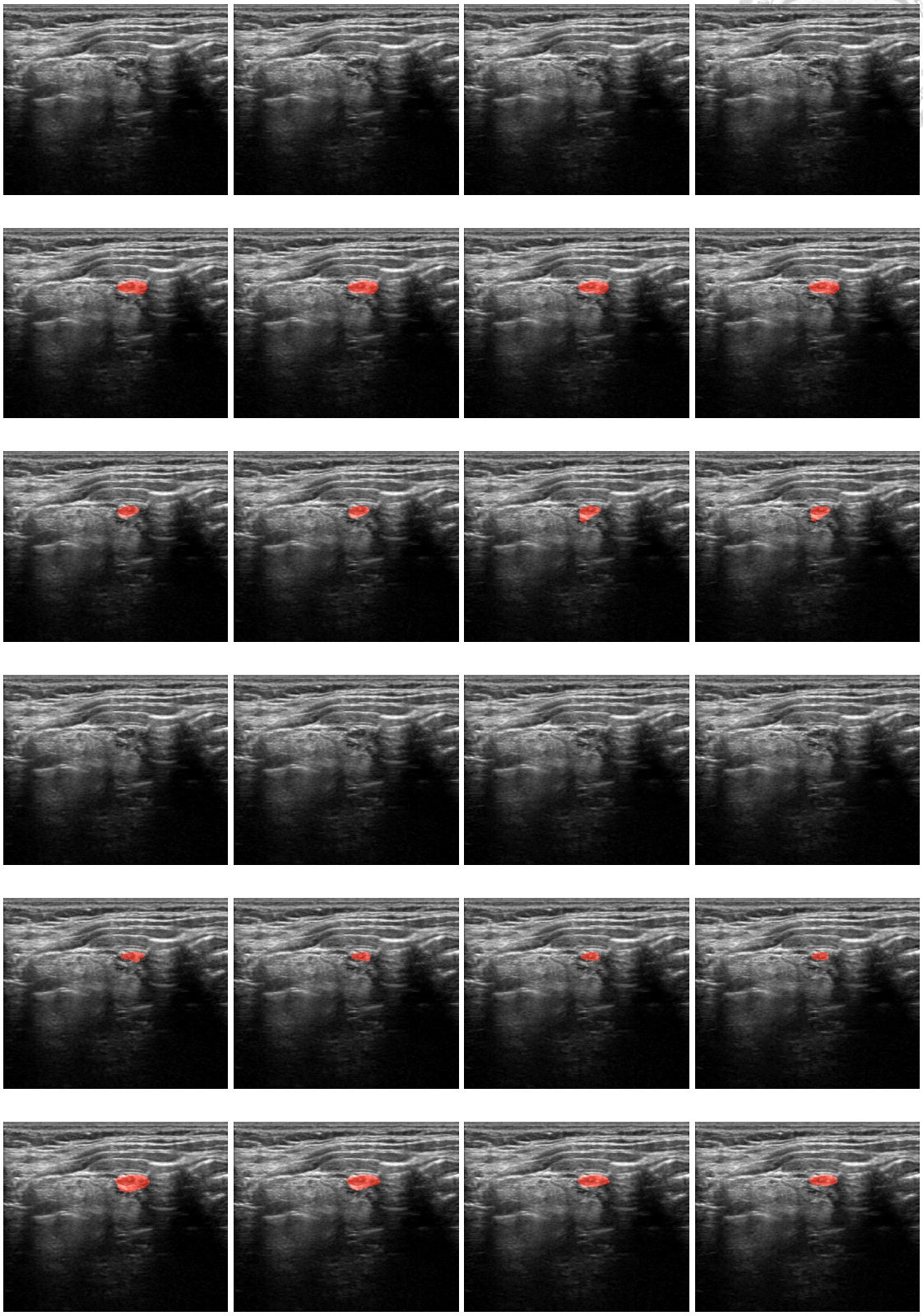


Figure 4.7: First to sixth columns in order are : original images, ground truth, UNet prediction, DeepLabV3 prediction, Swin-UNet prediction and our prediction. The experiment inference on Normal cases of Patient Throat Dataset. The target is tiny.

doi:10.6342/NTU202302561

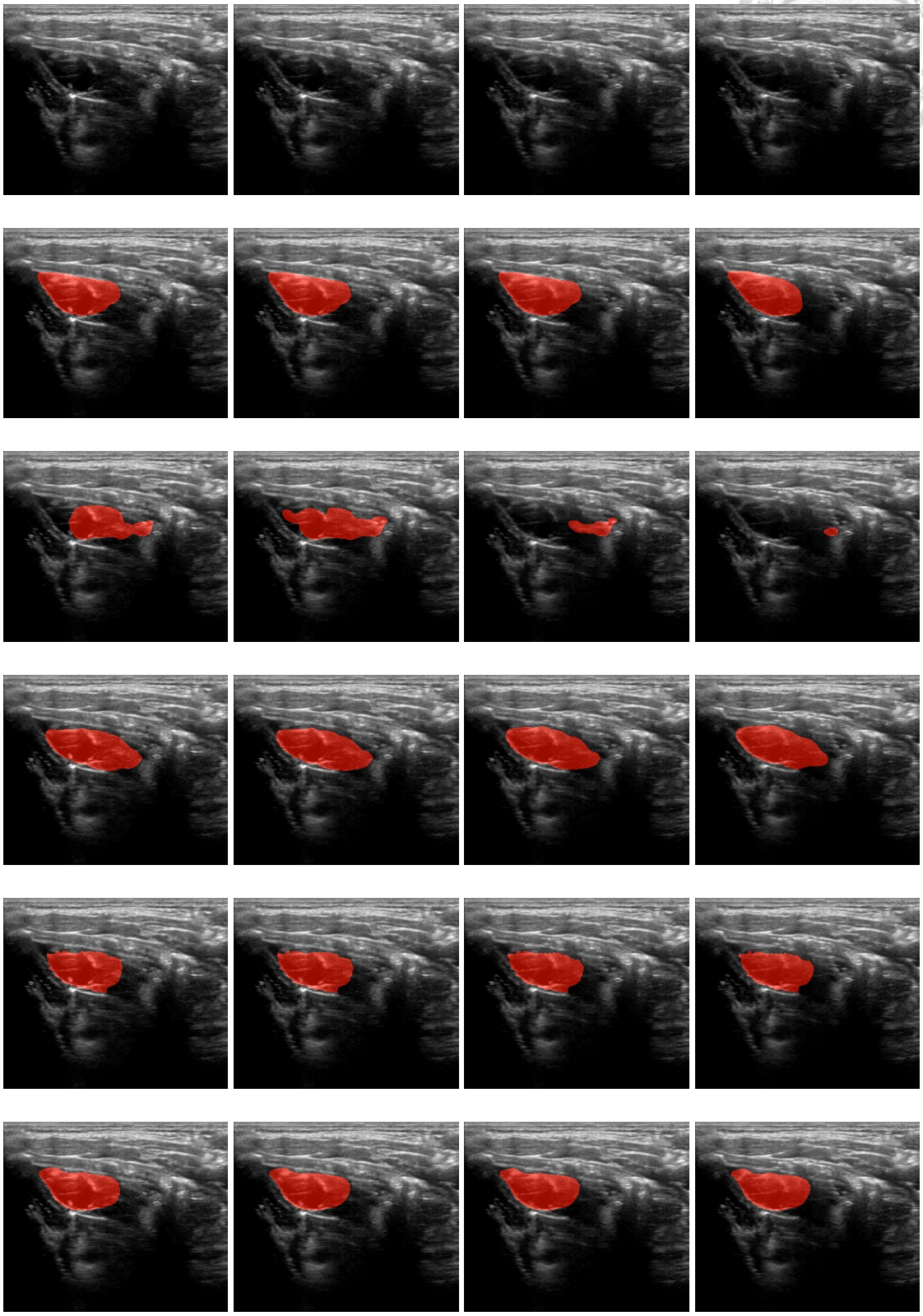


Figure 4.8: First to sixth columns in order are : original images, ground truth, UNet prediction, DeepLabV3 prediction, Swin-UNet prediction and our prediction. The experiment inference on Calcified cases of Patient Throat Dataset.

doi:10.6342/NTU202302561

Chapter 5 Conclusion

In this paper, we proposed a model with attention-based feature extractor and temporal module for HA ultrasound image sequences segmentation. To leverage the power of Transformer, we take Swin Transformer block in our attention-based feature extractor block for feature representation and long-range information learning. Besides, we utilize temporal module to better segment HA area like doctor who infer through ultrasound image sequences. With such a model, doctors can confirm the residual HA volume in the patients' throat and see if it is necessary to supplement HA without requiring invasive inspection methods.

doi:10.6342/NTU202302561

References

[1] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. Attention augmented convolutional networks. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3286–3295, 2019.

[2] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision, pages 205–218. Springer, 2022.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020.

[4] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018.

doi:10.6342/NTU202302561

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2625–2634, 2015.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[10] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for mobilenetv3. In ICCV, 2019.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic

doi:10.6342/NTU202302561

segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV), pages 565–571. Ieee, 2016.

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

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, 2018.

[17] X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015.

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

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In International conference on machine learning, pages 10347–10357. PMLR, 2021.

[20] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 32–42, 2021.

doi:10.6342/NTU202302561

[21] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision, pages 20–36. Springer, 2016.

[22] 洪商荃. 基於人工智慧分析玻尿酸體積於注射式喉成型手術後的降解情形. 2021.

doi:10.6342/NTU202302561