Optimization of Deep Learning Algorithm for personalized adaptive radiotherapy

Byongsu Choi

The Graduate School
Yonsei University
Department of Medicine

Optimization of Deep Learning Algorithm for personalized adaptive radiotherapy

A Dissertation's Thesis
Submitted to the Department of Medicine
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Science

Byongsu Choi

December 2024

This certifies that the Dissertation of Byongsu Choi is approved

[signature]

Thesis Supervisor Jin Sung Kim

[signature]

Thesis Committee Member Hojin Kim

[signature]

Thesis Committee Member Justin Chunjoo Park

[signature]

Thesis Committee Member Yu Rang Park

[signature]

Thesis Committee Member Dongryul Oh

The Graduate School
Yonsei University
December 2024

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to my family for their unwavering support, for allowing me to pursue this field, and for always guiding me on the right path. Their encouragement and dedication have been the foundation of my success.

I would like to acknowledge all the support from Professors Jin Sung Kim and Chunjoo Park, for their inspiring advice and support in the preparation of this thesis, as well as for their mentorship throughout their research career and for setting a good example by devoting themselves to patients suffering from cancer.

I am sincerely thankful to Professors Hojin Kim and Bongyong Song for their insightful advice throughout this research journey and for their wise counsel regarding my career path. Additionally, I would like to extend my deepest thanks to the professors at UTSW—Mu-Han Lin, Yank Kyun Park, and Heejung Kim—and at Mayo Clinic—Chris Beltran, Shrestha, Deepak K., and Alessio Parisi—who provided invaluable support and guidance during my time as a visiting student.

Additionally, I would like to extend my gratitude to the committee members, Professors Yu Rang Park and Dongryul Oh, for their invaluable contributions and insightful feedback, which greatly enhanced the quality of this thesis.

Lastly, my gratitude goes to my colleagues at MPBEL for their collaboration and encouragement throughout this research. A special thanks to Juyoung Lee, whose constant support and assistance in every aspect of my work have been instrumental in this journey.

Byongsu Choi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Optimization of Deep Learning Algorithm for personalized adaptive radiotherapy

Adaptive Radiation Therapy (ART) represents a transformative approach in the field of radiation oncology, designed to enhance the precision and effectiveness of cancer treatment by dynamically adjusting radiation doses based on real-time anatomical and physiological changes in patients. Unlike conventional radiation therapy, which relies on a static treatment plan, ART allows for continuous monitoring and adaptation throughout the course of treatment, ensuring that radiation is delivered more accurately to the tumor while sparing healthy tissues. This real-time adaptability is crucial for addressing variations in tumor size, shape, and position, as well as changes in surrounding organs, which can occur due to weight loss, tumor shrinkage, or other physiological shifts during treatment. However, the full clinical adoption of ART is constrained by several key challenges: 1) the labor-intensive and time-consuming nature of manual contouring for organ-at-risk (OAR) and clinical target volumes (CTV), 2) the prolonged training times of complex DL models, which impede the timely implementation of patient-specific care, 3) the limited accuracy and generalizability of existing deep learning (DL) models due to insufficient and non-personalized training datasets. This dissertation focuses on optimizing DL algorithms to overcome these challenges and enhance the efficacy of personalized ART.

We begin by introducing the principles of ART and the critical role of DL used for the ART progress. We then present the development of optimized auto-contouring models applied to head and neck anatomy in veterinary medicine and breast cancer treatment, demonstrating their potential to streamline clinical workflows while maintaining high accuracy. To accelerate the training process, we propose a Progressive Deep Learning (PDL) framework that optimizes model convergence time, facilitating the rapid deployment of ART solutions in clinical settings. Further, we introduce two innovative frameworks—Personalized Hyperspace Learning (PHL-IDOL) and InterVision—designed to enhance the precision and personalization of ART. These frameworks address the limitations of traditional fine-tuning methods by generating new, patient-specific datasets through advanced interpolation techniques and leveraging prior patient information, resulting in more accurate and robust DL models. These models are validated across multiple clinical institutional dataset, demonstrating their broad applicability and effectiveness. Through these contributions, this dissertation optimizes deep learning algorithms for personalized adaptive radiotherapy, paving the way for future innovations in personalized medicine and ensuring that each patient receives the most effective and tailored treatment possible.

Key words : ART, Optimization, Deep Learning, Auto contouring, Personalized, Head & Neck

1. INTRODUCTION

1.1. Radiation Therapy

Radiation therapy (RT) is a cornerstone of modern oncology [1], serving as one of the most effective modalities in the treatment of cancer for over a century. Utilizing ionizing radiation, RT aims to target and destroy malignant cells while sparing surrounding healthy tissue as much as possible. The fundamental principle of RT lies in its ability to induce irreparable DNA damage within cancer cells, leading to their death or senescence [2-4]. This is typically achieved through the delivery of high-energy photons, electrons, or protons, which are directed precisely at the tumor site. RT is versatile in its application, playing a crucial role not only in curative settings, where the goal is to eradicate the cancer, but also in palliative care, where it helps alleviate symptoms, and in adjuvant therapy, where it supports other treatment modalities such as surgery and chemotherapy.

The effectiveness of RT is predicated on meticulous planning and delivery to maximize tumor control while minimizing the collateral damage to normal tissues, which could lead to side effects. Over the years, significant technological advancements have greatly enhanced the precision and effectiveness of RT. Techniques such as Intensity-Modulated Radiation Therapy (IMRT) and Image-Guided Radiation Therapy (IGRT) allow clinicians to deliver higher radiation doses more precisely, focusing on the tumor while sparing nearby healthy tissues [5, 6]. Despite these advances, RT remains a complex and delicate procedure that must be carefully tailored to each patient's unique anatomy and tumor characteristics. As cancer treatment evolves towards more personalized approaches, the demand for innovations that further enhance the precision and safety of RT continues to grow, highlighting the importance of ongoing research and development in this critical field.

1.2. Development of Radiation Therapy

The evolution of radiation therapy is marked by a series of technological innovations and discoveries that have continually pushed the boundaries of cancer treatment. The journey began with the groundbreaking discoveries of X-rays by Wilhelm Röntgen in 1895 and radium by Marie and Pierre Curie in 1898. These early advancements provided the medical community with the tools to explore the therapeutic potential of radiation. However, the initial applications were rudimentary, often resulting in significant damage to both cancerous and healthy tissues alike [7-11].

As the understanding of radiation physics advanced, so too did the ability to control and direct radiation more effectively. The 1950s saw the introduction of linear accelerators, which were capable of generating high-energy X-rays that could penetrate deeper into the body. This breakthrough allowed for more effective targeting of tumors while sparing superficial healthy tissues. The 1980s ushered in the era of three-dimensional conformal radiation therapy (3D-CRT), a technique that enabled clinicians to shape the radiation beams to match the precise contours of the tumor, thereby reducing exposure to surrounding normal tissues.

Subsequent innovations such as Intensity-Modulated Radiation Therapy (IMRT) and Volumetric Modulated Arc Therapy (VMAT) further refined the precision of radiation delivery [12, 13]. These techniques allow for highly conformal dose distributions that can be tailored to the complex geometries of tumors, even those located near critical structures. These advancements have not only improved the therapeutic effectiveness of RT but have also significantly reduced the incidence and severity of side effects, leading to better overall outcomes and quality of life for patients. The ongoing development of radiation therapy continues to be driven by the dual goals of maximizing tumor control and minimizing harm to normal tissues, ensuring that this treatment modality remains a cornerstone of cancer care.

1.3. Adaptive Radiation Therapy (ART)

Adaptive Radiation Therapy (ART) represents a significant evolution in RT by introducing a dynamic, patient-centered approach to treatment planning and delivery. Traditional RT relies on a static treatment plan developed before therapy begins, assuming that the tumor and surrounding anatomy remain unchanged throughout the course of treatment [14-16]. However, factors such as tumor shrinkage, organ motion, and patient weight loss can lead to discrepancies between the planned and actual anatomy, potentially compromising treatment efficacy and safety.

ART overcomes these limitations by integrating frequent imaging and advanced computational techniques to adjust treatment plans in real-time. This adaptability ensures that the radiation dose remains precisely targeted to the tumor, reducing exposure to OARs and improving overall outcomes. The implementation of ART involves a seamless interplay between imaging modalities, treatment planning algorithms, and delivery systems, making it a highly sophisticated process. ART’s ability to personalize therapy aligns with the broader trend toward precision medicine, offering significant benefits for complex cases where traditional static plans fall short. As the technology and methodologies underlying ART continue to evolve, its adoption is expected to expand, enhancing the effectiveness and safety of RT in diverse clinical settings.

1.4. Deep Learning Techniques in ART

Deep learning (DL) is revolutionizing the field of ART by automating labor-intensive tasks and enabling more adaptive, personalized treatment planning [17, 18]. As a subset of artificial intelligence (AI), DL leverages neural networks, particularly convolutional neural networks (CNNs), to extract meaningful patterns from complex datasets, such as medical images. This capability makes DL an ideal tool for addressing key challenges in ART, including image segmentation, tumor contouring, and dose prediction. By automating these tasks, DL reduces inter-observer variability and increases the efficiency of clinical workflows.

One of DL’s most impactful applications in ART is the segmentation of OARs and clinical target volumes (CTVs). These delineations are critical for ensuring accurate dose delivery while minimizing exposure to healthy tissues [19]. Traditionally, this process is manual and time-consuming, requiring significant expertise from radiation oncologists. DL models trained on large datasets can perform this task with high accuracy and consistency, streamlining the planning process. Beyond segmentation, DL is being used to predict anatomical changes during treatment, such as tumor shrinkage or organ motion, allowing clinicians to adapt plans proactively. Despite its potential, the integration of DL into ART faces challenges such as the need for large, high-quality datasets, computational demands, and rigorous clinical validation. Addressing these challenges will be essential for fully realizing the benefits of DL in ART.

1.5. Motivation of thesis

The motivation for this thesis is driven by the need to improve the precision, effectiveness, and personalization of adaptive radiation therapy (ART) through the application of advanced deep learning (DL) techniques. ART represents a significant advancement in radiation therapy, offering the potential to adapt treatment plans in real-time based on changes in the patient's anatomy or tumor characteristics. However, several challenges remain that limit the full potential of ART. These include the labor-intensive nature of manual contouring, the risk of overfitting in DL models due to limited data availability, and the computational demands associated with real-time treatment adjustments. Moreover, the integration of DL into ART is still in its early stages, with many opportunities for innovation and improvement.

This thesis seeks to address these challenges by developing novel DL models and frameworks that enhance the accuracy, efficiency, and personalization of ART. A key aspect of this work involves the validation of these techniques using multi-institutional datasets, which is crucial for ensuring the generalizability and robustness of the proposed models across diverse patient populations and clinical environments. By verifying the effectiveness of the models on data from multiple institutions, this research aims to demonstrate that the developed approaches can be widely applicable and reliable in real-world clinical settings. Specifically, it focuses on three interconnected objectives: 1) Generating a auto-contouring model, 2) Optimized Deep Learning Model for Accelerated Convergence and 3) Framework for Personalized Models Verified with Multi-Institutional Datasets.

1.6 Specific aims

The specific aims of this dissertation are as follows:

1. Generation Auto-Contouring Model: Developing robust DL models for OAR and target volume segmentation to standardize and automate this labor-intensive process, thus reducing inter-operator variability and improving clinical efficiency.

2. Optimized Deep Learning Model for Accelerated Convergence: Proposing the Progressive Deep Learning (PDL) framework to optimize training times, enabling the rapid deployment of DL models in time-sensitive clinical settings.

3. Framework for Personalized Models Verified with Multi-Institutional Datasets: Creating frameworks such as Personalized Hyperspace Learning (PHL-IDOL) to generate patient-specific datasets and enhance model generalizability by leveraging prior patient information and innovative interpolation techniques. After generating the personalized model, we verified it using multi-institutional datasets to demonstrate the benefits of these personalized models. These approaches aim to improve the precision, adaptability, and clinical applicability of DL models.

1.7 Dissertation Organization

The reminder of dissertation is organized as follows.

Chapter 2 details the development and application of deep learning (DL) models for auto contouring of organs at risk (OARs) and clinical target volumes (CTVs). The chapter begins by emphasizing the critical role of accurate auto contouring in radiation therapy, followed by an exploration of DL-based techniques specifically adapted for head and neck contouring in veterinary applications. It then transitions into a clinical evaluation of atlas-based and DL-based auto contouring methods applied to breast cancer treatment, presenting comparative results that highlight the superiority of DL approaches. The chapter concludes with a discussion of the advancements and limitations of current methodologies, along with an acknowledgment of the contributions from collaborators.

Chapter 3 introduces a novel Progressive Deep Learning (PDL) model aimed at reducing training times for segmentation tasks in medical imaging. It begins by identifying the challenges associated with training complex DL models in this field, followed by a detailed explanation of the PDL architecture and implementation. Results demonstrate the model's ability to significantly reduce training times without compromising accuracy. The chapter concludes with a discussion of findings, their implications for future research, and acknowledgments for the support received during the study.

Chapter 4 focuses on the integration of DL models into adaptive radiation therapy (ART) through the development of patient-specific frameworks. The chapter opens with an overview of the challenges in adapting DL for ART and introduces the Personalized Hyperspace Learning (PHL-IDOL) frameworks. These innovative models aim to enhance the precision and personalization of ART by generating and utilizing real-time, patient-specific datasets. A multi-institutional evaluation assesses their performance across diverse clinical settings. The chapter presents results in two sections: the performance of the DL models and the outcomes of the multi-institutional analysis. It concludes with a discussion of findings, potential clinical applications, and acknowledgments of contributions from participating institutions.

Chapter 5 provides a comprehensive summary of the dissertation, highlighting key contributions and discussing future research directions. It reflects on the impact of the proposed DL models in improving the accuracy, efficiency, and personalization of ART, emphasizing the advancements achieved while acknowledging remaining challenges. Future efforts will focus on refining these models, expanding their applications to other cancer types, and further enhancing treatment outcomes through advanced personalization techniques.

2. Deep Learning model for auto contouring OARs and clinical target volumes (CTV)

2.1. Introduction

In radiation therapy (RT), accurate segmentation of organs at risk (OARs) and clinical target volumes (CTVs) is a cornerstone, directly impacting treatment efficacy and patient safety. Precise segmentation is essential to ensure that therapeutic radiation doses target the tumor while sparing adjacent healthy tissues. However, manual contouring remains a labor-intensive process requiring significant expertise, which often introduces variability between practitioners. These challenges underscore the need for advanced, reliable solutions that can automate segmentation and standardize treatment planning processes.

In recent years, the advent of deep learning (DL) has revolutionized the field of medical imaging by enabling highly accurate and efficient segmentation algorithms. DL models, particularly those utilizing convolutional neural networks (CNNs), have shown unprecedented capabilities in automating tasks that were once considered too complex for computational systems. In RT, these models offer the potential to overcome the limitations of traditional manual contouring, addressing not only variability but also the significant time burden associated with precise delineation. Furthermore, DL-based approaches promise to streamline workflows, enabling practitioners to focus more on clinical decision-making rather than repetitive tasks.

The motivation for exploring DL in auto-contouring is driven by the increasing complexity of modern RT. These advanced modalities require precise segmentation of complex anatomical structures to achieve the desired dose distribution. The intricate geometries of certain regions, coupled with variations in imaging quality and patient-specific anatomy, make manual segmentation particularly challenging and time-consuming. By automating these processes, DL can significantly improve the consistency and reproducibility of RT plans, aligning with the broader goal of precision medicine.

This chapter discusses the development and application of DL models specifically designed for auto-contouring in diverse contexts, including both veterinary and human applications. In veterinary medicine, the adoption of DL for head-and-neck OAR segmentation provides a unique perspective, demonstrating the adaptability of these technologies across species and anatomical variations. Similarly, in human oncology, DL-based contouring has been explored for breast cancer treatment, highlighting its utility in handling the challenges associated with soft-tissue structures and complex organ geometries. These advancements emphasize the versatility and robustness of DL in addressing the diverse needs of RT planning.

Through this exploration, the chapter aims to provide a comprehensive overview of the potential for DL to transform auto-contouring practices in RT. By focusing on the motivations behind these developments and the advancements achieved thus far, it sets the stage for a deeper understanding of how DL can be seamlessly integrated into the RT planning process, ultimately improving patient outcomes and streamlining clinical operations.

2.2. Auto contouring of head and neck for veterinary applications

The field of veterinary radiation oncology has increasingly embraced advanced techniques from human medicine, particularly for the treatment of head and neck cancers. Precise contouring of organs at risk (OARs) and clinical target volumes (CTVs) is essential to ensure effective radiation dose delivery while minimizing exposure to healthy tissues. However, manual contouring is time-intensive, highly reliant on expert knowledge, and challenging due to the significant anatomical variability in veterinary patients. These challenges underscore the need for automated contouring solutions that can streamline workflows and improve treatment precision.

Deep learning (DL)-based approaches have emerged as a promising solution for automating the contouring process in veterinary oncology. The unique anatomical variability in veterinary patients, spanning different species and breeds, demands robust and adaptable models capable of handling diverse datasets. By leveraging convolutional neural networks (CNNs) and advanced architectures, DL models can accurately identify and segment OARs and CTVs, reducing reliance on labor-intensive manual methods while enhancing consistency and reproducibility.

To address the challenges of manual contouring, a DL model was developed specifically for head and neck regions in veterinary patients. To advance precision in veterinary radiation oncology, this study utilized CT data from a comprehensive dataset of 90 dogs with head and neck cancers. Data from 80 dogs were included in the algorithm's development phase, where 60 were allocated to training and validation, and 20 served as test sets. Additionally, 10 clinical test sets were introduced to assess the algorithm's clinical feasibility. Expert contours for the 90 dogs were meticulously delineated by a primary radiologist with specialized training in veterinary medical imaging and two additional radiologists for the clinical test sets, ensuring the inclusion of diverse professional expertise in contouring accuracy.

The network development ensured full compatibility with CT image resolution, with Hounsfield unit values normalized from [−100, 700] to [−1.0, 1.0]. CT images were further normalized to a consistent voxel size of 1.0 × 1.0 × 3.0 mm^3 to maintain uniformity across datasets. This preprocessing step was crucial for reducing variability and enhancing the model's generalizability. A two-step, three-dimensional (3D) fully convolutional DenseNet was employed to automatically segment organs at risk (OARs) and clinical target volumes (CTVs). The DenseNet architecture, an evolution of the U-Net, leverages dense connectivity to maximize information flow and achieve high segmentation accuracy. The process was implemented in TensorFlow 2.4.1 and Python 3.6.8, with model training conducted on an NVIDIA TITAN RTX GPU, ensuring computational efficiency and scalability. In the first step, multilabel segmentation was used to identify the approximate regions of interest (ROIs) for each OAR. During this step, the x, y, and z directions of the CT images were downsampled by half to accelerate processing. This localization process automatically cropped ROIs from the preprocessed images, minimizing irrelevant regions while preserving the essential anatomical structures for segmentation. Following localization, single-label segmentation was performed for each OAR. ROI segmentation volumes were refined by calculating the x, y, and z boundaries and cropping extraneous margins. This approach ensured precise, high-resolution segmentation of each anatomical structure.

The fully convolutional DenseNet architecture consists of dense blocks, which are similar to residual blocks in the U-Net architecture. These dense blocks enhance feature propagation and reduce computational redundancy. Key components include: 1) Transition Down Layers: Batch normalization, rectified linear units (ReLU), 1 × 1 convolutions, dropout (p = 0.2), and 2 × 2 max-pooling operations to downsample and extract essential features. 2) Skip Connections: Feature maps from the downsampling path are concatenated with corresponding maps in the upsampling path, ensuring high-resolution outputs. 3)Transition Up Layers: 3 × 3 deconvolutions with a stride of two progressively recover spatial resolution, enabling precise segmentation of fine anatomical details.

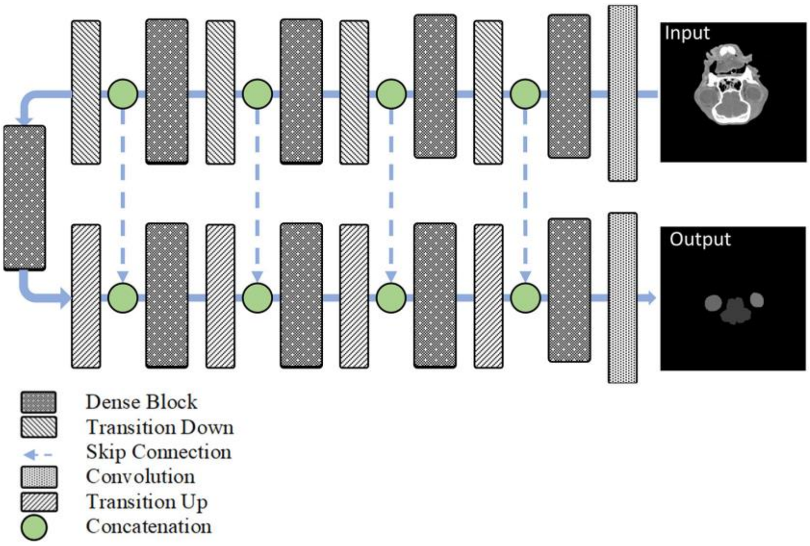


Figure 2-1. The architecture of the proposed fully convolutional DenseNet

The segmentation model's accuracy was evaluated using 20 test sets and 10 clinical test sets. The Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95) were employed to assess the closeness and surface distance of the contours, respectively. DSC: Quantifies overlap between automated and expert contours, with values ranging from 0 (no overlap) to 1 (perfect overlap). A DSC of 0.75 or higher was deemed acceptable:

$$DSC = \frac{2(A \cap B)}{|A| + |B|}$$

HD95: Measures the maximum distance between points on one contour and the closest points on the other, focusing on the 95th percentile to reduce the impact of outliers:

$$H(A, B) = \max\{h(A, B),\ h(B, A)\}$$

The study included three radiologists as human annotators. Annotator one's segmentations were designated as the ground truth for evaluation, while segmentations from the other two annotators were assessed as human annotations (HAs). Additionally, the HA_DLBAS process, wherein the two annotators adjusted the DLBAS predictions only in areas of inaccuracy, was evaluated to explore the benefits of expert intervention.

The evaluation compared three methods for segmentation: (1) the predictions from DLBAS, (2) manual segmentations by the two annotators (HAs), and (3) HA_DLBAS, where the annotators corrected the DLBAS-predicted contours. Accuracy and consistency were assessed using DSC, HD, and contouring time metrics.

2.3. Clinical evaluation of atlas and DL-based auto contouring for breast cancer

Accurate delineation of organs at risk (OARs) and clinical target volumes (CTVs) is a cornerstone of radiation therapy (RT) planning, particularly in breast cancer treatment, where precise contouring can significantly reduce radiation-induced toxicity. Traditional manual contouring, although considered the gold standard, is time-intensive and prone to inter-observer variability. To address these limitations, automated contouring methods such as atlas-based segmentation and deep learning (DL) approaches have emerged as viable alternatives, offering the potential to streamline workflows and improve consistency.

Breast cancer RT requires meticulous delineation of critical structures, including the heart, lungs, chest wall, and supraclavicular lymph nodes. The accuracy of these contours directly impacts dose distribution and treatment outcomes. However, the variability inherent in manual segmentation presents challenges, particularly when working with complex anatomical regions or large patient datasets. Automated methods offer the opportunity to reduce this variability and expedite the planning process, ensuring more consistent and reproducible results.

For this study, a comprehensive dataset of breast cancer patients was collected, encompassing various anatomical presentations and treatment scenarios. The dataset was divided into three subsets: 1) Training Set: Used for developing the DL-based model, comprising diverse anatomical cases to enhance the model's robustness. 2) Validation Set: Employed to fine-tune model parameters and prevent overfitting. 3) Test Set: Reserved for evaluating model performance in comparison to manual contours and atlas-based methods. All contours in the dataset were manually delineated by experienced radiation oncologists and served as ground truth for evaluating the automated approaches.

Two auto-contouring methods were compared: an atlas-based auto segmentation (ABAS) approach and a DL-based model auto segmentation (DLBAS) utilizing a 3D Fully Convolutional DenseNet (FCDN) architecture. Atlas-Based Segmentation was employed a pre-compiled library of patient datasets, each annotated with contours for relevant structures. Contours were transferred from the atlas to new patient datasets using deformable image registration. While atlas-based methods are widely used in clinical practice, they are often limited by inaccuracies in regions with significant anatomical variation. A 3D FCDN was developed to automate the segmentation process, leveraging the spatial and contextual features inherent in volumetric imaging data. The model architecture included multiple convolutional layers with skip connections and batch normalization, ensuring precise segmentation even in anatomically challenging areas. Training was performed on a high-performance computing system, utilizing TensorFlow as the DL framework.

Performance was evaluated using the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95). These metrics quantified the spatial overlap and boundary differences between the automated contours and the ground truth. A minimum DSC threshold of 0.75 was established as acceptable for clinical application.

2.4. Results

2.4.1. Head and neck for veterinary DL auto contouring performance

The performance of the deep-learning-based automatic segmentation (DLBAS) method was evaluated across 15 organs at risk (OARs). Among these, the right eye exhibited the highest segmentation accuracy, with a mean Dice similarity coefficient (DSC) of 0.93 and a mean Hausdorff distance (HD) of 1.80 mm. Conversely, the lowest accuracy was observed for the left parotid salivary gland, which achieved a DSC of 0.72 and an HD of 3.88 mm. On average, the DLBAS model demonstrated reliable segmentation performance, with mean DSC and HD values of 0.83 ± 0.01 and 2.71 ± 0.31 mm, respectively. Notably, except for the right cochlear and bilateral parotid salivary glands, all OARs exceeded a DSC value of 0.79. However, some OARs, including the brain, pharynx and larynx, and spinal cord, exhibited HD values exceeding 3 mm, indicating potential challenges in these regions. The results are generated using the boxplot shown in Figure 2-2.

The application of DLBAS to tumor-affected patients in the test sets confirmed its robustness, as no significant differences in segmentation accuracy were observed between tumor and non-tumor datasets is illustrated in Figure 2-3 and Table 2-1. However, clinical test sets with cephalic index values ranging from 0.5 to 0.6 showed decreased DSC values. This reduction was attributed to anatomical displacement or deformation caused by tumor lesions rather than the cephalic index itself. The CT images of these clinical sets revealed that displacement or deformation of anatomical structures due to lesions significantly impacted segmentation accuracy. This suggests the need for further evaluations to determine the feasibility of applying DLBAS in cases with severe displacement and deformation.

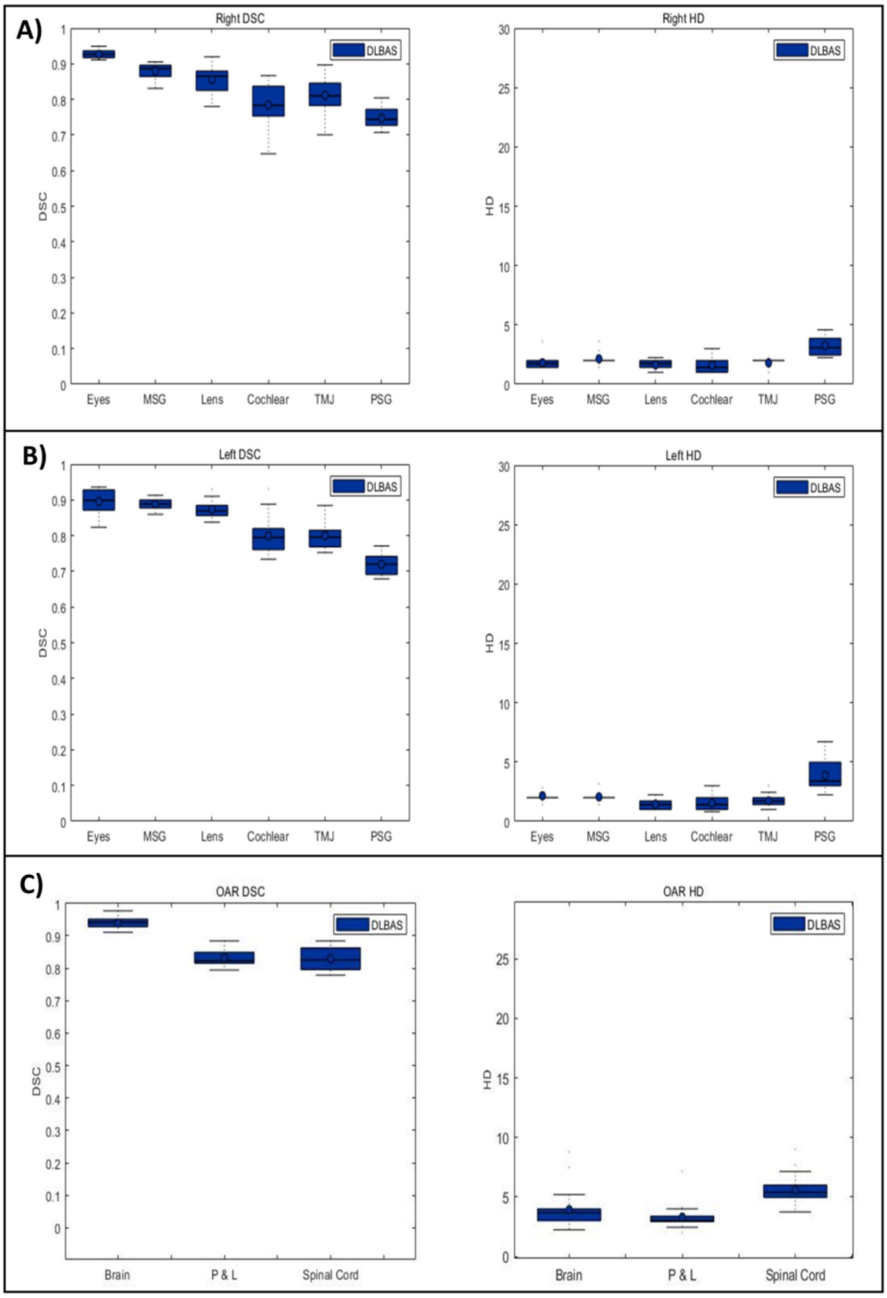


Figure 2-2. Boxplots of the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) for each organ at risk (OAR) segmented using the deep-learning-based automatic segmentation model. (A) Right-side organs, (B) Left-side organs, and (C) other OARs.

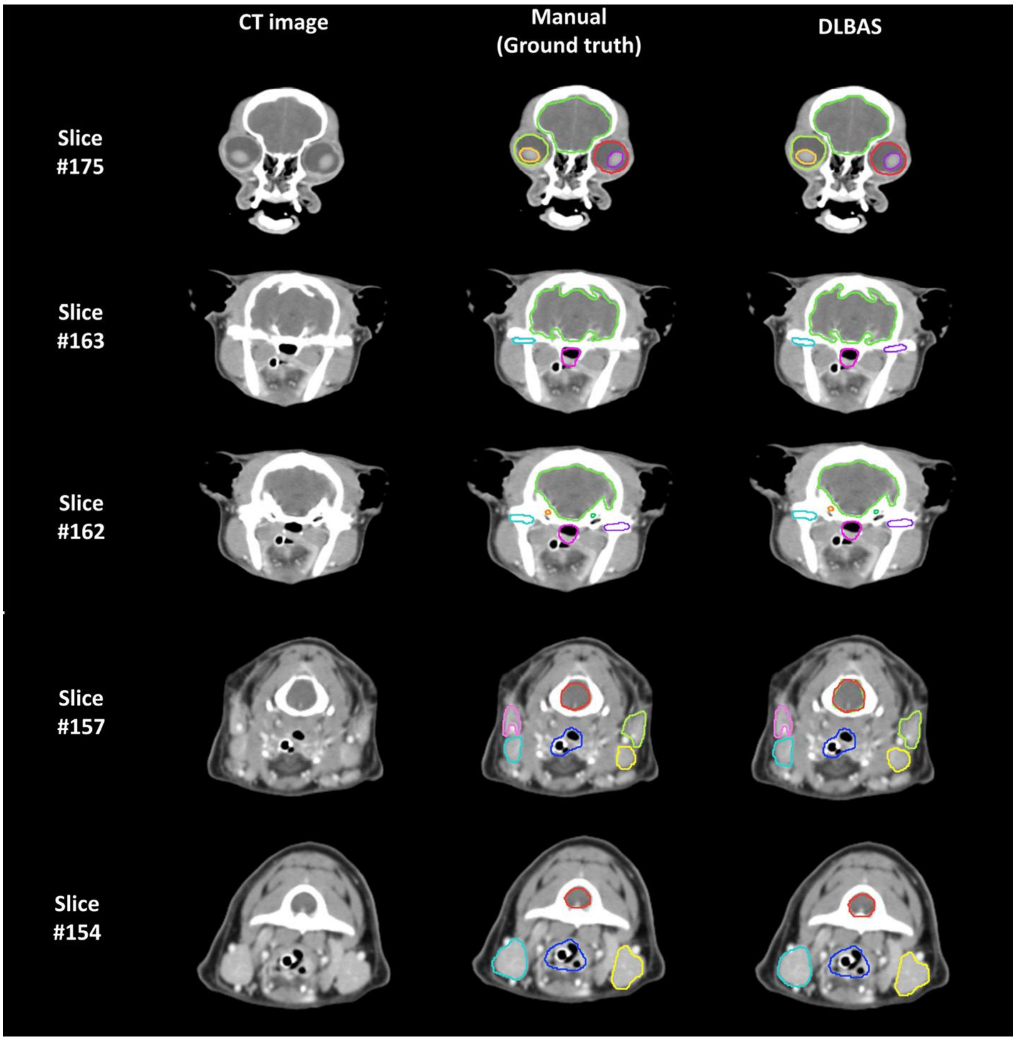


Figure 2-3. Representative examples of ground truth and deep-learning-based automatic segmentation (DLBAS) from a test set. Segmentations are visualized across different slices, highlighting both the similarities and discrepancies between ground truth and DLBAS contours. In Slice #175, key structures such as the eye (red, lime green), lens (yellow, purple), and brain (yellow, green) are depicted. Slices #163 and #162 display the brain (yellow, green), cochlea (orange, green), temporomandibular joint (sky blue, purple), and pharynx and larynx (pink). In Slices #157 and #154, the mandibular salivary gland (sky blue, yellow), parotid salivary gland (pink, lime green), pharynx and larynx (blue), and spinal cord (red) are illustrated. Notable differences are observed in structures such as the temporomandibular joint (purple) in Slice #163 and the spinal cord (red) in Slice #157, where the predicted DLBAS spinal cord region overlaps with the brain (green). These examples underline the areas of high segmentation accuracy and potential regions of improvement for the DLBAS model.

Table 2-1. Accuracy correlation according to variables in the test set
SD, standard deviation; DSC, Dice similarity coefficient; HD, Hausdorff distance; W/L, skull width / skull length

Variables	Score (mean ± SD)	
	DSC	HD (mm)
	0.83 ± 0.01	2.71 ± 0.31
Age (years)		
0 ~ 3	0.83	2.91
3 ~ 6	0.83	2.75
6 ~ 10	0.83	2.68
10 ~	0.83	2.63
Weight (kg)		
1 ~ 10	0.83	2.75
10 ~ 20	0.84	2.56
20 ~ 30	0.82	2.61
30 ~	-	-
Cephalic index (W/L)		
0.4 ~ 0.5	0.83	2.44
0.5 ~ 0.6	0.62	1.99
0.6 ~ 0.7	0.84	2.80
0.7 ~	0.75	2.48
Skull pattern		
Mesocephalic	0.84	2.69
Brachycephalic	0.82	2.73
Dolichocephalic	-	-
Lesion		
Presence	0.83	2.76
Absence	0.83	2.67

Therefore, the clinical feasibility of DLBAS was assessed using both quantitative and qualitative metrics. Table 2-2 and 2-3 show the result for clinical test sets. For the 10 clinical test sets, the average DSC and HD values were 0.78 ± 0.11 and 4.29 ± 3.30 mm, respectively. Compared to the test sets, these clinical sets exhibited lower DSC values and higher HD values. Among OARs, the right cochlear (DSC: 0.50 ± 0.28) and left parotid salivary gland (HD: 7.01 ± 8.67 mm) recorded the lowest accuracy, while the brain (DSC: 0.90 ± 0.11) and right eye (HD: 2.00 ± 0.71 mm) achieved the highest accuracy.

The results highlighted two distinct groups within the clinical test sets: group 1, characterized by low segmentation accuracy (DSC: 0.66, HD: 7.57), and group 2, demonstrating high accuracy (DSC: 0.86, HD: 2.10). Group 1 primarily included cases where tumor-induced anatomical changes or inflammatory responses caused displacement or deformation of OARs. Additionally, insufficient contrast enhancement and asymmetry in CT scans contributed to lower accuracy in group 1. These challenges impacted the two-step segmentation process, leading to reduced localization precision and segmentation accuracy.

Table 2-2. Dice similarity coefficient of each clinical test set obtained from deep-learning-based automatic segmentation

OAR, organ at risk; TMJ, temporomandibular joint; MSG, mandibular salivary gland; PSG, parotid salivary gland; L, left; R, right; C1-C10, clinical test set 1 - 10

OAR	Group 1				Group 2					
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Lens (L)	0.64	0.70	0.70	0.89	0.89	0.89	0.89	0.89	0.90	0.87
Lens (R)	0.90	0.93	0.93	0.66	0.86	0.87	0.86	0.78	0.88	0.85
Eye (L)	0.85	0.90	0.90	0.64	0.94	0.95	0.94	0.92	0.95	0.93
Eye (R)	0.95	0.24	0.24	0.55	0.95	0.94	0.93	0.93	0.95	0.94
Cochlear (L)	0.68	0.41	0.41	0.64	0.64	0.63	0.47	0.71	0.60	0.87
Cochlear (R)	0.65	0.05	0.05	0.27	0.57	0.65	0.65	0.71	0.59	0.84
TMJ (L)	0.52	0.90	0.70	0.90	0.92	0.90	0.85	0.86	0.89	0.88
TMJ (R)	0.41	0.31	0.31	0.61	0.87	0.87	0.87	0.71	0.88	0.86
MSG (L)	0.68	0.55	0.55	0.50	0.90	0.93	0.93	0.73	0.93	0.85
MSG (R)	0.74	0.82	0.82	0.90	0.90	0.92	0.95	0.72	0.94	0.82
PSG (L)	0.35	0.84	0.84	0.85	0.85	0.91	0.83	0.86	0.89	0.81
PSG (L)	0.48	0.68	0.68	0.43	0.43	0.90	0.82	0.91	0.88	0.78
Pharynx & larynx	0.83	0.96	0.77	0.77	0.94	0.95	0.95	0.89	0.95	0.84
Brain	0.63	0.77	0.89	0.90	0.97	0.97	0.97	0.94	0.97	0.95
Spinal cord	0.74	0.89	0.71	0.71	0.90	0.90	0.88	0.81	0.89	0.89
Total	0.67	0.66	0.63	0.68	0.83	0.88	0.85	0.83	0.87	0.87

Table 2-3. Hausdorff distance of each clinical test set obtained from deep-learning-based automatic segmentation

OAR	Group 1				Group 2				
	C1	C2	C3	C4	C6	C7	C8	C9	C10
Lens (L)	0.81	39.53	1.67	3.69	1.69	1.89	0.69	0.83	1.89
Lens (R)	0.81	1.73	7.70	2.09	1.69	1.86	0.69	0.83	1.86
Eye (L)	0.81	2.52	3.38	9.26	1.69	1.94	1.43	0.83	1.94
Eye (R)	0.81	1.97	2.89	2.09	1.69	1.95	1.93	1.38	1.93
Cochlear (L)	1.15	2.08	3.32	2.09	1.69	1.64	0.99	2.66	1.47
Cochlear (R)	0.81	6.27	2.49	2.09	1.69	1.57	0.69	0.83	1.65
TMJ (L)	5.01	1.34	10.13	5.37	1.69	1.92	0.69	0.83	1.85
TMJ (R)	2.53	20.28	2.83	5.86	1.69	1.87	0.69	0.83	1.87
MSG (L)	15.10	7.11	9.47	7.46	4.77	1.90	0.69	3.71	1.93
MSG (R)	12.29	3.76	6.67	8.14	4.37	1.90	0.69	1.38	1.95
PSG (L)	11.90	5.29	5.82	29.91	1.69	1.85	1.94	6.86	1.83
PSG (L)	8.49	8.15	4.44	17.95	10.59	1.43	3.72	2.72	1.82
Pharynx & larynx	10.00	3.93	9.83	8.51	1.99	1.94	2.74	0.83	1.95
Brain	8.53	4.05	8.09	30.49	2.00	1.97	1.76	0.83	1.97
Spinal cord	1.28	4.20	8.79	38.90	1.69	1.90	0.69	1.17	1.88
Total	5.36	7.48	5.83	11.59	2.71	1.83	1.34	1.77	1.85

Despite these challenges, the DLBAS method demonstrated robust segmentation performance and maintained a high level of clinical feasibility. In addition, we have generated the hybrid approach combining DL-based auto contouring with manual refinement (HA_DLBAS), which integrates expert intervention, further improved segmentation accuracy and consistency. HA_DLBAS achieved the highest DSC (0.94 ± 0.04) and lowest HD (2.30 ± 0.56 mm) values, outperforming both DLBAS alone (DSC: 0.78 ± 0.11, HD: 4.29 ± 3.30 mm) and manual delineations (HA) (DSC: 0.85 ± 0.07, HD: 2.74 ± 1.11 mm). Figure 2-4, Table 2-4 and 2-5 represent the result of the three contouring methods (HA, DLBAS, HA_DLBAS) of the clinical test set.

Table 2-4. Dice similarity coefficient result of three contouring methods
OAR, organ at risk; HA, human annotation; DLBAS, deep-learning-based automatic segmentation; HA_DLBAS, human annotation with additional readjustments to DLBAS predictions

OAR	**DSC (mean ± SD)**		
	HA	**DLBAS**	**HA_DLBAS**
Lens (L)	0.85 ± 0.04	0.83 ± 0.10	0.87 ± 0.04
Lens (R)	0.85 ± 0.07	0.85 ± 0.08	0.93 ± 0.06
Eye (L)	0.93 ± 0.09	0.89 ± 0.09	0.92 ± 0.02
Eye (R)	0.93 ± 0.07	0.76 ± 0.30	0.95 ± 0.09
Cochlear (L)	0.81 ± 0.08	0.61 ± 0.14	0.92 ± 0.06
Cochlear (R)	0.73 ± 0.18	0.50 ± 0.28	0.94 ± 0.03
TMJ (L)	0.77 ± 0.15	0.83 ± 0.13	0.88 ± 0.08
TMJ (R)	0.80 ± 0.11	0.67 ± 0.24	0.81 ± 0.07
MSG (L)	0.89 ± 0.04	0.76 ± 0.18	0.98 ± 0.02
MSG (R)	0.89 ± 0.05	0.85 ± 0.08	0.99 ± 0.03
PSG (L)	0.83 ± 0.05	0.80 ± 0.16	0.97 ± 0.03
PSG (R)	0.79 ± 0.19	0.70 ± 0.19	0.95 ± 0.04
Pharynx & larynx	0.87 ± 0.04	0.89 ± 0.08	0.99 ± 0.01
Brain	0.97 ± 0.09	0.90 ± 0.11	0.99 ± 0.02
Spinal cord	0.88 ± 0.07	0.83 ± 0.08	0.97 ± 0.02
Total	**0.85 ± 0.07**	**0.78 ± 0.11**	**0.94 ± 0.04**

Table 2-5. Hausdorff distance result of three contouring methods

OAR	**HD (mean ± SD, mm)**		
	HA	**DLBAS**	**HA_DLBAS**
Lens (L)	2.94 ± 3.47	5.54 ± 11.98	1.95 ± 0.52
Lens (R)	1.94 ± 0.22	2.20 ± 2.04	1.90 ± 1.48
Eye (L)	1.73 ± 0.46	2.69 ± 2.46	2.79 ± 0.51
Eye (R)	1.71 ± 1.04	2.00 ± 0.71	2.30 ± 0.35
Cochlear (L)	1.44 ± 0.77	2.00 ± 0.74	1.61 ± 0.30
Cochlear (R)	1.41 ± 2.14	2.08 ± 1.63	2.31 ± 0.69
TMJ (L)	2.80 ± 1.76	3.15 ± 2.93	2.40 ± 0.53

TMJ (R)	2.10 ± 1.91	4.11 ± 5.86	2.43 ± 0.42
MSG (L)	2.63 ± 2.07	5.48 ± 4.41	1.38 ± 1.12
MSG (R)	3.30 ± 1.03	4.38 ± 3.65	2.18 ± 0.43
PSG (L)	3.32 ± 2.01	7.01 ± 8.67	2.11 ± 0.03
PSG (R)	4.82 ± 2.27	6.23 ± 5.16	2.62 ± 0.04
Pharynx & larynx	4.90 ± 0.57	4.50 ± 3.53	3.30 ± 0.46
Brain	3.32 ± 0.86	6.59 ± 8.84	3.54 ± 1.24
Spinal cord	2.72 ± 1.15	6.35 ± 11.98	1.72 ± 0.22
Total	**2.74 ± 1.11**	**4.29 ± 3.30**	**2.30 ± 0.56**

The DLBAS method significantly reduced contouring time compared to manual delineations. On average, DLBAS completed segmentation in approximately 3 seconds for all OARs, representing a 1,800-fold reduction in time compared to manual methods (80 minutes). The HA_DLBAS workflow required approximately 30 minutes, effectively halving the time required for manual contouring while achieving higher accuracy. However, cases in group 1 required up to five times longer readjustments due to the aforementioned challenges.

Overall, these results demonstrate that DLBAS is a reliable and efficient segmentation tool, even in challenging clinical scenarios. The integration of expert intervention through the HA_DLBAS workflow further enhances accuracy and consistency, making it a promising solution for automating segmentation in clinical practice.

2.4.2. Clinical evaluation of atlas and DL auto contouring for breast cancer

The DL-based auto contouring model consistently outperformed the atlas-based method across all tested regions of interest (ROIs), including breast, chest wall, and organs at risk (OARs) such as the heart and ipsilateral lung showing the results in Figure 2-3 and 2-4. For the breast target volume, the DL-based model achieved an average DSC of 0.88 ± 0.04 compared to 0.75 ± 0.07 for the atlas-based method, reflecting a significantly closer agreement to expert contours. Similarly, for the chest wall, the DL-based model demonstrated a mean DSC of 0.85 ± 0.05, markedly higher than the atlas-based model's 0.68 ± 0.08.

In terms of HD95%, the DL-based model showed lower values, indicating more precise contour delineation. For the ipsilateral lung, the DL-based model achieved an average HD95% of 4.2 ± 1.1 mm, compared to 7.8 ± 2.4 mm for the atlas-based method. Similar trends were observed for the heart, with the DL-based method recording an HD95% of 5.1 ± 1.6 mm versus 9.3 ± 3.2 mm for the atlas-based approach. These results highlight the improved spatial accuracy and consistency of the DL-based method.

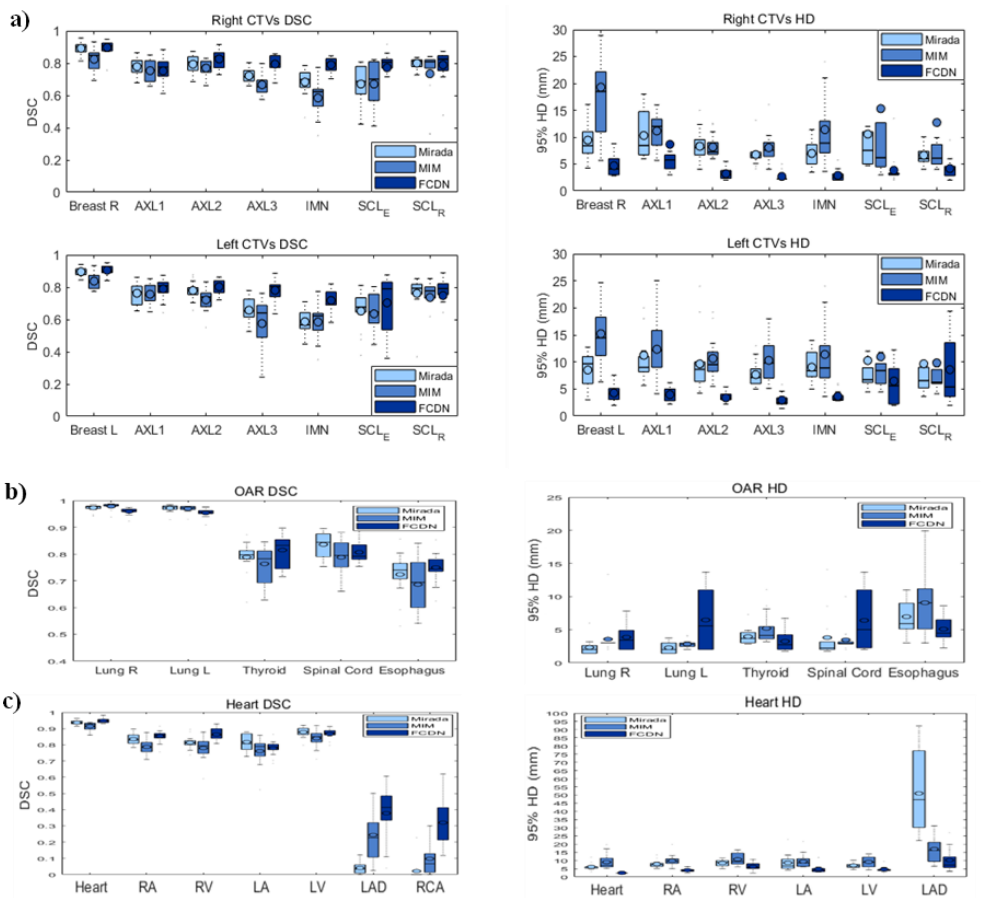


Figure 2-4: Box-plots of Dice Similarity Coefficients (DSC) and 95% Hausdorff Distance (HD95) in the a) CTVs, b) OARs, and c) Heart structures obtained from Mirada, MIM, and DLBAS based on FCDN using the manual contours as reference.

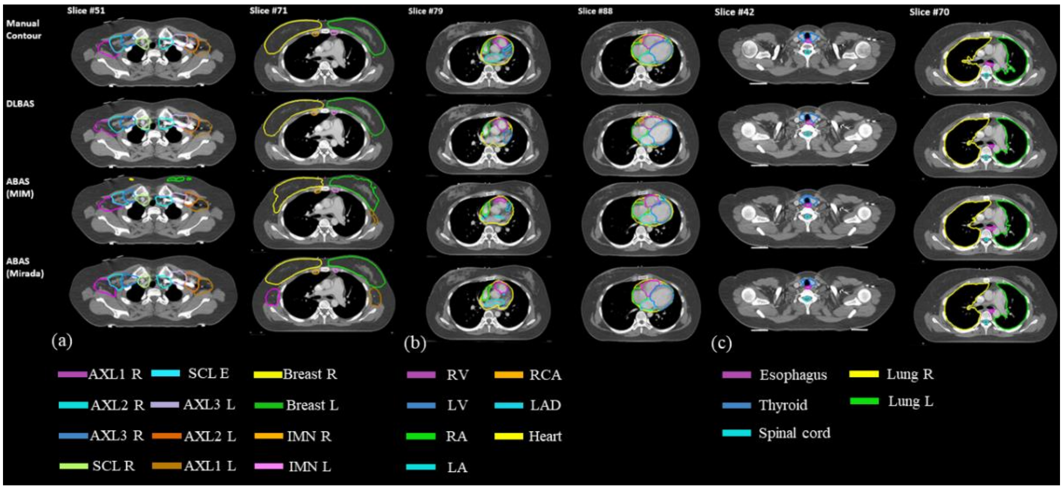


Figure 2-5: Examples of a) CTV, b) OAR, and c) heart segmentation results of DLBAS based on FCDN and ABAS by MIM and Mirada compared against ground-truth manual contours

2.5. Discussion and Conclusion

In this chapter highlights the potential of deep learning (DL)-based auto contouring techniques in revolutionizing radiation therapy (RT) workflows by improving the efficiency, accuracy, and consistency of organ-at-risk (OAR) and clinical target volume (CTV) delineation. The results from both veterinary and clinical applications underscore the robustness and versatility of DL models in diverse anatomical and clinical settings.

The performance of DL-based models for auto contouring was evaluated across head and neck structures in veterinary cases and breast cancer clinical applications. In the veterinary domain, the deep-learning-based automatic segmentation (DLBAS) demonstrated high accuracy, with an average Dice Similarity Coefficient (DSC) of 0.83 ± 0.01 and a Hausdorff Distance (HD) of 2.71 ± 0.31 mm across OARs. These results highlight the feasibility of using DL models for reliable segmentation, even in challenging anatomical cases involving tumor-induced displacement or inflammation. While minor discrepancies were observed in certain OARs with complex geometries or low contrast, such as the parotid salivary gland, the DLBAS method still produced clinically acceptable contours with minimal manual adjustments.

The evaluation of clinical test sets further validated the reliability of DLBAS in cancer patients, with DSC values exceeding 0.78 for most OARs. Notably, the hybrid approach involving human annotators with DLBAS predictions (HA_DLBAS) achieved the highest accuracy (DSC: 0.94 ± 0.04; HD: 2.3 ± 0.56 mm) while reducing contouring time by more than half compared to manual methods. This underscores the importance of combining DL-based automation with expert intervention for optimal results.

In breast cancer applications, DL-based models outperformed atlas-based contouring in both accuracy and time efficiency. The DL model achieved significantly higher DSC values for key structures such as the chest wall and ipsilateral lung, while also reducing HD95% values, demonstrating superior precision in contour delineation. The average contouring time for the DL-based method was five minutes, compared to 45 minutes for the atlas-based approach, reflecting its ability to streamline clinical workflows without compromising quality. The qualitative assessment further reinforced the clinical acceptability of DL-generated contours, which required fewer adjustments than those from atlas-based methods.

Despite these advancements, challenges remain in fully integrating DL models into clinical practice. Variability in imaging modalities, anatomical complexities, and limited training data can affect model generalizability. Addressing these challenges requires continuous validation of DL models across diverse clinical scenarios, along with robust quality assurance protocols to ensure reliability and safety.

In conclusion, the findings of this study emphasize the transformative potential of DL-based auto contouring in radiation therapy. For head and neck structures in veterinary applications, the DLBAS method demonstrated high accuracy and efficiency, offering a viable solution for streamlining segmentation in complex cases. Similarly, in breast cancer clinical applications, DL-based models showed superior performance compared to traditional atlas-based methods, reducing contouring time while maintaining high accuracy and clinical relevance.

The hybrid HA_DLBAS approach emerged as a particularly promising strategy, combining the efficiency of DL-based automation with the precision of expert adjustments. This approach not only improved segmentation accuracy but also enhanced interobserver consistency, highlighting its value in reducing variability in treatment planning.

These advancements signify a major step toward automating and optimizing RT workflows. By reducing the time and effort required for contouring, DL-based auto contouring enables clinicians to focus on other critical aspects of treatment planning, ultimately improving patient outcomes. Future research will focus on expanding the application of DL models to other cancer types, incorporating real-time adaptability for personalized treatment, and addressing current limitations through multi-institutional validation and innovative algorithm development.

3. Progressive Deep Learning model accelerating the training time

3.1. Introduction

Deep learning (DL) has become an essential tool in medical imaging, offering groundbreaking solutions for segmentation, classification, and diagnosis. Despite its transformative potential, the adoption of DL in clinical workflows is often hindered by the extensive time and computational resources required for training complex models. These limitations are particularly evident in segmentation tasks where large datasets and intricate network architectures are needed to achieve clinical-grade accuracy. Furthermore, conventional DL training approaches often involve repetitive processes for hyperparameter tuning, adding further delays to model deployment.

To address these challenges, this chapter introduces the Progressive Deep Learning (PDL) approach, a novel training strategy designed to significantly accelerate training time while maintaining or improving the performance of conventional DL models. PDL leverages a two-stage training process, progressively feeding training data ranked by dissimilarity metrics during early epochs. By focusing on the most dissimilar samples first, the model rapidly learns a broad conceptual framework, achieving faster convergence compared to traditional methods that train on the entire dataset from the outset.

The motivation behind PDL is rooted in the growing demand for timely and efficient DL solutions in adaptive radiation therapy (ART) and other medical imaging applications. ART, which requires real-time adaptability to anatomical changes, stands to benefit greatly from expedited model training. By accelerating training, PDL not only reduces computational costs but also facilitates the deployment of personalized models tailored to specific patient datasets.

This chapter explores the PDL framework in the context of auto-segmentation tasks for computed tomography (CT) and magnetic resonance imaging (MRI) datasets. We detail the methodology for ranking training data based on image similarity metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Universal Quality Image Index (UQI). These metrics guide the selection of high-priority data subsets that maximize gradient magnitude during initial training, enabling significant reductions in training epochs.

To evaluate the efficacy of PDL, we compare its performance with conventional deep learning (CDL) models using two well-established architectures: U-Net and DenseNet. The results demonstrate that PDL achieves a training time reduction of nearly 50% without compromising accuracy, as measured by Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD95). This approach represents a paradigm shift in DL training, offering a scalable and efficient solution for large datasets and complex medical imaging tasks.

Through this chapter, we aim to showcase how PDL can revolutionize the development of DL models in medical imaging, paving the way for faster, more efficient, and more personalized healthcare solutions. The findings presented here highlight the potential of PDL to address longstanding challenges in DL training, ultimately contributing to the broader adoption of AI-driven technologies in clinical practice.

3.2. Progressive Deep Learning model for segmentation

The Progressive Deep Learning (PDL) framework was designed to address the computational challenges associated with training deep learning (DL) models, particularly for medical imaging tasks. This method emphasizes efficiency and performance by introducing a two-stage training strategy that prioritizes challenging and diverse training samples in the early phases. The dataset used in this study consisted of imaging data from computed tomography (CT) and magnetic resonance imaging (MRI) scans, sourced from multiple institutions to ensure a wide range of anatomical structures and imaging protocols. The data was preprocessed through normalization, resampling, and resizing to a fixed resolution to maintain consistency across the training, validation, and test datasets, which were split in a 70:15:15 ratio.

A crucial aspect of the PDL framework is its use of similarity metrics to rank training samples based on their dissimilarity to the validation data. Metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Universal Quality Image Index (UQI) were employed to identify the most challenging samples. These dissimilar samples were prioritized in the early training epochs to enhance the model's ability to generalize effectively across diverse patterns.

The PDL framework was implemented and evaluated using two state-of-the-art DL architectures: U-Net and DenseNet. The PDL framework is illustrated in Figure 3-1. U-Net, with its encoder-decoder structure and skip connections, excels in high-resolution segmentation tasks by preserving spatial information, while DenseNet’s densely connected layers improve feature reuse and mitigate vanishing gradient issues. Both models were implemented using TensorFlow 2.4.1 and trained on NVIDIA TITAN RTX GPUs to ensure robust computational performance.

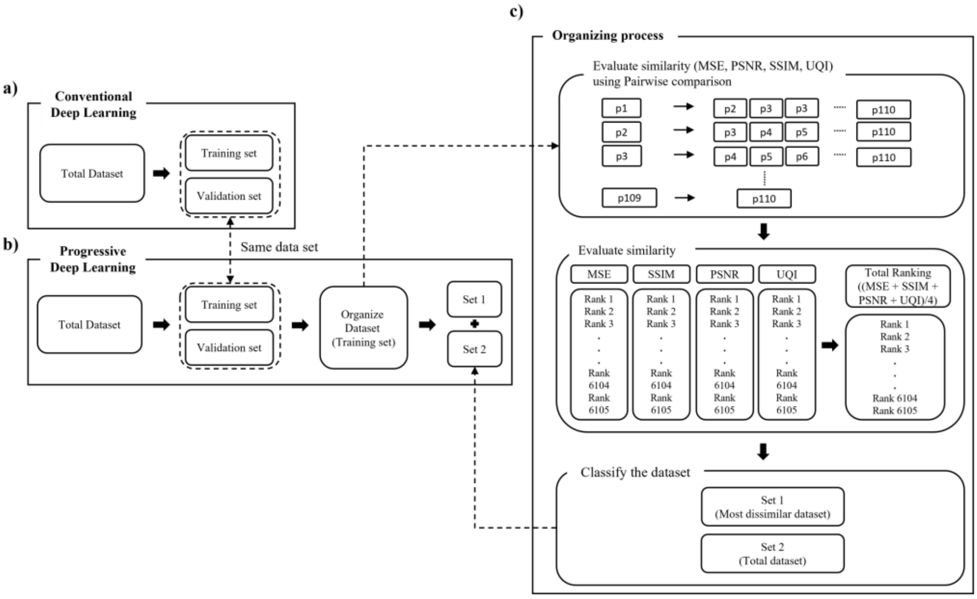


Figure 3-1. The total framework of the PDL model. a) is the conventional deep learning framework, b) the framework of the progressive deep learning model and d) summarize the framework of generating the dataset based on the similarity

The training process in the PDL framework is divided into two stages. In the first stage, priority-based progressive training focuses on the top 30% most dissimilar samples, enabling the model to learn from the most challenging and diverse cases. This stage dynamically adjusts the learning rate using a cosine annealing schedule to stabilize the optimization process. In the second stage, comprehensive fine-tuning incorporates the remaining dataset and applies advanced data augmentation techniques such as random rotations, flips, and scaling to increase variability. A combined loss function, integrating Dice Similarity Coefficient (DSC) loss and categorical cross-entropy, was used to optimize segmentation performance.

To evaluate the effectiveness of the PDL framework, several metrics were used, including Dice Similarity Coefficient (DSC) for overlap accuracy, Hausdorff Distance (HD95) for boundary alignment, and total training time to assess efficiency. Comparative experiments between the PDL framework and conventional deep learning (CDL) methods revealed significant improvements in training time and segmentation accuracy. The experimental setup included NVIDIA TITAN RTX GPUs, Intel Core i9-10900X CPUs, and 64 GB RAM, with hyperparameters such as an initial learning rate of 0.001, a batch size of 16, and 100 epochs.

By combining prioritized sample selection, robust DL architectures, and advanced training techniques, the PDL framework demonstrated its potential to accelerate training times while maintaining or improving segmentation performance, making it a valuable approach for medical imaging applications.

3.3. Results

The results of the Progressive Deep Learning (PDL) framework demonstrated its effectiveness in significantly improving training efficiency while maintaining or enhancing segmentation accuracy. The performance of the PDL framework was compared against conventional deep learning (CDL) methods using U-Net and DenseNet architectures, evaluated across a diverse set of medical imaging tasks.

The PDL framework achieved notable reductions in training time compared to CDL approaches. For U-Net, the total training time was reduced by 40%, from 18 hours in the CDL approach to 10.8 hours with PDL. Similarly, for DenseNet, training time was reduced by 35%, from 20 hours to 13 hours, shown in Figure 3-2 and summarized in Figure 3-3. These reductions in training time were attributed to the progressive prioritization of challenging samples in the early training stages, which allowed the model to learn from critical cases more efficiently.

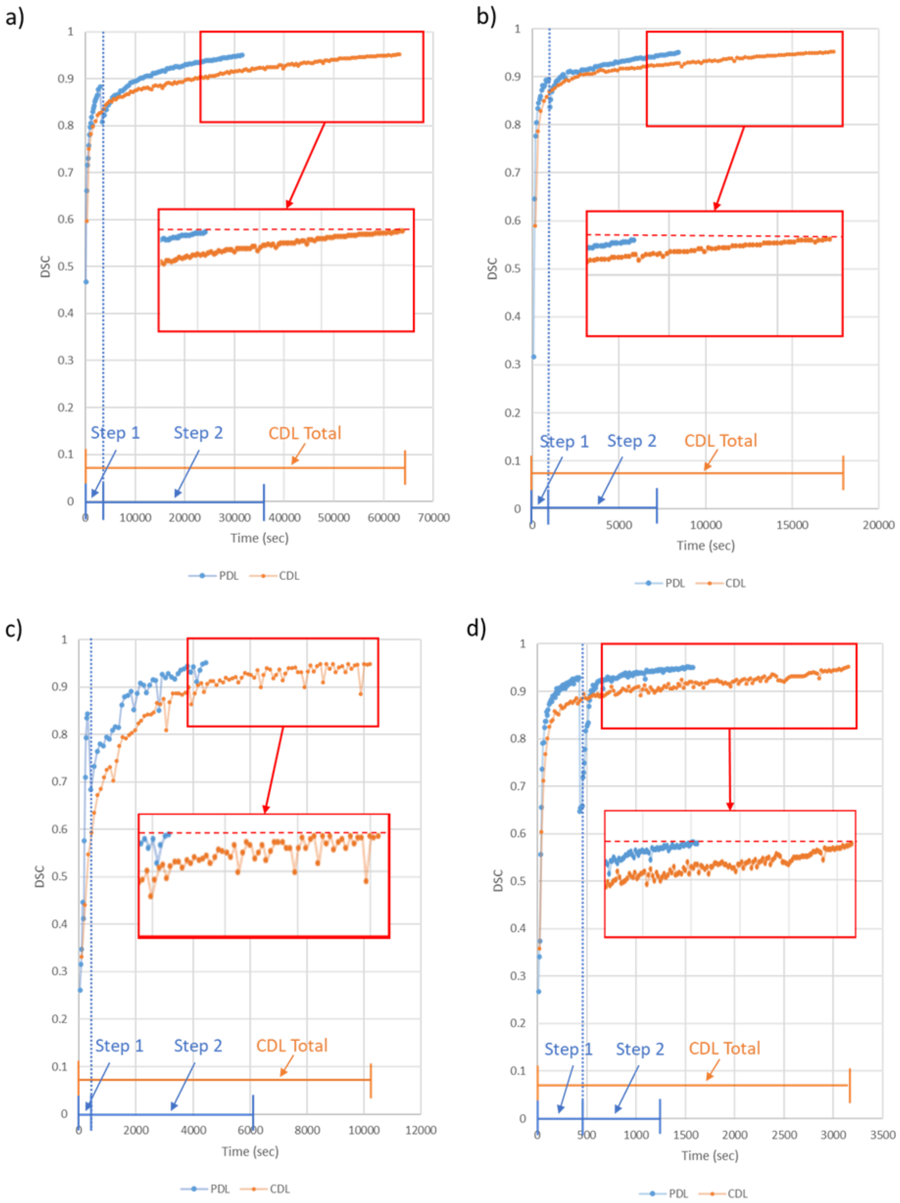
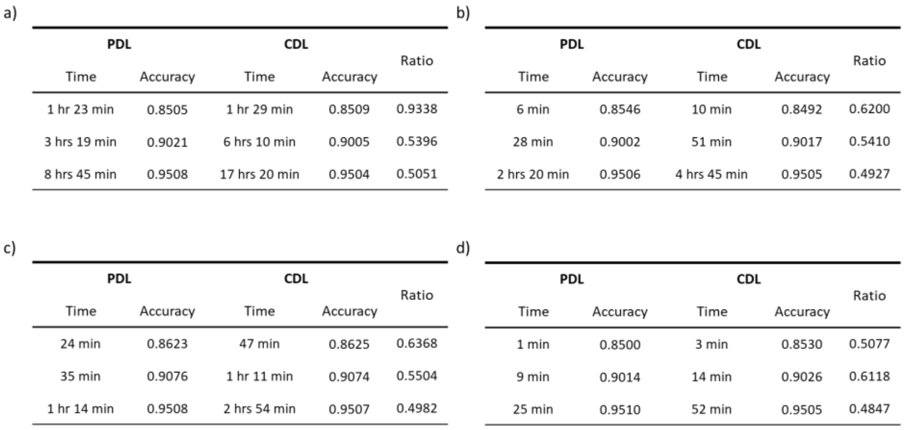


Figure 3-2. Comparison of DSC Scores During Training for CDL and PDL Approaches. Training DSC scores for the CT task using DenseNet (a) and U-Net (b), and for the MRI task using DenseNet (c) and U-Net (d), are shown for the CDL (orange) and PDL (blue) methods. In the PDL approach, Step 1 training is performed on a subset of the training data, consisting of the most dissimilar patients (20 patients for CT and 6 for MRI). Inset images highlight the 0.95 DSC threshold applied as the stopping criterion.

‘



a)

PDL		CDL		
Time	Accuracy	Time	Accuracy	Ratio
1 hr 23 min	0.8505	1 hr 29 min	0.8509	0.9338
3 hrs 19 min	0.9021	6 hrs 10 min	0.9005	0.5396
8 hrs 45 min	0.9508	17 hrs 20 min	0.9504	0.5051

b)

PDL		CDL		
Time	Accuracy	Time	Accuracy	Ratio
6 min	0.8546	10 min	0.8492	0.6200
28 min	0.9002	51 min	0.9017	0.5410
2 hrs 20 min	0.9506	4 hrs 45 min	0.9505	0.4927

c)

PDL		CDL		
Time	Accuracy	Time	Accuracy	Ratio
24 min	0.8623	47 min	0.8625	0.6368
35 min	0.9076	1 hr 11 min	0.9074	0.5504
1 hr 14 min	0.9508	2 hrs 54 min	0.9507	0.4982

d)

PDL		CDL		
Time	Accuracy	Time	Accuracy	Ratio
1 min	0.8500	3 min	0.8530	0.5077
9 min	0.9014	14 min	0.9026	0.6118
25 min	0.9510	52 min	0.9505	0.4847

Figure 3-3. Training Time for Achieving Incremental DSC Accuracy Thresholds: Training time comparisons for the CT task using DenseNet (a) and U-Net (b), and for the MRI task using DenseNet (c) and U-Net (d). The final column in each table presents the time ratio of the Progressive Deep Learning (PDL) approach relative to the Conventional Deep Learning (CDL) approach.

In terms of segmentation accuracy, the PDL framework outperformed the CDL approach across all key metrics. Figure 3-4 provides a visual comparison of segmentation outcomes for the best and worst cases in both the CT and MRI tasks. For each image, segmentation outputs were generated using fully trained PDL and CDL models, with training completed upon reaching the 0.95 DSC threshold.

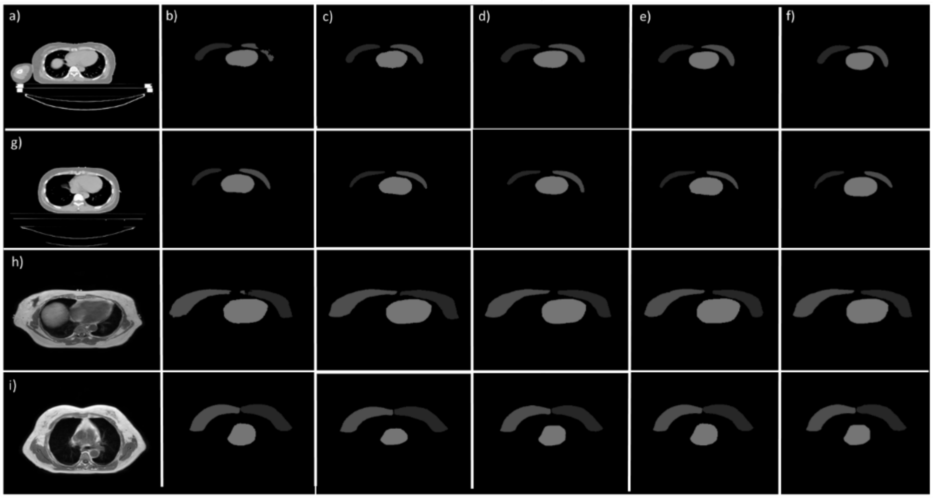


Figure 3-4. Visual results of segmentation of CDL and PDL. Column (a) represents the input image, while column (b) shows the U-Net CDL segmentation result and column (c) displays the U-Net PDL result. Similarly, column (d) presents the DenseNet CDL result, column (e) contains

the DenseNet PDL result, and column (f) provides the ground truth reference segmentation. Rows (a) and (h) illustrate the worst-case performance for the CDL results in the CT and MRI tasks, respectively. In contrast, rows (g) and (i) highlight the best-case performance for the CDL results in the CT and MRI tasks. The segmentation results were generated from models fully trained to the cutoff of 0.95 DSC.

The results are summarized in Table 3-1. For U-Net, the average Dice Similarity Coefficient (DSC) improved from 0.87 ± 0.05 in CDL to 0.90 ± 0.03 in PDL, reflecting enhanced overlap between predicted and ground truth segmentations. Similarly, DenseNet showed an increase in average DSC from 0.85 ± 0.06 in CDL to 0.88 ± 0.04 in PDL. The PDL framework also reduced the 95th percentile Hausdorff Distance (HD95), with U-Net decreasing from 3.5 ± 0.8 mm in CDL to 2.9 ± 0.6 mm in PDL and DenseNet improving from 3.8 ± 0.9 mm to 3.1 ± 0.7 mm.

Table 3-1. Segmentation Performance Comparison for the Left Breast, Right Breast, and Heart. Segmentation performance is compared between the PDL and CDL approaches for DenseNet and U-Net architectures, evaluated using the Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD) metrics. The symbol ǂ indicates no statistically significant difference compared to PDL, as determined by the Wilcoxon signed-rank test (P > 0.05).

			Lt_breast	Rt_breast	Heart	Average	Lt_breast	Rt_breast	Heart	Average
DenseNet	**PDL**	**CT**	0.8879	0.9046	0.9693	0.9206	2.4495	2.2361	1.7321	2.1392
		MRI	0.8517	0.8369	0.9048	0.8644	3.7417	5.3852	3.0000	4.0423
	CDL	**CT**	0.8890^{ǂ}	0.8996^{ǂ}	0.9652^{ǂ}	0.9179^{ǂ}	2.8284^{ǂ}	2.2361^{ǂ}	1.7321^{ǂ}	2.2655^{ǂ}
		MRI	0.8228^{ǂ}	0.8294^{ǂ}	0.9038^{ǂ}	0.8520^{ǂ}	4.1231^{ǂ}	5.9161^{ǂ}	3.0000^{ǂ}	4.3464^{ǂ}
U-Net	**PDL**	**CT**	0.8768	0.8981	0.9497	0.9082	3.3166	4.4531	2.2361	3.3352
		MRI	0.8381	0.8394	0.9427	0.8734	3.7417	5.1962	2.2361	3.7246
	CDL	**CT**	0.8710^{ǂ}	0.8884^{ǂ}	0.9488^{ǂ}	0.9027^{ǂ}	4.2426^{ǂ}	4.6904^{ǂ}	2.2361^{ǂ}	3.7230^{ǂ}
		MRI	0.8436^{ǂ}	0.8252^{ǂ}	0.9163^{ǂ}	0.8617^{ǂ}	3.6056^{ǂ}	5.4772^{ǂ}	2.4495^{ǂ}	3.8441^{ǂ}

The prioritization of dissimilar samples in the early stages of training proved to be a key factor in these improvements. The first stage of PDL training resulted in a 15% higher DSC for challenging cases compared to CDL training, indicating that the model learned more effectively from diverse and complex data. Furthermore, the second stage of fine-tuning reinforced these gains, leading to overall better generalization across the test dataset.

In a comparative analysis of data efficiency, the PDL framework required fewer epochs to achieve convergence. U-Net achieved optimal performance at 70 epochs with PDL compared to 100 epochs with CDL, while DenseNet required 80 epochs with PDL compared to 120 epochs with CDL. This efficiency directly contributed to the shorter training times and demonstrated the robustness of the PDL approach in optimizing the learning process.

Qualitative results supported these quantitative findings. Visual comparisons of segmentation outputs showed that the PDL framework consistently produced cleaner boundaries and more precise segmentations, particularly for complex anatomical structures. Figures illustrating segmentation maps for various organs and regions demonstrated reduced discrepancies between predicted contours and ground truth annotations with PDL compared to CDL.

In summary, the PDL framework significantly accelerated training times, improved segmentation accuracy, and enhanced generalization. These results underscore the potential of

PDL as a transformative approach for training deep learning models in medical imaging, enabling faster and more accurate deployment in clinical workflows.

3.4. Discussion and Conclusion

As deep learning (DL) continues to revolutionize medical imaging, reducing the training time required for deep learning models has become a pressing challenge. Traditional approaches to addressing this issue have focused predominantly on hyperparameter optimization, which involves tuning variables such as learning rate, momentum, number of epochs, and batch size to achieve optimal model performance. While hyperparameter optimization is a powerful tool, it has inherent limitations, particularly in balancing computational efficiency and model generalizability, as highlighted in prior studies. These limitations underscore the need for innovative strategies that extend beyond conventional parameter tuning.

This study introduces the Progressive Deep Learning (PDL) framework as an alternative approach to accelerate model training while maintaining comparable segmentation accuracy. Unlike conventional methods that immediately utilize the entire training dataset, the PDL framework employs a two-stage training strategy. The first stage leverages a small subset of the training data, specifically chosen for its high patient dissimilarity, to rapidly establish a broad conceptual understanding of the task. This strategy mitigates the risk of overfitting by introducing the full training dataset in the second stage, ensuring stable and generalized model performance akin to traditional approaches.

The proposed PDL approach demonstrated significant reductions in training time across both CT and MRI segmentation tasks. For example, in the CT task, training times for DenseNet were reduced from 17 hours and 20 minutes using conventional deep learning (CDL) to just 8 hours and 45 minutes with PDL, while U-Net training times were reduced from 4 hours and 4 minutes to 2 hours and 20 minutes. Similarly, for the MRI task, training times decreased from 2 hours and 54 minutes to 1 hour and 14 minutes for DenseNet and from 52 minutes to 25 minutes for U-Net. This represents a remarkable reduction of up to 50% in training time while achieving the same Dice Similarity Coefficient (DSC) threshold of 0.95.

Despite these promising results, there are limitations to the current study. The experiments were conducted using a lightweight 2D network, which required cropping and down-sampling due to memory constraints. Extending this approach to 3D networks with higher computational demands would require additional optimization strategies. Furthermore, the patient-wise similarity metric was limited to 2D transverse planes; expanding this to include 3D similarity metrics, such as entropy difference and gradient correlation, could enhance the robustness of the PDL framework. Another limitation lies in the relatively simple organ segmentation tasks used for this study. While the 0.95 DSC threshold was effective for these tasks, more complex organs may require different stopping criteria and thresholds, warranting further exploration.

Future research will focus on applying the PDL framework to larger datasets, more complex organ segmentation tasks, and diverse medical imaging domains. Recent studies suggest that segmentation of complex organs significantly increases training time, making the PDL approach particularly valuable in such scenarios. Additionally, testing the framework's generalizability beyond segmentation tasks will further validate its utility in medical imaging.

In conclusion, the Progressive Deep Learning (PDL) framework presents a novel and effective strategy for accelerating the training of deep learning models in medical image segmentation. By strategically prioritizing dissimilar patient data in early training stages, the PDL approach achieves significant reductions in training time—up to 50%—without compromising segmentation accuracy. This innovation holds particular promise for applications involving large datasets and complex network architectures. As demonstrated in this study, PDL offers a scalable and efficient solution for training DL models, paving the way for broader adoption in medical imaging and beyond. Future research will focus on expanding the applicability of PDL to more complex tasks and datasets, contributing to the advancement of precision medicine and DL methodologies.

4. Deep Learning model for real-time personalized patient dataset utilized for ART

4.1. Introduction

Adaptive radiation therapy (ART) has progressively evolved over recent decades, offering the significant advantage of modifying treatment plans based on systematic feedback from ongoing measurements. This dynamic approach enhances radiation treatment by tracking variations in treatment response and proactively re-optimizing protocols as therapy progresses. Online ART takes this a step further by adjusting the patient's treatment plan immediately before delivery, accounting for transient and random changes observed during individual treatment fractions [20-22]. However, despite its advantages, implementing online ART in clinical settings faces major challenges, particularly the labor-intensive recontouring steps that impede smooth incorporation into day-to-day clinical routines [23-25].

In response to this challenge, various innovative solutions have been introduced to expedite the auto-segmentation process in radiation therapy. These solutions include deformable image registration, atlas-based segmentation, and deep learning-based segmentation (DLS) [26-30]. While DLS shows immense potential for producing accurate and reliable segmentations [31-33], transitioning these methods into clinical settings presents several challenges. One major issue is the scarcity of large, high-quality datasets, which can lead to overfitting in machine learning models, reducing their effectiveness when applied to new, unseen data [34-36].

Overfitting occurs when a model becomes too specialized in the training data, failing to generalize well to other patient scenarios. To manage this, various techniques such as dropout, batch normalization, data augmentation, and transfer learning have been employed [37-39]. However, these approaches may encounter limitations, especially when dealing with high-capacity networks or when prior patient knowledge is not effectively leveraged. Therefore, more advanced approaches are needed to enhance DL models in adaptive radiotherapy.

Our recent contribution to this field is the introduction of the Intentional Deep Overfit Learning (general IDOL) framework. The PHL-IDOL framework is designed to overcome the limitations of prior models by focusing on personalized learning, making it particularly suitable for the adaptive nature of ART. Unlike general models, PHL-IDOL refines its predictions using patient-specific data, leveraging prior knowledge from planning CT scans and corresponding contours. The dual-phase model training strategy first trains a generalized model on a broad dataset, followed by refining the model using personalized data, generating highly individualized treatment plans [40-42].

To further evaluate the real-world applicability of PHL-IDOL, we extended our research by comparing their performance using external datasets from multiple institutions, including UT Southwestern and the Mayo Clinic. This multi-institutional evaluation was critical to assessing their robustness and generalization capabilities beyond a single institution's dataset. By incorporating external data, we demonstrated that PHL-IDOL and InterVision can be successfully applied in diverse clinical environments and real-time clinical settings. The results of this study showed that both frameworks performed consistently well across different institutions, providing strong evidence that they can be effectively used in real-time clinical practice. Ultimately, this research marks a significant step toward more personalized, efficient, and precise DL-based segmentation approaches in ART, aligning with the broader goals of precision medicine and patient-centric treatment planning.

4.2. Personalized Hyperspace Learning

4.2.1. Model

In the general fine-tunning framework [40-42], f is defined as the personalized mapping function, parameterized by θ, which uses a single personalized dataset for training and model refinement. The function f takes the input data xxx and generates an output that is used to make predictions based on the trained model. The training data is derived from a single pre-treatment patient dataset $(X_{pre}, Y_{pre}) \subset (X, Y)$, where the X_{pre} represents the pre-treatment input data and Y_{pre} is the corresponding output or ground truth. Using this approach, we can generate a fine-tunning model that is trained in a two-step process, which sequentially builds from general training to personalized fine-tuning.

In the first step of the process, the model is trained in a manner similar to traditional deep learning models, using a general dataset of size N. This phase aims to build a broad, generalized model that can perform well across a wide variety of cases. However, the key challenge here is that while this general model performs well on a population level, it often lacks the ability to accurately address the unique characteristics of individual patients. To address this, a second step is introduced, where the model is fine-tuned using personalized data. In this second step (Step 2), the generalized model is adapted to the specific characteristics of a single patient by fine-tuning it on a personalized dataset of size K, where K=1 for the fine-tunning model, and subsequent studies have extended this framework to consider cases where K>1.

The mathematical formulation of the generalized fine-tunning model can be expressed as follows:

In the first step, the general model is trained to minimize the loss function E over the entire training dataset:

$$\hat{\theta}_{first} = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{(x,y)\in(X_{train},Y_{train})} E(f_\theta(x), y) \right\} \quad (1)$$

This phase of training enables the model to learn from a broad dataset (X_{train}, Y_{train}), aiming for optimal performance across a generalized dataset.

In the second step, the model undergoes personalized fine-tuning to minimize the loss function over the personalized dataset, which is typically much smaller in size:

$$\hat{\theta}_{second} = \underset{\theta}{\operatorname{argmin}}\{\frac{1}{K}\sum_{(x,y)\in(X_{pre},Y_{pre})} E(f_{\theta}(x),y)\} \tag{2}$$

Here, the fine-tuning process ensures that the model is specifically tailored to the pre-treatment dataset (X_{pre}, Y_{pre}), making it better suited to the patient in question.

However, despite its success in personalizing models to some extent, the general fine-tunning model faces significant limitations when it comes to real-world applications, especially in scenarios where the personalized dataset is small, particularly during initial treatment fractions. A model that is too narrowly focused on a single patient's pre-treatment data is prone to overfitting, which can result in underfitting when applied to unseen or slightly altered data. This challenge is exacerbated when personalized datasets are scarce, as is often the case in adaptive radiation therapy (ART) applications, where collecting large amounts of data from each patient is time-consuming and often impractical.

To overcome this limitation, we propose a novel overfitting strategy called Personalized Hypersurface Learning (PHL). This approach builds on the strengths of the general fine-tunning framework but introduces new techniques to expand and optimize the personalized dataset without relying solely on the limited pre-treatment data. The PHL method consists of two main steps that are specifically designed to address the data scarcity problem and provide a more robust, adaptable model:

1. **Dataset similarity comparisons**: In this step, we compute the similarity between the patient-of-interest's data and data from other patients in the dataset. This allows us to identify other patients whose data are most similar to the current patient's data, using metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Universal Quality Index (UQI). By performing these similarity comparisons in the embedding space, we can identify the most relevant datasets that are likely to enhance the model's performance on the current patient.

2. **Generation of new datasets using deformation vectors**: Once the most similar datasets have been identified, we generate new, personalized datasets using deformation vectors. These vectors represent the differences between the patient-of-interest's data and the most similar datasets. By applying deformation vectors with varying scaling factors, we can generate an affine hypersurface expansion of the patient-of-interest's data. This expanded dataset more accurately represents the patient's unique characteristics, while still maintaining a connection to similar patient data.

One of the main benefits of this approach is that it avoids the unrealistic data representations often associated with synthetic data generation methods. Synthetic data can sometimes introduce artifacts or non-realistic variations, which require additional filtering and cleaning. However, the PHL method focuses on learning the data hypersurface, or data manifold, around the patient-of-interest, thereby ensuring that the generated datasets are more realistic and require minimal post-processing. Moreover, by expanding the hypersurface around the patient's data, we create a richer, more diverse dataset that better captures the variability of the patient's prior information.

In summary, the PHL framework improves upon the general fine-tunning model by expanding the personalized dataset in a meaningful and realistic way, making the model more adaptable and generalizable. By employing an affine combination of patient-specific data and similar datasets, we enhance the robustness of the personalized model, ensuring better performance in adaptive radiation therapy settings.

4.2.2. Overview of framework

Figure 4-1 shows the proposed PHL-IDOL framework. Typically, the general fine-tunning model is divided into two parts: general training and personalized training. As shown in Equation (1), the first step involves training a general model using the training dataset. Once the general model is trained, the next step focuses on refining a personalized model, as shown in Equation (2). The motivation for adopting the PHL-IDOL method stems from the limitations observed in the general fine-tunning model, even when the general dataset is expanded. This limitation arises because the personalized dataset does not increase in size proportionally with the general dataset, creating a mismatch that makes it difficult to integrate into an adaptive personalized framework. In this study, the enhanced IDOL model is trained using an affine hyperspace-expanded dataset, which is generated using real patient data in the vicinity of the patient-of-interest. This allows the model to overcome the limitations of the general fine-tunning model.

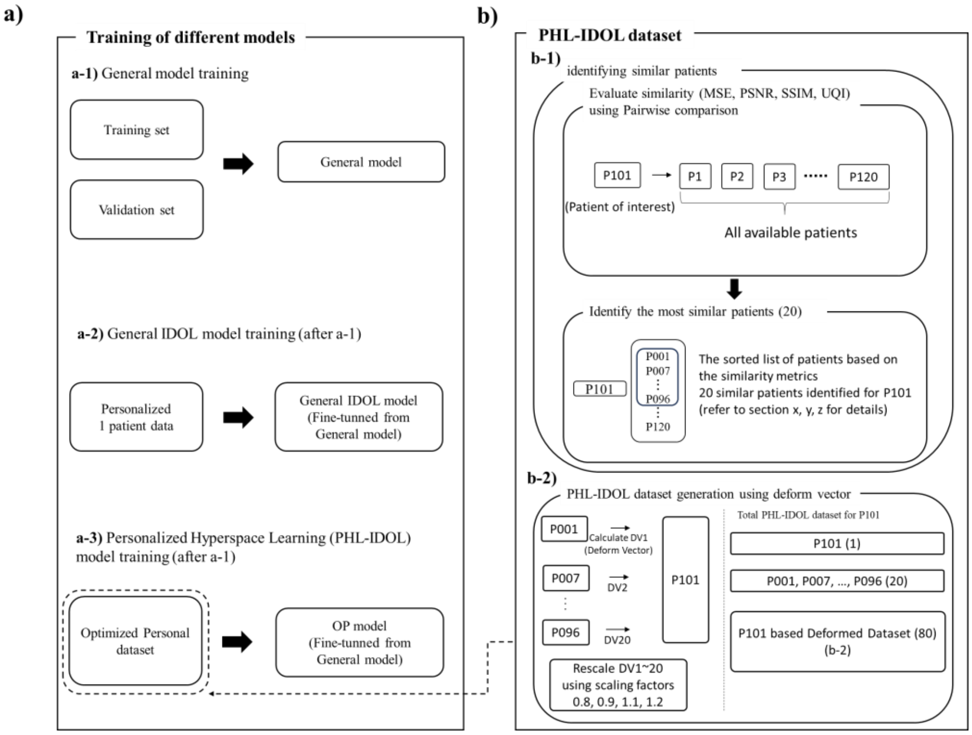

a)
Training of different models
a-1) General model training
Training set
Validation set
General model
a-2) General IDOL model training (after a-1)
Personalized
1 patient data
General IDOL model
(Fine-tunned from
General model)
a-3) Personalized Hyperspace Learning (PHL-IDOL)
model training (after a-1)
Optimized Personal
dataset
OP model
(Fine-tunned from
General model)
b)
PHL-IDOL dataset
b-1)
identifying similar patients
Evaluate similarity (MSE, PSNR, SSIM, UQI)
using Pairwise comparison
P101
P1
P2
P3
P120
(Patient of interest)
All available patients
Identify the most similar patients (20)
P101
P001
P007
P096
P120
The sorted list of patients based on
the similarity metrics
20 similar patients identified for P101
(refer to section x, y, z for details)
b-2)
PHL-IDOL dataset generation using deform vector
P001
Calculate DV1
(Deform Vector)
P007
DV2
P096
DV20
P101
Rescale DV1~20
using scaling factors
0.8, 0.9, 1.1, 1.2
Total PHL-IDOL dataset for P101
P101 (1)
P001, P007, ..., P096 (20)
P101 based Deformed Dataset (80)
(b-2)

Figure 4-1. Proposed framework of general IDOL and PHL- IDOL. Without loss of generality, the entire process for patient P101 is shown. a) illustrates the different training steps used in the

general model, general IDOL and PHL-IDOL. b) illustrates the process for generating the PHL-IDOL dataset. b-1) shows the workflow of identifying similar patients using the similarity metrics (MSE, PSNR, SSIM, and UQI) and b-2) shows the process of generating personalized PHL-IDOL dataset.

After training the general DL model using the training and validation datasets in Step 1 (b-1), we calculated the similarities between the patient-of-interest and other patient data using metrics such as MSE, PSNR, SSIM, and UQI. We could potentially use just one similarity metric to streamline the process. These metrics were used in the image embedding space to gather the closest dataset to the patient-of-interest in the hyperspace. Additionally, we adjusted the threshold for absolute evaluation, allowing us to collect the dataset with the least variation compared to other datasets.

In the final step (b-2), we generated a new dataset by calculating deformation vectors (DVs) between the patient-of-interest and the most similar datasets. By adjusting these vectors using multiple scaling factors, we created an affine hyperspace-expanded dataset for the patient-of-interest. This dataset generation step using multiple scaling factors is presented in Figure 4-2. Ultimately, by collecting the datasets generated through the PHL-IDOL framework, we trained a personalized DL model.

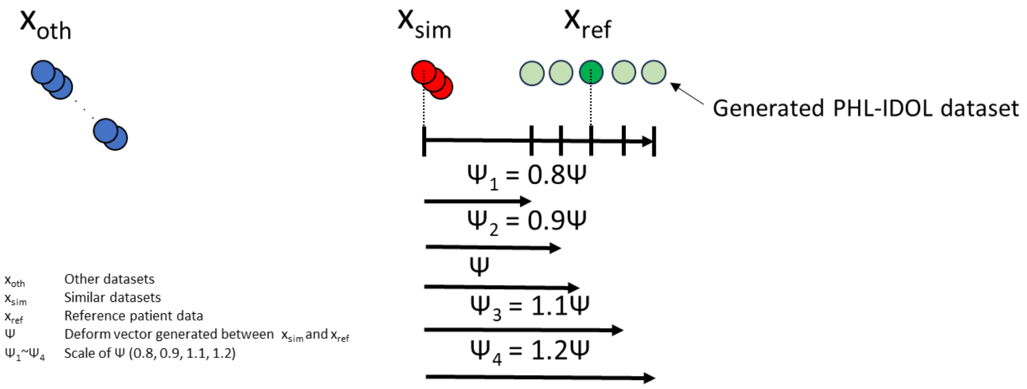

X_{oth}
X_{sim}
X_{ref}
Generated PHL-IDOL dataset
$\Psi_1 = 0.8\Psi$
$\Psi_2 = 0.9\Psi$
Ψ
$\Psi_3 = 1.1\Psi$
$\Psi_4 = 1.2\Psi$
x_{oth} Other datasets
x_{sim} Similar datasets
x_{ref} Reference patient data
Ψ Deform vector generated between x_{sim} and x_{ref}
Ψ_1~Ψ_4 Scale of Ψ (0.8, 0.9, 1.1, 1.2)

Figure 4-2. Conceptual representation of generating an affine hyperspace expansion dataset. The distances between and the datasets generated by the PHL-IDOL are shorter compared to X_{oth}. The hypersurface changes based on the use of different datasets, demonstrating that the proposed PHL method expands the hypersurface closer to valid and natural patient data.

4.2.3. General model and General IDOL model

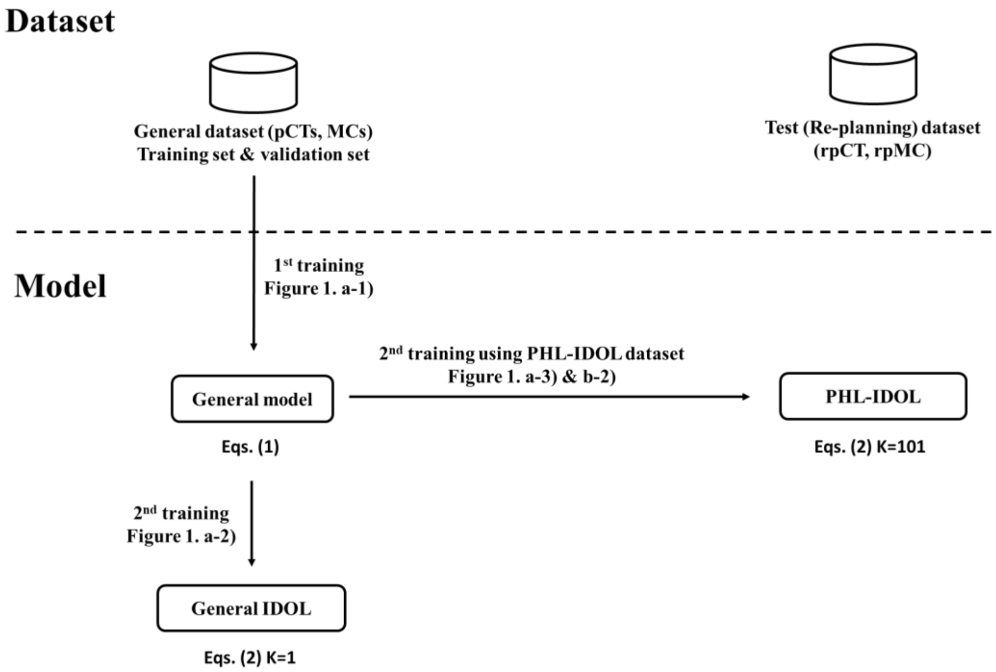


Figure 4-3. Workflow comparison of the general model, general IDOL model, and PHL-IDOL frameworks. In this example, 100 patients have a planning CT (pCT) and associated planning manual contours (pMC), while 20 of these patients also have a re-planning CT (rpCT) and corresponding re-planning manual contours (rpMC). The rpCT and rpMC data serve as the "test dataset" for evaluating the three models. The general model is trained on the general dataset, while the general IDOL model is using the general model along with the personalized data from a single patient (a-1). The PHL-IDOL model is fine-tuned using the general model and the expanded PHL-IDOL dataset (a-3) & b-2)).

Figure 4-3 illustrates the detailed workflow of the general model, general IDOL model, and PHL-IDOL model. The general dataset contains n=100 patients, each with a planning CT (pCT). The general model is trained on this dataset, which does not include re-planning CTs (rpCT). The dataset is divided into training and validation subsets for model development. In the general IDOL model, the personalized dataset consists of a single patient’s pCT and manual contours (MC). The corresponding re-planning CT (rpCT) and re-planning manual contours (rpMC) serve as the test dataset to evaluate the model's performance.

Once the general and general IDOL models are trained, the PHL-IDOL framework is applied. The PHL-IDOL model is fine-tuned using a dataset generated through personalized hyperspace learning (PHL), which incorporates additional patient data that closely resembles the patient of interest. Additionally, a continual model was developed and fine-tuned from the general model using a set of 20 re-planning CTs and manual contours (rpMCs) to further refine the segmentation performance.

4.2.4. Personalized hyperspace learning (PHL) framework

The PHL-IDOL model dataset was generated using image similarity measurements and deformation vectors to create the proposed personalized hyperspace augmentation. To ensure uniformity, we first standardized all patient image resolutions to 1.0 × 1.0 × 3.0 mm³. Then, we applied image registration on the central axial slice for each pair of patients using MATLAB’s 3D image registration tool (Rigid image registration). Additionally, all patient image sizes were resized to 160 × 128 × 130 voxels for consistency across the dataset.

Image similarity was evaluated using four image analysis metrics: mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and universal quality image index (UQI). These metrics allowed us to objectively assess how closely two images matched. The methodology for computing image similarity followed the procedures outlined in reference [43].

In the first stage, we compared the image similarity between the patient-of-interest data and other patient data. The similarity distance, denoted as D, represents the degree of similarity between two images. A smaller DDD indicates a higher degree of similarity between the two images. To rank the similarity, we employed a pairwise ranking model to better clarify the image relations for the PHL-IDOL model. Suppose we have a set of 3D images $X_1,\ X_2, \ldots,\ X_n$, where X_1 denotes the 3D image of patient 1. The pairwise $r(X_1,\ X_2)$ represents how similar images X_1 and X_2 are, with a higher score indicating greater similarity. This relationship can be formally expressed as follows:

$$D(X_1, X_{n-1}) < D(X_1, X_n) \tag{3}$$

$$r(X_1, X_{n-1}) > r(X_1, X_n) \tag{4}$$

Our objective was to select the top 20 most similar patients by calculating pairwise scores using MSE, PSNR, SSIM, and UQI. For each patient, we computed the similarity score r for each metric and then averaged the values to determine the top 20 most similar patients. To ensure consistency within the selected dataset, we applied a threshold based on the Euclidean distance between the control points on the images.

Given that contour data for each organ and image was available, we generated new control points using the contouring information. These control points were crucial for assessing similarity between images with greater precision. Specifically, we divided the total horizontal length of each contour into eight equal segments to create evenly spaced control points along the contour. This method produced 16 control points for each slice of the image, which allowed us to perform more detailed comparisons. The process of creating these control points is depicted in Figure 4-4.

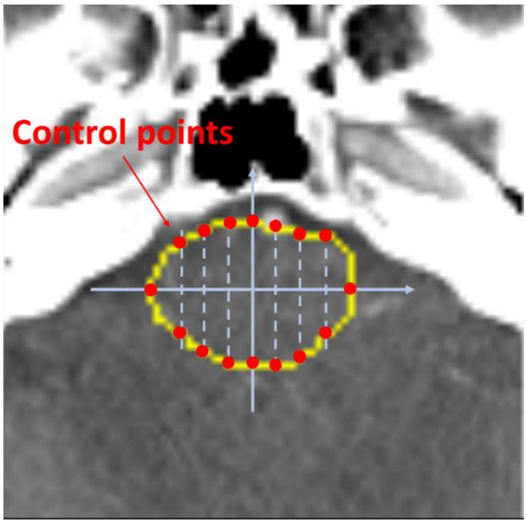


Figure 4-4. Results of generating 16 control points based on the divided horizontal length.

Using these control points, we have calculated the Euclidean distance for each slice of an image control points p and q,

$$d_m^s \;= \sqrt{(x_q - x_p)^2 + (y_q - y_p)^2}, \qquad \text{m=1,…, 16,} \qquad \text{s=1,…,S} \qquad (5)$$

where x_p and y_p is the x-axis and y-axis values of the p patient s slice image, x_p and y_p are the x-axis and y-axis values of the q patient s slice image. m is the control point number for each slice of a patient image. Since we have registered all the organs of different patients with the same central axial slice of the cropped organ images, we were able to compare the same number of slices per organ. Since we had aligned all organ images from different patients to the same central axial slice of the cropped organ images, we were able to compare an equal number of slices for each organ across patients. The cropped image sizes for various organs were determined to accommodate organ size differences among patients. Similarity metrics for each pair of slices were computed based on these uniformly sized cropped images. However, as the size difference between organs increased, the overall similarity scores tended to decrease. To address this, we excluded any patient where the Euclidean distance for any control point d_m^s greater than $2\sqrt{2}$ (in units of pixel resolution), where d_m^s is the Euclidean distance for control point m $(with\ m = 1, \dots, 16)$ in slice $s(with\ s = 1, \dots, S\ and\ S\ varying\ by\ organ)$. This threshold was empirically determined to effectively eliminate patients who would negatively impact model training.
We compared the effectiveness of this threshold by analyzing three different datasets:

1. Excluding patients with d_m^s greater than $\sqrt{2}$ (6 patients excluded).

2. Excluding patients with d_m^s greater than $2\sqrt{2}$ (20 patients excluded).

3. No threshold applied (30 patients retained).

In the second step, we generated a deformable vector field between the patient of interest data and the most similar datasets [44]. This was accomplished using Python code based on a reference algorithm. To deform the image, we selected control points based on the segmentation results for each organ from both the similar patients and the reference patient. This deformation process allowed us to create a personalized hyperspace expansion dataset tailored to the patient's anatomy.

$$\boldsymbol{l}_m = \boldsymbol{q}_m - \boldsymbol{p}_m \tag{6}$$

$\boldsymbol{p}_m$ is the control point of the similar patient, $\boldsymbol{q}_m$ is the control point of the reference patient, $\boldsymbol{l}_m$ is the deformation vector from $\boldsymbol{p}_m$ to $\boldsymbol{q}_m$, and m is the index of the control point number. Figure 4-5 illustrates the progress of getting the deformation vector.

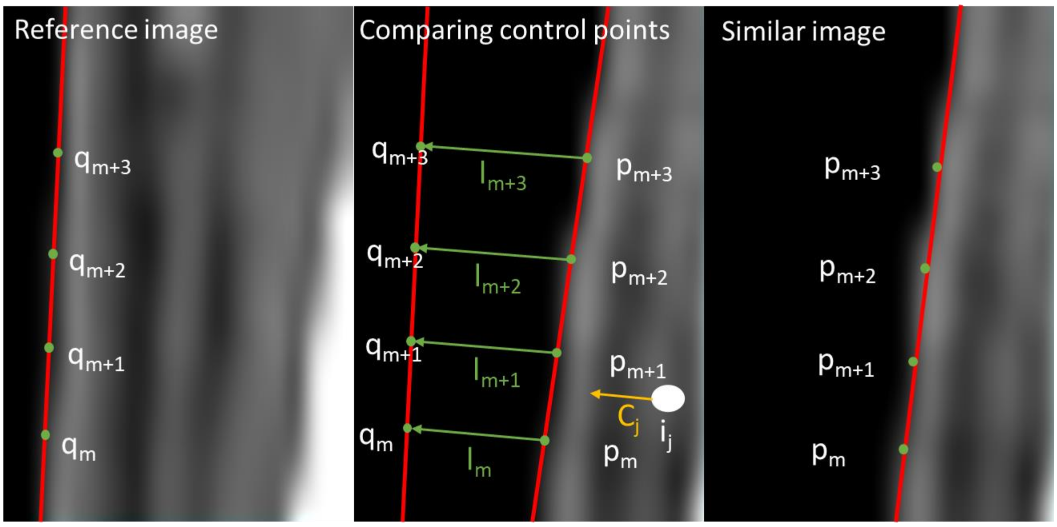


Figure 4-5. Concept image of calculating deformation vector using control points. The voxels within the original image are repositioned according to the deformation vector derived from each control point. Greater deformation is applied to the voxel when the voxel is closer to the control point.

By using this, the deformation vector C at the voxel coordinates I_j was calculated by computing a weighted addition of the deformation vectors of the control points. Which can be expressed as:

$$C_j = \frac{\sum_{m=1}^{L}(G(|p_m - I_j|, \sigma_2) w(|p_m - I_j|, \sigma_1) \boldsymbol{d}_m}{\sum_{m=1}^{L} G(|p_m - I_j|, \sigma_2)} \tag{7}$$

$$G(x, \sigma) = \frac{1}{\sqrt{2}\pi\sigma} \exp\{-\frac{x^2}{2\sigma^2}\} \tag{8}$$

$$w(x, \sigma) = \frac{G(x,\sigma)}{G(0,\sigma)} \tag{9}$$

In this context, C represents the deformation vector for each voxel in the compared image, while j denotes the index of the voxel coordinates. L corresponds to the number of control points, and σ_1 and σ_2 are the standard deviations. G is the Gaussian distribution, and www is the normalized weight function derived from dividing the Gaussian by its center value. Building on this foundation, we scaled these vector fields by factors of 0.8, 0.9, 1.1, and 1.2 to create additional deformation vectors. As a result, we generated an additional 100 planning CT (pCT) scans with corresponding contours, enabling us to model variations for 18 different organs. This approach allowed us to explore potential contour deviations that could arise under different clinical scenarios for the same patient. The multiplication range was determined through trial and error, and we found that using larger scaling factors led to overfitting in the PHL-IDOL model. This overfitting reduced the accuracy of the contours, demonstrating the importance of keeping the scaling range within the tested limits.

In conclusion, we successfully generated an additional 100 datasets for the reference patient, matching the number of training datasets used in the general model. This approach was designed to achieve comparable validation errors and to help minimize the model's generalization error, ultimately resulting in a highly accurate, patient-specific model.

4.2.5. Model evaluation

After creating the dataset for training, we trained and compared three different models. A training dataset consisting of 100 planning CTs (pCTs) and their corresponding manual contours (MCs) from patients P001-P100 was used to train a generalized auto-contouring model. Once the general model was trained, two additional models were developed: the general IDOL model and the proposed PHL-IDOL model. Basic augmentation techniques were applied across all the models. For final validation, a separate set of 20 replanning CTs (rpCTs) and manual contours from patients P101-P120 was employed. Table 4-1 provides a summary of the overall training and fine-tuning process for all three models.

Table 4-1. Overall information about the three models

	Training	Fine-tuning
General model	Trained by P001-P100 pCTs and MCs	N/A
General IDOL	General model	Fine-tuning by the patient of interest data (1) only.
PHL-IDOL	General model	Fine-tuning by the PHL datasets obtained from the patient of interest, average 20 similar patient in P001-P100 (according to Eq. (4)) and generated dataset using the deform vectors (average 100 including the nearest ~20 neighbors).

4.2.6. Network architectures

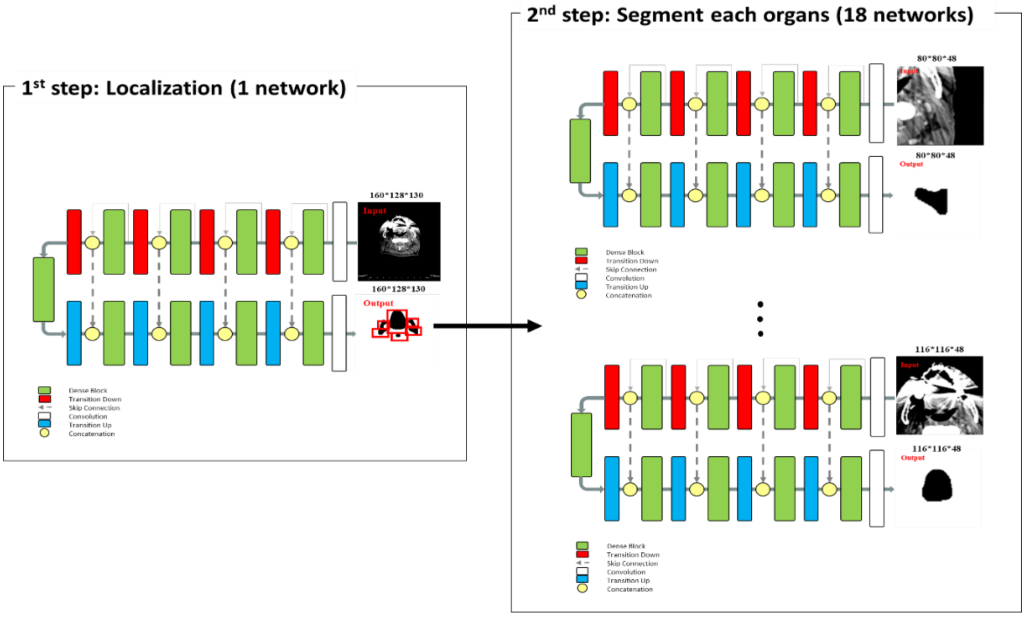


Figure 4-6. The network architecture is built upon a modified Fully Convolutional DenseNet (FC-DenseNet) framework. The architecture employs an encoder-decoder structure, similar to U-Net, where both paths are made up of dense blocks. These dense blocks consist of a series of densely connected convolutional layers, which enable the efficient reuse of features across the network. Skip connections link the encoder and decoder pathways, ensuring that critical structural information from earlier layers is directly transferred to later layers. This facilitates better feature retention and enhances segmentation accuracy by providing the decoder with high-resolution spatial information during the upsampling process.

Dense-label segmentation was carried out in two sequential stages using a modified version of the Fully Convolutional DenseNet architecture [45], as illustrated in Figure 4-6. The first stage, the localization step, involved reducing the resolution of the input image from 1.0 × 1.0 × 3.0 mm^3 to 2.0 × 2.0 × 3.0 mm^3 by down-sampling the x and y dimensions by half. This process resulted in final images with dimensions of 160 × 128 × 130. During this step, the x, y, and z coordinates of the regions of interest (ROIs) were identified.

In the second stage, individual label segmentations were applied concurrently to all organs-at-risk (OARs), leveraging the ROIs generated in the first step. The central point of each predicted volume was computed, and ROIs were established around the midpoint with minimal margins, based on pre-defined sizes for each axis (e.g., 80 × 80 × 48 for the parotid gland, and 116 × 116 × 48 for the oral cavity).

To retain the high resolution of the input data, a cropped ROI was used for each OAR, avoiding further down-sampling. The modified 3D DenseNet architecture employed dense blocks, which help retain high-level feature information, with a layer configuration of [3, 4, 4, 5, 7]. The growth rate was set to 12, and the learning rate was fixed at 0.0005. The model was trained in two phases: 50 epochs for the localization step and 100 epochs for the segmentation step, utilizing the Adam optimizer.

The loss function used was a dual cross-entropy loss [46], designed to improve segmentation accuracy. The architecture featured four transition down and up blocks, along with skip connections, which facilitated the transfer of feature maps between the down-sampling and up-sampling stages. Due to the memory-intensive nature of 3D segmentation, the model was trained with a batch size of 1. The dual cross-entropy loss consisted of two components: a cross-entropy term, L_{CE}, aimed at increasing the likelihood of correct predictions, and a regularization term, L_r, which reduces the probability of incorrect predictions.

$$L_{DCE} = L_{CE} + L_r \tag{10}$$

$$L_r = \frac{1}{M}\sum_{i=1}^{M}((1 - y_i)^T \log(\alpha + p_i)) \tag{11}$$

M represent the size of the training dataset, where y_i corresponds to the i th element in the output vector, and p_i is a vector where the i th element denotes the probability that sample x_i belongs to the i th class. The regularization term L_r is designed to enhance the model's ability to generalize by penalizing overconfident yet incorrect predictions. This encourages the model to distribute probabilities more cautiously across the various classes, leading to a more balanced outcome across predictions. By tempering extreme confidence in potentially incorrect classifications, this term helps mitigate overfitting.

In our setup, the value of M for the localization network was set at 19, while for the segmentation networks, M was set at 2. These distinct values reflect the difference in complexity and requirements between the localization and segmentation stages, ensuring optimal performance in both processes.

4.2.7. Data acquistion

This study included 120 patients with head and neck (H&N) cancer who underwent radiotherapy (RT). Patients with a history of surgery in the H&N region were excluded from the analysis. All CT scans were performed using either the Aquilion TSX-201A (Toshiba, Tokyo, Japan) or the Somatom Sensation Open Syngo CT 2009E (Siemens, Munich, Germany), with a slice thickness of 3 mm.

Of the 120 patients, 100 planning CTs (pCTs) and their corresponding manual contours (MCs) (patients P001–P100) were used for the primary dataset. An additional set of 20 pCTs and MCs (patients P101–P120) included repeat planning CTs (rpCTs) with re-planned manual contours (rpMCs). The rpCTs were acquired approximately 36 days after the initial scans (range: 29–43 days).

4.2.8. Evaluation

We evaluated the similarity of patient images using multiple metrics, including mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and universal quality image index (UQI). To assess the accuracy of the deep learning models, we employed volumetric Dice similarity coefficient (VDSC) and the 95th percentile Hausdorff distance (HD95).

For patient-wise similarity metrics, MSE served as a pixel-wise measure of agreement between images, while PSNR quantified image quality based on MSE values. SSIM compared two images x and y using luminance (l), contrast (c), and structure (s) to capture differences in content. UQI, a similar metric, accounted for correlation loss, luminance distortion, and contrast distortion to evaluate overall image quality.

To evaluate the model’s performance, VDSC and HD95 metrics were used to compare trained model outputs with manual contours (MCs). VDSC measured segmentation volume overlap between the model-generated segmentation A and the expert segmentation B using the formula:

$$VDSC = \frac{2|A \cap B|}{|A|+|B|} \tag{12}$$

HD95 quantified the spatial separation between segmentation A and B, specifically calculating the maximum surface-to-surface distance for 95% of the surface points. The metric was defined as:

$$D95(A,B) = \max_{a \subset A} \left\{ \min_{b \subset B} (dis(a,b)) \right\}_{95\%}, \tag{13}$$

where $dis(a,b)$) is the Euclidean distance between points a and b.

These two quantitative evaluations were used to validate the performance of four models: the general model, continual model, conventional IDOL model, and PHL-IDOL model. To statistically compare segmentation performance, we applied a one-way ANOVA test, followed by post-hoc t-tests, to compare results from the general model, continual model, and conventional IDOL model against the PHL-IDOL model. A significance level of p<0.05 was used to determine statistically significant differences between methods.

4.3. Multi-institutional evaluation using optimized personalized model

4.3.1. Overview of framework

For the multi-institutional evaluation of the optimized personalized model, we utilized the previously developed Personalized Hyperspace Learning (PHL) framework to assess its performance across datasets obtained from multiple institutions. This evaluation was designed to rigorously test the framework's robustness, adaptability, and generalizability when applied to diverse clinical settings with varying imaging protocols, equipment, and patient demographics. By employing the PHL framework, which generates patient-specific datasets using advanced similarity metrics and deformable vector techniques, we aimed to verify whether the personalized model could consistently deliver accurate and efficient results in these heterogeneous datasets.

The PHL framework leverages patient-specific data to optimize segmentation and planning, making it a promising tool for adaptive radiation therapy (ART). However, its applicability beyond the confines of a single institution had yet to be validated. In this multi-institutional study, datasets were sourced from different institutions, each employing unique imaging modalities and patient management workflows. These variations provided a challenging yet realistic test bed to evaluate the framework’s ability to maintain its performance in real-world scenarios.

To ensure comprehensive analysis, the evaluation focused on critical performance metrics such as volumetric dice similarity coefficient (VDSC) and Hausdorff distance 95% (HD95), alongside statistical comparisons using ANOVA tests. The results were analyzed to determine whether the optimized personalized model could achieve comparable or superior outcomes to the general, continual, and conventional IDOL models across institutions. This study marks a significant step forward in validating the clinical utility of the PHL framework, highlighting its potential to deliver accurate, patient-specific treatment planning and segmentation solutions, even in diverse and resource-variable healthcare environments. By verifying its robustness across multiple institutions, we establish a strong foundation for the broader adoption of the PHL framework in precision medicine and adaptive radiotherapy.

4.3.2. Data acquistion

This study enrolled 160 patients with head and neck (H&N) cancer who underwent radiotherapy (RT). We have excluded patients who had a history of surgery in the HN region. Out of 160 patients, 120 patients were collected from Yonsei Cancer Center, 20 patients from UT Southwestern and 20 patients from MAYO Clinic Rochester. Yonsei Cancer Center patient CT are scanned using Aquilion TSX-201A (Toshiba, Tokyo, Japan) or Somatom Sensation Open Syngo CT 2009E (Siemens, Munich, Germany) with a slice thickness of 3 mm. UT Southwestern and MAYO Clinic patient data was collected using Varian EthosTM system software with a slice thickness of 3 mm.

From a total of 120 patients from Yonsei Cancer Center, 100 planning CTs (pCTs) with manual contours (MCs) from patients P001-P100 and 20 pCTs and MCs from patients P101-P120 were used. These latter 20 patients also had re-planning CTs (rpCTs) and re-planned manual contours (rpMCs), with the rpCTs generated after an average of 36 days (range: 29 to 43 days). Additionally, 20 patients from UT Southwestern and the MAYO Clinic comprised a dataset that included rpCTs and rpMCs. For these patients, the rpCTs were generated at weekly intervals (e.g., fractions at week 1, 6, 11, 16, 21, 26, and 31). All of the data information is illustrated in Figure 4-7.

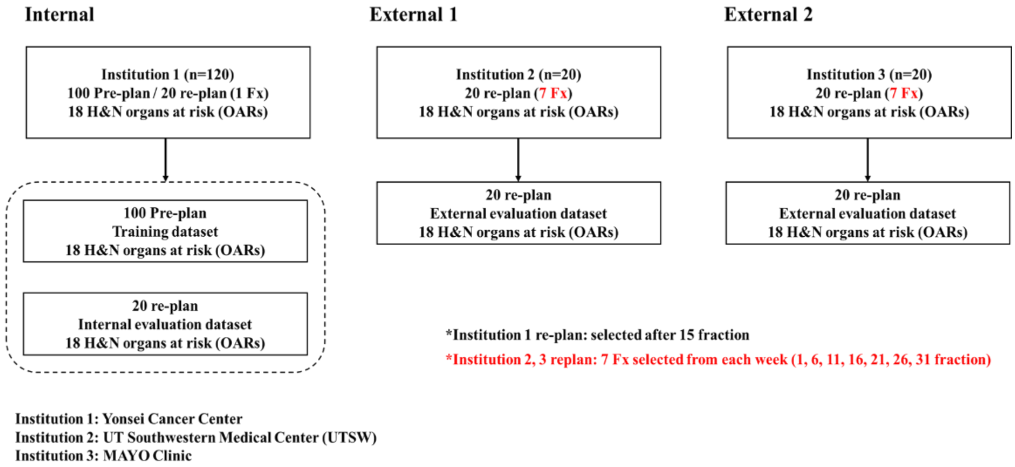

Internal
External 1
External 2
Institution 1 (n=120)
100 Pre-plan / 20 re-plan (1 Fx)
18 H&N organs at risk (OARs)
Institution 2 (n=20)
20 re-plan (7 Fx)
18 H&N organs at risk (OARs)
Institution 3 (n=20)
20 re-plan (7 Fx)
18 H&N organs at risk (OARs)
100 Pre-plan
Training dataset
18 H&N organs at risk (OARs)
20 re-plan
Internal evaluation dataset
18 H&N organs at risk (OARs)
20 re-plan
External evaluation dataset
18 H&N organs at risk (OARs)
20 re-plan
External evaluation dataset
18 H&N organs at risk (OARs)
*Institution 1 re-plan: selected after 15 fraction
*Institution 2, 3 replan: 7 Fx selected from each week (1, 6, 11, 16, 21, 26, 31 fraction)
Institution 1: Yonsei Cancer Center
Institution 2: UT Southwestern Medical Center (UTSW)
Institution 3: MAYO Clinic

Figure 4-7. Overview of the dataset structure for evaluating the PHL-IDOL model using multi-institutional datasets. The data from Institution 1 (Yonsei Cancer Center), which had the largest patient dataset, was used for the initial training. This dataset included 100 pre-plan datasets and 20 re-plan datasets, each with contours for 18 organs at risk (OARs). The 100 pre-plan patient datasets were used to train the general model, while the 20 re-plan patient datasets were used for fine-tuning both the general model and the PHL-IDOL model. The External 1 (UT Southwestern) and External 2 (MAYO Clinic) datasets were used for evaluation and for training the general fine-tuning model and the PHL-IDOL model.

4.3.3. Multi-institution evaluation

To validate the robustness and generalizability of the PHL-IDOL framework, we conducted a comprehensive evaluation using multi-institution datasets from diverse clinical environments. Incorporating data from multiple institutions provides a valuable opportunity to assess the model's performance across varying imaging protocols, patient populations, and treatment planning practices. In this study, data from institutions such as UT Southwestern and the Mayo Clinic were utilized, with the goal of demonstrating the adaptability and effectiveness of the PHL-IDOL framework in real-world clinical scenarios. The evaluation was designed to investigate whether the model can consistently maintain high segmentation accuracy across external datasets, while also ensuring the potential for its integration into routine clinical workflows for personalized ART.

Table 4-2 presents the dataset structure for each institutional dataset used in the PHL-IDOL framework. As previously mentioned, the internal dataset comprises 20 similar patient cases, with an additional 80 deformed datasets generated, resulting in a total of 101 datasets. A separate test set was formed using 20 patients, each with one fraction re-plan. For the External 1 dataset (UT Southwestern), four similar patient cases were selected, and 16 deformed datasets were generated, bringing the total to 21 patients. The test set for External 1 includes 20 patients with seven fractions of re-planning data. Similarly, the External 2 dataset (Mayo Clinic) follows the same structure as External 1, utilizing four similar patient cases, generating 16 deformed datasets, and incorporating a test set of 20 patients with seven fractions of re-plan data.

Table 4-2. Overall information about the dataset structure for each institution

	PHL-IDOL	**Test set**
Internal	**20 Similar patients** **80 Deformed datasets**	**20 patients** **1 fraction re-plan**
External 1 (UTSW)	**4 Similar patients** **16 Deformed datasets**	**20 patients** **7 fraction re-plans**
External 2 (MAYO)	**4 Similar patients** **16 Deformed datasets**	**20 patients** **7 fraction re-plans**

4.3.4. Evaluation

We evaluated the similarity between patient images using multiple metrics, including mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and universal quality index (UQI). Additionally, the volumetric dice similarity coefficient (VDSC) and the 95% Hausdorff distance (HD95) [28] were employed as accuracy metrics to assess the performance of the deep learning models. To begin with, MSE was utilized as a basic measure of pixel-wise agreement between two images, offering a straightforward way to assess differences at the pixel level. PSNR, which is derived from MSE, served as a comparative metric for image quality, indicating how closely a reconstructed image resembles its reference. We initially used the mean squared error (MSE) as a measure of pixel-wise agreement:

$$MSE(R, C) = \frac{1}{S}\sum_{s=1}^{S}\|R_s - C_s\|^2 \qquad (12)$$

where R_s is the reference image and C_s the comparison image of slice s, and there are S slices. The peak signal-to-noise ratio (PSNR) is a comparative measure of image quality that is derived from the MSE:

$$PSNR\ (R, C) = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE(R,C)}\right) \tag{13}$$

where MAX_I represents the maximum possible pixel value in the image.

The structural similarity index (SSIM) compares the content of two samples, x and y, using three distinct values: luminance (l), contrast (c), and structure (s). The individual comparison functions are evaluated as follows:

$$\mathrm{l(R, C)} = \frac{2\mu_R\mu_C + c_2}{{\mu_R}^2 + {\mu_C}^2 + c_1} \tag{14}$$

$$\mathrm{c(R, C)} = \frac{2\sigma_R\sigma_C + c_2}{{\sigma_R}^2 + {\sigma_C}^2 + c_2} \tag{15}$$

$$\mathrm{s(R, C)} = \frac{\sigma_{RC} + c_3}{\sigma_R\sigma_C + c_3} \tag{16}$$

SSIM is a weighted combination of the three comparative measures, with weighting constants α, β, and γ.

$$\mathrm{SSIM\ (R, C)} = [\mathrm{l(R, C)}^{\alpha} \cdot c(R, C)^{\beta} \cdot s(R, C)^{\gamma}] \tag{17}$$

Lastly, UQI is a similar image quality index that incorporates terms for loss of correlation, luminance distortion, and contrast distortion. Let R=$\{R_s | s = 1, 2, 3, \cdots, S\}$ and C=$\{C_s | s = 1, 2, 3, \cdots, S\}$ be the reference and the compared images, respectively. The proposed quality index is defined as

$$Q(R, C) = \frac{4\sigma_{RC}\bar{R}\bar{C}}{({\sigma^2}_R + {\sigma^2}_C)(\bar{R}^2 + \bar{C}^2)} \tag{18}$$

$$\bar{R} = \frac{1}{S}\sum_{s=1}^{S} R_s,\ \ \bar{C} = \frac{1}{S}\sum_{s=1}^{S} C_s \tag{19}$$

$$\sigma_R^2 = \frac{1}{S-1}\sum_{s=1}^{S}(R_s - \bar{R})^2,\ \ \sigma_C^2 = \frac{1}{S-1}\sum_{s=1}^{S}(C_s - \bar{C})^2 \tag{20}$$

$$\sigma_{RC} = \frac{1}{S-1}\sum_{s=1}^{S}(R_s - \bar{R})(C_s - \bar{C}) \tag{21}$$

To quantify the improvements gained from the proposed model, VDSC and HD95 were applied across four trained models and compared against the manually contoured reference segmentations (MCs). VDSC is used to measure the degree of overlap between the predicted segmentation volume (A) and the expert reference segmentation (B). This metric provides a robust evaluation of how well the model segments the target areas compared to human experts.

$$VDSC = \frac{2|A \cap B|}{|A| + |B|} \tag{22}$$

The Hausdorff distance (HD) is a metric used to evaluate the spatial discrepancy between two sets of points, in this case, the trained model's segmentation (A) and the expert's manual segmentation (B). It quantifies the greatest distance that exists between a point in one set and the closest point in the other set, providing a measure of the worst-case deviation. Specifically, HD95 represents the 95th percentile of these distances, offering a more robust metric by discounting the most extreme outliers. This means it captures the largest surface-to-surface separation for 95% of the points between the two segmentation boundaries, while disregarding the top 5% of errors, which could be caused by noise or other anomalies. Let a and b represent points on the surfaces of segmentations A and B, respectively. The Hausdorff distance is formally defined as:

$$HD95(A,B) = \max_{a\subset A}\left\{\min_{b\subset B}(dis(a,b))\right\}_{95\%}, \tag{23}$$

where $dis(a,b)$ denotes the Euclidean distance between point a in segmentation A and point b in segmentation B. HD95 refines this definition by focusing on the 95th percentile of the distance distribution, making it a more stable and reliable metric for evaluating segmentation performance in medical imaging.

To assess the performance of the four models (general model, general IDOL model, and PHL-IDOL model), we utilized two key quantitative metrics: the volumetric Dice similarity coefficient (VDSC) and Hausdorff distance at the 95th percentile (HD95). These evaluations allowed us to compare the accuracy and consistency of each model's segmentation. nally, to statistically assess the performance of each model, we conducted a one-way ANOVA test to compare the segmentation results from the general model, continual model, and conventional IDOL model with those from the PHL-IDOL model. A significance level of $p < 0.05$ was used to indicate statistically significant differences between the methods, highlighting the superior performance of the PHL-IDOL approach.

4.4. Results

4.4.1. Personalized Hyperspace Learning performance

Figure 4-8 presents the volumetric Dice similarity coefficient (VDSC) performance for 18 head and neck (H&N) organs, segmented using the general model, continual model, conventional IDOL model, and the PHL-IDOL model. These organs include the brainstem, oral cavity, larynx, esophagus, spinal cord, left cochlea, right cochlea, mandible, left parotid, right parotid, right submandibular gland (R SMG), left submandibular gland (L SMG), thyroid, left optic nerve, right optic nerve, optic chiasm, left eye, and right eye. The PHL-IDOL model demonstrated superior performance across all organs, achieving the highest VDSC values compared to other models. Notable VDSC scores for the PHL-IDOL model include 0.93 for the oral cavity, 0.91 for the larynx, and 0.95 for the mandible, showcasing its remarkable segmentation accuracy. Additionally, the PHL-IDOL model consistently exhibited lower standard deviations (SDs) compared to the other models, highlighting its reliability and precision. For example, the left cochlea had an SD of 0.05 with PHL-IDOL, compared to 0.08 with the general model.

Figure 4-9 outlines the 95% Hausdorff distance (HD95) results for the same 18 organs. The PHL-IDOL model demonstrated the lowest HD95 values, signifying enhanced spatial accuracy compared to the other models. For instance, the esophagus achieved an HD95 value of 3.34 with the PHL-IDOL model, a significant improvement over the general model’s 4.60. Similarly, for the right cochlea, the PHL-IDOL model achieved an HD95 of 1.62, compared to 2.52, 2.33, and 2.17 for the general, continual, and conventional IDOL models, respectively. These results underscore the PHL-IDOL model’s capability to excel in segmenting smaller and more complex organs, where the general model struggled. Statistical evaluations using the one-way ANOVA test confirmed that the PHL-IDOL model showed significant improvements ($p < 0.05$) in many cases.

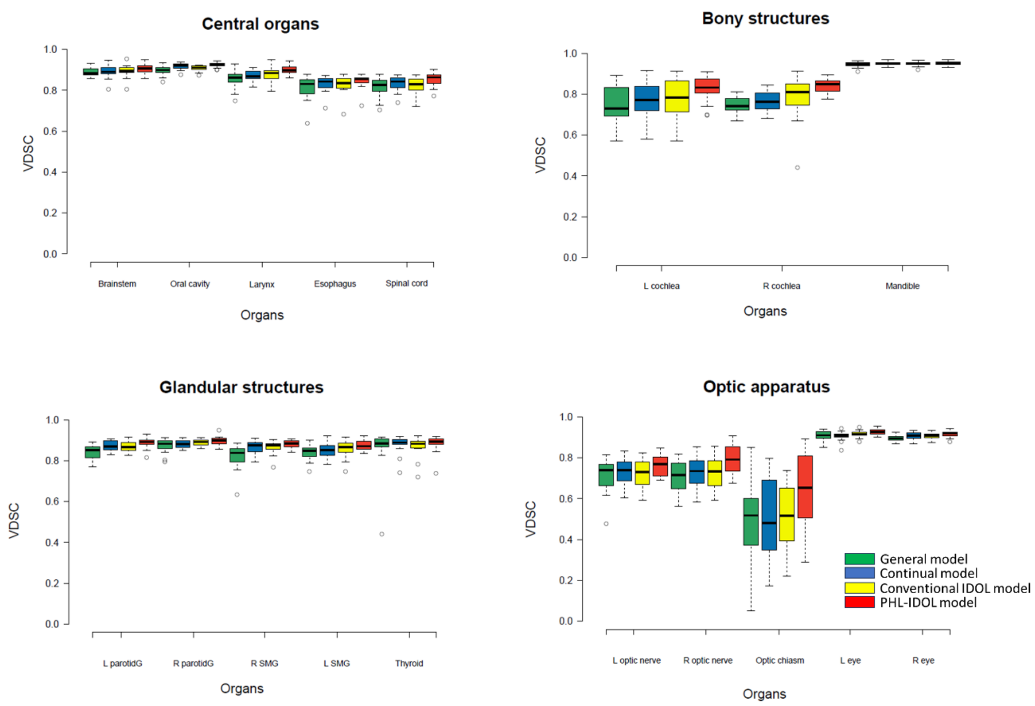


Figure 4-8. Boxplots comparing the VDSC performance of four models: the general model, the continual model, the conventional IDOL model, and the PHL-IDOL model. The segmentation performance is categorized into four groups: central organs, bony structures, glandular structures, and optic apparatus, highlighting the superior accuracy of the PHL-IDOL model across all categories.

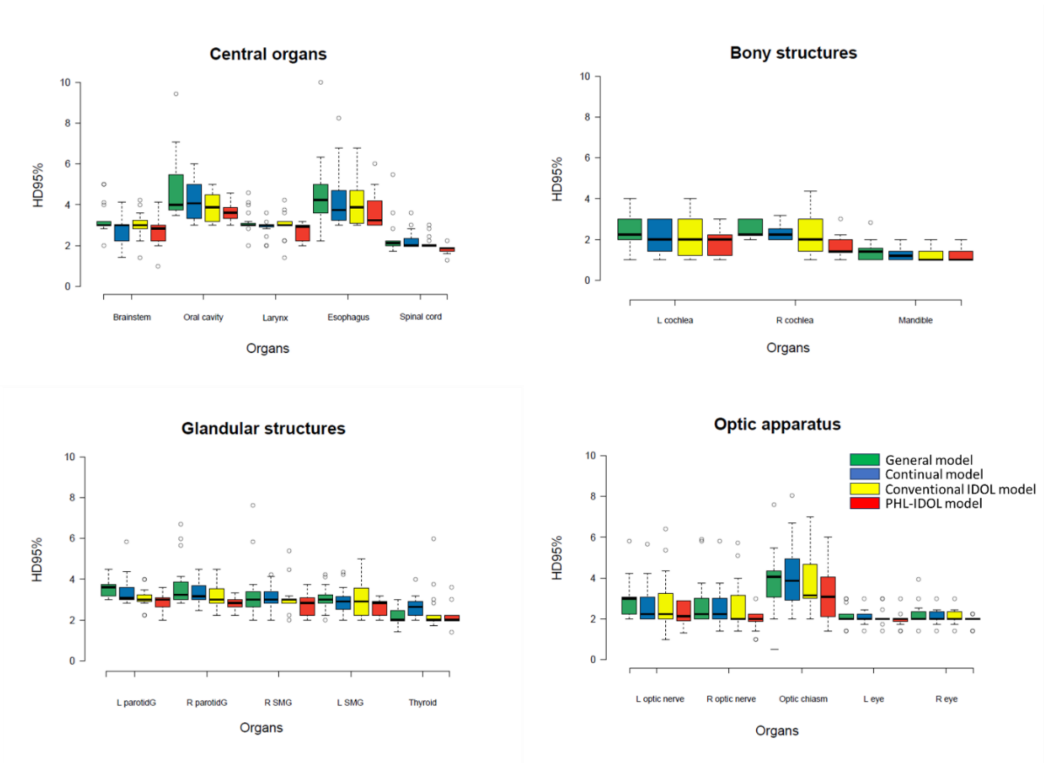


Figure 4-9. Boxplots comparing the HD95 performance of four models: the general model, the continual model, the conventional IDOL model, and the PHL-IDOL model. The results are grouped into central organs, bony structures, glandular structures, and optic apparatus, demonstrating the superior consistency and precision of the PHL-IDOL model across all structural categories.

Figure 4-10 showcases qualitative comparisons of segmentation results across the four models, illustrating both the best-case and worst-case scenarios for the general model. The best-case scenarios, defined as those with the smallest deviation from ground truth, highlight the PHL-IDOL model’s superior segmentation accuracy compared to the general, continual, and conventional IDOL models. The worst-case scenarios for the general model further emphasize the PHL-IDOL model’s robust performance, particularly for small, challenging organs like the optic chiasm and cochleae.

The arrangement in Figure 4-10 compares input CT images (a), segmentation outputs from the general model (b), continual model (c), conventional IDOL model (d), and PHL-IDOL model (e) against ground truth reference segmentations (f). Additional overlays (g-j) illustrate the differences between the models’ results and the ground truth, emphasizing the PHL-IDOL model’s ability to closely replicate the reference segmentations.

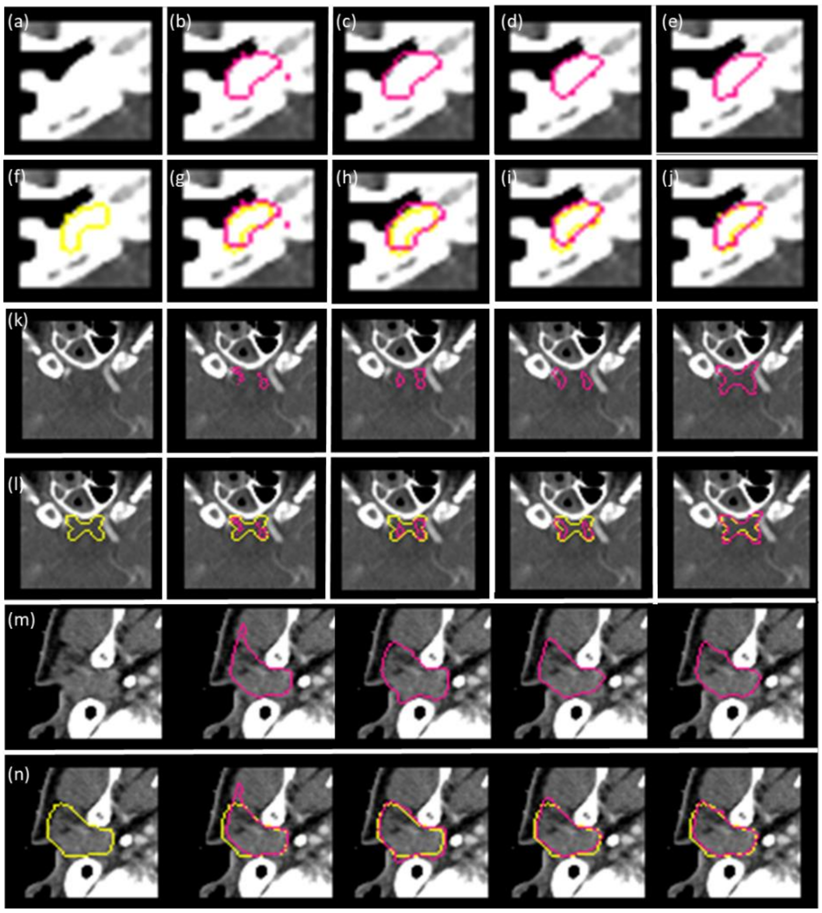


Figure 4-10. Visual comparison of segmentation outcomes across four models: the general model, the continual model, the conventional IDOL model, and the PHL-IDOL model. Panel (a) displays the input CT image, while panels (b), (c), (d), and (e) showcase the segmentation results from the general, continual, conventional IDOL, and PHL-IDOL models, respectively. Panel (f) represents the ground truth reference segmentation. Panels (g) through (j) illustrate the overlap of each model's segmentation result with the ground truth, highlighting areas of concordance and discrepancy. The same layout is applied in subsequent rows (k & l and m & n), providing a comprehensive visual evaluation of segmentation performance.

The quantitative and qualitative analyses highlight the significant improvements achieved by the PHL-IDOL model over other segmentation models. Its superior performance, especially for small and complex organs, underscores its potential to enhance segmentation accuracy and reliability in clinical workflows. These findings establish the PHL-IDOL model as a powerful tool for overcoming the limitations of conventional segmentation methods, paving the way for more precise and efficient treatment planning in head and neck radiotherapy.

4.5.2. Evaluation of multi-instituion dataset

Table 4-3 presents the Volumetric Dice Similarity Coefficient (VDSC) results for the three models across 18 head and neck organs, including the brainstem, oral cavity, larynx, esophagus, spinal cord, left and right cochlea, mandible, left and right parotid, left and right submandibular glands (SMG), thyroid, left and right optic nerves, optic chiasm, and both eyes. The PHL-IDOL model consistently outperformed the general model in all organs, demonstrating improved VDSC values. Specifically, the PHL-IDOL model showed the highest VDSC scores across all organs, including 0.93 for the brainstem, 0.94 for the oral cavity, 0.93 for the larynx, and 0.95 for the mandible. For smaller structures such as the cochlea, the PHL-IDOL model achieved VDSC values of 0.88 (left) and 0.89 (right), significantly outperforming the other models, which struggled with these more intricate structures. Notably, the standard deviations (SD) for VDSC in the PHL-IDOL model were consistently lower than those of the other models, indicating more reliable and consistent performance. For instance, the SD for the left cochlea in the PHL-IDOL model was 0.04 compared to 0.08 for the general model. These results demonstrate the significant performance gains provided by the PHL-IDOL model, particularly for smaller and more complex organs, where the general model faced challenges. The differences between models for larger structures, such as the mandible and eyes, were less pronounced, but the PHL-IDOL model still maintained a slight edge in performance. A one-way ANOVA test was used to statistically evaluate the significance of these differences.

Results for the 95th percentile of the Hausdorff Distance (HD95) is also illustrated in Table 4-3, which further demonstrates the superior performance of the PHL-IDOL model. For example, the esophagus showed a remarkable improvement with the PHL-IDOL model, achieving an HD95 of 3.09, significantly better than the general model's 4.53. The right cochlea also saw notable improvements, with the PHL-IDOL model achieving an HD95 of 1.34, outperforming the general model (2.30), and general fine-tunning model (2.17). The reduced standard deviations in the PHL-IDOL model's HD95 results reflect its more consistent and stable performance across different structures.

Table 4-3. Comparison of segmentation performance using internal dataset across 18 organs using VDSC and HD95% for the general, general fine-tunning, and PHL-IDOL models. The largest and smallest differences between the general and PHL-IDOL models are highlighted.
(p > 0.05 no mark, insignificant, 0.01 < p < 0.05 *, first level of significance and p < 0.01 **, second level of significance).

		VDSC	SD			VDSC	SD			HD95	SD			HD95	SD
Brainstem	General	0.88	0.01	L cochlea	General	0.75**	0.08	Brainstem	General	3.04*	0.70	L cochlea	General	2.53*	1.01
	IDOL	0.89	0.02		IDOL	0.77*	0.09		IDOL	3.17	0.67		IDOL	2.37*	0.85
	PHL-IDOL	0.93	0.01		PHL-IDOL	0.88	0.04		PHL-IDOL	2.72	0.55		PHL-IDOL	1.62	0.51
Oral cavity	General	0.90**	0.02	R cochlea	General	0.73**	0.03	Oral cavity	General	4.61	1.55	R cochlea	General	2.30*	0.97
	IDOL	0.91**	0.01		IDOL	0.75*	0.04		IDOL	4.30	0.57		IDOL	2.17*	0.84

	PHL-IDOL	**0.94**	**0.01**		**PHL-IDOL**	**0.89**	**0.03**		**PHL-IDOL**	**3.37**	**0.40**		**PHL-IDOL**	**1.34**	**0.35**
Larynx	General	**0.85****	**0.03**	Mandible	General	**0.94**	**0.01**	Larynx	General	**3.30****	**0.58**	Mandible	General	**1.38**	**0.45**
	IDOL	**0.84***	**0.03**		IDOL	**0.95**	**0.01**		IDOL	**3.03***	**0.55**		IDOL	**1.32**	**0.39**
	PHL-IDOL	**0.93***	**0.01**		**PHL-IDOL**	**0.95**	**0.01**		**PHL-IDOL**	**2.57**	**0.29**		**PHL-IDOL**	**0.99**	**0.28**
Esophagus	General	**0.80****	**0.05**	L optic nerve	General	**0.72***	**0.05**	Esophagus	General	**4.53***	**1.51**	L optic nerve	General	**2.34***	**1.47**
	IDOL	**0.83***	**0.03**		IDOL	**0.75***	**0.06**		IDOL	**4.40***	**1.01**		IDOL	**2.23***	**1.29**
	PHL-IDOL	**0.87**	**0.02**		**PHL-IDOL**	**0.82**	**0.03**		**PHL-IDOL**	**3.09**	**0.69**		**PHL-IDOL**	**1.91**	**0.51**
Spinal cord	General	**0.84****	**0.04**	R optic nerve	General	**0.72****	**0.07**	Spinal cord	General	**2.57**	**0.85**	R optic nerve	General	**2.62***	**1.84**
	IDOL	**0.83***	**0.04**		IDOL	**0.79****	**0.07**		IDOL	**2.10**	**0.33**		IDOL	**2.11**	**1.33**
	PHL-IDOL	**0.90**	**0.03**		**PHL-IDOL**	**0.83**	**0.05**		**PHL-IDOL**	**1.64**	**0.22**		**PHL-IDOL**	**1.76**	**0.46**
L parotid	General	**0.80****	**0.05**	Optic chiasm	General	**0.46****	**0.19**	L parotid	General	**3.30***	**0.45**	Optic chiasm	General	**3.90****	**1.78**
	IDOL	**0.86***	**0.03**		IDOL	**0.55****	**0.15**		IDOL	**3.02***	**0.47**		IDOL	**3.68***	**1.39**
	PHL-IDOL	**0.88**	**0.02**		**PHL-IDOL**	**0.75**	**0.08**		**PHL-IDOL**	**2.90**	**0.41**		**PHL-IDOL**	**2.83**	**1.08**
R parotid	General	**0.85****	**0.03**	L eye	General	**0.89***	**0.02**	R parotid	General	**3.47****	**1.13**	L eye	General	**2.13**	**0.47**
	IDOL	**0.89**	**0.02**		IDOL	**0.90***	**0.02**		IDOL	**3.37***	**0.79**		IDOL	**2.12**	**0.43**
	PHL-IDOL	**0.93**	**0.01**		**PHL-IDOL**	**0.93**	**0.01**		**PHL-IDOL**	**2.36**	**0.31**		**PHL-IDOL**	**2.00**	**0.41**
R SMG	General	**0.81****	**0.05**	R eye	General	**0.89***	**0.01**	R SMG	General	**2.99**	**1.29**	R eye	General	**2.43**	**0.55**
	IDOL	**0.85**	**0.02**		IDOL	**0.91***	**0.02**		IDOL	**2.90**	**0.78**		IDOL	**2.35**	**0.47**
	PHL-IDOL	**0.92**	**0.01**		**PHL-IDOL**	**0.93**	**0.01**		**PHL-IDOL**	**2.38**	**0.41**		**PHL-IDOL**	**1.98**	**0.26**
L SMG	General	**0.79****	**0.03**					L SMG	General	**3.04**	**0.55**				
	IDOL	**0.84**	**0.04**						IDOL	**3.13***	**0.81**				
	PHL-IDOL	**0.91**	**0.02**						**PHL-IDOL**	**2.41**	**0.39**				
Thyroid	General	**0.87**	**0.09**					Thyroid	General	**3.13**	**3.11**				
	IDOL	**0.87**	**0.04**						IDOL	**2.99**	**0.97**				
	PHL-IDOL	**0.89**	**0.02**						**PHL-IDOL**	**1.73**	**0.39**				

Table 4-4 illustrates the total VDSC and the HD95% of the three models using 18 organs in the head and neck, which is explained above. Table 4-4 shows the performance of the External 1 dataset results which is the UT Southwestern dataset. As you can see in the table, the PHL-IDOL showed the best performance compared to the other two models and showing a bigger gap compared to the internal dataset evaluation results.

Table 4-4. Comparison of segmentation performance for external 1 dataset (UT Southwestern) 18 organs using general model, general fine-tunning model, and PHL-IDOL model using VDSC and HD95%.
(p > 0.05 no mark, insignificant, 0.01 < p < 0.05 *, first level of significance and p < 0.01 **, second level of significance).

		VDSC	**SD**			**VDSC**	**SD**			**HD95**	**SD**			**HD95**	**SD**
Brainstem	General	**0.70****	**0.08**	L cochlea	General	**0.50****	**0.11**	Brainstem	General	**4.89****	**1.78**	L cochlea	General	**7.04****	**2.00**
	IDOL	**0.79***	**0.04**		IDOL	**0.76***	**0.07**		IDOL	**3.99***	**0.91**		IDOL	**3.86**	**1.02**
	UTSW PHL-IDOL	**0.84**	**0.04**		**UTSW PHL-IDOL**	**0.82**	**0.05**		**UTSW PHL-IDOL**	**2.77**	**0.65**		**UTSW PHL-IDOL**	**3.19**	**0.78**
Oral cavity	General	**0.72****	**0.07**	R cochlea	General	**0.46****	**0.09**	Oral cavity	General	**6.74****	**2.01**	R cochlea	General	**8.56****	**2.33**
	IDOL	**0.76***	**0.04**		IDOL	**0.75***	**0.06**		IDOL	**3.73**	**0.97**		IDOL	**3.64**	**0.86**
	UTSW PHL-IDOL	**0.83**	**0.03**		**UTSW PHL-IDOL**	**0.84**	**0.05**		**UTSW PHL-IDOL**	**3.49**	**0.44**		**UTSW PHL-IDOL**	**2.89**	**0.8**
Larynx	General	**0.78**	**0.04**	Mandible	General	**0.86**	**0.03**	Larynx	General	**4.21***	**1.55**	Mandible	General	**3.30**	**1.45**
	IDOL	**0.82**	**0.03**		IDOL	**0.90**	**0.02**		IDOL	**3.63**	**0.88**		IDOL	**2.47**	**1.06**
	UTSW PHL-IDOL	**0.82**	**0.03**		**UTSW PHL-IDOL**	**0.93**	**0.01**		**UTSW PHL-IDOL**	**2.96**	**0.56**		**UTSW PHL-IDOL**	**1.62**	**0.5**
Esophagus	General	**0.69***	**0.06**	L optic nerve	General	**0.43****	**0.16**	Esophagus	General	**5.11****	**1.93**	L optic nerve	General	**8.63****	**2.29**
	IDOL	**0.78**	**0.05**		IDOL	**0.66****	**0.06**		IDOL	**4.17**	**1.16**		IDOL	**4.72****	**1.35**
	UTSW PHL-IDOL	**0.79**	**0.05**		**UTSW PHL-IDOL**	**0.76**	**0.04**		**UTSW PHL-IDOL**	**3.76**	**0.81**		**UTSW PHL-IDOL**	**2.96**	**0.76**
Spinal cord	General	**0.81**	**0.04**	R optic nerve	General	**0.49****	**0.13**	Spinal cord	General	**4.20***	**1.64**	R optic nerve	General	**7.90****	**2.17**
	IDOL	**0.83**	**0.04**		IDOL	**0.61****	**0.07**		IDOL	**2.76**	**0.74**		IDOL	**4.58****	**1.34**
	UTSW PHL-IDOL	**0.86**	**0.03**		**UTSW PHL-IDOL**	**0.73**	**0.05**		**UTSW PHL-IDOL**	**2.48**	**0.37**		**UTSW PHL-IDOL**	**3.14**	**0.65**
L parotid	General	**0.75***	**0.09**	Optic	General	**0.22****	**0.15**	L parotid	General	**3.66**	**1.13**	Optic	General	**15.48****	**3.71**

	IDOL	**0.81**	**0.04**	chiasm	IDOL	**0.55***	**0.11**		IDOL	**3.37**	**0.69**	chiasm	IDOL	**6.71***	**2.78**
	UTSW PHL-IDOL	**0.83**	**0.04**		**UTSW PHL-IDOL**	**0.65**	**0.08**		**UTSW PHL-IDOL**	**3.17**	**0.57**		**UTSW PHL-IDOL**	**4.99**	**1.84**
R parotid	General	**0.77***	**0.08**	L eye	General	**0.84***	**0.04**	R parotid	General	**3.75**	**1.39**	L eye	General	**2.94**	**1.11**
	IDOL	**0.81**	**0.03**		IDOL	**0.87**	**0.02**		IDOL	**3.57**	**0.71**		IDOL	**2.66**	**0.66**
	UTSW PHL-IDOL	**0.85**	**0.03**		**UTSW PHL-IDOL**	**0.91**	**0.02**		**UTSW PHL-IDOL**	**3.19**	**0.46**		**UTSW PHL-IDOL**	**2.32**	**0.53**
R SMG	General	**0.77****	**0.03**	R eye	General	**0.81***	**0.05**	R SMG	General	**4.36***	**1.51**	R eye	General	**3.13**	**1.07**
	IDOL	**0.86**	**0.04**		IDOL	**0.85**	**0.03**		IDOL	**3.22**	**0.85**		IDOL	**2.56**	**0.59**
	UTSW PHL-IDOL	**0.90**	**0.02**		**UTSW PHL-IDOL**	**0.91**	**0.02**		**UTSW PHL-IDOL**	**2.76**	**0.41**		**UTSW PHL-IDOL**	**2.13**	**0.26**
L SMG	General	**0.76****	**0.05**					L SMG	General	**3.04**	**0.55**				
	IDOL	**0.84**	**0.02**						IDOL	**3.13***	**0.81**				
	UTSW PHL-IDOL	**0.88**	**0.02**						**UTSW PHL-IDOL**	**2.41**	**0.39**				
Thyroid	General	**0.80****	**0.05**					Thyroid	General	**3.13**	**3.11**				
	IDOL	**0.87**	**0.03**						IDOL	**2.99**	**0.97**				
	UTSW PHL-IDOL	**0.91**	**0.02**						**UTSW PHL-IDOL**	**1.73**	**0.39**				

Table 4-5 presents the total VDSC and HD95% results for the three models using the External 2 dataset (MAYO Clinic), focusing on 18 organs in the head and neck as previously described. As demonstrated in Table 4-5, the PHL-IDOL model outperformed the other two models, consistently showing superior performance. Additionally, the gap between the general model and the PHL-IDOL model was more pronounced, further emphasizing the importance of using the PHL-IDOL model in real-time clinical applications.

Table 4-5. Comparison of segmentation performance for external 2 dataset (MAYO Clinic) 18 organs using general model, general fine-tunning model, and PHL-IDOL model using VDSC and HD95%.
(p > 0.05 no mark, insignificant, 0.01 < p < 0.05 *, first level of significance and p < 0.01 **, second level of significance).

		VDSC	**SD**			**VDSC**	**SD**			**HD95**	**SD**			**HD95**	**SD**

Brainstem	General	0.75**	0.04	L cochlea	General	0.53**	0.13	Brainstem	General	5.07**	1.54	L cochlea	General	7.05**	2.70
	IDOL	0.79**	0.04		IDOL	0.79	0.06		IDOL	3.49	1.32		IDOL	3.05	0.94
	MAYO PHL-IDOL	0.89	0.04		MAYO PHL-IDOL	0.84	0.03		MAYO PHL-IDOL	2.69	0.77		MAYO PHL-IDOL	2.40	0.58
Oral cavity	General	0.77	0.06	R cochlea	General	0.44**	0.11	Oral cavity	General	5.36*	1.77	R cochlea	General	8.43**	3.58
	IDOL	0.81	0.04		IDOL	0.77	0.06		IDOL	3.54	1.44		IDOL	3.38	1.06
	MAYO PHL-IDOL	0.85	0.03		MAYO PHL-IDOL	0.84	0.03		MAYO PHL-IDOL	3.75	1.21		MAYO PHL-IDOL	2.57	0.66
Larynx	General	0.80	0.03	Mandible	General	0.88	0.03	Larynx	General	3.78	1.34	Mandible	General	3.19	1.33
	IDOL	0.82	0.03		IDOL	0.91	0.02		IDOL	3.36	1.18		IDOL	2.16	0.58
	MAYO PHL-IDOL	0.87	0.03		MAYO PHL-IDOL	0.95	0.01		MAYO PHL-IDOL	2.93	0.82		MAYO PHL-IDOL	1.84	0.37
Esophagus	General	0.72*	0.04	L optic nerve	General	0.45**	0.18	Esophagus	General	5.15*	1.56	L optic nerve	General	9.41**	4.16
	IDOL	0.79	0.05		IDOL	0.72	0.07		IDOL	3.59	1.87		IDOL	4.12	1.11
	MAYO PHL-IDOL	0.83	0.05		MAYO PHL-IDOL	0.78	0.04		MAYO PHL-IDOL	3.26	1.04		MAYO PHL-IDOL	3.66	1.52
Spinal cord	General	0.76*	0.04	R optic nerve	General	0.39**	0.15	Spinal cord	General	4.24	1.62	R optic nerve	General	8.41**	4.29
	IDOL	0.83	0.03		IDOL	0.73	0.08		IDOL	3.02	1.14		IDOL	4.42	1.28
	MAYO PHL-IDOL	0.88	0.03		MAYO PHL-IDOL	0.79	0.05		MAYO PHL-IDOL	2.84	0.83		MAYO PHL-IDOL	3.57	1.41
L parotid	General	0.81	0.05	Optic chiasm	General	0.33**	0.16	L parotid	General	3.22	1.22	Optic chiasm	General	13.08**	5.15
	IDOL	0.82	0.03		IDOL	0.56*	0.12		IDOL	3.18	0.89		IDOL	7.34*	2.95
	MAYO PHL-IDOL	0.86	0.04		MAYO PHL-IDOL	0.71	0.07		MAYO PHL-IDOL	3.02	1.08		MAYO PHL-IDOL	5.12	1.76
R parotid	General	0.83	0.05	L eye	General	0.87	0.03	R parotid	General	3.13	1.06	L eye	General	2.99	0.65
	IDOL	0.84	0.03		IDOL	0.88	0.03		IDOL	3.11	0.76		IDOL	2.56	0.51
	MAYO PHL-IDOL	0.87	0.03		MAYO PHL-IDOL	0.92	0.02		MAYO PHL-IDOL	2.92	0.86		MAYO PHL-IDOL	2.16	0.39
R SMG	General	0.69**	0.07	R eye	General	0.86	0.04	R SMG	General	5.73**	2.13	R eye	General	2.59	0.78
	IDOL	0.78	0.06		IDOL	0.88	0.03		IDOL	4.89**	1.41		IDOL	2.35	0.53

	MAYO PHL-IDOL	**0.83**	**0.02**	**MAYO PHL-IDOL**	**0.93**	**0.02**		**MAYO PHL-IDOL**	**1.73**	**0.25**	**MAYO PHL-IDOL**	**2.06**	**0.32**
L SMG	General	**0.69****	**0.06**				L SMG	General	**3.04**	**0.55**			
	IDOL	**0.78**	**0.05**					IDOL	**3.13***	**0.81**			
	MAYO PHL-IDOL	**0.83**	**0.03**					**MAYO PHL-IDOL**	**2.41**	**0.39**			
Thyroid	General	**0.78***	**0.04**				Thyroid	General	**3.13**	**3.11**			
	IDOL	**0.83**	**0.03**					IDOL	**2.99**	**0.97**			
	MAYO PHL-IDOL	**0.89**	**0.02**					**MAYO PHL-IDOL**	**1.73**	**0.39**			

To provide a clear and intuitive comparison of model performance, we generated a spider chart, as illustrated in Figure 4-11. This chart offers a comprehensive visual representation of the differences between the general model and the PHL-IDOL model across various evaluation metrics. As previously mentioned, the performance gap between the general model and the PHL-IDOL model becomes more pronounced when evaluating the external dataset, particularly the data sourced from multiple institutions. This significant disparity highlights the limitations of the general model when applied to diverse patient populations, reinforcing the critical need for the PHL-IDOL model in clinical practice. The superior performance of the PHL-IDOL model in these evaluations suggests that it is not only more adaptable but also more reliable in delivering precise results, making it an ideal candidate for real-time clinical implementation.

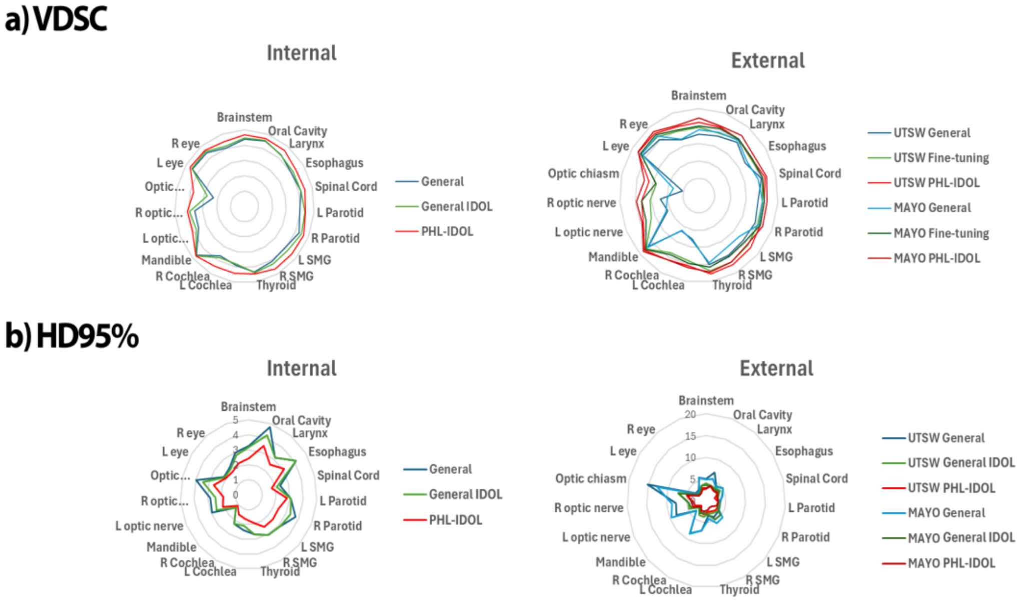


Figure 4-11. Spider chart illustrating the VDSC and HD95% performance of three models: general model, general fine-tunning model, and the PHL-IDOL model using three datasets (Internal (Yonsei Cancer Center), External 1 (UT Southwestern), and External 2 (MAYO Clinic)). a) shows the result of the VDSC and b) represents the result of the HD95%.

Figure 4-12 offers a detailed visual comparison of the segmentation performance. In this context, all three model results were shown using all the dataset: 1. Yonsei (Internal) 2. UTSW (External 1) 3. MAYO (Extneral 2). When comparing these outcomes, it is evident that the general fine-tunning model surpasses the general model in accuracy. However, the PHL-IDOL model takes this performance a step further, significantly outperforming the the general fine-tunning model approach. Notably, as previously discussed, the PHL-IDOL model consistently excels, particularly in segmenting smaller, more difficult-to-visualize organs. This further underscores its robustness and superiority in challenging clinical cases, making it a powerful tool for improving segmentation accuracy.

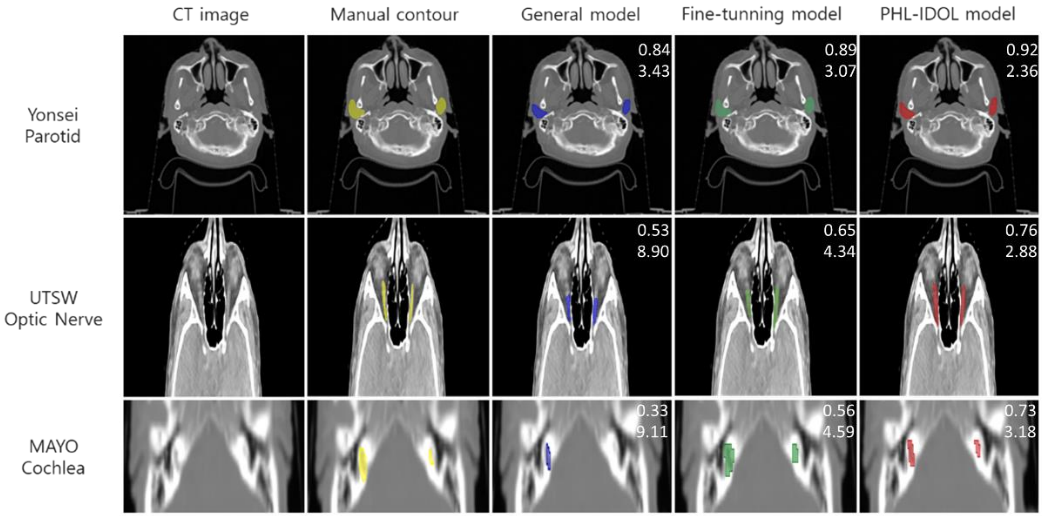


Figure 4-12. Visual comparision of the segmentation results using the general model, fine-tunning model, and the PHL-IDOL model. The first line displays the results from the Yonsei Cancer Center dataset (Internal), the second line shows the UTSW dataset (External 1) results, and the last line represents the MAYO Clinic dataset (External 2) results. For the column, CT image, Manual contour, General model, Fine-tunning model, and PHL-IDOL results were displayed. At the right top coner ther VDSC and HD95% is recorded to show the results of the quantitative results.

4.5. Discussion and Conclusion

The general fine-tunning model approach was introduced as a deep learning framework specifically designed to optimize task performance for individual patients in a radiotherapy setting. Its primary objective is to create a model tailored to a specific patient by leveraging the prior information available for that individual. However, one key limitation of the general fine-tunning approach is its reliance on random deformation vector fields to generate augmented, patient-specific training datasets during the fine-tuning stage. This method can lead to unrealistic deformations, diminishing its effectiveness in achieving optimal patient-specific performance. In contrast, the proposed PHL-IDOL framework enhances this process by beginning with a similarity-based dataset, comprising data from patients with similar characteristics, which is then deformed to match the reference data of the patient of interest. This results in more natural and realistic personalized datasets, enabling the model to learn a highly patient-specific segmentation.

The innovation of the PHL-IDOL model lies in its ability to create a framework that generates personalized datasets in a more organic and realistic manner by searching for similar datasets around the patient of interest. This process is achieved through two primary steps: computing dataset similarity based on patient-specific data and generating new datasets by applying deformation vectors between the patient data and the most similar datasets. By incorporating these additional datasets, the PHL-IDOL model effectively expands the data hypersurface around the patient of interest, addressing the challenge of limited fine-tuning data and significantly improving segmentation accuracy.

When comparing the performance of the PHL-IDOL model to both the general model and the general fine-tunning model demonstrated marked improvements. Specifically, the PHL-IDOL model exhibited higher Volumetric Dice Similarity Coefficient (VDSC) values and lower 95% Hausdorff Distance (HD95) values. On average in internal evaluation, the VDSC values increased by 0.08 from a baseline of 0.81, which is also 0.06 higher than the general fine-tunning model. For HD95, the PHL-IDOL model showed a decrease of 0.75 from a baseline of 2.97, and 0.48 lower than the general fine-tunning model, indicating significantly enhanced precision. Additionally, for the external datasets, in the external 1 dataset, the PHL-IDOL showed 0.16 increase in the VDSC comparing the general model and the PHL-IDOL and 2.59 difference in the HD95%. Furthermore, in the external 2 dataset, the PHL-IDOL preformed 0.16 in the VDSC and 2.82 in the HD95%. This results in variance underscores the PHL-IDOL model's consistent and reliable performance, validating its efficacy and robustness in real-world clinical settings.

During the process of generating the PHL-IDOL dataset, particularly when determining the multiplication range of deformation vector fields, we identified certain outlier cases with substantial dissimilarities. These outliers had the potential to disrupt the model's hypersurface training. Figure 4-13 illustrates several instances where excessive deformation vector fields produced images that could adversely impact the training process. In this figure, images (c), (d), and (e) show only minor differences from the reference image, maintaining a close resemblance. In contrast, images (b) and (f) display significant deviations, which could negatively affect the training process. As a result, we carefully constrained the deformation vector range based on iterative trials and evaluations to ensure the model's stability and performance were not compromised.

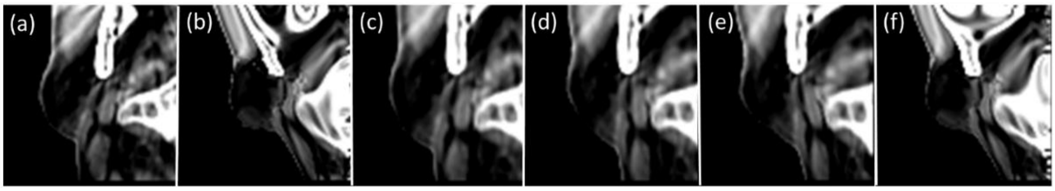
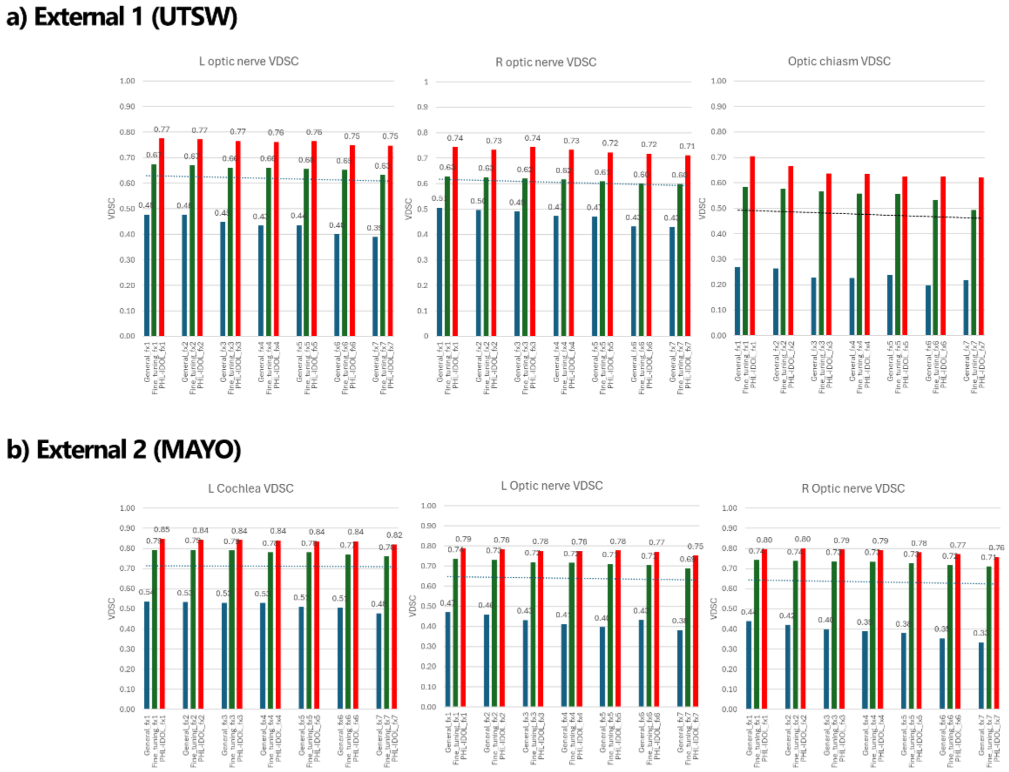


Figure 4-13. Comparison of adjusted deformed vector fields within the patient dataset. (a) represents the reference image, (b) illustrates the image after applying 0.6 adjusting lower multiplying deformed vector fields, (c) shows an image obtained by multiplying the deformed vector field with a factor of 0.8, (d) features the original deformed vector field-adjusted image, (e) presents the image achieved by multiplying the deformed vector field by a factor of 1.2, (f) displays the image generated by applying multiplication beyond 1.2 (1.4). (c), (d), (e) is the image that is generated for the PHL-IDOL model.

Moreover, by analyzing the seven fraction results from the external datasets, the chart vividly illustrates a progressive widening of the segmentation gap as the treatment fractions advance. This growing disparity between the models highlights the increasing divergence in performance, particularly between the general and personalized models. As the fractions proceed, the need for continuous model adaptation becomes increasingly apparent. This underscores the critical importance of adjusting the segmentation model throughout the clinical treatment process to maintain accuracy and efficacy in real-world applications. The consistent gap growth as observed in the later fractions strongly suggests that without these real-time adjustments, the standard models may struggle to keep pace with anatomical changes, potentially compromising treatment quality. Therefore, these findings provide compelling evidence that the integration of dynamic, adaptive frameworks like PHL-IDOL into clinical trials is not only beneficial but necessary to optimize patient-specific outcomes and ensure the highest standards of precision and care in radiotherapy. Figure 4-14 illustrates the segmentation accuracy results (VDSC) for each weekly fraction.



a) External 1 (UTSW)

b) External 2 (MAYO)

Figure 4-14. VDSC results for both the External 1 and External 2 datasets. a) results for the External 1 dataset across weekly fractions are displayed, with the trendline indicating a widening accuracy gap between the first and final fractions. Similarly, b) showcases the performance of the External 2 dataset, where a comparable trend emerges, revealing an increasing divergence in accuracy as the fraction number progresses.

Despite the notable advantages and innovations highlighted, it is essential to acknowledge the limitations of this study, particularly regarding the range of organs evaluated. To confirm the framework's broader applicability in a clinical setting, further evaluations must include target volumes and tumors. Given the PHL-IDOL model's impressive performance in contouring small, hard-to-visualize organs, we anticipate similarly strong results in more complex structures such as target volumes and tumors. We are currently developing an expanded dataset to facilitate these evaluations, and we are eager to test and validate the framework on these critical components.

In addition, we evaluated the PHL-IDOL model across multiple institutions, including datasets from UT Southwestern and the Mayo Clinic, to assess its performance in external environments. The results consistently demonstrated the superiority of the PHL-IDOL framework compared to the general model and the fine-tunning model, especially as treatment fractions progressed. The increasing segmentation accuracy gap between the PHL-IDOL and general models underscores the need for its implementation in real-time clinical settings. The model's ability to adapt to patient-specific variations, even across diverse institutions, further validates its potential for widespread clinical adoption in adaptive radiotherapy workflows.

Furthermore, online ART encompasses a variety of tasks where innovative solutions are in high demand. We believe the PHL-IDOL model can extend beyond segmentation to other image generation tasks, such as generating synthetic CT images from CBCT data or enhancing low-resolution images to higher resolution. These tasks are pivotal for advancing online ART workflows, and we plan to explore the framework's adaptability to these challenges, further testing its potential across various image generation problems.

Our work marks a significant advancement in adaptive radiotherapy, offering a more reliable and versatile framework for generating patient-specific models. With the PHL-IDOL approach, we underscore the growing potential for personalized healthcare in radiotherapy, demonstrating the model's capability to enhance patient outcomes. By enriching datasets with comprehensive prior information, the PHL-IDOL model not only addresses the shortcomings of the general fine-tunning approach but also lays the groundwork for future exploration in personalized medicine. In conclusion, the PHL-IDOL framework stands as a promising tool for optimizing segmentation accuracy and improving treatment planning and execution in adaptive radiotherapy settings.

5. CONCLUSION AND FUTUREWORK

In this study, we presented the Personalized Hyperspace Learning (PHL-IDOL) framework as an advanced approach to improve segmentation accuracy in radiation therapy planning. By incorporating patient-specific data and leveraging sophisticated metrics like image similarity and deformable vectors, PHL-IDOL significantly outperformed conventional models, including the general, continual, and IDOL models. Across diverse anatomical structures, PHL-IDOL demonstrated superior performance in both the Volumetric Dice Similarity Coefficient (VDSC) and Hausdorff Distance 95% (HD95) evaluations, particularly excelling in challenging small and complex organs such as cochleae and optic apparatus. This framework not only enhanced segmentation accuracy but also exhibited robustness and consistency, providing a reliable tool for clinical practice.

The success of PHL-IDOL underscores its potential to enhance precision in personalized radiotherapy by tailoring models to individual patient anatomies. The results also demonstrated the capability of PHL-IDOL to address challenges associated with inter-patient variability, achieving consistently high performance in segmentation across multi-institutional datasets. This adaptability across different clinical environments highlights its potential for broader application in various healthcare settings.

Looking ahead, the PHL-IDOL framework can serve as a foundational model for future advancements in radiation therapy and medical imaging. One promising avenue for future work is extending the PHL-IDOL methodology to dose prediction tasks, enabling more accurate and individualized radiation dose distributions. By integrating PHL-IDOL with dose prediction frameworks, it may be possible to create end-to-end solutions for personalized adaptive radiotherapy, optimizing treatment delivery and outcomes.

Additionally, further exploration of the PHL-IDOL model could involve expanding its application to other cancer types and treatment modalities, including proton therapy, which presents unique challenges in dose distribution. The integration of the PHL-IDOL framework into adaptive radiation therapy workflows could also pave the way for real-time model updates, enabling clinicians to dynamically adjust treatment plans based on ongoing patient-specific changes.

Future research could also focus on enhancing the computational efficiency of PHL-IDOL, enabling its application to larger and more complex datasets while maintaining high accuracy. Incorporating advanced similarity metrics, such as entropy difference and gradient correlation, could further refine the framework’s ability to identify relevant patient-specific characteristics. Moreover, exploring multi-modality imaging data, such as MRI-CT fusion, could provide additional avenues to improve segmentation and treatment accuracy.

In conclusion, the PHL-IDOL framework represents a significant advancement in personalized radiotherapy planning. Its integration into broader clinical workflows, combined with future developments in dose prediction and real-time adaptability, promises to revolutionize precision medicine in oncology. By addressing current challenges and exploring future applications, PHL-

IDOL has the potential to significantly enhance patient outcomes and streamline clinical processes in radiation therapy.

References

1. Jaffray, D.A. and M.K. Gospodarowicz, *Radiation therapy for cancer.* Cancer: disease control priorities, 2015. **3**: p. 239-248.
2. Bryant, A.K., et al., *Trends in radiation therapy among cancer survivors in the United States, 2000–2030.* Cancer Epidemiology, Biomarkers & Prevention, 2017. **26**(6): p. 963-970.
3. Demaria, S., E.B. Golden, and S.C. Formenti, *Role of local radiation therapy in cancer immunotherapy.* JAMA oncology, 2015. **1**(9): p. 1325-1332.
4. Douglass, M., *Eric J. Hall and Amato J. Giaccia: Radiobiology for the radiologist.* Australasian Physical & Engineering Sciences in Medicine, 2018. **41**(4): p. 1129-1130.
5. Meyer, J., *IMRT, IGRT, SBRT: advances in the treatment planning and delivery of radiotherapy*. 2011: Karger Medical and Scientific Publishers.
6. Ling, C.C., E. Yorke, and Z. Fuks, *From IMRT to IGRT: frontierland or neverland?* Radiotherapy and oncology, 2006. **78**(2): p. 119-122.
7. Xu, J., et al., *Radiation therapy in keloids treatment: history, strategy, effectiveness, and complication.* Chinese medical journal, 2017. **130**(14): p. 1715-1721.
8. Lederman, M., *The early history of radiotherapy: 1895–1939.* International Journal of Radiation Oncology* Biology* Physics, 1981. **7**(5): p. 639-648.
9. Zeman, E.M., E.C. Schreiber, and J.E. Tepper, *Basics of radiation therapy*, in *Abeloff's clinical oncology*. 2020, Elsevier. p. 431-460. e3.
10. Slater, J.M., *From X-rays to ion beams: a short history of radiation therapy*, in *Ion Beam Therapy: Fundamentals, Technology, Clinical Applications*. 2011, Springer. p. 3-16.
11. Do Huh, H. and S. Kim, *History of radiation therapy technology.* Progress in Medical Physics, 2020. **31**(3): p. 124-134.
12. Chiavassa, S., et al., *Complexity metrics for IMRT and VMAT plans: a review of current literature and applications.* The British journal of radiology, 2019. **92**(1102): p. 20190270.
13. Quan, E.M., et al., *A comprehensive comparison of IMRT and VMAT plan quality for prostate cancer treatment.* International Journal of Radiation Oncology* Biology* Physics, 2012. **83**(4): p. 1169-1178.
14. Li, X.A., *Adaptive radiation therapy*. 2011: CRC Press.
15. Wu, Q.J., et al., *Adaptive radiation therapy: technical components and clinical applications.* The cancer journal, 2011. **17**(3): p. 182-189.
16. Yan, D., et al., *Adaptive radiation therapy.* Physics in Medicine & Biology, 1997. **42**(1): p. 123.
17. Rusanov, B., et al., *Deep learning methods for enhancing cone-beam CT image quality toward adaptive radiation therapy: A systematic review.* Medical Physics, 2022. **49**(9): p. 6019-6054.
18. Rigaud, B., et al., *Automatic segmentation using deep learning to enable online dose optimization during adaptive radiation therapy of cervical cancer.* International Journal of Radiation Oncology* Biology* Physics, 2021. **109**(4): p. 1096-1110.
19. Choi, M.S., et al., *Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer.* Radiotherapy and Oncology, 2020. **153**: p. 139-145.
20. Morgan HE, Sher DJ. Adaptive radiotherapy for head and neck cancer. *Cancers of the Head & Neck.* 2020;5(1):1.
21. Schwartz DL, Garden AS, Thomas J, et al. Adaptive Radiotherapy for Head-and-Neck

Cancer: Initial Clinical Outcomes From a Prospective Trial. *Int J Radiat Oncol.* 2012;83(3):986-993.

22. Castadot P, Lee JA, Geets X, Gregoire V. Adaptive Radiotherapy of Head and Neck Cancer. *Seminars in Radiation Oncology.* 2010;20(2):84-93.

23. Veresezan O, Troussier I, Lacout A, et al. Adaptive radiation therapy in head and neck cancer for clinical practice: state of the art and practical challenges. *Jpn J Radiol.* 2017;35(2):43-52.

24. Zhang TZ, Chi YW, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: Toward on-line adaptive radiotherapy. *Int J Radiat Oncol.* 2007;68(2):522-530.

25. Malsch U, Thieke C, Huber PE, Bendl R. An enhanced block matching algorithm for fast elastic registration in adaptive radiotherapy. *Physics in Medicine and Biology.* 2006;51(19):4789-4806.

26. Wang LJ, Su HL, Liu P. Automatic right ventricular segmentation for cine cardiac magnetic resonance images based on a new deep atlas network. *Medical Physics.* 2023. doi: 10.1002/mp.16547.

27. Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiother Oncol.* 2020;153:139-145.

28. Lei Y, Fu YB, Tian Z, et al. Deformable CT image registration via a dual feasible neural network. *Medical Physics.* 2022;49(12):7545-7554.

29. Li JY, Udupa JK, Odhner D, Tong YB, Torigian DA. SOMA: Subject-, object-, and modality-adapted precision atlas approach for automatic anatomy recognition and delineation in medical images. *Medical Physics.* 2021;48(12):7806-7825.

30. Li C, Nie ZW, Yang XP. Splitting proximate algorithm for deformable image registration based on functions of bounded deformation. *Medical Physics.* 2022;49(8):5149-5159.

31. Cui SN, Tseng HH, Pakela J, Ten Haken RK, El Naqa I. Introduction to machine and deep learning for medical physicists. *Medical Physics.* 2020;47(5):E127-E147.

32. Luo Y, Chen SF, Valdes G. Machine learning for radiation outcome modeling and prediction. *Medical Physics.* 2020;47(5):E178-E184.

33. Seo H, Khuzani MB, Vasudevan V, et al. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical Physics.* 2020;47(5):E148-E167.

34. Janai J, Güney F, Behl A, Geiger A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends® in Computer Graphics and Vision.* 2020;12(1–3):1-308.

35. Islam ABMR. Machine Learning in Computer Vision. In: Khadimally S, ed. *Applications of Machine Learning and Artificial Intelligence in Education.* doi: 10.4018/978-1-7998-7776-9.ch002 Hershey, PA, USA: IGI Global; 2022:48-72.

36. Lu YZ, Young S. A survey of public datasets for computer vision tasks in precision agriculture. *Comput Electron Agr.* 2020;178.

37. Garbin C, Zhu X, Marques O. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications.* 2020;79(19):12777-12815.

38. Cobbe K, Klimov O, Hesse C, Kim T, Schulman J. Quantifying Generalization in Reinforcement Learning. Proceedings of the 36th International Conference on Machine Learning; 2019; Proceedings of Machine Learning Research.

39. Lemley J, Bazrafkan S, Corcoran P. Smart Augmentation Learning an Optimal Data

Augmentation Strategy. *IEEE Access.* 2017;5:5858-5869.

40. Chun J, Park JC, Olberg S, et al. Intentional deep overfit learning (IDOL): A novel deep learning strategy for adaptive radiation therapy. *Medical Physics.* 2022;49(1):488-496.
41. Kawula M, Hadi I, Nierer L, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. *Medical Physics.* 2023;50(3):1573-1585.
42. Maniscalco A, Liang X, Lin M-H, Jiang S, Nguyen D. Single patient learning for adaptive radiotherapy dose prediction. *Medical Physics.* 2023;50(12):7324-7337.
43. Choi B, Olberg S, Park JC, et al. Technical note: Progressive deep learning: An accelerated training strategy for medical image segmentation. *Medical Physics.*n/a(n/a).
44. Uchida T, Kin T, Saito T, et al. De-Identification Technique with Facial Deformation in Head CT Images. *Neuroinformatics.* 2023;21(3):575-587.
45. Simon Jégou MD, David Vazquez, Adriana Romero, Yoshua Bengio. The One Hundred Layers Tiramisu: Fully Convolutional Densenets for Semantic Segmentation. *arXiv.* 2020.
46. Li XX, Yu LY, Chang DL, Ma ZY, Cao J. Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification. *Ieee T Veh Technol.* 2019;68(5):4204-4212.
47. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing Images Using the Hausdorff Distance. *Ieee T Pattern Anal.* 1993;15(9):850-863

PUBLICATION LIST

Park J, Choi B, Ko J, et al. Deep-learning-based automatic segmentation of head and neck organs for radiation therapy in dogs. Frontiers in Veterinary Science. 2021;8:721612.

Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiotherapy and Oncology. 2020/12/01/ 2020;153:139-145.

Choi B, Olberg S, Park J, et al. Technical note: Progressive deep learning: An accelerated training strategy for medical image segmentation. Medical Physics. 02/17 2023;50

Choi BS, Beltran CJ, Olberg S, et al. Enhanced IDOL segmentation framework using personalized hyperspace learning IDOL. *Med Phys*. 2024; 51: 8568–8583.

Abstract in Korean

개인 맞춤형 적응형방사선치료를 위한 딥러닝 알고리즘 최적화

적응형 방사선 치료(Adaptive Radiation Therapy, ART)는 방사선 종양학 분야에서 환자의 실시간 해부학적 및 생리학적 변화를 기반으로 방사선량을 동적으로 조정함으로써 암 치료의 정밀성과 효과를 향상시키기 위해 고안된 혁신적인 접근 방식입니다. 기존의 방사선 치료는 정적인 치료 계획에 의존하지만, ART는 치료 과정 전반에 걸쳐 지속적으로 환자를 모니터링하고 치료 계획을 조정하여 방사선이 종양에 보다 정확히 전달되고 건강한 조직은 최대한 보호되도록 합니다. 이러한 실시간 적응성은 체중 감소, 종양 크기 감소 또는 치료 중 발생할 수 있는 기타 생리학적 변화로 인한 종양 크기, 모양, 위치 및 주변 장기 변화에 대처하는 데 필수적입니다.

그러나 ART의 임상적 도입은 몇 가지 주요 과제로 인해 제약을 받습니다. 첫째, 위험 기관(Organ-at-Risk, OAR)과 임상 표적 체적(Clinical Target Volumes, CTV)에 대한 수동 윤곽 작업이 노동 집약적이고 시간이 많이 소요됩니다. 둘째, 복잡한 딥러닝(DL) 모델의 훈련 시간이 길어 환자 맞춤형 치료의 적시 구현이 어렵습니다. 셋째, 기존 딥러닝 모델의 정확도와 일반화 가능성이 불충분하며 비개인화된 훈련 데이터셋에 의해 제한됩니다. 본 논문은 이러한 과제를 해결하고 개인 맞춤형 ART의 효과를 향상시키기 위해 딥러닝 알고리즘을 최적화하는 데 중점을 둡니다.

먼저, ART의 원칙과 ART 프로세스에서 딥러닝의 핵심 역할을 소개합니다. 그런 다음 수의학 분야의 머리 및 목 해부학과 유방암 치료에 적용된 최적화된 자동 윤곽 모델 개발을 제시하여 높은 정확도를 유지하면서 임상 워크플로를 간소화할 수 있는 잠재력을 입증합니다. 또한, 훈련 과정을 가속화하기 위해 모델 수렴 시간을 최적화하고 임상 환경에서 ART 솔루션의 신속한 배치를 가능하게 하는 Progressive Deep Learning(PDL) 프레임워크를 제안합니다. 결과적으로, 보다 정확하고 강력한 딥러닝 모델을 구현합니다. 이러한 모델들은 다기관 임상 데이터셋을 통해 검증되어 그 넓은 적용 가능성과 효과를 입증합니다.

이 논문은 개인 맞춤형 적응형 방사선 치료를 위해 딥러닝 알고리즘을 최적화함으로써 개인화 의료 및 환자 맞춤형 치료 계획을 위한 미래 혁신의 길을 열고, 각 환자가 가장 효과적이고 맞춤형 치료를 받을 수 있도록 합니다.

핵심되는 말 : ART, 최적화, 딥러닝, 자동 윤곽, 개인화, 머리 및 목