UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Advancing medical image segmentation via pseudo-labeling of public datasets

Author:
Roman MISHCHENKO

Supervisor:
Dmytro FISHMAN

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2023

Declaration of Authorship

I, Roman MISHCHENKO, declare that this thesis titled, “Advancing medical image segmentation via pseudo-labeling of public datasets” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“Keep the candle above your head,
until your hand becomes tired —
that’s a whole life. Night is not enough.”

Vasyl Stus

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Advancing medical image segmentation via pseudo-labeling of public datasets

by Roman MISHCHENKO

Abstract

Our study explores the difficulties and possible resolutions in the domain of medical image segmentation, with a special emphasis on utilizing unlabeled public datasets to improve tumor segmentation. We suggest a strategy that incorporates pseudo-labeling methodologies with real-world data to enhance the learning potential of segmentation models. Yet, the findings imply that while improvements in model performance exist, they are not substantial. The research underscores the paramount importance of data quality over quantity, emphasizing that image characteristics influence the effectiveness of the process more than the total number of images.

Contents

List of Figures

List of Tables

Dedicated to the Armed Forces of Ukraine

Chapter 1

Introduction

In the field of medical conditions, cancer stands as one of the most all-round and significant health challenges faced by humanity. Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells in the body. It arises when the body's normal control mechanisms stop working, and old cells do not die, forming a mass of tissue called a tumor. Not all tumors are cancerous; benign tumors do not spread to other body parts and are not life-threatening. There are numerous types of cancer, including breast, lung, skin, kidney, and prostate cancer. Despite its severity, advancements in early detection and treatment have significantly improved survival rates for many types of cancer. A key factor in these advancements is the strategic importance of radiation therapy planning. This process is critical as it significantly influences the effectiveness of the treatment, ensuring optimal therapeutic benefit for patients while reducing unnecessary damage to their health. Utilizing imaging techniques such as computed tomography (CT) scans, oncology specialists can accurately pinpoint the cancer's location, allowing for precision in radiation delivery. This careful planning and precision not only enhance the effectiveness of the therapy but also help reduce potential side effects.

1.1 Motivation

Since speed and accuracy are highly important for successful treatment, Computer Vision and Machine Learning algorithms have become popular for medical image segmentation [12]. Accurate automatic segmentation of computed tomography images significantly speeds up the work of radiologists, improves the precision of radiation therapy planning, facilitates quicker analysis to reduce the potential for human error in the diagnostic process, and curtails the overall cost of healthcare [17, 6]. Additionally, reducing errors in CT image segmentation can increase trust in the results and reduce the need for additional studies. However, there are still several challenges associated with tumor segmentation that need to be addressed effectively. The first is the lack of labeled data; in many cases, obtaining a large amount of manually annotated data is difficult or expensive, making it challenging to constantly improve the quality of segmentations. The second is the opportunity to handle the new data. The heterogeneity of medical images poses a significant challenge in developing universally applicable models for image segmentation. The two primary factors to consider are imaging conditions and inter-patient variability. The former includes discrepancies arising from the brand and model of the imaging equipment, imaging protocols, patient positioning, etc. On the other hand, inter-patient variability accounts for the significant anatomical differences and disease variety among patients. For instance, the size and shape of organs can greatly vary between individuals. Furthermore, tumors exhibit a vast range of shapes, fluctuate greatly in size, and can appear in various body locations - from organs like lungs or kidneys to tissues such as bone or skin, which collectively underscores the complexity of accurate tumor detection and treatment. Therefore, it can be difficult for the segmentation model to be effectively used on new, previously unencountered data.

1.2 Proposed solution

The main challenge in the field of medical segmentation is the lack of labeled data [17, 14]. This problem arises due to confidentiality limitations within the medical field and the expenses associated with data labeling. Typically, to address this problem, methods aimed at extracting more information from the data or augmenting labeled data are employed [16, 17]. However, in this study, our emphasis is not on data augmentation. Instead, we aim to investigate the advantages of using additional unlabeled real-world data. We will particularly focus on the potential benefits of pseudo-labeling this unlabeled data as a strategy to improve the learning capabilities of our models. Nonetheless, this procedure won't replace the existing augmentation techniques since they could be applied in cooperation; instead, it offers an additional way to enhance the quality of segmentations. Therefore, our research questions in this work can be summarized as follows:

1. Does pseudo-labeled data benefit the performance of models in the task of medical image segmentation?
2. How does the incorporation of different unlabeled datasets affect the performance of the models?

The remainder of this thesis is structured as follows. Chapter 2 provides a comprehensive review of the medical image segmentation field and the progress made in leveraging unlabeled data through semi-supervised methods. In Chapter 3, we will detail our suggested solution, the data employed, and the metrics used for evaluation. Subsequently, Chapter 4 will discuss the experiments conducted, each presented in the order they were carried out, together with their respective results. Our findings and conclusions will be summarized in the final chapter.

Chapter 2

Related Work

This chapter will cover the context and review recent progress in the Medical Image Segmentation field. As part of this review, we will discuss methods that make use of unlabelled data to improve model training. We will also elaborate on current challenges in the medical image segmentation area related to or adjacent to our goals of creating artificial labels for public datasets and using them for training.

2.1 Medical Images Segmentation

Medical image segmentation is an essential task in medical practice, the primary focus of which is to increase the reliability of results during disease diagnosis. During this process, human experts assign a label to each pixel in the image. This label determines to which class or category that pixel belongs. Based on these labeled images, medical workers can plan appropriate treatment for each patient individually. Recently, Machine Learning algorithms and Computer Vision approaches have gained popularity in image analysis due to their ability to predict pixel-wise labels [8]. A prime example of this is U-Net, a convolutional neural network (CNN) specifically designed for the task of medical image segmentation [12]. This kind of CNN can be utilized for the segmentation of various organ types. However, due to the variations in organ structures and the complexities of various illnesses, it remains a challenge to enhance the effectiveness of the segmentation models.

Most modern segmentation models are designed as fully supervised, implying that these approaches rely on labeled data during the model development process. This usage of labeled data essentially means that each instance in the training dataset comes with a corresponding label that denotes the correct output for the segmentation. Thus, a fully supervised method involves training a machine learning model with a set of input-output pairs where both the input features and the output target or labels are known. In this context, training involves using these labels as objectives, which the models are programmed to identify and learn. The main advantage of fully supervised methods is that they tend to be very accurate, as they can leverage the available label information to guide the learning process effectively. Nevertheless, the primary problem of the existing fully supervised machine learning models is a strong dependence on the quality and the number of labels. This obstacle could lead to overfitting if the model is not properly regularized during the training. This is extremely vital when the segmentation model applies to new unseen data.

Since medicine is a strictly regulated field, accessing medical data can be challenging due to its sensitive nature. Additionally, acquiring high-quality labeled data is expensive and time-consuming because it requires manual work by experts. While large volumes of unlabeled data are comparatively easier to access, their usage is often limited due to the absence of comprehensive labels.

2.2 Semi-supervised learning

Semi-supervised learning has become one of the most efficient ways to improve the quality of medical image segmentation with a small amount of labeled data. By learning the structure and distribution of data, semi-supervised approaches enhance the precision of the model, acting as a link between supervised learning, which demands a lot of labeled data, and unsupervised learning, which operates solely on unlabeled data.

Modern semi-supervised learning approaches are built on the assumption that data follows a certain distribution, which can result in data points being closely positioned or forming distinct clusters. Based on this assumption, unlabeled data can be grouped with closely related labeled data and used for further studies [1, 6]. By associating unlabeled data with corresponding labeled clusters, semi-supervised learning allows for an expanded training dataset. Thus, leveraging this strategy can potentially enhance model accuracy and robustness.

Semi-supervised approaches often aim to leverage the underlying connections within data to enhance the accuracy of the model's predictions. By exploiting the structure in the unlabeled data, these models strive to learn a better representation of the underlying distribution, thereby enhancing their ability to generalize from the limited labeled data. This aligns with the Manifold hypothesis, which suggests that high-dimensional data, such as images, tends to exist on a lower-dimensional manifold within the overall space. The hypothesis essentially posits that real-world high-dimensional data, despite its complexity, often follows simpler structures when observed from a certain perspective. This implies that the crucial characteristics and relationships of the data can be effectively captured using lower-dimensional representation. By focusing on this lower-dimensional manifold, semi-supervised approaches can more effectively capture the latent patterns in the data.

Semi-supervised methods often utilize pseudo-labeling techniques to leverage value from unlabeled data. Pseudo-labeling involves assigning artificial labels to unlabeled data by using the model's predicted output as a form of weak supervision 2.1. While using pseudo-labels can help extract additional information from unlabeled data, it may also lead to the amplification of errors if the pseudo-labels are incorrect. Therefore, to prevent the potential amplification of errors, consistency regularization is commonly used in conjunction with model tuning.

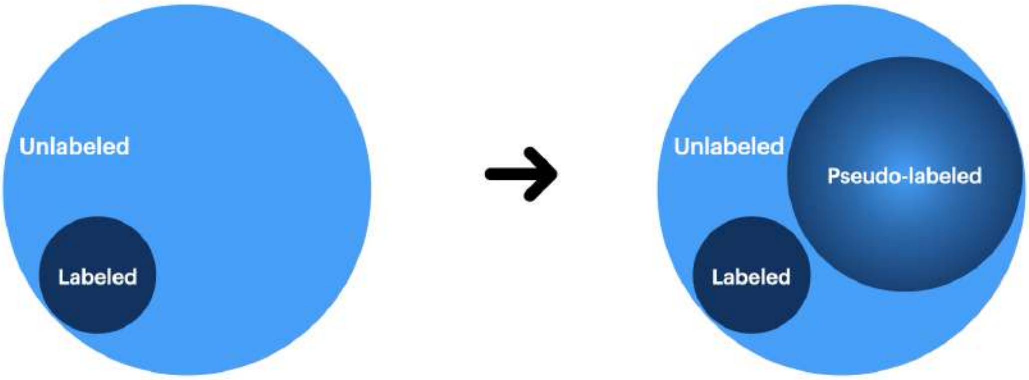


FIGURE 2.1: Unlabeled data utilization diagram. Aggregation of additional pseudo-labeled data from the unlabeled

Convolutional Neural Networks have become an efficient method for medical image segmentation [8, 17, 4]. They have been adopted for semi-supervised approaches with two techniques: self-training – iterative model retraining with the expansion of training data with their predictions, and co-training – separate data for two or more classifiers, each of which uses additional training data given by the other classifier [17, 1].

The Teacher-Student models framework is also often used in semi-supervised approaches. In this method, the Teacher model is first trained on the labeled data and then used to generate pseudo-labels for the unlabeled data. The Student model is then trained on the combined labeled and unlabeled data using the predictions from the Teacher model as soft targets. This collaborative dynamic allows the Student model to gain insights from the Teacher model's knowledge and apply it to its own learning process. Hence, even with limited access to labeled data, the Student model can enhance its learning capabilities. Through this strategy, the Teacher-Student model framework showcases a resourceful approach to making the most of available data. This approach was used by Zhen Zhao et al. to propose AugSeg – a two-branch Teacher-Student method that adopts different data augmentation techniques [16]. This results in a more accurate and robust model than models trained only on the limited labeled data.

There are different approaches to semi-supervised learning depending on the type of data being used. However, this work will not focus on those methods, but it is worth mentioning them to provide additional context on the topic:

- Graph-based methods leverage the data's graph structure to learn from labeled and unlabeled data. Individual nodes of the graph could represent labeled or unlabeled data, where labeled data is used to propagate labels across the graph. Graph-based methods follow the assumption that similar data points are likely to have similar labels. This approach was used by Anca Ciurte et al. for ultrasound data by making use of a graph of intensity patches as an image representation [6].

- Kernel-based methods can use a small amount of labeled data to identify discrete clusters that may contain unlabeled data. Chen Qin et al. showed that brain image segmentation data might not follow a specific distribution but tends to have consistent classifications [11]. In the Kernel-based approach, the kernel function transforms the input data into a higher-dimensional feature space, allowing the algorithm to learn a non-linear decision boundary and use the unlabeled data in the process.

2.3 Research Gap

Due to the limited availability of medical image data, augmented data is often used in medical image segmentation research. Researchers typically use methods such as random perturbations, including Gaussian noise and randomized transformations, to create more data and enhance model robustness. While these methods can generate more data effectively, they do not consistently enhance segmentation quality. This inconsistency arises because the model might learn from data that it is unlikely to encounter in real-world situations. Furthermore, these methods may not scale well due to the high diversity of medical images. To address these issues, researchers often use weak augmentations to generate new labeled data for better exploitation of labeled information. A recent study by Zhen Zhao et al. used adaptive label-injecting augmentation that can make full use of labeled data to aid the training on unlabeled samples, thereby improving image segmentation model performance [16].

Given that these semi-supervised approaches still require a certain amount of labeled data, determining the ideal size of the labeled dataset becomes crucial. In recent studies, researchers have conducted experiments to explore the varying proportions of unlabeled data in relation to labeled data [10]. Yuexiang Li et al. have shown that a semi-supervised approach can produce results that are close to a fully supervised approach, with less than 50% of labeled data, but cannot outperform it if all data is labeled [10]. This research emphasizes the adaptability of semi-supervised learning in managing diverse datasets effectively, given its less stringent dependency on labeled data. The importance of such scalability becomes more apparent when considering the resource constraints and the balance needed between the availability of labeled data and computational efficiency.

Therefore, semi-supervised learning offers a trade-off between the accuracy of supervised learning and the scalability of unsupervised learning, which derives its scalability from its ability to learn from vast amounts of unlabeled data. By effectively using more of the available data, semi-supervised methods can gain a broader understanding of the underlying structures and patterns in the data. This, in turn, can lead to more robust and generalizable models, expanding the potential and capabilities of semi-supervised learning techniques in various domains.

While the semi-supervised approach might be crucial for improving CT scan segmentation, it is still unclear how it can make use of additional real-world data, which can have imperfections and distortions. Given that the impact of incorporating additional real-world data into the semi-supervised training process has not been fully investigated, it may be beneficial to examine its effect on the model's performance. This is crucial as this data could contain different disease forms, imperfections, and distortions. Moreover, utilizing semi-supervised techniques in this manner could open up additional opportunities for the application of unlabeled data. This suggests the potential for a second use case where unlabeled data is effectively employed.

Chapter 3

Proposed Approach

In this section, we outline the origin of the data and the methods employed for processing it, as well as the training of various models. We then present the suggested schema for generating and handling pseudo-labels. In conclusion, we discuss the evaluation metrics employed to assess the outcomes of our experiments.

3.1 Datasets Description

Our first objective is to select appropriate datasets for our experiments. Our main focus will be on medical images that incorporate kidneys and tumors 3.1 because this type of data has a large dataset with high-quality labels. Additionally, we can adapt existing datasets that usually are not used for kidney segmentation but may contain them. The datasets used in this study are publicly available for non-commercial use, which facilitates progress and innovation in the field of medical image segmentation.

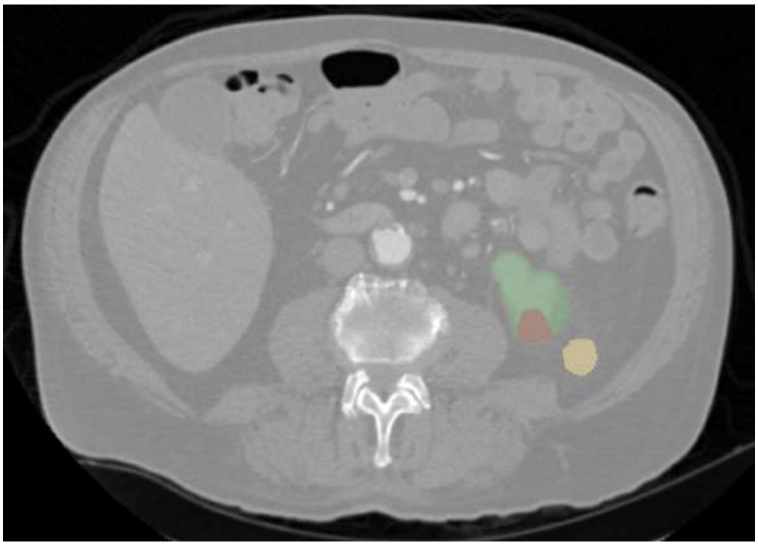


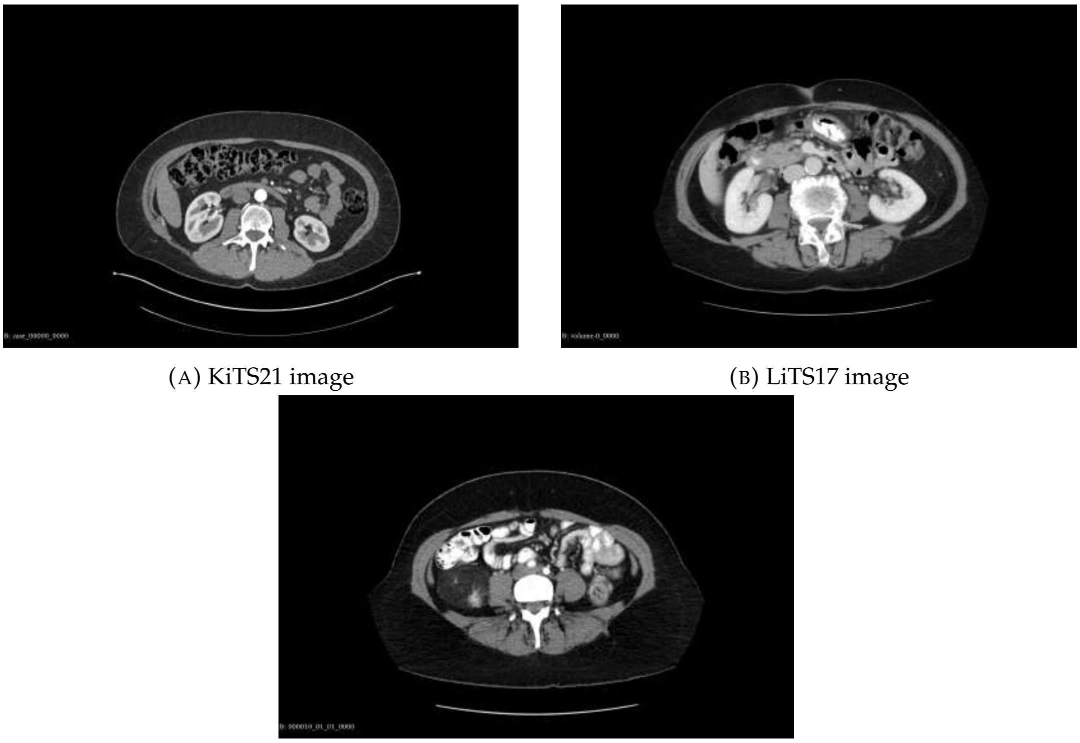
FIGURE 3.1: CT scan segmentation. Green – kidney, red – cyst, yellow – tumor

The following is the list of datasets used in our work:

- **KiTS21** This dataset contains 300 CT scans where kidney, tumor, and cyst were manually segmented. The KiTS21 patient population group contains individuals who received partial or complete kidney removal due to a suspected renal cancer diagnosis from 2010 to 2020 at either a medical facility of M Health Fairview or Cleveland Clinic [7]. Given the large number of images and the high standard of annotations in this dataset, it will serve as an excellent base for the initial training of the model and provide a reliable benchmark for subsequent comparisons. In addition, this dataset boasts an extra set of 100 test cases specifically reserved for performance evaluation purposes. These test cases are intentionally excluded from the training process, ensuring that the models are evaluated on unseen data. By utilizing these test cases, we can accurately measure the performance of the models.

- **LiTS17** This dataset was originally designed for liver segmentation training tasks. The image data for the LiTS challenge is collected from seven clinical sites worldwide. This dataset contains 131 CT scans. The studied cohort covers diverse types of liver tumor diseases, including primary tumor disease and secondary liver tumors [3]. Because this dataset contains a substantial number of image slices within images, they have the potential to include kidney structures, making the dataset a suitable choice for training models focused on kidney segmentation. Furthermore, some cases from this dataset contain tumors in the kidneys.

- **DeepLesion** This dataset contains over 10000 studies from 4,427 unique patients. These images have been collected in the institute's Picture Archiving and Communication Systems (PACS) over close to two decades [15]. However, not all studies include slices with kidneys. If we pick appropriate cases, we will get only 344 cases. DeepLesion also does not include segmentation labels for organs. However, it contains key slices with bounding boxes. Which may help measure the correctness of generated pseudo-labels for tumors.



(A) KiTS21 image

(B) LiTS17 image

(C) DeepLesion image

FIGURE 3.2: Example of images from the KiTS21, LiTS17, and DeepLesion dataset.

3.2 Training Framework

We propose to use the semi-supervised approach to training segmentation models that utilizes a fully-supervised model, also known as the Teacher model, to create pseudo-labels for public datasets. This process of creating pseudo-labels allows for the use of unlabelled data in the model training by combining the pseudo-labels with the unlabelled data into a more extensive and diverse dataset. The new augmented dataset, which includes labeled and unlabelled data with pseudo-labels, is then used to train the Student model. The Student model learns from the Teacher model by incorporating the information contained in the pseudo-labels, thus potentially improving performance. This approach allows the Student model to benefit from the knowledge and expertise of the Teacher model while also leveraging the information contained in the unlabelled data.

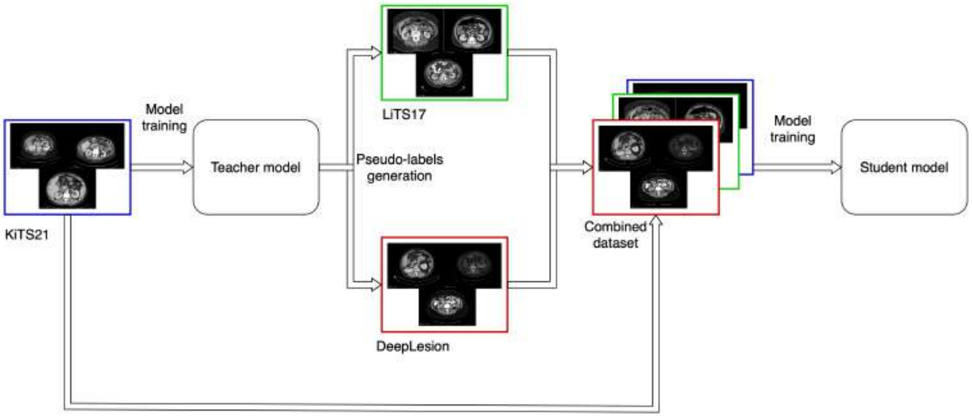


FIGURE 3.3: Illustration of proposed training framework. Started from the training of the Teacher model with the KiTS21 dataset, then generating pseudo-labels for LiTS17 and DeepLesion dataset, followed by merging datasets and training the Student model.

To prove this hypothesis, we need to measure the performance of the Student model by comparing it to the Teacher. We used the Dice score, which is commonly used in medical image segmentation tasks, as the evaluation metric [8, 1, 17, 9, 2]. It measures the similarity between the predicted segmentation mask and the target (ground-truth) segmentation mask by computing the ratio of the area of their overlap to the total area.

$$Dice = \frac{2|target \cap prediction|}{|target| + |prediction|} \tag{3.1}$$

A Dice score of 1 indicates perfect overlap between the predicted and ground-truth segmentation masks, while a 0 indicates no overlap. By employing two distinct types of Dice score, namely pixel Dice for individual scan slices and voxel Surface Dice for the segmented surface, we can acquire a more comprehensive and intricate understanding of the segmentations generated, thereby capturing a richer set of information.

$$SurfaceDice = \frac{2|\partial target \cap \partial prediction|}{|\partial target| + |\partial prediction|} \tag{3.2}$$

where $\partial target$ and $\partial prediction$ represent the surfaces of the structures of target and prediction surfaces.

Our primary goal is to investigate the effect of incorporating new data pseudo-labeled by the existing model into the training process. However, due to the different sources of data, the resulting dataset might exhibit an imbalance in segmentation types and image quality. Due to this disparity in data, our focus would not be on striking a balance between labeled and unlabeled data to reduce the quantity of labeled data in the training process. Rather, we will direct our efforts toward assessing the effectiveness of semi-supervised and fully supervised methodologies.

We used the KiTS21 to train the Teacher model. With this model, we can generate pseudo-labels for LiTS17 and DeepLesion datasets. At this point, segmentation errors are irrelevant since measuring them without ground truth labels is impossible. However, if the model has low accuracy, errors in generated pseudo-labels would affect the Student model outcomes.

3.3 Model Architecture

Generating pseudo-labels is a crucial part of our objective. We decided to use the nnU-Net framework for this goal. nnU-Net is a semantic segmentation method that automatically adapts to a given dataset, it analyzes the provided training cases and automatically configures a matching U-Net-based segmentation pipeline [8]. nnU-Net is designed to perform semantic segmentation tasks and can effectively process 3D images, accommodating varying input modalities and channels. It can understand the voxel spacings and anisotropies of the images, even for imbalanced classes. Since nnU-Net relies on supervised learning, it is perfect for Teacher model training and adjusting the Student model for further comparison. The nnU-Net framework is structured to train a collection of five models (folds) utilizing distinct data divisions for training and validation data. This method, referred to as Cross-Validation, results in enhanced performance due to the disagreements among folds. To guarantee optimal results, it is crucial to assess the best Teacher model's performance and employ all five folds of the model in a unified ensemble. This ensures sufficient confidence in generating pseudo-labels. The folds architecture visualization is provided in Appendix A. The performance of the nnUNet model is progressively enhanced through the iterative application of Dice and cross-entropy loss functions. The loss functions quantify how well the prediction of a model aligns with the actual data, thereby optimizing its overall model's performance. The combined use of both loss functions allows for a more comprehensive and effective optimization strategy. For segmentation tasks on 3D CT scans, the 3D variant of nnU-Net uses 3D convolutional layers to process the volumetric data. The 3D nnU-Net architecture consists of the following components:

- Contracting path (3D encoder): This part of the network captures the context in the input 3D CT scans. It consists of a series of 3D convolutional layers, followed by batch normalization and ReLU activation functions. After each convolutional block, there is a 3D max-pooling operation that reduces the spatial dimensions and increases the number of feature channels.

- Expanding path (3D decoder): This part of the network focuses on precise segmentation, and it aims to recover the spatial information lost during the encoding process. It consists of a series of 3D up-convolutional (transpose convolutional) layers, followed by batch normalization and ReLU activation functions. After each up-convolutional block, there is a concatenation operation with the corresponding feature maps from the contracting path to provide high-resolution features.

- Skip connections: These are the connections between the contracting and expanding paths, which help retain high-resolution information and improve localization accuracy. These connections are also implemented using 3D operations to maintain consistency with the volumetric data.
- Final layer: The final layer of the network is a 1x1x1 3D convolution followed by a softmax activation function to produce the final segmentation map with the same size as the input 3D CT scan.
- Multi-class segmentation: To segment kidneys, cysts, and tumors simultaneously, the model's output layer will produce multiple segmentation maps, one for each class (i.e., kidney, cyst, and tumor). This can be achieved by having multiple channels in the output layer corresponding to the number of classes.

Chapter 4

Experiments and results

In this chapter, we will discuss the experiments conducted and how their results were interpreted, beginning with the naive training approach and then moving on to generating pseudo-labels. Our training strategies, implemented using multiple datasets, will be outlined following the approach described in Chapter 3. In the following segment, we will reveal the outcomes of the suggested method. We will begin by discussing the creation of pseudo-labels, followed by a more detailed examination of the impact of varying data sources. We have categorized the segmentation outcomes into three groups: kidneys, mass (encompassing cysts and tumors), and tumors. Our aim in doing so is to underscore the unique advantages and disadvantages of each model.

4.1 Training configurations

To explore the impact of different training approaches on our model's performance, we divided our experiments into several configurations, each with a unique combination of data and training strategies. By testing multiple configurations, we hope to gain a deeper understanding of our approach results with different conditions and identify the most effective settings for our task.

The following is the list of configurations used in our work:

- **Naive training.** Naive training is the first configuration, mainly used to train the Teacher model and generate pseudo-labels for unlabeled data. The training process begins with data preprocessing, where the CT scans from the KiTS21 dataset are normalized to a consistent intensity range and augmented with rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, and other augmentation technics to increase the variability of the training data and improve model generalization. The training process involves feeding the input CT scans into the model, which learns to extract relevant features automatically and generate segmentation masks that delineate the kidney organs, cysts, and tumors. The model's performance is iteratively optimized by Dice and the cross-entropy loss function [8].

- **Experiment with the LiTS17 dataset.** To preprocess the LITS17 dataset for subsequent analysis, we utilize a pre-trained with the KiTS21 dataset Teacher model to generate pseudo-labels for 131 cases within the dataset. These pseudo-labels specifically identify kidneys, cysts, and tumors in the images. Subsequently, we merge the original images from the LITS17 dataset with the generated pseudo-labels and commence the training process with this combined dataset which includes a fusion of the KiTS21 dataset.

- **Experiment with the DeepLesion dataset.** Similar to the LITS17 dataset, we gather pseudo-labels for 344 cases from the DeepLesion dataset. The training process of feeding the model with data remains the same, with all parameters kept constant in order to measure the influence of different data added to model training. Since DeepLesion and LiTS contain different cases, we can also understand which data is most beneficial for semi-supervised learning by comparing the models' Dice scores.

- **Experiment with a combined dataset.** This experiment involves merging the KiTS21, LITS17, and DeepLesion datasets. By integrating data from various datasets, the model can potentially benefit from a more extensive range of anatomical variations of organs, tumor patterns, and imaging characteristics, resulting in improved segmentation accuracy and generalization capability.

- **Transfer learning experiment.** The main goal of this experiment is to provide an additional baseline to compare our proposed approach. This experiment employs a simple transfer learning approach of fine-tuning existing models from one domain with data from another. As a starting point, we chose a model trained on the original LITS17 dataset with two segmentation classes: liver and lesion. The aim of this experiment is to leverage the knowledge gained from the original LITS17 dataset by fine-tuning an existing model with the weights of this model on the KiTS21 dataset. This could potentially improve the performance of the nnU-Net model on the KiTS21 dataset by transferring the learned features and representations from the LITS17 dataset.

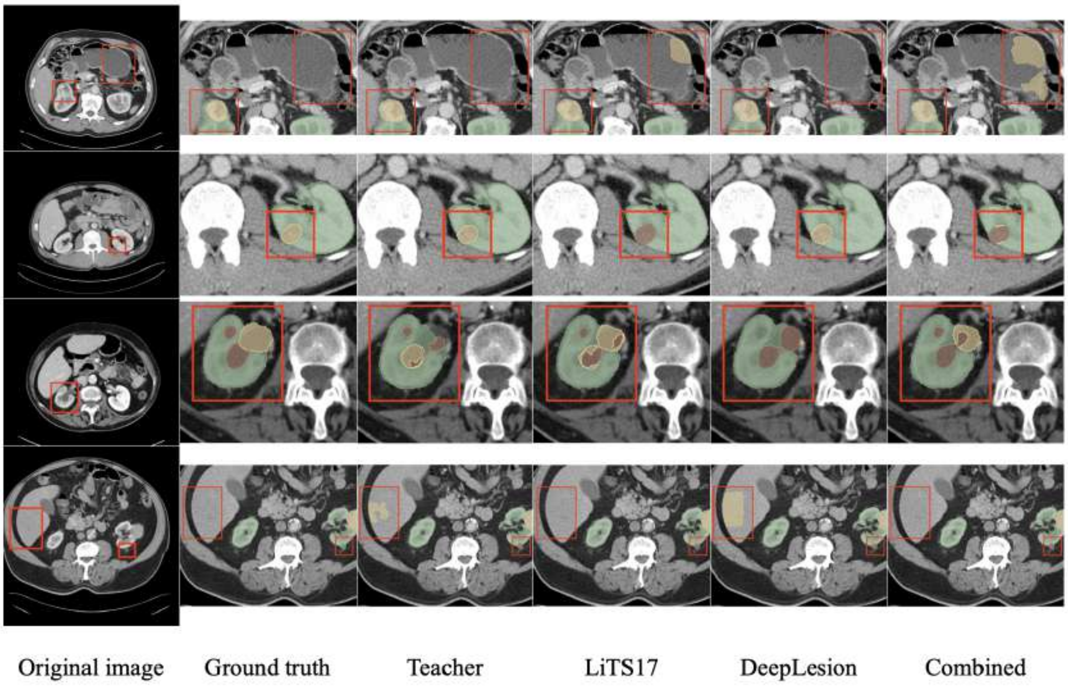


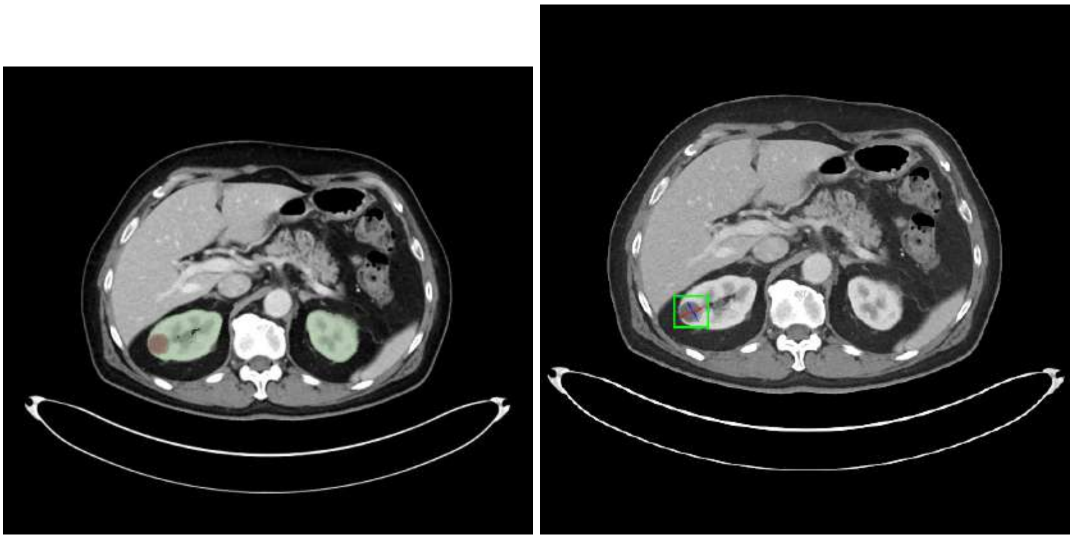
FIGURE 4.1: Contrasting segmentations of kidneys (green), cysts (red), and tumors (yellow) produced by different models - the Teacher model, LiTS17, DeepLesion, and Combined Student models - on the training dataset while also including the Ground Truth and Original image for reference.

Once the segmentations for the training set are obtained, it is evident that Student models with various data combinations yield different outcomes 4.1. In some instances, the segmentations more closely resemble the ground truth labels. However, there are cases where the Student models produce inferior results. The primary distinction typically lies in cyst and tumor segmentations, while kidney segmentations tend to be more similar.

4.2 Naive Training

Once the initial phase of training the Teacher model with the KiTS21 dataset is completed, we then generate pseudo-labels. With this completed, we are now ready to review, scrutinize, and assess the results. The performance and results of the Teacher model will serve as a reference point, establishing a baseline against which we can compare and measure the effectiveness of subsequent Student models and the entire pseudo-labeling approach overall.

To assess the accuracy of generated pseudo-labels on the DeepLesion dataset with the Teacher model, we took the provided bounding boxes of tumors on the key slices of CT scans and compared them to the bounding boxes of our pseudo-labels, measuring the level of agreement between them 4.2. A main challenge in this evaluation is aligning the key slice index with the 3D pseudo-label. This issue arises due to the fact that certain instances within the DeepLesion dataset lack multiple slices within the scan. Our qualitative examination of the pseudo-labels produced for the DeepLesion dataset indicates that they commonly align with bounding boxes. However, several instances within the dataset were not labeled. This indicates that the Teacher model cannot handle this data.



(A) Teacher model prediction

(B) Key slice

FIGURE 4.2: Teacher model result comparison with Key slice.

4.3 Experiments results

In order to accurately assess the performance of the trained models, we will utilize the KiTS21 test set, comprising 100 CT scans that were not part of the training process. Each kidney/cyst/tumor instance in the test set has been labeled multiple times by different labels. These multiple labels were used to generate plausible complete labels for each patient [7]. This enables us to evaluate the effectiveness of the ensemble of folds on previously unencountered data by implying two approaches:

- **Sampled.** In this approach, the evaluation is conducted using plausible complete labels for each patient, generated by sampling from the multiple labels provided by different annotators.

- **Majority vote.** This evaluation method involves computing an aggregated majority voting segmentation from the multiple labels provided by different annotators.

Metric	Sampled Dice	Majority vote Dice	Sampled surface Dice	Majority vote Surface Dice
Combined	0.8753	0.8780	0.7864	0.7942
DeepLesion	0.8747	0.8773	0.7878	0.7958
LiTS17	0.8867	0.8894	0.7982	0.8063
Transfer learning	0.8858	0.8887	**0.8039**	**0.8125**
Teacher	**0.8873**	**0.8902**	0.8035	0.8117

TABLE 4.1: Average Dice and Surface Dice scores on the KiTS21 Test dataset. The Teacher model yields the best results for the Dice scores, while the Transfer Learning experiment excels in the Surface Dice scores. The top-performing results are accentuated with bold text, and those that ranked second-best are identified by underlining.)

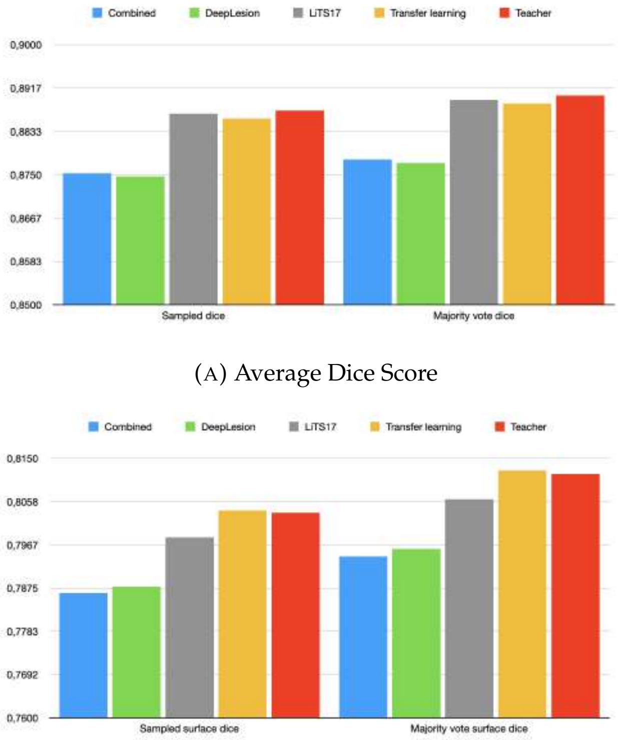
Based on 4.1, it is evident that the LiTS17 dataset experiment yields superior average scores in comparison to the DeepLesion and Combined datasets. Additionally, the Combined dataset experiment demonstrates marginally improved outcomes relative to the DeepLesion dataset experiment. However, both significantly lag behind the LiTS17 experiment. This proves that the quality of data selected for generating pseudo-labels is highly important for model training. By examining the average Dice score, we can observe that the LiTS17 experiment outperforms Transfer Learning in both Sampled and Majority vote Dice categories.

Nevertheless, the Teacher model exhibits the highest average Dice score, suggesting that the additional data acquired during the training process in semi-supervised experiments did not necessarily benefit the model 4.3a and, as shown DeepLesion experiment might harm the model's performance.

Analyzing the Surface Dice, it is apparent that the Teacher model outperforms the Student models, while the Transfer Learning experiment achieves superior Surface Dice results 4.3b. Our results show that the semi-supervised method has a higher degree of overlap between the predicted segmentation and the ground truth segmentation, but it performs less well in capturing the surface alignment. Thus, the model from the LiTS17 experiment performs well in terms of overall spatial overlap but may struggle to accurately capture the boundaries of the segmented regions.

The findings highlight that merely employing pseudo-labeling for data does not guarantee the enhanced performance of the trained model. The success of this method depends on the quality of the data rather than the size of the dataset.

Exploring the results more thoroughly, our aim is to assess the outcomes for each separate segmentation category. The specific attention to each category allows us to identify specific strengths and weaknesses within each category, providing a more



(A) Average Dice Score

(B) Average Surface Dice Score

FIGURE 4.3: Average Dice and Surface Dice scores on the KiTS21 Test dataset for Teacher and Student models with different data combinations. Y-axis for the Dice score 4.3a starts from 0.85, and for the Surface Dice 4.3b from 0.76. In terms of average Dice scores, the Teacher model outperforms all others, followed closely by the LiTS17 Student model. However, when considering average Surface Dice scores, the Transfer Learning model is superior, with the Teacher model securing the second-best position.

nuanced understanding of the data. By focusing on each segmentation group individually, we note that the LiTS17 experiment attains the top scores for Sampled and Majority vote Dice for the kidney 4.2. However, the Teacher model retains its superiority in terms of Surface Dice for the same organ.

Metric	Sampled kidney Dice	Majority vote kidney Dice	Sampled kidney surface Dice	Majority vote kidney surface Dice
Combined	0.9714	0.9747	0.9455	0.9580
DeepLesion	0.9711	0.9744	0.9470	0.9595
LiTS17	**0.9742**	**0.9775**	0.9490	0.9615
Transfer learning	0.9719	0.9753	<u>0.9507</u>	<u>0.9635</u>
Teacher	<u>0.9734</u>	<u>0.9768</u>	**0.9518**	**0.9644**

TABLE 4.2: Kidney segmentation scores on the KiTS21 Test dataset. The LiTS17 Student model has the best Dice scores, while the Teacher model yields the best Surface Dice scores. The top-performing results are accentuated with bold text, and those that ranked second-best are identified by underlining.

Examining the mass (cyst and tumors) segmentation outcomes 4.3, the Teacher model continues to display the most favorable results, suggesting that the combination of datasets in semi-supervised experiments lacked sufficient cases involving cysts. Nonetheless, the findings show that the LiTS17 experiment outperforms the DeepLesion in terms of results.

Metric	Sampled mass Dice	Majority vote mass Dice	Sampled mass surface Dice	Majority vote mass surface Dice
Combined	0.8573	0.8597	0.7354	0.7410
DeepLesion	0.8559	0.8582	0.7356	0.7416
LiTS17	0.8647	0.8671	0.7439	0.7499
Transfer learning	0.8673	0.8700	0.7540	**0.7606**
Teacher	**0.8691**	**0.8718**	**0.7542**	0.7605

TABLE 4.3: Mass segmentation scores on the KiTS21 Test dataset. The Teacher model yields the best Dice scores for mass segmentations, while the Transfer Learning experiment slightly outperforms the others in the Majority Vote Surface Dice score. The top-performing results are accentuated with bold text, and those that ranked second-best are identified by underlining.

Given that our main objective is to investigate the influence of integrating pseudo-labels into the training procedure on tumor segmentations, it becomes essential to individually examine the results of tumor segmentations. A review of the tumor segmentation scores 4.4 reveals that the LiTS17 experiment slightly outperforms the Teacher model in terms of Dice scores. However, this improvement is heavily reliant on the quality of the data, and the combination of datasets with pseudo-labels should be balanced to include various segmentation types. Under these conditions, Student models may yield better Dice scores, but they might still face challenges in accurately capturing the context of the surface.

Metric	Sampled tumor Dice	Majority vote tumor Dice	Sampled tumor surface Dice	Majority vote tumor surface Dice
Combined	0.7972	0.7995	0.6783	0.6835
DeepLesion	0.7970	0.7992	0.6808	0.6863
LiTS17	**0.8212**	**0.8236**	0.7018	0.7076
Transfer learning	0.8183	0.8209	**0.7071**	**0.7133**
Teacher	0.8195	0.8221	0.7044	0.7103

TABLE 4.4: Tumor segmentation scores on the KiTS21 Test dataset. The LiTS17 Student model stands out with the highest Dice scores, whereas the Transfer Learning experiment excels in Surface Dice. Coming in second in all categories, the Teacher model also demonstrates commendable performance for tumor segmentations. The top-performing results are accentuated with bold text, and those that ranked second-best are identified by underlining.

4.4 Consistency experiments

The verification of our derived results is an essential part of our study. In this segment, our goal is to carry out further experiments to affirm the robustness of the solution we propose. By incorporating diverse training data in these experiments, we seek to verify the outcomes we attained in the preceding section. Moreover, these experiments may offer a significant understanding of the process of pseudo-labeling and model training.

4.4.1 Custom split experiments

Due to data limitation constraints, testing our method on various datasets is not possible. However, we have the option to form a custom split of the KiTS21 dataset and repeat the training procedures using it. A custom split, in this context, means dividing the KiTS21 dataset into distinct subsets specifically tailored for training and testing our model. This type of division allows us to control and manipulate the data exposure during the model training, thereby providing us with unique insights about its behavior. For this purpose, we have crafted a random split where 250 instances of the dataset are used for training and the remaining 50 for testing. Using this custom training dataset, we then proceed with the Teacher model training and generate pseudo-labels for the LiTS17 dataset. We decided to exclude the DeepLesion dataset from this experiment, as trials involving it produced inferior results compared to those using the LiTS17 dataset. The latter demonstrated superior performance in our other semi-supervised experiments. Following the integration of pseudo-labels with the training set, we proceed with training a new Student model.

To understand the impact of pseudo-labels on model training, we opted to carry out an additional transfer learning experiment. In this particular experiment, apart from using pseudo-labeled data, the Student model also inherited the weights from the Teacher model, serving as a form of weak supervision. The primary objective of this experiment is to determine whether the incorporation of the Teacher model's weights into the Student's training process would be beneficial.

In order to assess the outcome of these experiments, we utilize 50 testing images and generate predictions using both the Teacher and Students models. By contrasting these predictions with the actual ground truth labels, we obtain the Dice scores, which are detailed in 4.5.

Metric	Kidney Dice	Mass Dice	Tumor Dice	Kidney Surface Dice	Mass Surface Dice	Tumor Surface Dice
Teacher	**0.9543**	**0.8237**	**0.8357**	**0.9091**	**0.7124**	0.7165
LiTS17	0.9524	0.8136	0.8262	0.9064	0.6997	0.7052
LiTS17 + Transfer learning	0.9514	0.8181	0.8350	0.9062	0.7071	**0.7180**

TABLE 4.5: Average Dice scores for custom split experiments. The Teacher model surpasses others in nearly all categories, with the singular exception of Tumor Surface Dice. In this category, the LiTS17 combined with the Transfer Learning experiment demonstrates superior performance. The top-performing results are accentuated with bold text, and those that ranked second-best are identified by underlining.

The outcomes of these experiments indicate that regardless of the split variation, our proposed method maintains a consistent performance. Upon examining the results from the Transfer learning experiment, it is evident that the differences are not significant. However, the Surface Dice scores suggest that this model training approach may be more efficient in understanding the context of the surface. A visual assessment of the model's segmentations leads to a comparable conclusion. In the majority of cases, segmentations are alike, with minor variations observed in some images.

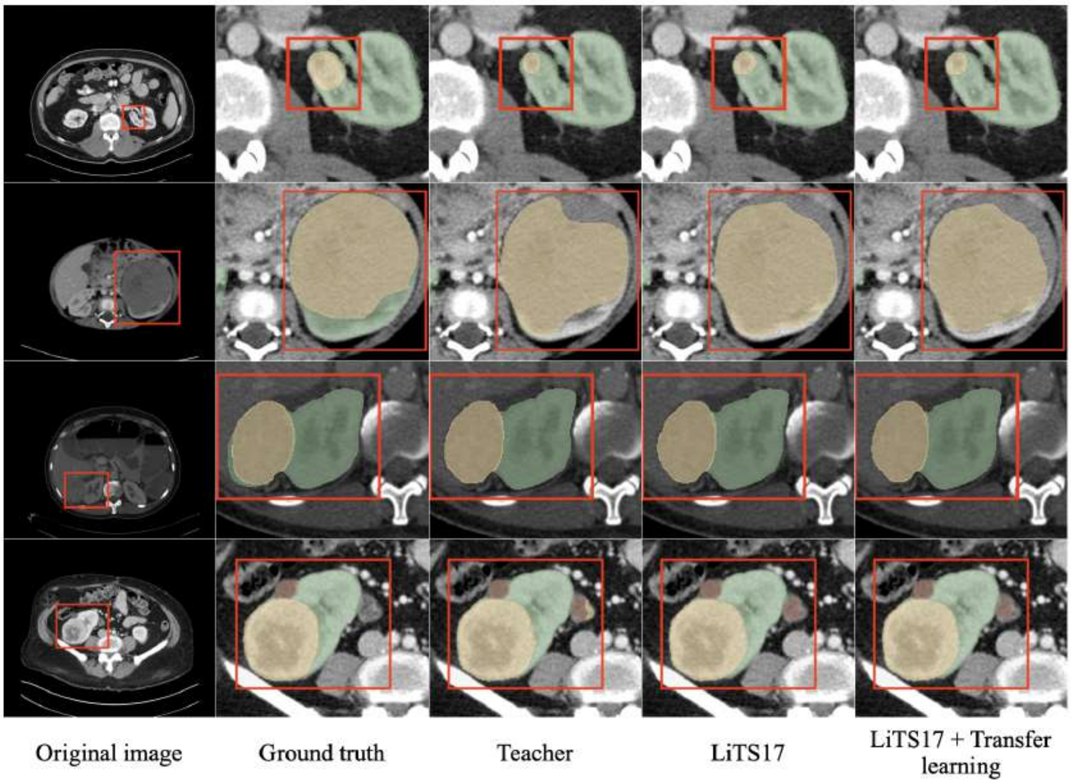


FIGURE 4.4: Contrasting segmentations of kidneys (green), cysts (red), and tumors (yellow) produced by different models - the Teacher model, LiTS17, and LiTS17 + Transfer learning Student models - on the custom split testing dataset while also including the Ground Truth and Original image for reference.

4.4.2 Diverse organs analysis

An alternative method to evaluate our approach involves choosing a different organ type for model training. These experiments primarily aim to assess the scalability of our method across various types of segmentations. Given that the LiTS17 dataset has already been utilized in prior experiments, we opted to conduct an additional experiment focusing on liver organ segmentations. In this experiment, we train the Teacher model using the original LiTS17 ground truth labels and subsequently generate pseudo-labels for the KiTS21 dataset. Using this combined dataset, we then proceed to train the Student model. These models have two types of segmentation: liver and lesion, where lesion refers to an abnormal change in tissue or organ structure.

The obtained results reveal that the Student model underperforms compared to the Teacher model when it comes to lesion segmentations 4.6. A plausible explanation for this could be that the KiTS21 dataset was primarily created for kidney segmentations and may not encompass instances with liver lesions. Additionally, given the extensive size of the liver image, the overall efficiency of these models may be compromised due to the restricted field of view of the architecture, which in turn may not gather ample contextual information.

Metric	Liver Dice	Lesion Dice
Teacher	**0.9742**	**0.7720**
KiTS21	0.9708	0.7053

TABLE 4.6: Liver experiment average results on the training set. The Teacher model yields the top performance for both categories.

Visual comparison of the segmentations on the training dataset also confirms that in the case of liver segmentations, the Teacher model outperforms the Student 4.5. Comparing the Dice score for each individual medical image, there were almost no improvement from the Student model. This points out that further research on the different organ types with different dataset combinations is needed.

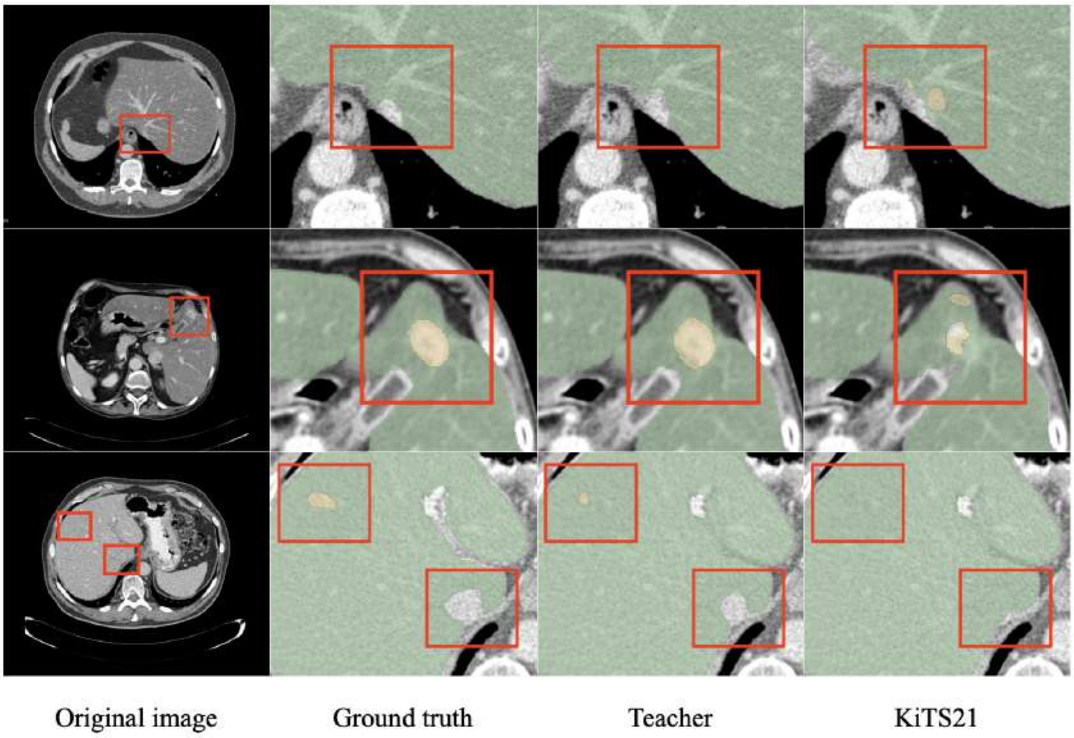


FIGURE 4.5: Contrasting segmentations of the liver (green) and lesion (yellow), produced by the Teacher model and KiTS21 Student model - on the LiTS17 dataset while also including the Ground Truth and Original image for reference.

Chapter 5

Discussion

In reference to the research questions articulated in Section 1, we carried out a series of experiments to comprehend the impact of using unlabeled data on the training process. While our principal emphasis was on enhancing the quality of tumor segmentations, we also assessed the models' performance for kidney and mass segmentations. This enabled us to conduct a comprehensive evaluation of our findings. Additionally, we collected data on how our proposed method interacts with various organ segmentations, thereby offering more insight into its scalability.

5.1 Findings analysis

Upon more in-depth analysis of the obtained results, it becomes apparent that introducing additional pseudo-labeled data to the training process yields varying outcomes. Given that the pseudo-labels were produced by the same Teacher model, the only variation lay in the data applied. Consequently, the different outcomes could be attributed to two factors:

- The Teacher model might not be capable of handling the data for generating pseudo-labels. This issue could stem from disparities with the training data, suggesting that the Teacher model was overfilled with the KiTS21 dataset and thus faced difficulties with segmentation when dealing with data from another source. As a result, any errors generated during the pseudo-labels creation were magnified during the Student model's training, leading to a drop in performance.
- The datasets may contain images of widely varying quality. Images from the KiTS21, LiTS17, and DeepLesion datasets were collected for differing purposes, utilizing different methodologies and sourced from different origins, which may lead to notable differences among them. These differences could be reflected in aspects such as resolution, contrast level, number of slices, and others. Nevertheless, a more in-depth exploration into how these image characteristics affect model performance warrants further investigation.

The disparity in the Dice scores across different segmentation categories reveals an issue related to the imbalance in the datasets used for training the Student models. A detailed examination of kidney, mass, and tumor segmentations shows that the Dice scores for mass segmentations experience the most significant drop. This issue primarily stems from a low prevalence of cases featuring cyst segmentations in the scans sourced from the LiTS17 and DeepLesion datasets. Therefore, each fold from the model ensemble receives less data with cyst segmentation during the training process. This scarcity of specific data leads to a lack of variety in the training samples, inhibiting the model's ability to learn effectively. Future research efforts might need to consider strategies to handle such data imbalances better. This could involve sourcing more diverse datasets or implementing techniques to synthetically balance the data distribution in existing datasets.

After examining the additional experiments aimed at validating the results of the suggested approach, it becomes clear that the Teacher-Student framework produces consistent outcomes under various conditions. This indicates the reliability of this approach. However, the experiments involving liver segmentation once again underscored the significance of the used data. Given that the KiTS21 dataset wasn't intended for liver segmentations, its inclusion in the training process, along with the generated pseudo-labels, doesn't enhance the performance of the model. This highlights the crucial role that appropriately tailored datasets play in ensuring optimal model performance. In future research, it might be beneficial to adapt the framework to handle less-than-ideal dataset conditions to ensure the efficiency of this approach across different segmentation tasks.

5.2 Limitations

While this study provides valuable insights into the utilization of pseudo-labeled data in the training process, it's important to acknowledge certain limitations that may have influenced the findings. These limitations do not invalidate the results but rather provide a context within which the findings should be interpreted.

Firstly, in reference to the reported imbalance in the final datasets, it's important to acknowledge our lack of extensive expertise in the medical field, as we are not healthcare professionals. As a result, we are not in a position to inject our personal biases into the data or methodically arrange the pseudo-labeled data to pinpoint the appropriate cases within the dataset.

Secondly, given the absence of ground truth labels in the pseudo-labeled data, we were unable to gauge the quality of the generated labels. As a result, any errors that arose during the pseudo-labeling process were subsequently ingrained into the training of the Student model.

Lastly, one of the limitations we faced pertained to the time constraints for the experiments. Given that the nnU-Net model operates as an ensemble consisting of five folds, the training process for each fold is time-intensive due to our work with 3D medical images. This resulted in significant time investment for the full training process, which inherently restricted the number of experiments we could conduct within our project timeline.

5.3 Future research

Our acquired results indicate that there is potential for additional research in this field. The starting point could be more experiments aimed at comparing our proposed approach. Initiating additional experiments involving various types of segmentations, including lungs, breast, prostate, and more, could be a valuable next step. These investigations can bring insights into the scalability of our approach and assist in determining its suitable applications. Moreover, exploring these different types of segmentations could provide a more comprehensive understanding of the pseudo-labeling approach across varied medical imaging contexts. It may also offer insights into the potential modifications or adjustments needed to enhance the model's performance for each specific segmentation task.

Another possible avenue could be the generation of CT scans using generative networks such as Vox2Vox, which could supply supplementary training data [5]. With this synthesized data, we can conduct experiments to compare the model's performance against its performance on real-world data. Furthermore, exploring how the model performs with synthetic versus real data could offer interesting insights into its robustness and adaptability. Moreover, it might also provide a way to augment dataset and alleviate some of the issues related to data imbalance or scarcity.

Another strategy to enhance the usage of pseudo-labels involves incorporating an additional preprocessing step to grade the pseudo-labels based on their quality. This method, designed for 2D images, was employed by Yuchao Wang et al. as a means to ascertain the reliability of generated labels and prioritize the trustworthy labels during training [13]. Adding a similar layer of evaluation could significantly improve the reliability of the medical image segmentation model by ensuring that it is trained primarily on high-quality, dependable pseudo-labels. Additionally, this could help reduce error propagation from the Teacher to the Student model, leading to better overall performance.

Chapter 6

Conclusions

To summarize, our thesis focused primarily on incorporating pseudo-labeled data sourced from public datasets into the model training process. This approach has provided us with substantial insights and understandings. We can group these insights and outcomes into a series of key takeaways that reflect the core findings of our research.

- **Pseudo-labels are readily adopted into the model training process.**

  Incorporating pseudo-labeled real-world data into the training process of the segmentation model is a straightforward process. Nonetheless, this method does not offer effective control over the segmentation distribution, which could potentially lead to imbalances in the final dataset.

- **Pseudo-labeled data does improve the segmentation model's performance, but the effect is not significant.**

  The enhancements in segmentations, as quantified by the Dice scores, fluctuate based on the distinct pseudo-labeled data combinations used. These findings also indicate that models trained with pseudo-labeled data may be less effective at capturing surface details. However, minor advancements in tumor segmentation were observed in one of the experiments.

- **Quality over quantity.**

  The quality of the data combined with pseudo-labels carries more weight than the sheer quantity of cases. Our exploratory efforts using varied data combinations revealed that datasets boasting a higher image quality are of greater significance. The detailed information in each image slice, the clarity provided by high resolution, and the distinctiveness afforded by strong contrast play a crucial role in the effectiveness of the training process. This reinforces the importance of quality over quantity in the context of image-based data for model training.

- **The weak supervision derived from the Teacher model's weights outperforms the sole addition of pseudo-labeled data.**

  Our consistency experiments demonstrated that when the Student model integrates the weights of the Teacher model, it provides an additional boost to the model's performance as opposed to simply expanding the data through pseudo-labeling.

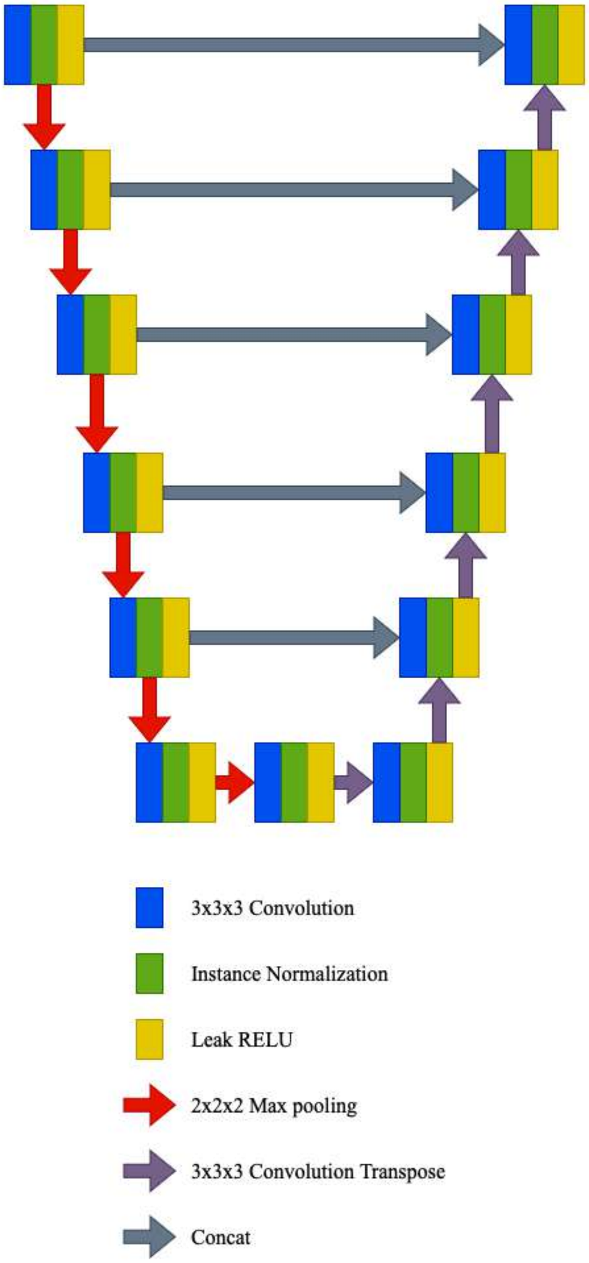
- **Maintaining the balance in the training dataset is crucial for various types of segmentation.**

  Our experiments show that even with high-quality images, the Student model's performance may be harmed for some categories. A balanced dataset could be a soulution that ensures that the model is well-equipped to identify and learn different categories of the data. It helps prevent bias towards any particular type of segmentation, thereby improving the overall performance and accuracy of the model in diverse scenarios. Given that combining data with pseudo-labeled datasets can potentially introduce imbalances, it becomes crucial to apply preprocessing measures to the images to guarantee the robustness of the model.

While the improvement in the model's performance may seem modest, we still deem the use of unlabeled data through the generation of pseudo-labels in the training process as effective. Nevertheless, there is room for further exploration of pseudo-labeling. Its potential applications in the field could be vast, from improving model robustness to possibly enabling models to understand complex patterns in data, presenting an interesting path for future research.

Appendix A

nnU-Net Model’s Fold Architecture



Bibliography

[1] W. Bai et al. “Semi-supervised learning for network-based cardiac mr image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*. 2017, pp. 253–260.

[2] Hritam Basak et al. “An Embarrassingly Simple Consistency Regularization Method for Semi-Supervised Medical Image Segmentation”. In: (2022).

[3] Patrick Bilic et al. “The Liver Tumor Segmentation Benchmark (LiTS)”. In: Jan. 2019.

[4] Shuqing Chen et al. “Towards Automatic Abdominal Multi-Organ Segmentation in Dual Energy CT using Cascaded 3D Fully Convolutional Network”. In: (2017).

[5] Marco Domenico Cirillo, David Abramian, and Anders Eklund. *Vox2Vox: 3D-GAN for Brain Tumour Segmentation*. 2020.

[6] A. Ciurte et al. “Semi-Supervised Segmentation of Ultrasound Images Based on Patch Representation and Continuous Min Cut”. In: *PLoS ONE* 9.7 (2014), e100972.

[7] Nicholas Heller et al. *KiTS21 Challenge*. 2021.

[8] F. Isensee, P.F. Jaeger, S.A.A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nat Methods* 18 (2021), pp. 203–211.

[9] X. Li et al. “Transformation- Consistent Self-Ensembling Model for Semisupervised Medical Image Segmenta- tion”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2 (2021), pp. 523–534.

[10] Y. Li et al. “Self-Loop Uncertainty: A Novel Pseudo-Label for Semi-supervised Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Vol. 12261. Lecture Notes in Computer Science. Cham: Springer, 2020.

[11] C. Qin et al. “A Semi-supervised Large Margin Algorithm for White Matter Hyperintensity Segmentation”. In: *Machine Learning in Medical Imaging*. Ed. by L. Wang et al. Vol. 10019. Lecture Notes in Computer Science. Cham: Springer, 2016.

[12] 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*. Ed. by N. Navab et al. Vol. 9351. Lecture Notes in Computer Science. Cham: Springer, 2015.

[13] Yuchao Wang et al. *Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels*. 2022.

[14] Yingda Xia et al. “Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation”. In: *Medical Image Analysis* 65 (2020).

[15] Ke Yan et al. "DeepLesion: Automated Deep Mining, Categorization and Detection of Significant Radiology Image Findings using Large-Scale Clinical Lesion Annotations". In: (Oct. 2017).

[16] Zhen Zhao et al. "Augmentation Matters: A Simple-yet-Effective Approach to Semi-supervised Semantic Segmentation". In: (2022).

[17] Y. Zhou et al. "Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training". In: *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*. 2019, pp. 121–140.