Deep Learning-Based Cascade 3D Kidney Segmentation Method

by

Zixin Hao

ORCID: 0009-0005-8004-4613

A thesis submitted in total fulfillment for the
degree of Master of Information and Technology

in the
Faculty of Engineering and Information Technology
School of Computing and Information Systems
THE UNIVERSITY OF MELBOURNE

June 2024

THE UNIVERSITY OF MELBOURNE

Abstract

Faculty of Engineering and Information Technology
School of Computing and Information Systems

Doctor of Philosophy

by Zixin Hao
ORCID: 0009-0005-8004-4613

Renal tumors, prevalent malignancies within the urinary system, necessitate early diagnosis and precise tumor localization for effective patient treatment and prognosis. This research focuses on automating the analysis of renal tumors in abdominal CT images by initiating semantic segmentation of the kidney. It facilitates automated decision-making processes in diagnosis and treatment. U-Net is highly regarded in this field for its ability to train high-precision models using relatively small datasets. So, this research focuses on utilizing the 3D U-Net model. To address challenges such as inadequate edge detection and the segmentation of small objects, the study employs a cascade 3D U-Net architecture. Given the clinical demand for high efficiency and lightweight models, this framework incorporates residual blocks to enhance model convergence speed and overall performance. Various training configurations, alongside effective pre-processing and post-processing strategies, were explored to ensure accurate segmentation of renal and renal tumors. Using the KiTS2019 challenge data, the method achieved 23th place on the leader board as of May 2024 with the following scores: an average of 0.9159, a Kidney Dice of 0.9794, and a Tumor Dice of 0.8524. The findings from this research underline the efficacy of the enhanced cascade 3D U-Net model, demonstrating significant improvements in the precision of renal and renal tumor segmentation.

Keywords: Machine Learning, Abdominal 3D Medical Imaging Segmentation, Image Analysis, Deep Learning, U-net.

Declaration of Authorship

I certify that:

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- the thesis is 6272 words in length (excluding text in images, table, bibliographies and appendices).

Signed: Zixin Hao

Date: 02/06/2024

Acknowledgements

I would like to extend my sincere thanks to my project advisor, Professor Brian Chapman, for his invaluable guidance and constructive suggestions throughout this project. His earnest responsibility and expert advice have been crucial to my research progress and personal development. I am also grateful to Hao Xu, who provided insightful information during our discussions.

Contents

List of Figures

List of Tables

Abbreviations

ROI	**R**egions **O**f **I**nterest
CNN	**C**onvolutional **N**eural **N**etworks
FCN	**F**ully **C**onvolutional **N**etworks
nnU-Net	**n**o-**n**ew**UNet**
HUs	**H**ounsfield **U**nit**s**
IN	**I**nstance **N**ormalization
SGD	**S**tochastic **G**radient **D**escent

Chapter 1

Introduction

1.1 Research Background and Implications

Kidney cancer ranks among the most prevalent cancers. In 2018, there were over 400,000 new cases of kidney cancer diagnosed worldwide, resulting in more than 175,000 deaths [1]. Up until 2023, the incidence of kidney cancer has been rising by approximately 1% annually for both men and women. However, the mortality rate from kidney cancer has been decreasing by about 2% per year [2]. One contributing factor could be the increased use of abdominal imaging for various unrelated medical reasons, leading to more frequent incidental discoveries of asymptomatic renal tumors. As a result, a larger number of tumors are being detected while they are still small and localized, which is believed to enhance the overall survival rates for the disease [3]. This shows that early detection of tumors and timely treatment is important.

Extensive evidence suggests that kidney tumors are predominantly benign, particularly when detected at a small size [4]. Nevertheless, metastatic kidney cancer remains highly lethal. The challenge of preventing kidney tumors from progressing to metastatic cancer is substantial, primarily due to the trade-off between the risk of disease onset and the costs associated with over-treatment. Although some experts propose that a kidney mass biopsy might resolve this dilemma in therapeutic decision-making, its practical application is limited by inadequate sensitivity and an inability to predict the tumor's future development. Consequently, its adoption in clinical practice is minimal [5].

Increasingly, the term "radiome" is gaining recognition as a strong quantitative predictor in cancer outcomes. In radiomics, precise spatial delineation of image structures is crucial due to the time-consuming nature and significant variability of manual delineation, which can affect algorithm sensitivity [6]. Moreover, to safeguard patients' postoperative well-being, physicians frequently opt for kidney-sparing surgeries, which rely on meticulous preoperative assessments of tumor size, shape, and location to improve surgical results [7]. Thus, developing accurate automatic methods for semantic segmentation is of great interest. Given the scarcity of medical resources, the development of automatic segmentation technologies using artificial intelligence is crucial. These technologies not only reduce human error but also improve precision and efficiency, easing medical burdens and benefiting patients.

1.2 Literature Review

Diagnosis of kidney tumors typically involves analyzing abdominal CT images, with segmentation being crucial for further diagnosis and treatment. Current methods for medical image segmentation can be divided into traditional techniques and approaches based on deep learning.

1.2.1 Traditional Segmentation Method

Earlier developed traditional image segmentation methods are both simple and effective, aiming to extract key features from images to enhance the efficiency of image analysis. These methods are typically based on threshold, edge, graph theory, clustering, and region, and have numerous applications in medical image segmentation [8]. Given the extensive range of methods, this review will focus on a selection of the most representative techniques.

Threshold-based segmentation method: It segments Regions Of Interest(ROI) based solely on pixel values, ignoring spatial correlations. Thresholding, known for its speed, simplicity, and interactivity, is one of the earliest widely-used segmentation methods. The success of this method largely depends on selecting the right threshold. For example, R. Helen et al. used an improved Otsu method for segmenting lung

parenchyma in CT images [9]. However, these results are often coarse and noise-prone, typically used in pre- or post-processing steps in medical image segmentation.

Region-based segmentation method: The method divides images into blocks based on similarity criteria, primarily include region growing [10] and region splitting and merging [11]. For instance, X. Zhang et al. utilized a bidirectional region growing approach for CT brain segmentation [12]. The efficacy of these methods critically depends on the choice of initial seed points, limiting their application on a larger scale. And it is sensitive to noise, which is easy to lead to regional vacancy.

Edge-based segmentation method: In an image, if a significant disparity in gray scale values exists between a pixel and its adjacent pixels, this pixel is likely located at a boundary. Detecting these boundary pixels and connecting them can establish edge contours, effectively segmenting the image into distinct regions. Common edge detection differential operators include Roberts [13], Sobel [14], and Canny [15]. M.N. Saad and colleagues have enhanced segmentation in X-ray imaging of lung parenchyma by integrating edge detection algorithms with morphological processing techniques [16]. Edge detection approaches often fail to fully capture foreground objects, and reconstructing these contours requires significant computational effort. Thus, these methods are rarely applied to three-dimensional segmentation scenarios.

1.2.2 Deep Learning Methods

Due to difficulties in feature representation, image segmentation remains one of the most challenging tasks, especially in extracting features from medical images that are blurry, noisy, and low in contrast. With the advancement of deep learning, medical image segmentation no longer requires manual feature engineering. Deep learning encompasses numerous sub-fields, such as weak supervision and unsupervised learning, each further divided into more specific methods including various backbone networks, loss functions, network blocks, etc. In this section, we primarily focus on the most popular branch within supervised learning: Convolutional Neural Networks (CNN).

Image semantic segmentation aims to achieve pixel-level classification of images. To achieve this, encoder-decoder architectures have been proposed, such as Fully Convolutional Networks (FCN)[17], U-Net [18], Deeplab [19], etc. In these structures, the

encoder is generally used for feature extraction, while the decoder reconstructs the features to their original spatial dimensions and produces the final segmentation result. The first high-impact encoder-decoder architecture was U-Net, proposed by Ronneberger et al.

U-Net is a CNN designed for medical image segmentation. It uses a symmetrical structure and jumping connections to effectively deal with noise and fuzzy boundaries in medical images. By integrating feature maps of different resolutions and employing skip connections, the model effectively combines low-level features with high-level features, this kind of network can capture the details of the image more accurately and better solve the problem of medical image segmentation. Currently, U-Net is widely regarded as the standard for medical image segmentation tasks and has spurred numerous significant advancements.

In medical imaging, such as CT and MRI, data often exists in 3D volume form, thus the use of 3D convolutional kernels can more effectively capture the high-dimensional spatial correlations of the data. Building on this concept, Çiçek et al. expanded the U-Net architecture for 3D data applications, introducing the 3D U-Net that directly processes 3D medical data [20]. However, due to computational resource limitations, 3D U-Net includes only three down-sampling steps, which hampers its ability to extract deep-layer image features, limiting its segmentation accuracy. Additionally, Milletari et al. introduced a similar architecture, V-Net, which incorporates residual connections to design a deeper network structure (four down-samplings) and thus achieves superior performance [21]. Similarly, using residual connections in 3D networks, Yu et al. [22] developed Voxresnet, and Xiao et al. [23] designed Res-UNet. However, these 3D networks typically face challenges with high computational costs and GPU memory usage due to their large number of parameters.

Cascade models of 3D are popular for enhancing accuracy by using multiple sequential models. These models fall into frameworks like coarse-fine segmentation, where two 3D networks are used: the first performs broad segmentation, and the second refines this output. For example, Christ et al. developed a cascaded approach where an initial FCN broadly segments the liver, and a second FCN focuses on detailed liver-tumor segmentation [24]. Similarly, Yuan et al. applied a basic FCN for general liver segmentation

and then used a more detailed FCN to refine this for precise liver-tumor segmentation, incorporating enhanced areas for even greater accuracy [25].

Many studies often focus on modifying U-Net architectures to suit specific tasks. However, Isensee et al. [26] pointed out in their paper that excessive manual adjustments to network structures might lead to overfitting on specific datasets. To address this issue, they introduced a framework named no-newUNet (nnU-Net), which automates the adjustment of hyperparameters based on the characteristics of each dataset, thus eliminating the need for manual intervention. The nnU-Net emphasizes efficient preprocessing, strategic training, smart inference techniques, and advanced post-processing. This approach suggests that while complex network designs often rely on empirical experience lacking theoretical interpretability, they also pose a higher risk of overfitting. Conversely, a finely tuned simple model can outperform most modified variants, significantly influencing the direction of their research.

1.3 Research Content and Thesis Arrangement

1.3.1 Main Research Content

The objective of this study is to propose a series of improved segmentation schemes for application in 3D U-net, guided by the results of the KiTS19 challenge, aimed at segmenting kidneys and kidney tumors in CT images. The main research content of this paper encompasses three aspects:

Firstly, it investigates issues related to the imbalance between foreground and background. The kidneys occupy only a small area of the abdomen, and kidney tumors are even smaller. This imbalance between the kidney, tumor, and background makes training segmentation models challenging.

Secondly, it explores how to increase the depth and receptive field of the network with limited memory resources. The computational cost and GPU memory consumption of 3D images are relatively high. High memory usage limits the depth and receptive field of the network, which are two critical factors for performance improvement.

Thirdly, the study focuses on optimizing network performance for kidney segmentation tasks based on existing 3D segmentation models, particularly adjustments beyond

the network structure. Examples include pre-processing (resampling), post-processing, training settings (loss, data augmentation), inference (patch-based strategies), and other factors.

1.3.2 Main Innovation Point

The research findings of this paper feature the following innovative points:

- The 3D U-net network was modified by incorporating residual blocks and deep supervision. This adaptation, along with external parameter configurations, enables the use of a deeper network within limited memory resources, achieving a larger receptive field and enhancing segmentation accuracy.
- A cascade pipeline was introduced, employing a cascading scheme in the task of kidney segmentation, which mitigates the problem of foreground and background imbalance and improves the precision of tumor segmentation.
- A novel method for delineating regions of interest was proposed, capable of excluding more irrelevant areas, thereby conserving memory and enhancing computational speed.

1.3.3 Organizational structure and chapter arrangement

The thesis is composed of five chapters

Chapter 1 introduces the research background and significance of medical image segmentation, and explains the importance of medical image segmentation in the medical field. This paper introduces the historical research process and research status of medical image segmentation, and analyzes the advantages and disadvantages of existing methods.

Chapter 2 delves into the pre-processing methods commonly utilized in medical imaging, followed by a discussion of classification evaluation techniques. These topics serve as the theoretical foundation for the subsequent chapters, providing essential context and methodologies that underpin further exploration and application within the field.

Chapter 3 is dedicated to a comprehensive presentation of the proposed cascade 3D U-net approach, detailing the pipeline and network architecture extensively.

Chapter 4 explores the experimental framework and the outcomes obtained, providing a detailed discussion and analysis of the results, thereby confirming the effectiveness of the proposed method.

Chapter 5 summarizes the main content and contributions of this paper, evaluating its strengths and weaknesses. It also outlines future research directions and emerging topics in the field of abdominal medical image segmentation.

Chapter 2

Background Theory

2.1 Introduction to CT Imaging of the Abdomen

The liver, kidneys, and their tumors have complex boundaries within the abdominal cavity. The abundance of overlapping organs in the projection complicates tumor localization. Therefore, CT imaging is commonly used for precise localization in clinical diagnosis.

CT is an imaging technique that utilizes X-rays to generate images based on the varying degrees of X-ray transparency of different body tissues, ultimately producing a visual result. The density of these tissues is expressed in Hounsfield Units (HUs), reflecting the absorption levels of X-rays by specific structures within the body. The range extends from +1000 HU (high density), 0 HU (water density), to -1000 HU (low density), where -1000 HU represents air and +1000 HU represents cortical bone. In our experimental method, outliers are sheared out prior to regularization to avoid interference.

2.1.1 CT Image Pre-processing Technologies

The three most important pre-processing techniques for CT images in this research are as follows: Data normalization, Adjustment of window width and level, and Resampling.

1. **Data normalization** can eliminate dimensional differences between data and make different features comparable, which is even more important in the processing of medical images [27].

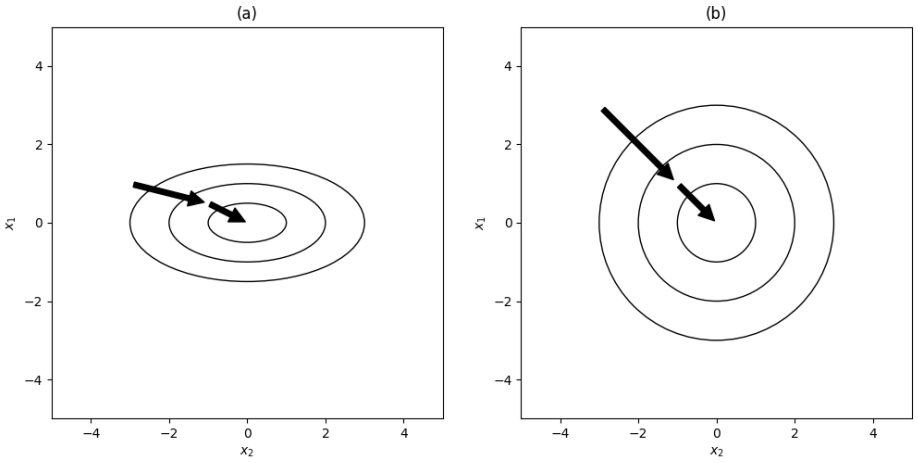

FIGURE 2.1: The Impact of Data Normalization on Gradient Updating. (a) Without Normalization; (b) With Normalization

In medical imaging data, the distribution of pixel values is often influenced by various factors such as scanning equipment, scanning protocols, the physical condition of the scanned individual, gender, age, ethnicity, and even the scanning environment. These variations in data distribution can create challenges in the model training process, affecting convergence and efficiency. As illustrated in Figure 2.1, assume x_1 and x_2 are two sets of input data. In Figure 2.17(a), when the distributions of x_1 and x_2 differ significantly, the update rates for both are not synchronized, requiring more iterations to find the optimal solution. There is also a risk of getting stuck in local optima during the optimization process. However, when x_1 and x_2 are normalized to the same scale, the update rates become more consistent, facilitating easier and more efficient gradient descent to find the optimal solution. In this thesis, I'm using Z-score normalization.

2. **The CT window level and window width** can be adjusted to clearly display specific organ regions. Window width represents the range of grayscale values in the image; a wider window width encompasses a larger range of grayscale values, and vice versa. The window level is the midpoint of the window width range and is typically set to the specific grayscale value of the tissue being observed, thereby highlighting particular tissue structures. Increasing the window width reveals more detailed information about various tissues and organs, but reduces contrast, potentially impairing diagnostic capability. Conversely, a narrower window width

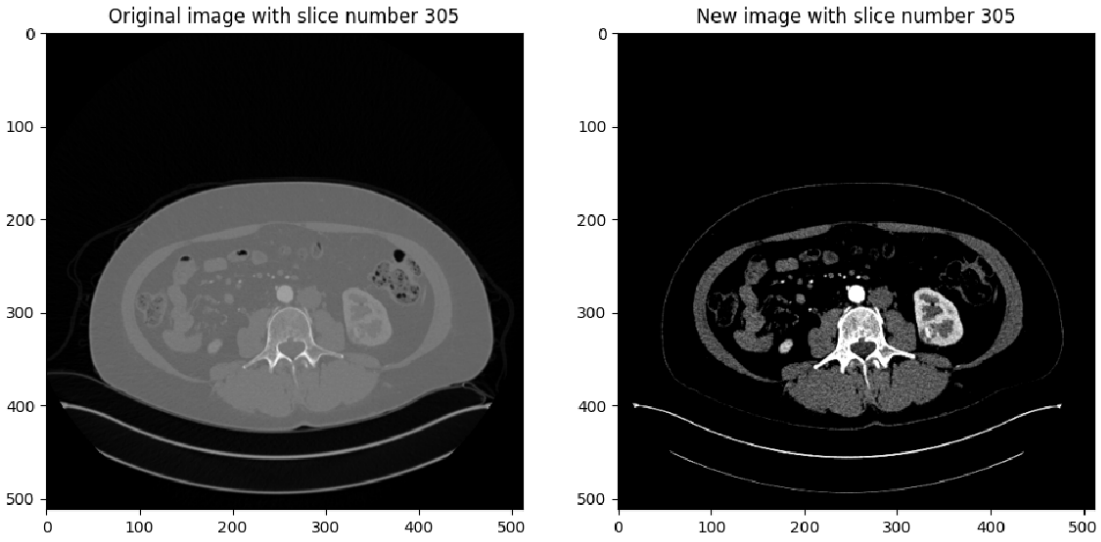

FIGURE 2.2: Original CT Image and the Pre-processed CT Image

reduces the number of displayed structures but enhances contrast, aiding in clearer observation of the anatomical details.

In practical applications, the appropriate window width and level should encompass the variability of the observed tissue's CT values. For instance, the CT value of the liver ranges from 50-70 HU, and that of the kidneys ranges from 40-50 HU. For clearer diagnostic imaging of abdominal organs, adjustments to the window width and level are often made to facilitate image segmentation. As shown in the Figure 2.2, this allows for more accurate differentiation of the liver, kidneys, and other organ types.

3. Different scanning devices and scanning protocols may produce CT images with different resolutions, which is a unique challenge for medical imaging data. **Resampling normalizes** these images to a uniform resolution and voxel size, making subsequent image processing and analysis more consistent. Improved model stability and compatibility. The resampling technique can also upsample or downsample the image according to the requirements, such as reducing the amount of data and improving the computational efficiency. This technique will be used in the experiments in this paper. The principle is to change the resolution and size of the image through spatial coordinate transformation and interpolation methods (such as nearest neighbor interpolation, bilinear interpolation, cubic interpolation, etc.) for further processing and analysis, and at the same time reduce artifacts and noise introduced by resampling through anti-aliasing and denoising processing.

2.2 Medical Image Evaluation Matrix and Loss Function

The Dice coefficient is a gold standard to measure segmentation [28]. Its calculation formula is shown in Equation 2-17. It ranges from 0 to 1, with values closer to 1 indicating more accurate segmentation.

$$Dice = \left(\frac{2 \times (V_{seg} \cap V_{gt})}{V_{seg} + V_{gt}} \right) \times 100\% \tag{2.1}$$

There are two commonly used methods for calculating the Dice metric. The first method applies the Dice calculation process to all segmentation targets and then averages the Dice values obtained for each target. This approach can cause segmentation targets with larger volumes to have a significant impact on the overall Dice value. The second method treats the Dice coefficient as an evaluation metric for each individual segmentation target. First, the Dice values for each target are calculated separately, and then the average of these Dice values is taken. This method is suitable for multi-categorical classification or cases where there are large differences in the sizes of the segmentation targets. In this study, the second Dice calculation method is adopted.

The Dice coefficient has several variants that are widely studied and applied in the field of medical image segmentation and computer vision to suit different needs and scenarios. Other commonly used variants are as follows:

- The Generalized Dice Coefficient addresses the class imbalance problem by assigning different weights to each class [28]. It is defined as:

$$GDC = \frac{2 \sum_{l=1}^{L} w_l \sum_{i=1}^{N} (p_{il} \cdot g_{il})}{\sum_{l=1}^{L} w_l \sum_{i=1}^{N} (p_{il} + g_{il})} \tag{2.2}$$

  where w_l is the weight for class l, and p_{il} and g_{il} represent the prediction and ground truth for class l at voxel i.

- The Squared Dice Loss is an improved version of the Dice loss that squares the overlap part to reduce the vanishing gradient problem [21]:

$$Dice_{squared} = \frac{2 \sum (P \cap G)}{P^2 + G^2} \tag{2.3}$$

  where P and G denote the prediction and ground truth segmentations.

- The Tversky Index is a generalization of the Dice coefficient, with parameters to control the penalties for false positives and false negatives [29]:

$$Tversky(P, G; \alpha, \beta) = \frac{|P \cap G|}{|P \cap G| + \alpha|P \setminus G| + \beta|G \setminus P|} \tag{2.4}$$

  When $\alpha = \beta = 0.5$, the Tversky index reduces to the Dice coefficient.

- The Surface Dice Coefficient focuses on the matching degree of boundaries in image segmentation [30]. It is defined as:

$$\text{Surface Dice} = \frac{2 \cdot |A_\delta \cap B_\delta|}{|A_\delta| + |B_\delta|}$$

  where A_δ is the set of boundary points within the threshold distance from the segmentation result, and B_δ is the set of boundary points within the threshold distance from the ground truth.

Chapter 3

Methods

3.1 Cascade Method Pipline

Compared to 2D neural networks, processing 3D data presents challenges such as high computational cost and significant memory consumption. Even when computational resources are not a limiting factor, researchers still seek lightweight, fast, and accurate methods. So, this thesis proposes a cascaded segmentation framework for the MICCAI 2019 Kidney Tumor Segmentation Challenge (KiTS 2019). The framework is shown in Figure 3.1.

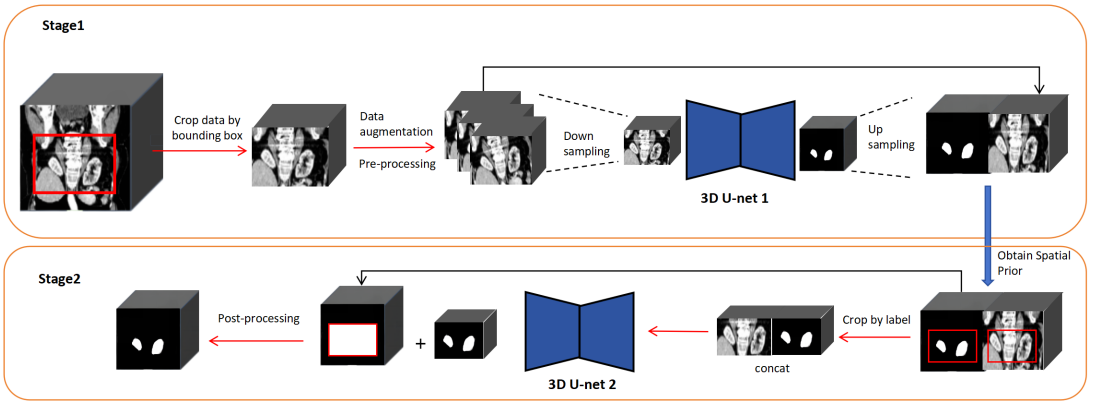

FIGURE 3.1: The Pipline of the Method

In Stage 1, the process begins by cropping the data using a bounding box to isolate the region of interest, aiming to reduce irrelevant background and focus on the area

likely containing the kidney and tumor. After cropping, the data undergoes augmentation and pre-processing to enhance the model's generalization capabilities by creating a more diverse training set. Next, the pre-processed data is downsampled to reduce computational complexity and memory usage. This downsampled data is then fed into the first 3D U-Net model (3D U-Net 1) for initial segmentation, which identifies the kidney and tumor regions. The output is subsequently upsampled to restore the original resolution, ensuring spatial details are preserved. The initial segmentation result provides a spatial prior, offering valuable information on the likely locations of the kidney and tumor, which guides the subsequent stage.

In Stage 2, the spatial information derived from Stage 1 is utilized to enhance segmentation accuracy. As illustrated in Figure 3.1, the image is cropped using the initial segmentation labels, which allows the model to concentrate on a more precise area. This targeted approach reduces the search space and significantly improves segmentation precision. The cropped region, along with the initial segmentation results, is then input into a second 3D U-Net model for detailed analysis. This stage aims to achieve a more accurate and precise delineation of the kidney and tumor. Finally, post-processing techniques are applied to the output to correct any remaining artifacts or errors, thereby ensuring higher accuracy and overall quality.

This two-stage cascaded approach systematically refines the segmentation results at each step, thereby significantly improving the accuracy and robustness of kidney tumor delineation. More details will be given below.

3.2 Model

The detailed architecture of the 3D U-Net models used in our cascaded segmentation framework is depicted in Figure 3.2. The improved model adds residual blocks and deep supervision [31], and uses instance regularization and LeakyReLU [32] as activation functions.

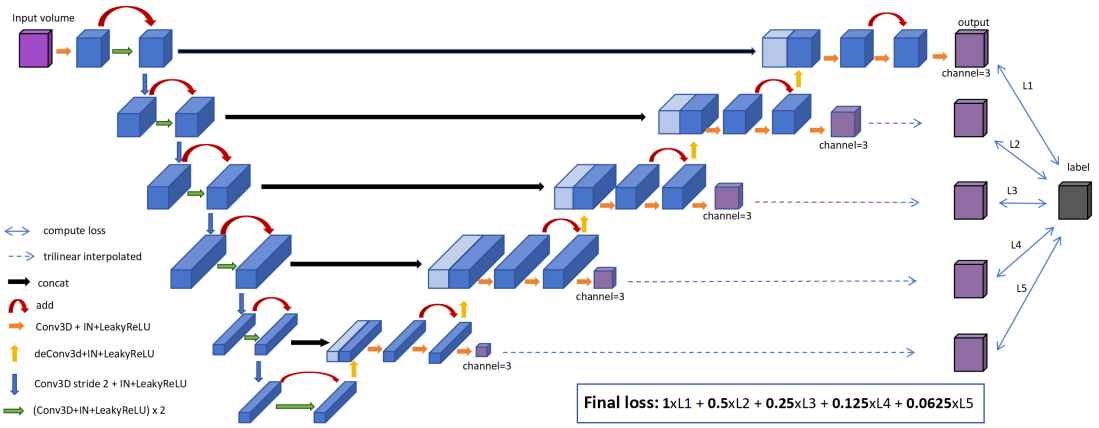

FIGURE 3.2: The 3D segmentation model based on 3D U-net used in our pipline

The input volume undergoes a series of 3D convolutions, each followed by Instance Normalization (IN) [33] and LeakyReLU activation function to extract low-level features. **The encoding pathway** involves:

- **Conv3D + IN + LeakyReLU**: Extracts features and introduces non-linearity.
- **Downsampling (Conv3D stride 2 + IN + LeakyReLU)**: Halves the spatial dimensions, capturing more abstract features while reducing computational complexity.

The bottleneck layer contains the most abstracted features, serving as a bridge between the encoder and decoder.

The decoder restores the spatial resolution through:

- **Upsampling (deConv3D + IN + LeakyReLU)**: Increases spatial dimensions, combining high-level and detailed features.
- **Skip Connections (concat)**: Merges upsampled feature maps with corresponding encoder feature maps to preserve spatial context and detail, essential for accurate segmentation.

The final output is a segmented volume with the same spatial dimensions as the input, but with a specified number of channels corresponding to the segmented classes. The model employs a **multi-scale loss computation strategy**, which calculates losses at

multiple levels of the network (L1, L2, L3, L4, L5). These intermediate losses are then weighted and combined to form the final loss. This multi-scale approach ensures that the network learns to make accurate predictions at various resolutions, enhancing its robustness and ability to generalize.

Due to the considerable depth of the network, residual blocks have been incorporated into both the encoder and decoder to prevent gradient vanishing. Additionally, to enhance the regularization effect and facilitate more efficient gradient propagation, deep supervision has been integrated into the model. Instance normalization is employed to reduce dependency on batch size, thus mitigating the impact of small batches necessitated by limited GPU resources. To prevent the issue of ReLU dying, LeakyReLU is introduced, allowing a small non-zero gradient when the input is negative, thereby preventing neuron inactivity (becoming inactive for all inputs). These enhancements are implemented to improve the efficiency and generalization capability of the model.

3.3 Evaluation Matrix

In kidney segmentation, accuracy often suffers due to errors in boundary predictions, and it is crucial to address the imbalance between the foreground and background. To solve these two problem, the evaluation metric used in this study is a combination of the Surface Dice(for boundary predictions) and Generalized Dice coefficients introduced in Chapter 2. By averaging these two Dice scores, the final Dice score provides a balanced assessment, guiding the model's improvement.

3.4 Preprocessing and Postprocessing Methods

This section will introduce the main preprocessing and postprocessing methods in the cascade method pipline.

Bounding box: This paper proposes a method to remove more irrelevant areas in bounding boxes, improving image reading and computational efficiency. There are two existing methods:

1. Manually specifying a fixed-size cuboid as the bounding box at a fixed location to crop the foreground region.

2. Cropping the foreground region for each 3D image based on the label.

The first method results in a relatively large ROI because the positions of each patient's kidneys in the image are not exactly the same, especially in the z-axis direction (assuming the ROI of image A is higher and that of image B is lower, the bounding box needs to be expanded to cover the entire ROI area of both images). This method results in a bounding box that is too large, with a lot of unnecessary space not being cropped out. The second method improves this issue, but the cropped result sizes vary, making it difficult to directly input them into the network or perform patch training.

Therefore, combining the advantages of these two methods, the new method will determine a fixed bounding box size that can cover all target areas in each training CT image, with the center of the bounding box being determined by the label in each image. This approach sufficiently removes excess regions while ensuring consistent image sizes. This method indeed significantly removes unnecessary regions; however, the prerequisite for using it is the availability of a label, whether it is the ground truth or the prediction generated in stage 1 of our method.

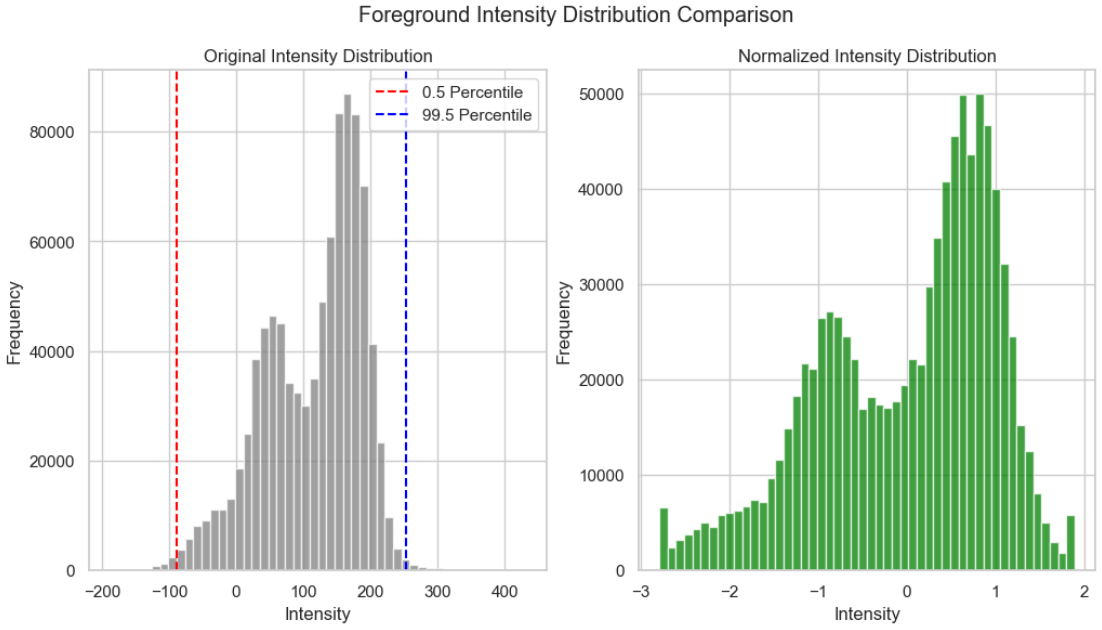

FIGURE 3.3: Foreground Intensity Distribution Comparison for a example case

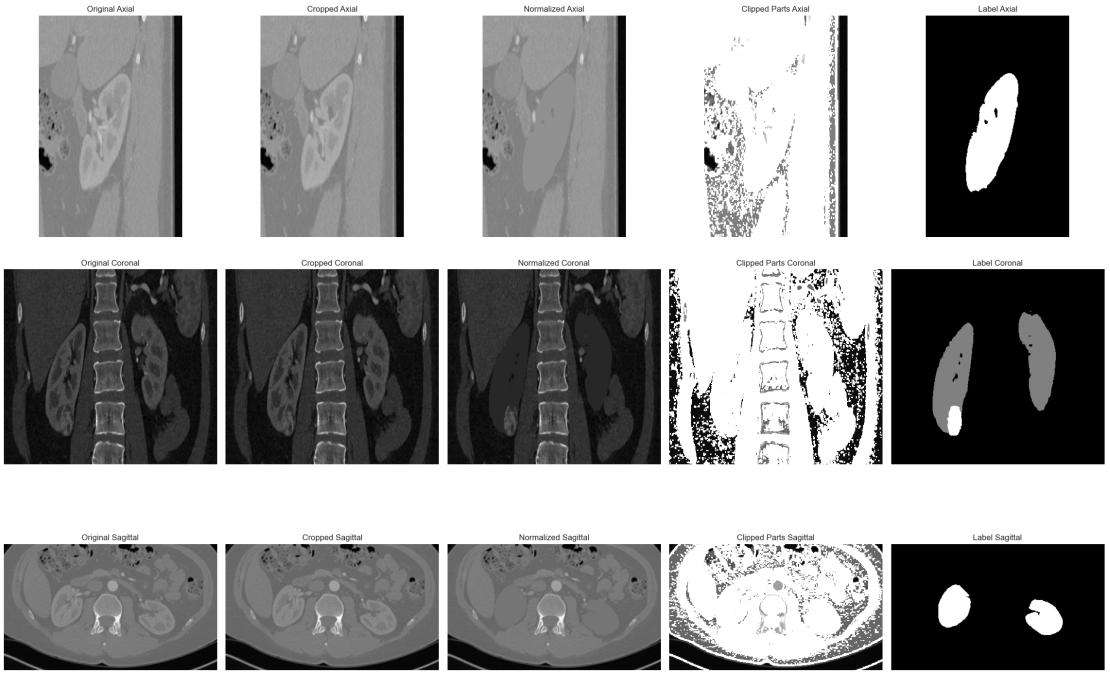

FIGURE 3.4: Visualization of the Normalization

Normalization: The intensity values of each case were clipped to the 0.5th and 99.5th percentiles of the foreground region intensity values in the training set, as shown in Figure 3.3, retaining only the data between the red and blue lines. This avoids the interference of outliers, such as medical devices that sometimes have metal in the body, which are much denser than human tissue. Then, intensity values were normalized by subtracting the mean and then dividing by the standard deviation of the foreground region intensity. The distribution of the normalized intensity data is shown in the second image of Figure 3.3. From the fourth column of Figure 3.4, it can be seen that scaling the intensity removes some noise, and these extreme values might affect the normalization effect. Comparing the third and the last columns, it is evident that the normalized images are more favorable for segmentation and closer to the label.

Resampling: To enhance the network's ability to learn spatial semantic information, all patient data were resampled to align with the median voxel spacing of their respective datasets. For the image data, third-order spline interpolation was employed, resulting in volumes being resampled to $0.78 \times 0.78 \times 1.0$ mm, while nearest neighbor interpolation was applied to the corresponding segmentation masks.

Postprocessing: Using connected component analysis, remove all connected components except for the largest one to reduce false positive regions.

Chapter 4

Experiment and Analysis

4.1 Dataset Introduction

The KiTS19 dataset used in this research, published as part of the MICCAI 2019 Kidney Challenge, is a authoritative public dataset [34]. Since its introduction, nearly all state-of-the-art segmentation network models have utilized this dataset for training and testing network performance. To this day, submissions to the dataset's website continue to increase. The dataset is comprised of 210 cases. Each scan contains two kidneys and one or more kidney tumors. Both the original scan images and the manually annotated gold standard images are stored in NIfTI format, which is a three-dimensional data

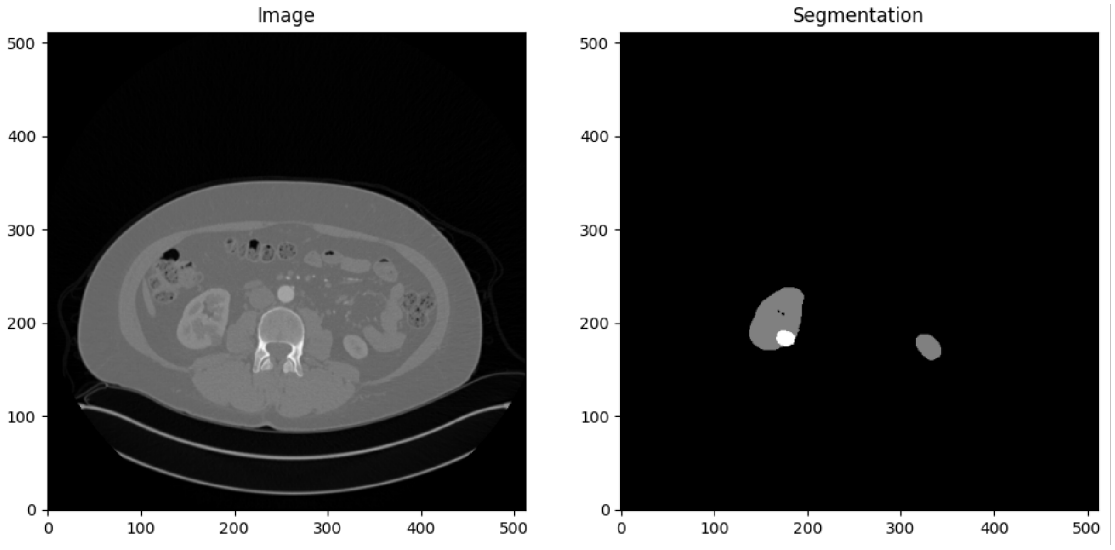

FIGURE 4.1: A slice example of Case 0

type (num_slices, height, width), where the num_slices parameter corresponds to the axial view of the slice data. In the experiments of this study, I divided it into training datasets of 190 cases and validation dataset of 20 cases. Figure 4.1 is an example of a slice of data.

4.1.1 Data Statistics Analysis

The data file contains some metadata, with two particularly important ones for segmentation tasks: voxel spacing information and image size. The former is used for resampling and normalization; the latter is for making it compatible with network models and pre-processing methods.

TABLE 4.1: Statistical Spacing Values of Dimensions in KiTS19 Dataset

Statistic	**X-axis (mm)**	**Y-axis (mm)**	**Z-axis (mm)**
Mean Spacing	0.80	0.80	3.18
Median Spacing	0.78	0.78	3.00
Max Spacing	1.04	1.04	5.00
Min Spacing	0.44	0.44	0.50
Q1 Spacing	0.72	0.72	1.25
Q3 Spacing	0.87	0.87	5.00

Table 4.1 shows the statistical spacing for each dimension in the KiTS19 dataset. The Z-axis spacing varies significantly, ranging from 0.50mm to 5.00mm, indicating variability due to different scanning instruments or modes. The X-axis and Y-axis spacings are more consistent, ranging from 0.44mm to 1.04mm. The interquartile range (Q1 and Q3) reveals that Z-axis spacing is concentrated at higher values (Q3 is 5.00mm). This information will be used to optimize subsequent image processing.

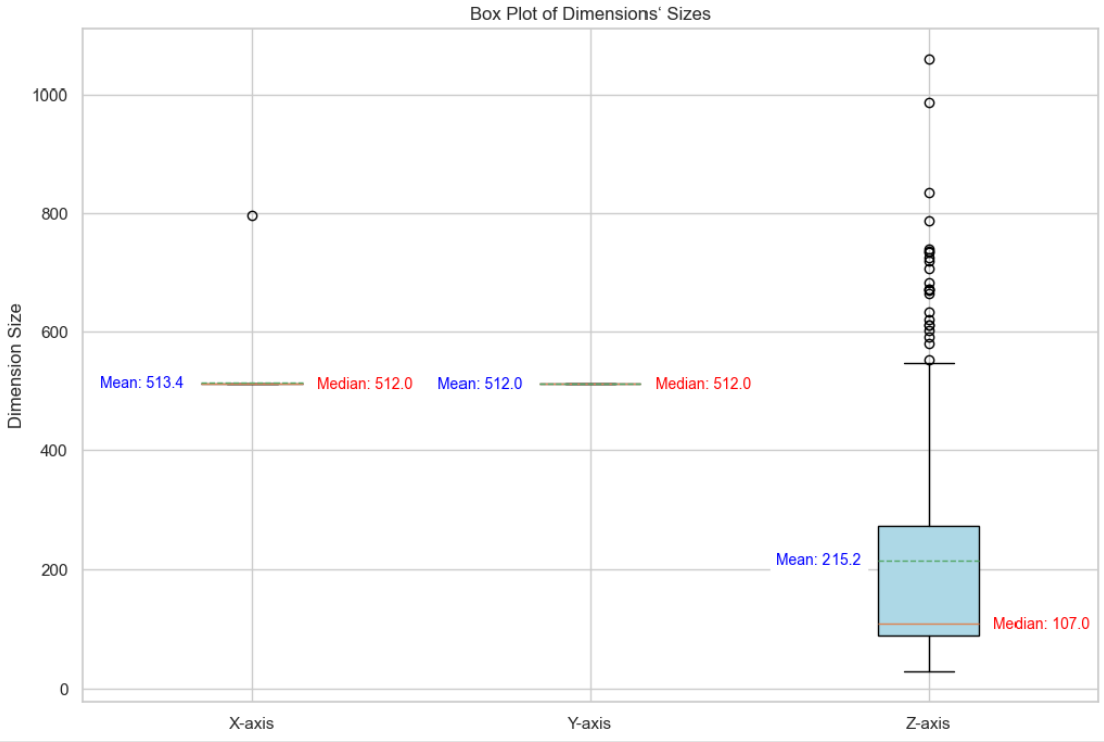

FIGURE 4.2: Box Plot of the Dimensions' Sizes for KiTS19

From Figure 4.2, it can be seen that the X-axis and Y-axis dimensions are consistent and nearly identical for all data, except for one outlier. We excluded this outlier in our experiments as it would affect pre-processing operations like normalization and density adjustment. The Z-axis dimensions are more widely distributed.

4.1.2 Dataset Analysis

Limitations: The KiTS19 challenge dataset only includes patients treated within a single healthcare system, resulting in a high concentration of data from one geographic area [35]. This concentration might limit the **generalizability** of techniques created with this data to different areas globally. However, the dataset benefits from diverse imaging protocols and scanner types, as preoperative studies were conducted at various institutions. Additionally, **the KiTS19 dataset is retrospective**, and the training and test sets were divided randomly. Consequently, there are concerns that changes in data distribution over time might reduce the effectiveness of these methods on future prospective data.

Bias: These biases may affect the representativeness of the dataset and the generalizability of the research results, but it is difficult to determine and verify whether these biases have a substantial impact on the research results.

- **Patient Selection Bias:** The dataset includes only patients who did not opt out of research data usage, excluding those who chose to protect their privacy. This may result in a sample that does not fully represent all patients with renal malignancies.
- **Imaging Data Bias:** Only patients with available contrast-enhanced imaging data are included. This excludes those without such imaging, potentially affecting data comprehensiveness and representativeness.

Errors: Manual annotation of segmentation labels is prone to errors. Due to the presence of some errors, I believe that when the evaluation metrics (such as the Dice coefficient) exceed a certain threshold (such as 97%), efforts to further improve accuracy may become less meaningful. This is because it would become too difficult to distinguish between the errors made by the model and its performance that has already reached or surpassed human levels. Whether to pursue higher accuracy should depend on the specific application and requirements. In fields like clinical practice, excessively high accuracy may be unnecessary and waste resources.

4.2 Experiment Setup

In experiments, we utilized a cloud server instance equipped with PyTorch 2.1.0 and Python 3.10 on Ubuntu 22.04. The hardware configuration included a single NVIDIA RTX 4090 GPU with 24GB of memory, 16 virtual CPUs from an Intel(R) Xeon(R) Gold 6430 processor, and 120GB of RAM. The setup also included CUDA 12.1 for GPU acceleration.

4.3 Training Program

This research conducted experiments on the KiTS19 dataset using a patch-based training method combined with a sliding window inference approach. The training configuration was carefully designed to maximize model performance while avoiding unnecessary resource expenditure. The experiment selected a batch size of 2 to balance data processing capacity with memory constraints and computational efficiency. Throughout the training process, the experiment dynamically adjusted the learning rate, maintaining a minimum threshold of 1e-06. To manage the learning rate effectively, the experiment employed a scheduler with a tolerance of 0.001 and a patience period of 30 epochs, enabling appropriate adjustments during performance plateaus to prevent overfitting and enhance training robustness. Training was conducted over 350 epochs, providing the model with sufficient iterations to converge and generalize effectively.

The optimization strategy utilized the Stochastic Gradient Descent optimizer, configured with a momentum of 0.99 to accelerate convergence and stabilize the training process. Additionally, the experiment applied a weight decay of 0.00003 to regularize the model, mitigating overfitting by penalizing large weights.

Data augmentation can enhance the model's generalization capabilities. Various augmentation techniques, including rotation, scaling, gamma correction, and mirroring, were incorporated. These augmentations simulated different variations and distortions in the input data, improving the model's robustness to real-world variations.

- **Rotation, scaling, and gamma correction:** Change the orientation of the image by rotating, resize the image by scaling, and improve the brightness and contrast of the image with gamma correction to optimize the visual effect.
- **Mirroring operations:** Horizontal and vertical flipping images are included for data augmentation.

4.4 Experiment Result and Analysis

The experiment conducted on the KiTS19 dataset involved evaluating various kidney and tumor segmentation methods, with results summarized in Table 4.2. The table

TABLE 4.2: Comparison of Kidney and Tumor Segmentation Methods. Res indicates Residuals Module; DS indicates deep supervision

Methods/(Team)	**Kidney Dice**	**Tumor Dice**	**Mean Dice**	**Online Leaderboard Rank**
		Baselines		
nnUnet	97.95	85.54	91.75	7
(Isensee F et al.)	97.37	85.09	91.23	64(Placed 1st in 2019)
(Xiaoshuai Hou et al.)	96.74	84.54	90.64	106(Placed 2nd in 2019)
Vanilla 3D U-Net	97.02	81.76	89.39	289
		Proposed Method (PM)		
PM	97.94	85.24	91.59	23
PM without Res	97.54	84.63	91.09	73
PM without Res and DS	97.39	84.39	90.89	86

compares baseline methods with the proposed method (PM) and its variants (without some modules).

To provide a comprehensive comparison, I have listed the performance of the currently most popular and state-of-the-art method, nnU-Net, alongside a standard 3D U-Net (since the proposed method is based on 3D U-Net), as well as the scores and rankings of the top two participants in the 2019 competition. This will thoroughly showcase the performance of the proposed method.

PM achieved a Kidney Dice of 97.94, a Tumor Dice of 85.24, and a Mean Dice of 91.59, resulting in a 23rd place ranking. This demonstrates PM's competitive performance and effectiveness.

To further understand the contributions of different components, two variants of PM were evaluated. The first variant, PM without Rcs, showed a slight decrease in performance with a Kidney Dice of 97.54, Tumor Dice of 84.63, and Mean Dice of 91.09, ranking 73rd. This indicates that removing the Rcs component negatively impacts tumor segmentation accuracy.

The second variant, PM without Rcs and DS, showed further performance degradation, achieving a Kidney Dice of 97.39, Tumor Dice of 84.39, and Mean Dice of 90.89, ranking 86th. This suggests that both Rcs and DS components are essential for maintaining high segmentation performance.

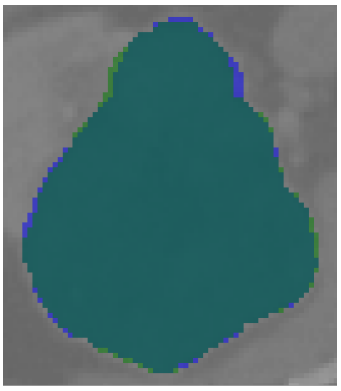

FIGURE 4.3: Segmentation Visualization of a Sample Case. Blue is ground truth label, green is the predicted label

The segmentation accuracy of PM still needs improvement at the edges. From the visualized segmentation results in Figure 4.3, it is evident that most of the accuracy loss occurs at the edges. However, it seems unavoidable due to the inherent errors people make when drawing the ground truth. This issue is common among most teams. The necessity of solving this problem should be judged in conjunction with clinical needs.

Chapter 5

Conclusion and Future

5.1 Summary of the Entire Work

This study concentrates on segmenting kidneys and renal tumors in abdominal CT scans, tackling the challenge of foreground-background imbalance. The goal is to enhance the depth and receptive field of networks within limited memory resources and to optimize the performance of kidney segmentation tasks based on existing 3D segmentation models. Traditional 3D-Unet models are not precise enough for target segmentation. Despite improvements in accuracy with model development, the models have become increasingly complex and computationally intensive (e.g., segmentation models incorporating transformers). The main contributions of this paper are as follows:

This paper provides a comprehensive review of the development of a specific branch in this field, focusing on the evolution of the 3D U-Net in deep learning. It introduces abdominal CT imaging and some important preprocessing techniques, along with background theories. Additionally, it covers the Medical Image Evaluation Matrix and loss functions.

A cascade pipeline is proposed, where the first stage performs coarse segmentation to determine the target location, and the obtained prior information is passed to the second stage for fine segmentation. This approach balances accuracy with limited computational resources. Auxiliary processing, such as using bounding boxes to remove irrelevant regions, improves model efficiency and indirectly addresses the foreground-background imbalance. An improved 3D U-Net is also proposed, incorporating residual blocks and

deep supervision to prevent gradient vanishing, allowing the model to increase depth and receptive field appropriately while enhancing generalization ability. Experimental results validate the effectiveness of the proposed modules. During the experiments, it was found that edge accuracy loss was significant. To mitigate this issue, the model used a combination of Surface Dice (for boundary predictions) and Generalized Dice coefficients as the loss function.

The paper also provides an in-depth analysis of the dataset and discusses the necessity of pursuing higher accuracy, considering that the inherent error might exceed the required improvement in accuracy.

5.2 Future Work

We plan on enhancing our segmentation techniques in future endeavors, with the aim to address challenges that persist after optimization and to venture into new realms. One major aspect we are looking into is the enhancement of edge accuracy and segmentation precision. This will be achieved through advanced loss functions and optimization techniques, in addition to other model upgrades that we plan on implementing.

For tasks that have already achieved high accuracy, such as kidney segmentation where the Dice coefficient is high, it should be beneficial to use error comparison as an evaluation metric. By having two groups annotate the ground truth, if the method's error with ground truth 1 is less than the error between ground truth 1 and ground truth 2, then the improvement can be considered effective.

We also plan to leverage the efficiency of EfficientNet and a large amount of training data by integrating it with our encoder to further enhance model performance.

Looking ahead, with the development of large models and drawing on the progress in the NLP field, we predict that "intermediate tasks" will gradually be phased out, and end-to-end tasks/systems will become the main trend. Therefore, we will focus on developing more integrated end-to-end segmentation systems to improve overall task performance and user experience.

Given the rapid advancement in large language models (LLMs) and the current scarcity and inaccuracy of annotated data, exploring zero-shot learning applications is necessary.

This involves achieving efficient model training and application with little to no annotated data. Through these measures, we hope to further enhance the robustness and generalization ability of our models, validating our methods on larger and more diverse datasets to verify their applicability across various populations and imaging conditions.

Bibliography

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, Ahmedin Jemal, et al. Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48, 2023.

[3] Yukio Homma, Kazuki Kawabe, Tadaichi Kitamura, Yoji Nishimura, Mitsuru Shinohara, Yasushi Kondo, Isao Saito, Shigeru Minowada, and Yasuyuki Asakage. Increased incidental detection and reduced mortality in renal cancer—recent retrospective analysis at eight institutions. *International journal of urology*, 2(2):77–80, 1995.

[4] Carmen Sebastià, Daniel Corominas, Mireia Musquera, Blanca Paño, Tarek Ajami, and Carlos Nicolau. Active surveillance of small renal masses. *Insights into imaging*, 11:1–18, 2020.

[5] John T Leppert, Janet Hanley, Todd H Wagner, Benjamin I Chung, Sandy Srinivas, Glenn M Chertow, James D Brooks, Christopher S Saigal, Urologic Diseases in America Project, et al. Utilization of renal mass biopsy in patients with renal cell carcinoma. *Urology*, 83(4):774–780, 2014.

[6] Rongjie Liu, Hesham Elhalawani, Abdallah Sherif Radwan Mohamed, Baher Elgohari, Laurence Court, Hongtu Zhu, and Clifton David Fuller. Stability analysis of ct radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clinical and Translational Radiation Oncology*, 21:11–18, 2020.

[7] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.

[8] KKD Ramesh, G Kiran Kumar, K Swapna, Debabrata Datta, and S Suman Rajest. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(27):e6–e6, 2021.

[9] R Helen, N Kamaraj, K Selvi, and V Raja Raman. Segmentation of pulmonary parenchyma in ct lung images based on 2d otsu optimized by pso. In *2011 international conference on emerging trends in electrical and computer technology*, pages 536–541. IEEE, 2011.

[10] Alain Tremeau and Nathalie Borel. A region growing and merging algorithm to color segmentation. *Pattern recognition*, 30(7):1191–1203, 1997.

[11] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[12] Xiaoli Zhang, Xiongfei Li, and Yuncong Feng. A medical image segmentation algorithm based on bi-directional region growing. *Optik*, 126(20):2398–2404, 2015.

[13] Azriel Rosenfeld. The max roberts operator is a hueckel-type edge detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):101–103, 1981.

[14] Wenshuo Gao, Xiaoguang Zhang, Lei Yang, and Huizhong Liu. An improved sobel edge detection. In *2010 3rd International conference on computer science and information technology*, volume 5, pages 67–71. IEEE, 2010.

[15] Li Er-Sen, Zhu Shu-Long, Zhu Bao-shan, Zhao Yong, Xia Chao-gui, and Song Li-hua. An adaptive edge-detection method based on the canny operator. In *2009 International Conference on Environmental Science and Information Application Technology*, volume 1, pages 465–469. IEEE, 2009.

[16] Mohd Nizam Saad, Zurina Muda, Noraidah Sahari Ashaari, and Hamzaini Abdul Hamid. Image segmentation for lung region in chest x-ray images using edge detection and morphology. In *2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014)*, pages 46–51. IEEE, 2014.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[19] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[20] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[21] Fausto Milletari, Nassir Navab, and Seyed-Ahmad 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.

[22] Hao Chen, Qi Dou, Lequan Yu, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, 2016.

[23] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018.

[24] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016:*

19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, pages 415–423. Springer, 2016.

[25] Wei Tang, Dongsheng Zou, Su Yang, and Jing Shi. Dsl: Automatic liver segmentation with faster r-cnn and deeplab. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part II 27*, pages 137–147. Springer, 2018.

[26] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[28] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[29] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[30] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[31] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. Pmlr, 2015.

[32] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[34] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[35] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020.