Learning with Uncertainty in Medical Image Segmentation

by

Sukesh ADIGA VASUDEVA

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, AUGUST 03, 2024

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Herve Lombaert, Thesis Supervisor
Department of Software and IT Engineering, École de technologie supérieure
Department of Computer Engineering and Software Engineering, Polytechnique Montréal

Mr. Jose Dolz, Co-Supervisor
Department of Software and IT Engineering, École de technologie supérieure

Mr. Jean-Marc Lina, Chair, Board of Examiners
Department of Electrical Engineering, École de technologie supérieure

Mr. Christian Desrosiers, Member of the Jury
Department of Electrical Engineering, École de technologie supérieure

Mr. Juan Eugenio Iglesias, External Examiner
Athinoula A. Martinos Center for Biomedical Imaging, MGH & Harvard Medical School

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JULY 12, 2024

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Herve Lombaert and Prof. Jose Dolz, for their invaluable guidance and support throughout this four-and-a-half-year journey. I am deeply thankful for the opportunities they provided me by accepting me as their student. I am grateful to have them as my supervisors for all the assistance and encouragement in every up and down stage of this journey. Their expertise and countless discussions have been instrumental in shaping my understanding of the Machine learning and Medical Imaging field. Their constructive feedback has significantly enhanced my writing and presentation abilities. The resources they provided at ETS and the freedom they granted me were crucial in enabling me to contribute to research and develop this thesis to its current state. I am truly grateful for all the opportunities provided by their mentorship during this pivotal phase of my career and life. Additionally, I would like to sincerely thank my master's supervisor, Prof. Jayanthi Sivaswamy, for showing me the essence of research.

I am also immensely thankful to the members of my thesis committee, Prof. Jean-Marc Lina, Prof. Christian Desrosiers, and Prof. Juan Eugenio Iglesias, for evaluating my thesis and providing their valuable insights.

I extend my appreciation to my colleagues at Shapets, Livia, and Neuro-iX. I want to first thank my Shapets labmates Karthik, Melanie, Pierre, Benoit, Mathilde, and Arash for all the memorable conversations, activities, and enriching lab and journal club meetings. Additionally, I want to thank Akhil, Saypra, Bala, Shambhavi, Sajjad, Hoel, Malik, Jerome, and Ziko for generously sharing their experiences and knowledge, spreading positivity, and offering assistance. Special thanks to Saypra, Bala, and Raghav for exploring restaurants around Montreal with me and motivating me to stay physically fit. Though I may inadvertently overlook some names, I am grateful to the many members of LIVIA and Neuro-iX. I am also glad to have had several inspiring and thought-provoking conversations with Prof. Christian Desrosiers and Prof. Sylvain Bouix throughout this journey.

At the beginning of my Ph.D. journey, Karthik and Raghav made my transition to the new city as seamlessly as possible. They have been close friends and later became flatmates. Always ready for technical discussions, arguments, or offering valuable suggestions, they pushed me beyond my boundaries. I would also like to thank Chetan for sharing his experience and knowledge. Thank you, Vasudha, for all your kindness and friendship. I want to thank Bala, a neighbor and fellow Ph.D. student, for his encouragement and insights into the industry. Additionally, I must thank Samruddhi, Rutuja, Sumedh, Nehal, and Smitha for all those trips, conversations beyond research, board games, and food during weekends in Montreal.

Finally, I extend heartfelt thanks to my family for their unconditional love, sacrifices, encouragement, and understanding throughout this journey. Their unwavering support has been my source of strength, especially when I lost something precious. I am indebted to my parents beyond words for never doubting my decisions and always offering kindness. I am grateful to my brother, Nagaraj, whose advice and mentorship have been invaluable to both my research and my life. Thank you, Vindhya, for the moral support that has helped me in every endeavor. My adorable nieces, Aadhya and Abhijna, always bring joy and smiles during difficult times. Last but not least, I am thankful to my wife, Anusha, who joined this journey midway and wholeheartedly supported me daily despite being far away, helping me believe in myself during challenging situations.

Apprentissage avec incertitude dans la segmentation d'images médicales

Sukesh ADIGA VASUDEVA

RÉSUMÉ

La segmentation d'images est essentielle dans de nombreuses applications cliniques et de recherche, telles que la caractérisation des maladies, la planification chirurgicale, les mesures diagnostiques et l'analyse des formes. Cependant, la délimitation manuelle prend du temps, peut nécessiter une expertise et est sujette à la variabilité. Les algorithmes automatisés offrent une solution à ces limitations, facilitant ainsi le flux de travail clinique et de recherche. De récentes techniques basées sur l'apprentissage profond ont permis de fournir avec succès une segmentation automatisée de haute qualité, utilisant généralement une quantité substantielle de données étiquetées. Cependant, les étiquettes peuvent être ambiguës ou peu fiables. Cette thèse s'attaque à ces défis avec pour objectif principal de développer des outils sensibles à l'incertitude qui peuvent aider à la formation de réseaux de segmentation d'images. En particulier, le premier objectif propose une stratégie d'étiquetage souple basée sur l'intensité pour s'attaquer aux ambiguïtés potentielles dans l'annotation. Le deuxième objectif présente une estimation de l'incertitude tenant compte de l'anatomie pour guider le réseau de segmentation sous une supervision limitée. Le troisième objectif propose une représentation basée sur l'attention pour une segmentation faiblement supervisée. Les résultats de ces objectifs de recherche ont donné lieu à trois revues, deux publications de conférence évaluées par des pairs et un court article de conférence. Les contributions de chaque objectif de recherche sont résumées ci-dessous.

Dans le premier objectif, nous proposons une approche de lissage des étiquettes géodésiques qui capture les détails d'intensité de l'image dans le processus d'étiquetage souple. Les intensités de l'image transmettent des informations qui pourraient clarifier les ambiguïtés potentielles dans l'annotation. Cependant, les méthodes d'étiquetage souple existantes ne reposent que sur des masques de segmentation, ignorant le contexte d'image sous-jacent associé à l'étiquette. Nous exploitons la transformation de distance géodésique pour capturer les variations d'intensité entre les pixels. Les cartes générées modifient les étiquettes dures pour obtenir de nouvelles étiquettes souples basées sur l'intensité. Les étiquettes souples géodésiques résultantes modélisent mieux les relations spatiales et par classe car elles capturent les variations des gradients d'image à travers les classes et l'anatomie. Les avantages de nos étiquettes souples géodésiques basées sur l'intensité sont évalués sur trois ensembles divers de jeux de données de segmentation accessibles au public. Nos résultats expérimentaux montrent que la méthode proposée améliore systématiquement la précision de la segmentation par rapport aux techniques d'étiquetage souple de pointe en termes de similarité de Dice et de distance de Hausdorff.

Le deuxième objectif vise à estimer l'incertitude en exploitant la représentation anatomiquement consciente pendant l'entraînement du réseau de segmentation dans des conditions semi-supervisées. Plus précisément, une représentation anatomiquement consciente est d'abord apprise pour modéliser les masques de segmentation disponibles. La représentation apprise mappe une prédiction de segmentation dans une segmentation anatomiquement plausible. L'écart

par rapport à la segmentation plausible aide à estimer les cartes d'incertitude au niveau des pixels sous-jacentes. Ces cartes filtrent les régions cibles non fiables pour guider le réseau de segmentation. La méthode proposée estime par conséquent l'incertitude en utilisant une seule inférence à partir de notre représentation, réduisant ainsi le calcul total pendant l'entraînement par rapport aux approches existantes tenant compte de l'incertitude. Nous évaluons notre méthode sur deux ensembles de données de segmentation accessibles au public. Notre approche anatomiquement consciente améliore la précision de la segmentation par rapport aux méthodes semi-supervisées de pointe en termes de deux mesures d'évaluation couramment utilisées.

Enfin, le troisième objectif propose d'apprendre une représentation dynamique basée sur l'attention pour l'analyse d'images médicales. En particulier, une représentation est apprise en intégrant un module d'attention dans un réseau d'intégration. Ce mécanisme d'attention intégré fournit un aperçu visuel direct des caractéristiques discriminantes du réseau d'intégration. De plus, un seul apprenant métrique est inadéquat pour apprendre une variété d'attributs d'objet dans les images, tels que la couleur, la forme ou les artefacts. Au lieu de cela, plusieurs apprenants métriques pourraient aider à apprendre différents aspects de ces attributs dans les sous-espaces d'une intégration globale. Cependant, le nombre d'apprenants doit être trouvé empiriquement pour chaque nouvel ensemble de données. Nous présentons donc un apprenant de sous-espace dynamique, qui supprime la nécessité de connaître *apriori* le nombre d'apprenants dans l'approche à apprenants multiples. Les avantages de notre représentation dynamique basée sur l'attention sont évalués dans l'application de la segmentation faiblement supervisée, du regroupement d'images et de la récupération d'images. Notre méthode fournit une carte d'attention directement pendant l'inférence pour illustrer l'interprétabilité visuelle des caractéristiques d'intégration. Ces cartes d'attention proposent des étiquettes proxy, améliorant la précision de segmentation jusqu'à 15% dans le score Dice par rapport aux techniques d'interprétation de pointe. De plus, notre méthode obtient des résultats compétitifs par rapport à l'approche d'apprentissage multimétrique et surpasse considérablement le réseau de classification en termes de scores de clustering et de récupération sur trois ensembles de données de référence publics différents.

Les travaux de recherche décrits dans cette thèse font progresser la segmentation des images médicales en supervision complète, semi-faible et faible. Nos étiquettes souples basées sur l'intensité améliorent la segmentation, en particulier dans les régions difficiles. Notre approche d'estimation de l'incertitude tenant compte de l'anatomie utilise efficacement une annotation limitée, réduisant ainsi le besoin d'étiquetage extensif. L'approche de représentation basée sur l'attention fournit une organisation structurée des données et une interprétabilité visuelle, permettant une segmentation avec uniquement des étiquettes au niveau de l'image. Cette thèse présente de nouveaux outils qui aident les cliniciens et les chercheurs en fournissant une délimitation plus rapide, cohérente et précise des objets cibles.

Mots-clés: Étiquetage souple, Incertitude anatomique, Apprentissage semi-supervisé, Apprentissage métrique, Apprentissage faiblement supervisé, Segmentation d'image

Learning with Uncertainty in Medical Image Segmentation

Sukesh ADIGA VASUDEVA

ABSTRACT

Image segmentation is vital in many clinical and research applications, such as disease characterizations, surgical planning, diagnostic measurements, and shape analysis. However, manual delineation is time-consuming, may require expertise, and is subject to variability. Automated algorithms offer a solution to these limitations, thereby assisting clinical and research workflow. Recent deep learning-based techniques have successfully provided high-quality automated segmentation, generally using a substantial amount of labeled data. However, the labels can be ambiguous or unreliable. This thesis tackles these challenges with the primary objective of developing uncertainty-aware tools that can aid in training image segmentation networks. Particularly, the first objective proposes an intensity-based soft labeling strategy to tackle potential ambiguities in the annotation. The second objective presents an anatomically-aware uncertainty estimation to guide the segmentation network under limited supervision. The third objective proposes an attention-based representation for weakly supervised segmentation. The findings from these research objectives have resulted in three journals, two peer-reviewed conference publications, and a short conference article. The contributions of each research objective are summarized below.

In the first objective, we propose a Geodesic Label Smoothing (GeoLS) approach that captures image intensity details within the soft labeling process. The image intensities convey information that could clear potential ambiguities in the annotation. However, existing soft-labeling methods rely only on segmentation masks, ignoring the underlying image context associated with the label. We leverage the geodesic distance transform to capture the intensity variations between pixels. The generated maps modify the hard labels to obtain new intensity-based soft labels. The resulting geodesic soft labels better model spatial and class-wise relationships as they capture the variations of image gradients across classes and anatomy. The benefits of our intensity-based geodesic soft labels are assessed on three diverse sets of publicly accessible segmentation datasets. Our experimental results show that the proposed method consistently improves the segmentation accuracy compared to state-of-the-art soft-labeling techniques in terms of the Dice similarity and Hausdorff distance.

The second objective aims to estimate uncertainty by leveraging anatomically-aware representation during training of segmentation network under semi-supervised settings. Specifically, an anatomically-aware representation is first learned to model the available segmentation masks. The learned representation maps a segmentation prediction into an anatomically plausible segmentation. The deviation from the plausible segmentation aids in estimating the underlying pixel-level uncertainty maps. These maps filter the unreliable target regions to guide the segmentation network. The proposed method consequently estimates the uncertainty using a single inference from our representation, reducing the total computation during training compared to existing uncertainty-aware approaches. We evaluate our method on two publicly

available segmentation datasets. Our anatomically-aware approach improves the segmentation accuracy over the state-of-the-art semi-supervised methods in terms of two commonly used evaluation measures.

Finally, the third objective proposes to learn an attention-based dynamic representation for medical image analysis. Particularly, a representation is learned by integrating an attention module into an embedding network. This integrated attention mechanism provides a direct visual insight into the discriminative features of the embedding network. Furthermore, a single metric learner is inadequate for learning a variety of object attributes in images, such as color, shape, or artifacts. Instead, multiple metric learners could aid in learning different aspects of these attributes in subspaces of an overarching embedding. However, number of learners is to be found empirically for each new dataset. We, therefore, present a dynamical subspace learner, which removes the need to know *apriori* the number of learners in the multiple learners approach. The benefits of our attention-based dynamic representation are evaluated in the application of weakly supervised segmentation, image clustering, and image retrieval. Our method provides an attention map directly during inference to illustrate the visual interpretability of the embedding features. These attention maps offer proxy labels, improving the segmentation accuracy by up to 15% in the Dice score compared to state-of-the-art interpretation techniques. Moreover, our method achieves competitive results compared to the multiple metric learner approach and significantly outperforms the classification network in terms of clustering and retrieval scores on three different public benchmark datasets.

The research work described in this thesis advances medical image segmentation across full, semi, and weak supervision. Our intensity-based soft labels enhance the segmentation, especially in challenging regions. Our anatomically-aware uncertainty estimation approach effectively uses limited annotation, reducing the need for extensive labeling. The attention-based representation approach provides structured data organization and visual interpretability, enabling segmentation with only image-level labels. This thesis presents new tools that assist clinicians and researchers by providing faster, consistent, and accurate delineation of target objects.

Keywords: Soft labeling, Anatomically-aware Uncertainty, Semi-supervised Learning, Metric learning, Weakly supervised learning, Image Segmentation

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

BraTS	Brain Tumor Segmentation
CAM	Class Activation Map
CE	Cross-Entropy
CNNs	Convolutional Neural Networks
CRF	Conditional Random Fields
CT	Computed Tomography
DAE	Denoising Autoencoder
DL	Deep Learning
DML	Deep Metric Learning
DSC	Dice Score Coefficient
FCN	Fully Convolutional Network
FRQNT	Fonds de Recherche du Québec Nature et Technologies
GAP	Global Average Pooling
GDT	Geodesic Distance Transform
GGD	Generalized Geodesic Distance
HD	Hausdorff Distance
JBHI	Journal of Biomedical and Health Informatics
LS	Label Smoothing

MCDO	Monte-Carlo Dropout
MedIA	Medical Image Analysis
MELBA	Machine Learning for Biomedical Imaging
MICCAI	Medical Image Computing and Computer-Assisted Intervention
MIDL	Medical Imaging with Deep Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NMI	Normalized Mutual Information
NSERC	Natural Sciences and Engineering Research Council of Canada
OH	One Hot
ReLU	Rectified Linear Unit
SDM	Signed Distance Map
SGD	Stochastic Gradient Descent
SSL	Semi-Supervised Learning
WSS	Weakly Supervised Segmentation

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathcal{D}$	training dataset consists of set of 2D or 3D images
$(\mathbf{X}^i, \mathbf{Y}^i)$	i^{th} image and its corresponding ground truth in dataset
$\hat{\mathbf{Y}}^i$	model prediction for an input image $\mathbf{X}^i$
$\mathbf{P}^i$	predicted probability for an input image $\mathbf{X}^i$
$y_{v,c}$	ground truth value at a pixel or voxel v for a class c
$\hat{y}_{v,c}$	predicted segmentation value at a pixel or voxel v for a class c
$p_{v,c}$	predicted probability value at a pixel or voxel v for a class c
C	number of classes
Ω	spatial image domain
$\mathcal{L}$	loss function
$\mathcal{S}$	set of seed points
$\mathcal{D}_L$	set of labeled dataset
$\mathcal{D}_U$	set of the unlabeled dataset
$\mathbf{U}^i$	uncertainty map for an image $\mathbf{X}^i$
θ	network parameters
N_l	training samples in labeled dataset
N_u	training samples in unlabeled dataset
H	threshold value

d	embedding size
$d(\cdot, \cdot)$	distance metric
$A(\cdot)$	attention module
$\mathbf{A}^i$	attention map for an image $\mathbf{X}^i$
$S(\cdot)$	feature extractor

INTRODUCTION

0.1 Anatomical Representation to Imaging

Anatomy has been studied for centuries to interpret the structures and physiology of the human body. Early methods involved direct observation of dissection and vivisection of animal and human bodies, providing valuable insight into the fundamentals of anatomy. These observations subsequently enabled the creation of anatomical drawings (Keele, 1964; Vesalius, 1543) and atlases (Braune, 1872) (Fig. 0.1a-b). Such illustrations became vital tools for visual understanding and studying anatomy in medical education as well as in traditional surgery. The advent of photography later facilitated surgeons' ability to document diseases and anatomy more accurately for clinical case studies. The discovery of X-rays by Wilhelm Roentgen in 1895 (Roentgen, 1931) showed a noninvasive way to capture internal anatomy (Fig. 0.1c). Such imaging technology quickly evolved, becoming a standard tool for medical diagnosis by offering a noninvasive visual representation of anatomy.

Modern medical imaging has since grown remarkably by incorporating cutting-edge technologies in healthcare (Bradley, 2008; Wolbarst, Capasso & Wyant, 2013). Over the decades, many imaging techniques, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), optical imaging (including tomography and microscopy), and ultrasonography, have been developed and transformed the landscape of medical imaging domain. These techniques generate an image by gathering measurements of an object of interest through advanced sensors. The source used in these measurements varies across the broad spectrum of electromagnetic waves, such as X-ray in radiography, radio frequency in MRI, sound waves in ultrasonography, and visible light in dermatoscopy. Various sources lead to visualizations of different anatomical structures, tissues, tumors, and bones, providing increasingly detailed and two- or three-dimensional images. The different types of images are typically referred to as modalities (Suetens, 2017). Multi-modal imaging captures different

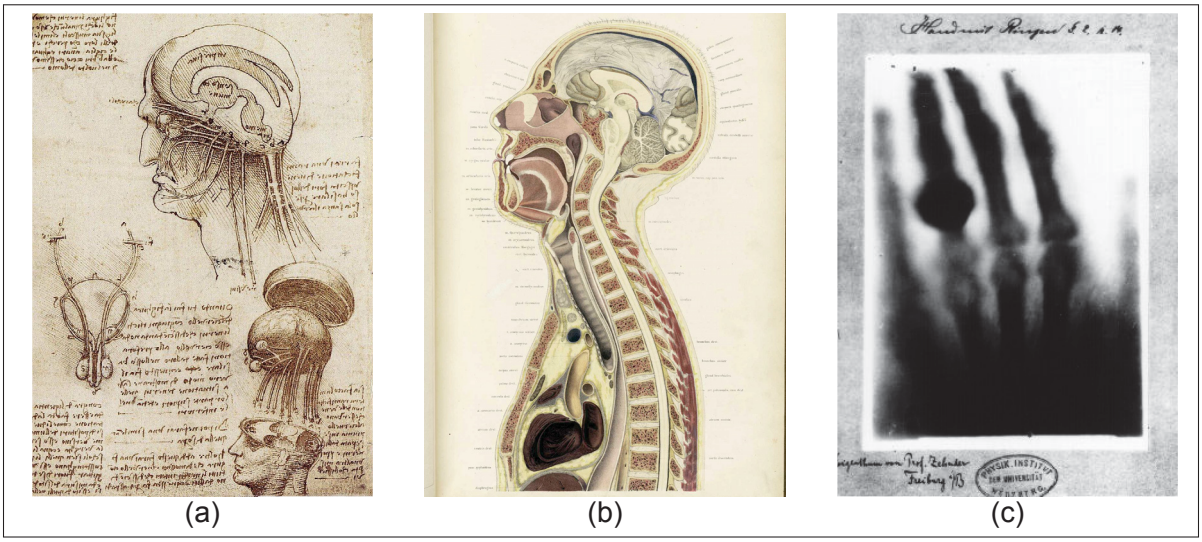

Figure 0.1 Illustration of human anatomical structures from drawing to imaging. (a) a drawing of the brain and skull by Leonardo da Vinci (1452–1519), (b) a cross-section anatomical atlas by Wilhelm Braune (1831-1892), and (c) the first X-ray imaging of a hand by Wilhelm Roentgen (1895).
Taken from (a) Wikipedia contributors (2024a), (b) Braune (1872), and (c) Wikipedia contributors (2024b)

tissue characteristics using multiple sources at the same location. Examples of different imaging modalities are shown in Figure 0.2. These varying image modalities have significantly enhanced diagnostic capabilities, enabling the detection and characterization of various medical conditions (Suetens, 2017; Taylor, 1996; Van Ginneken, Schaefer-Prokop & Prokop, 2011).

0.2 Medical Image Analysis

Medical images possess rich information about the patient's health in a noninvasive manner. These images are assessed to extract meaningful information to quantify a disease, plan treatment, and monitor diverse medical conditions (Duncan & Ayache, 2000). The benefits of images are reflected in a growing number of imaging exams performed in healthcare (Alexander, McGill, Tarasova, Ferreira & Zurkiya, 2019; Richards, Maskell, Halliday & Allen, 2022; Smith-Bindman *et al.*, 2019). It is estimated that about 4.2 Billion imaging exams are performed per year globally

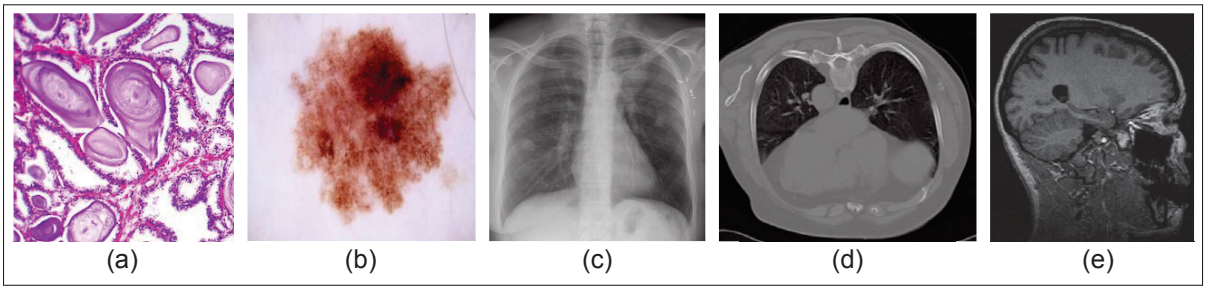

Figure 0.2 Examples of imaging modalities that capture different regions of human anatomy. (a) a high-resolution breast histology image, (b) a microscopic image of a skin lesion, (c) a grayscale chest X-ray image, and (d)-(e) 2D slices from the 3D volume of abdominal CT and brain MR images, respectively.
Taken from (a) Aresta *et al.* (2019), (b) Combalia *et al.* (2019), and (c), (d), (e) Suetens (2017)

(Mahesh, Ansari & Mettler Jr, 2022), showing that medical imaging has been a well-established tool in modern healthcare systems.

Over the years, the growth of imaging has been associated with advancements in hardware and digital technology. However, image analysis yet relies on trained clinicians or radiologists. Analyzing images solely by human experts is laborious, expensive, and prone to errors. For instance, manual delineation of organs or tumors can take hours or even days for a single patient (Shi *et al.*, 2022). It is critical in several time-sensitive clinical examinations, such as interventions, treatments, screening, computer-aided diagnosis, and prognosis. Moreover, manual analysis by human experts is limited to meet the rapidly growing pace of image-based examination[1] (Konstantinidis, 2023; Sokolovskaya *et al.*, 2015). These examinations will significantly burden the healthcare system, affecting delayed diagnosis and contributing to errors (Winder, Owczarek, Chudek, Pilch-Kowalczyk & Baron, 2021). The expensive nature of manual analysis also hinders scaling for screening programs and large-scale studies. Also, there exists variability in image analysis by multiple experts, which leads to ambiguity, causing delayed or missed diagnosis (Becker *et al.*, 2019; Hsieh *et al.*, 2022).

[1] https://www.rsna.org/news/2022/may/global-radiologist-shortage

Automated computation algorithms have the potential to address the limitations mentioned earlier by assisting radiologists or clinicians (Hardy & Harvey, 2020; Langlotz, 2019). Such algorithms provide faster and more consistent inference than humans, which may aid clinicians in reducing analysis time and likely decreasing errors due to workload (Burton, Albur, Eberl & Cuff, 2019). These algorithms naturally scale the analytical demands for screening programs and large-scale studies.

The origin of an automated image analysis tool can be traced back to 1970, when a computer algorithm semi-automatically delineated a left ventricle, enabling the direct quantification of ejection fraction in the heart region (Strauss, Zaret, Hurley, Natarajan & Pitt, 1971). Such computer algorithms were gradually adopted into clinical systems. Duncan & Ayache (2000) summarizes the initial advancements in various image analysis tasks within the medical field.

In the early stages, automated algorithms primarily relied on handcrafted features, demonstrating promising results in various analysis tasks (Heimann & Meinzer, 2009; Van Leemput, Maes, Vandermeulen & Suetens, 1999). Such models need to be optimized for each image or task, resulting in a slower analysis. Data-driven models overcome these limitations by learning relationships between images and desired output for a given task. Earlier methods relied on statistical learning to design such models (Learned-Miller, 2005). Recent advancements in deep learning and computational capabilities have enabled the modeling of complex nonlinear relationships within the data (Goodfellow, Bengio & Courville, 2016; Prince, 2023; Zhang, Lipton, Li & Smola, 2023). These models learn the features directly from images and achieve state-of-the-art performance on diverse medical image analysis tasks (Ayache & Duncan, 2016; Litjens *et al.*, 2017; Zhou, Greenspan & Shen, 2023).

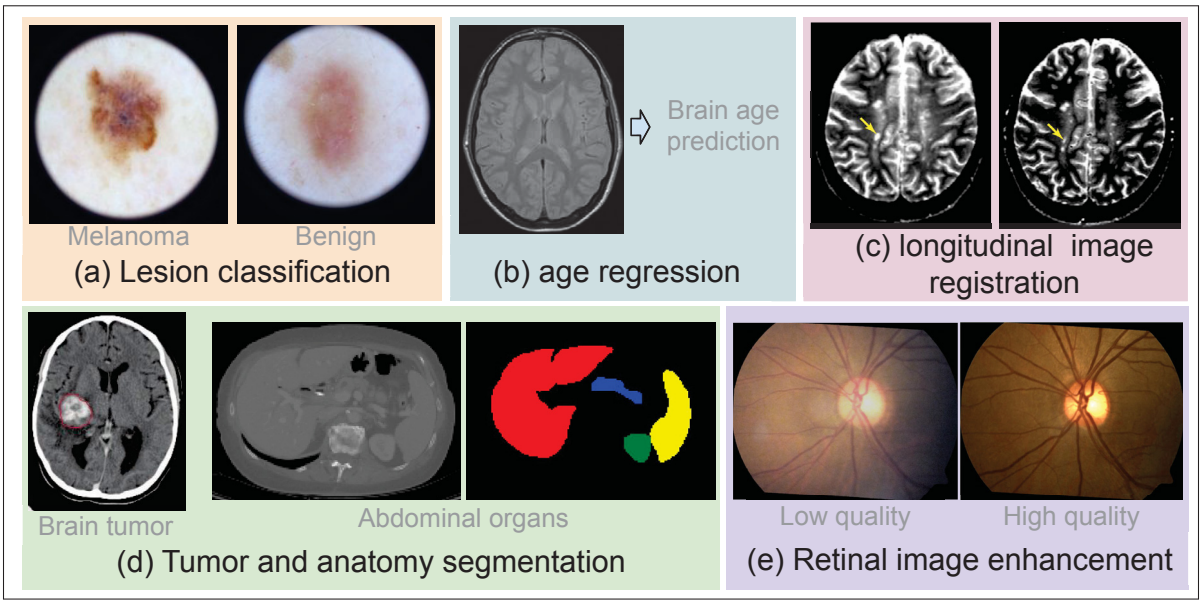

Figure 0.3 Examples of medical image analysis tasks. (a) skin lesion classification as melanoma or benign, (b) brain age prediction via regression, (c) alignment of brain images, (d) delineation of a brain tumor and abdominal organs, and (e) enhancement of retinal image. Taken from (a) Combalia *et al.* (2019), (b), (c), (d-left) Suetens (2017), and (d-right) Ma *et al.* (2021b), and (e) Adiga (2019)

0.3 Common tasks in Medical Image Analysis

The specific medical image analysis tasks involve classification, regression, registration, segmentation, and image enhancement (Zhou *et al.*, 2023). For instance, a classification task is a prediction of a category at *image-level*, which is helpful in detection and screening systems. A continuous value is predicted in the regression task, such as a volume or age prediction from an image. Some analysis requires aligning two or more images at *pixel-level*, called registration. Medical images are also degraded due to noise or acquired at low quality (for low dose purposes), where enhancement tasks aid by mapping them to high-quality images. Delineating an anatomy or a pathology at *pixel-level* is critical in diagnosis and treatment, which is referred to as image segmentation. Examples of medical image analysis tasks are depicted in Fig 0.3. Among these tasks, segmentation is crucial in downstream clinical applications, such as volume measurement, planning radiation therapy, monitoring of disease progression, or cell counting. Figure 0.4

broadly summarizes the use of image segmentation in downstream clinical applications. This thesis mainly focuses on the medical image segmentation task under different data settings, which are discussed in the next section.

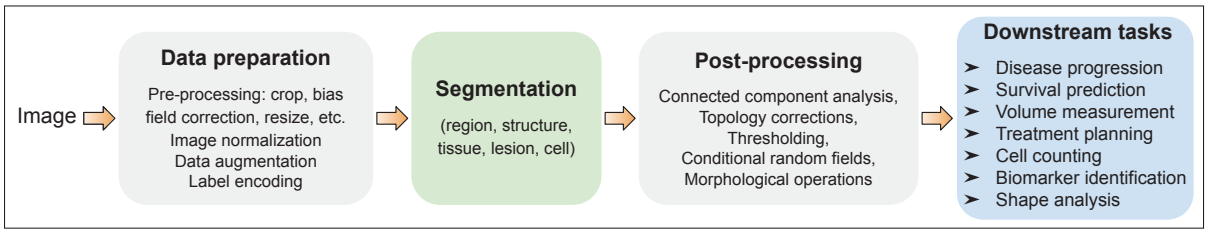

Figure 0.4 Illustration of how segmentation is utilized in medical image analysis applications. It involves data preparation, segmentation, and post-processing of segmented masks, which are subsequently used in downstream clinical applications

0.4 Challenges and Motivation

The collection of medical imaging data is increasing due to advances in imaging techniques (Li, Zhang, Müller & Zhang, 2018b; Zhang & Metaxas, 2016) as well as collaborative initiatives (Oakden-Rayner, 2020) in the community, e.g., UK biobank [2], ADNI [3] and grand challenges [4]. An increasing number of medical images, coupled with the labor-intensive nature of manual analysis and the lack of radiologists, emphasize the need to develop automated image analysis tools. Many classical algorithms are computationally complex to analyze such a growing scale of images. Recent advancements in computer vision and deep learning show a promising direction in handling large-scale data for most visual tasks. Nevertheless, these deep models are often driven by substantial amounts of annotated data.

The collection of medical imaging data varies in terms of both targeted imaging and annotations. According to estimates by the World Health Organization (WHO), there are approximately 2

[2] https://www.ukbiobank.ac.uk/imaging-data/

[3] http://adni.loni.usc.edu/

[4] A grand challenge is a platform for end-to-end development of machine learning solutions in biomedical imaging. https://grand-challenge.org/

million types of medical devices spanning over 7000 categories [5], contributing a diverse array of image collections. For instance, low-income countries often employ affordable imaging devices, where image quality varies. Such diverse image collections pose challenges in building segmentation tools for different data scenarios. Furthermore, obtaining annotation is labor-intensive, expensive, and demands expert knowledge of medical data. In addition, the segmentation task requires pixel-wise or voxel-wise annotations, amplifying the complexity of obtaining precise annotations. Consequently, various annotation types are leveraged for segmentation tasks, including image-level, point-based, scribble, and bounding box annotations. Depending on the type and amount of annotations, different machine learning techniques are employed, including fully supervised, semi-supervised, or weakly supervised techniques. For instance, fully supervised methods leverage image-annotation pairs to train segmentation models. These learning techniques are effective when extensive labeled data is accessible. However, obtaining a large number of annotated data can be challenging. A semi-supervised approach tackles such issues by combining unlabeled data with a small amount of labeled data. When sparse annotations such as image tags, points, or scribble are available, a weakly supervised method can be developed for the segmentation task.

Nevertheless, the annotations can be unreliable for various reasons, resulting in suboptimal training of segmentation models. For instance, the annotation can be ambiguous in challenging regions. These ambiguities originate from poorly defined image intensities due to low contrast, variations in image acquisition, partial volume effect, or motion artifacts. In scenarios with limited annotation, pseudo labels are derived from model predictions of unlabeled data, or proxy labels are formed using saliency maps from a network trained with weak labels such as image tags. These generated labels are employed in the training of segmentation models. The reliability of these generated labels significantly influences the effectiveness of learning the segmentation in such cases.

[5] https://www.who.int/health-topics/medical-devices

0.5 Research Objectives and Contributions

In the previous section, we highlighted the general challenges of learning medical image segmentation models. The main objective of this thesis is to tackle these challenges by developing a set of uncertainty-aware tools that can aid in training image segmentation networks. As the challenges vary with different labeling scenarios, we address the main objective with three specific objectives. The first objective proposes to develop an intensity-based soft labeling strategy to tackle potential ambiguities in the annotation. The second objective is to build an anatomically-aware representation for uncertainty estimation in order to guide the segmentation network training under limited supervision. The third objective proposes to learn an attention-based representation that provides reliable proxy labels for weakly supervised segmentation tasks. The specific details of these three objectives of this thesis are as follows:

1. **Intensity-based soft labeling for image segmentation:** The first objective of this thesis is to integrate image information in soft labeling for image segmentation. In conventional segmentation approaches, annotation masks are typically encoded in the form of hard labels. Such encoding lacks inter-class relationships in the image and spatial relationships between a given pixel and its neighbors. These relationships are essential in image segmentation, as pixel-level prediction depends on its neighbors. Soft-label assignments alleviate these limitations of hard labels. However, existing soft-labeling methods rely only on annotation masks to train a model. These approaches do not provide reliable soft labels as they ignore the underlying image context associated with the label. The image intensities convey information that could help to clear potential uncertainties in the annotation. The proposed work incorporates the image intensity information within the soft labeling process. The resulting intensity-based soft labels better model spatial and class-wise relationships by capturing the variations of image gradients across anatomy and labels. The empirical results validate the benefits of using intensity information in our soft labeling for segmentation

tasks. This research contributes new intensity-based soft labels, offering potential solutions for applications facing challenges in annotation due to ambiguities in image intensities across labels.

2. **Anatomically-aware uncertainty for semi-supervision:** The second objective of the thesis is to learn an anatomically-aware uncertainty estimation for semi-supervised segmentation. Current semi-supervised methods leverage unlabeled images by generating pseudo labels from model predictions or regularizing their model predictions during the training process. The reliability of predictions is critical in training such models. The uncertainty-aware approaches address this issue by guiding the model with reliable target regions. However, the existing uncertainty methods rely on multiple inferences from model predictions, which is computationally expensive. Moreover, these uncertainty maps capture pixel-wise disparities and lack anatomical knowledge of the data. We present a method that learns an anatomically-aware representation from the available segmentation labels. The learned representation will provide the uncertainty maps to guide the training of the segmentation model. The representation enables the uncertainty estimates using a single inference, thereby minimizing the total computation. The results from various experiments validate the benefits of our anatomical-aware uncertainty for image segmentation under semi-supervised settings. The proposed anatomically-aware approach effectively leverages the limited labels with enhanced segmentation accuracy, reducing the annotation cost.

3. **Attention-based representation for weak-supervision:**

   The third objective of the thesis is to learn an attention-based representation that provides proxy labels for weakly-supervised segmentation. Existing weakly-supervised methods employ different types of supervision, including image-level labels, points, scribbles, or bounding boxes. An image-level label is commonly employed as it is one of the inexpensive forms of weak supervision. Current methods based on such supervision produce saliency regions from the image classification network using class activation maps or attention

maps. The generated saliency maps are subsequently used as proxy labels for semantic segmentation. However, these salient maps mainly focused on the most discriminant areas. This research uses a deep metric learning technique to obtain reliable saliency regions from an embedding network. The proposed attention-based representation method dynamically learns embedding space using multiple learners and directly provides visual attention maps. The generated maps act as proxy labels for weakly supervised segmentation. The experiments highlight the effectiveness of our proxy labels obtained from attention-based representation for the image segmentation task. The representation is also validated for clustering and retrieval tasks.

Overall, this thesis contributes towards improving image segmentation across different levels of supervision through three research objectives. The overview of thesis contributions is depicted in Fig. 0.5. The background Chapter provides a detailed review of various segmentation regimes and related work required to understand this thesis.

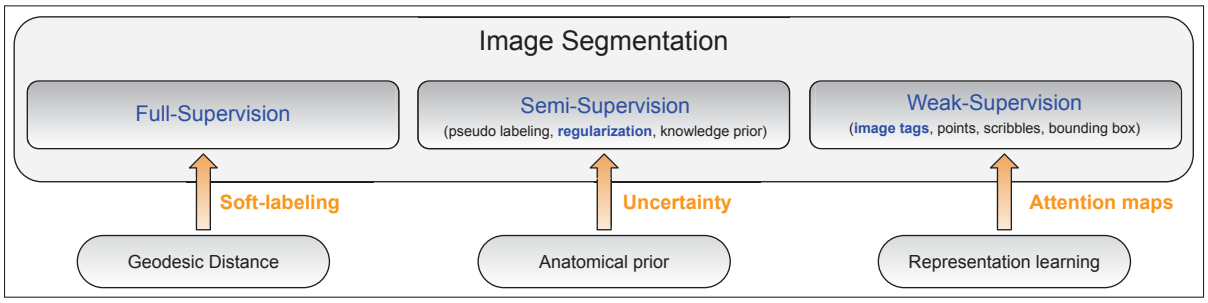

Figure 0.5 **Outline of thesis contributions.** This thesis explores different supervision employed in image segmentation. The research objectives aim to leverage various cues, highlighted in orange, to enhance the different segmentation regimes outlined in blue

0.6 Thesis Outline

The organization of the work reported in the thesis is described in this section. This introductory chapter provided an overview of medical image analysis, challenges in segmentation methods,

motivation, and research contribution of this thesis. **Chapter 1** presents the literature on the state-of-the-art methods in different types of image segmentation and other related areas required for the thesis, such as distance transform, anatomical prior, and representation learning techniques. **Chapter 2** presents our first research work on intensity-based soft labeling for image segmentation. The content of this chapter corresponds to the journal article *"GeoLS: an Intensity-based, Geodesic Soft Labeling for Image Segmentation"* submitted to the Journal of Machine Learning for Biomedical Imaging (MELBA), one of the emerging journals in the field of medical image analysis. An initial article of this work was published at the Medical Imaging with Deep Learning (MIDL) conference and presented as an oral talk. **Chapter 3** introduces the anatomically-aware uncertainty estimation for semi-supervised segmentation. This chapter corresponds to the journal article entitled *"Anatomically-aware Uncertainty for Semi-supervised Image Segmentation"* published in the Journal of Medical Image Analysis (MedIA), recognized as one of the premier journals within the community. A part of this work was initially published in Medical Image Computing and Computer-Assisted Intervention (MICCAI), a leading conference in the field. **Chapter 4** presents an attention-based representation that dynamically learns embedding space and provides attention maps for weakly supervised segmentation. The content presented in this chapter corresponds to the journal article titled *"Attention-based Dynamic Subspace Learners for Medical Image Analysis"* published in the Journal of Biomedical and Health Informatics (JBHI), considered one of the top journals in medical image analysis. This journal article was also presented as a short paper at the Medical Imaging with Deep Learning (MIDL) conference. Finally, the **Conclusion Chapter** summarizes the works and discusses its limitations, recommendations, and future scope of the presented work.

0.7 Published Work

Findings in this thesis have led to the following publications.

– Journals:

1. **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve. "GeoLS: an Intensity-based Geodesic Soft Labeling for Image Segmentation". *Submitted to Journal of Machine Learning for Biomedical Imaging (MELBA) - 2024*.

2. **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve. "Anatomically-aware Uncertainty for Semi-supervised Image Segmentation". *Medical Image Analysis (MedIA) - 2023*.

3. **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve. "Attention-based Dynamic Subspace Learners for Medical Image Analysis". *IEEE Journal of Biomedical And Health Informatics (JBHI) - 2022*.

– Conferences:

1. **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve. "GeoLS: Geodesic Label Smoothing for Image Segmentation". *International Conference on Medical Imaging with Deep Learning (MIDL) - 2023*.

2. **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve. "Leveraging Labeling Representations in Uncertainty-based Semi-supervised Segmentation". *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) - 2022*.

– Short papers:

1. **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve. "Attention-based Dynamic Subspace Learners". *International Conference on Medical Imaging with Deep Learning (MIDL) - 2022*.

Other Publications:

Apart from the aforementioned publications, I had opportunities to collaborate with a few other publications during the course of my doctoral journey.

1. Murugesan Balamurali, **Adiga Vasudeva Sukesh**, Liu Bingyuan, Lombaert Herve, Ben Ayed Ismail, Dolz Jose. "Trust your neighbours: Penalty-based constraints for model Calibration". *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) - 2023*.

2. Chauvin Laurent, **Adiga Vasudeva Sukesh**, Dolz Jose, Lombaert Herve, Toews Matthew. "A Large-scale Neuroimage Analysis using Keypoint Signatures: UK Biobank". *International Conference on Organization for Human Brain Mapping (OHBM) - 2020.*

0.8 Code Availability

Each work in this thesis is implemented in Python programming language with the PyTroch library (Paszke *et al.*, 2019) [6]. The code and scripts are available in the following links:

- Intensity-based soft labeling for image segmentation:
  https://github.com/adigasu/GeoLS

- Anatomically-aware uncertainty for semi-supervision:
  https://github.com/adigasu/Anatomically-aware_Uncertainty_for_Semi-supervised_Segmentation

- Attention-based representation learners for weak-supervision:
  https://github.com/adigasu/Dynamic_subspace_learners

[6] https://pytorch.org/

CHAPTER 1

BACKGROUND

Overview

Segmentation is essential in numerous image analysis and computer vision applications, spanning various domains such as medical imaging, machine vision, scene understanding, video surveillance, autonomous driving, and augmented reality. This thesis focuses on medical imaging applications, as segmentation plays a pivotal role in our healthcare system. For instance, it precisely delineates organs or tumors, aids in measuring diagnostic information such as volume (e.g., monitoring atrophy) or ejection fraction in cardiology, and assists in planning radiotherapy and surgeries. This chapter provides an overview of prominent segmentation approaches from traditional to deep learning techniques, different types of supervision used, and a discussion of a few related literature.

1.1 Automated Image Segmentation

Segmentation consists of dividing an image into distinct regions so that pixels with similar characteristics are assigned the same class label. This technique is often used to detect objects, structures, landmarks, boundaries, or anomalies in an image. For example, a clinician desires to delineate organs or tumors, as in Fig. 1.1, for pre-operative planning, volume quantification, tumor screening, or survival prediction. The quality of these delineations is crucial for such tasks, as false or incorrect labeling could lead to a wrong diagnosis, treatment, or analysis. Since manual delineation is expensive, automated segmentation algorithms are preferred to provide high-quality per-pixel delineation to assist the clinician. Over the years, several segmentation algorithms have been developed by leveraging domain-specific knowledge to address segmentation problems in medical images. This section briefly reviews image segmentation methods, spanning from classical to deep learning-based approaches.

Figure 1.1 Example of images and their corresponding segmentation ground truth of (a) abdominal organs and (b) brain tumors, obtained from clinical experts

1.1.1 Traditional segmentation approaches

Early segmentation techniques relied upon simple image properties or characteristics to partition an image into meaningful regions. These image properties include intensity, texture, shape, color, and spatial relationships, which are leveraged in various ways. A brief discussion of some of these methods follows.

Thresholding is the simplest way of segmenting an image by selecting a threshold value to divide the object of interest from their background (Otsu, 1979; Sezgin & Sankur, 2004). Edge-based approaches employ filters such as Sobel or Canny Edge detector (Canny, 1986) to detect and link the image edges with object boundaries. In clustering methods, pixels are grouped based on their distance to cluster centers (Coleman & Andrews, 1979). These approaches often use k-means clustering with image features, such as intensity, color, texture, and location, to measure the distance (Achanta *et al.*, 2012). Region-based approaches iteratively merge or split pixels or regions based on similarity in image properties (Gould, Gao & Koller, 2009). These methods can be bottom-up, where pixels are merged, e.g., region growing (Adams & Bischof, 1994) or top-down, where regions are split, e.g., region merging (Nock & Nielsen, 2004), watershed method (Vincent & Soille, 1991) to obtain the segmented regions.

A graph partitioning approach represents an image using graph theory, where each image element, such as a pixel or superpixel (i.e., groups of pixels), is a node, and their relationships

form edges. The goal is to segment the graph into disjoint subsets that correspond to meaningful regions in the image (Boykov, Veksler & Zabih, 2001; Peng, Zhang & Zhang, 2013). For instance, (Shi & Malik, 2000) cuts the graph based on similarity within regions as well as dissimilarities across different regions. Grady (2006) assigns a label to each node based on its likelihood to the seed points during a random walk. Alternatively, a deformable method evolves iteratively to identify object boundaries, where internal forces encourage smoothness and external forces attract the contour towards object boundaries (Chan & Vese, 1999; Kass, Witkin & Terzopoulos, 1988; Xu, Pham & Prince, 2000).

Conversely, early machine learning methods employ handcrafted features extracted from images to train a classifier, where the model learns to predict pixels or groups of pixels into different categories (Bezdek, Hall & Clarke, 1993; Hall *et al.*, 1992). Feature extraction techniques include histograms, wavelet transforms, filters, or local binary patterns. Training data is then prepared with pairs of feature vectors and corresponding class labels, which are subsequently used to train the machine learning classifiers, such as support vector machines (Wang, Wang & Bu, 2011), decision trees (Shotton, Johnson & Cipolla, 2008), random forests (Breiman, 2001; Schroff, Criminisi & Zisserman, 2008), and k-nearest neighbors (Shen, Spann & Nacken, 1998).

In summary, traditional methods offer simplicity and interpretability for tasks where the data is not too complex. However, their performance may be limited in scenarios where data is highly variable or noisy. Additionally, some methods can be time-consuming and require domain expertise. The inference can also be slow due to iterative computation in a few techniques. Therefore, it is challenging to determine a single segmentation algorithm that generalizes across tasks and datasets.

1.1.2 Towards Deep learning-based segmentation

Over the last decade, Deep Learning (DL) techniques have been growing in popularity, showcasing promising developments in language processing, audio analysis, and computer vision tasks,

including semantic segmentation (Long, Shelhamer & Darrell, 2015; Milletari, Navab & Ahmadi, 2016; Ronneberger, Fischer & Brox, 2015). The success of DL methods lies in their ability to extract complex patterns from extensive datasets automatically.

The fundamental principles of deep learning mirror those of traditional neural networks. A neural network consists of layers of neurons followed by activation functions, e.g., sigmoid or rectified linear unit (Nair & Hinton, 2010). These neurons are usually fully connected with the next layer, forming an input layer, an output layer, and a set of intermediate layers. Such arrangement of these layers allows them to learn hierarchical features directly from the images, capturing subtle patterns and non-linear relationships within the data. Integrating convolutional layers forms Convolutional Neural Networks (CNNs) further improved the ability to learn patterns from both the input image and intermediate feature maps (LeCun *et al.*, 1989). A deep neural network encompasses numerous such layers and neurons, thereby acquiring an increased capacity for learning complex representations of data (LeCun, Bengio & Hinton, 2015). In addition, the pooling operation reduces feature dimensionality by preserving semantically similar features, whereas normalization layers (Ioffe & Szegedy, 2015) stabilize the training process by ensuring features have similar distributions. These key components in deep learning have substantially enhanced the training of networks. For instance, the CNNs in classification networks often use a combination of these layers followed by fully connected (or dense) layers for output class prediction, e.g., AlexNet (Krizhevsky, Sutskever & Hinton, 2012), VGG (Simonyan & Zisserman, 2015), ResNet (He, Zhang, Ren & Sun, 2016), DenseNet (Huang, Liu, Van Der Maaten & Weinberger, 2017).

Image segmentation associates a label to every pixel in an input image, requiring precise spatial alignment between an input image and model output. Earlier classification CNNs are not directly suitable for segmentation tasks due to the inclusion of pooling layers, which downsample the input image. Therefore, a mechanism to reverse this downsampling process is needed to obtain spatially aligned segmentation output. Up-sampling and transposed convolutional layers

address this by expanding the spatial resolution of downsampled feature maps, facilitating pixel-to-pixel learning. For instance, Fully Convolutional Networks (FCN) (Long *et al.*, 2015) employ up-sampling on all pooling outputs to the original spatial dimension and combine them to learn dense segmentations. The FCN also replaces fully connected layers in the classification network with convolutional blocks, which allows learning and inferring an arbitrary input size. Based on this idea, various segmentation architectures have been developed in the vision and medical community, such as U-Net (Çiçek, Abdulkadir, Lienkamp, Brox & Ronneberger, 2016; Ronneberger *et al.*, 2015), V-Net (Milletari *et al.*, 2016), DeepLab (Chen, Papandreou, Kokkinos, Murphy & Yuille, 2017), SegNet (Badrinarayanan, Kendall & Cipolla, 2017).

The U-Net architecture (Ronneberger *et al.*, 2015) is widely used in medical image segmentation applications. It consists of encoder and decoder blocks with fully convolutional layers, as shown in Fig 1.2. The encoder block reduces the resolutions like in classification networks, whereas the decoder block gradually increases the spatial resolutions to obtain an output segmentation that matches the input image size. In addition, skip connections are used between the encoder-decoder blocks to enable a direct flow of information and help preserve high-resolution features from the earlier layers. Since medical data are often 3-dimensional (3D) volumes (e.g., CT and MR scans), the standard U-Net architecture can be trivially extended to 3D to leverage volumetric information (Çiçek *et al.*, 2016; Milletari *et al.*, 2016). This extension has proven advantageous for numerous segmentation tasks (Isensee, Jaeger, Kohl, Petersen & Maier-Hein, 2021; Ma *et al.*, 2021a; Mehta & Arbel, 2018; Schlemper *et al.*, 2019; Wang *et al.*, 2022b).

1.2 Learning Techniques in Image Segmentation

The training and evaluation of the segmentation model can vary with the specific objectives and applications (Minaee *et al.*, 2021). Based on the task, the most common categories are semantic segmentation, instance segmentation, and panoptic segmentation. *Semantic segmentation* aims to assign each pixel in an image to one of the known classes. In contrast, *instance*

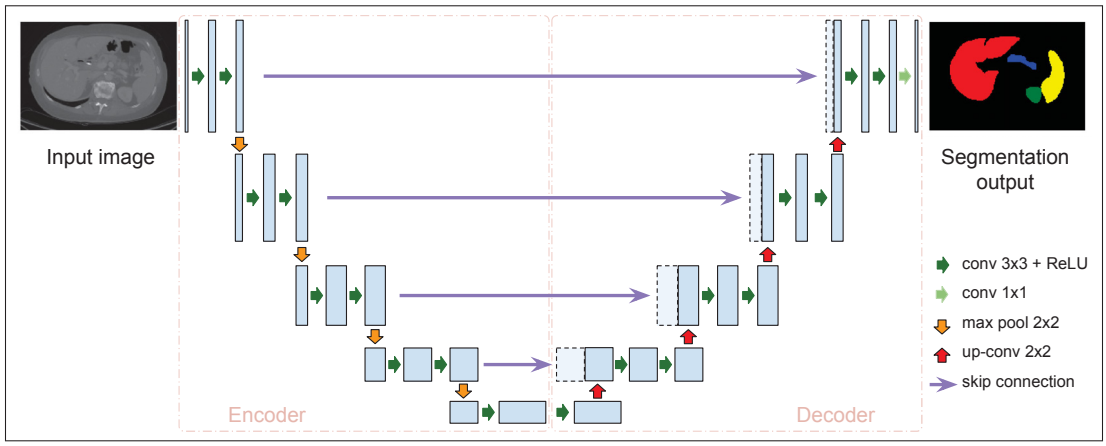

Figure 1.2 Illustration of U-Net, an encoder-decoder style segmentation architecture. Adapted from Ronneberger *et al.* (2015)

segmentation involves every pixel classification and also identifies individual objects within an image. Specifically, pixels belonging to the same class but different instances are assigned unique identifiers. The *panoptic segmentation* combines semantic and instance segmentation, providing a complete segmentation map by identifying all objects in individual and background classes. This thesis focuses on semantic segmentation for medical image analysis applications.

Before delving into the learning techniques, we establish the notation of semantic segmentation settings. Since semantic segmentation involves pixel-to-pixel learning, it typically relies on a substantial amount of image-label pairs (as in Fig 1.1). Let us assume a training dataset with N_l labeled samples, which is denoted as $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{Y}^i)\}_{i=1}^{N_l}$, where $(\mathbf{X}^i, \mathbf{Y}^i)$ is the i-th image-label pair with input image, $\mathbf{X}^i \in \mathbb{R}^{\Omega}$, and corresponding label, $\mathbf{Y}^i \in \{1, ..., C\}^{\Omega}$, having C classes including background class. Note that the label is often employed as a one-hot representation, i.e., $\mathbf{Y}^i \in [0, 1]^{C \times \Omega}$. These pairs have spatial domain Ω, which can be 2-D or 3-D. In a fully-supervised scenario, the model is trained with a large dataset and often yields high-quality segmentation results. Training and assessing the segmentation model needs objective function and evaluation measures, which are described next.

1.2.1 Common objective functions and evaluation measures in segmentation

Objective functions: The objective or loss function measures the difference between the network predictions and the ground truth annotations, which assists in learning the network parameters. The typical loss function employed in image segmentation includes the cross-entropy loss, the Dice loss (Milletari *et al.*, 2016; Sudre, Li, Vercauteren, Ourselin & Jorge Cardoso, 2017), or a combination of both losses.

The cross-entropy (CE) measures the discrepancy between the predicted probability distribution and the ground truth. For a C-class segmentation, the CE loss function at a pixel is defined as:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{1.1}$$

where y_c and p_c are the ground truth and the predicted probability values for a class c. The final loss is subsequently averaged over all the pixels and all images in a batch to optimize the network. The CE loss suffers from a class imbalance issue, which is prominent in medical image segmentation as the background region is dominant compared to the foreground regions.

In contrast, the Dice loss measures the overlap between the predicted probability and the ground truth mask. The generalized Dice loss (Milletari *et al.*, 2016; Sudre *et al.*, 2017) for a given class is defined as:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_v y_v p_v + \epsilon}{\sum_v y_v + \sum_v p_v + \epsilon}, \tag{1.2}$$

where v represents voxel or pixel in the spatial image domain Ω, and ϵ is a small constant added for numerical stability. The total loss is computed by averaging across all classes and all samples. Unlike the CE loss, the Dice loss addresses the class imbalance implicitly (Sudre *et al.*, 2017). Nevertheless, the disadvantage of the Dice loss is that its gradients can be unstable, especially for small segmentations, which can affect convergence. Therefore, the Dice and CE losses are often combined to leverage their respective benefits (Ma *et al.*, 2021a; Taghanaki *et al.*, 2019).

Model evaluation:

The evaluation measures estimate the quality of model predictions. The predictions are often evaluated with overlap and distance-based measures for the segmentation task. For instance, the Dice Similarity Coefficient (DSC) measures the degree of overlap between the ground truth mask and the predicted mask. For a given class, the DSC is defined as

$$DSC(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{2 \sum_v y_v \hat{y}_v + \epsilon}{\sum_v y_v + \sum_v \hat{y}_v + \epsilon}, \tag{1.3}$$

where y_v and $\hat{y}_v$ are values at voxel v in the ground truth ($\mathbf{Y}$) and the predicted segmentation mask ($\hat{\mathbf{Y}}$). The final Dice score is obtained by averaging the scores across all classes and samples, similar to the Dice loss. This score ranges between $[0, 1]$, where 0 signifies no overlap between the masks and 1 indicates perfect overlap. On the other hand, segmentation boundaries are measured with a distance-based measure. For instance, the Hausdorff Distance (HD) measures the maximum shortest distance between two point sets coming from the boundary of ground truth and predicted mask (Huttenlocher, Klanderman & Rucklidge, 1993). Suppose $\mathcal{A}$ and $\mathcal{B}$ are point sets from ground truth and predicted mask, respectively. The HD is given as

$$HD(\mathcal{A}, \mathcal{B}) = \max \left\{ \max_{a \in \mathcal{A}} d(a, \mathcal{B}), \max_{b \in \mathcal{B}} d(\mathcal{A}, b) \right\}, \tag{1.4}$$

where d is the minimum distance from the boundary pixel a or b to the entire set $\mathcal{B}$ or $\mathcal{A}$, respectively. Since HD is sensitive to outlying points, the 95^{th} percentile of the histogram of shortest distances is often used instead. The HD is more informative for the measurement of small or thin structures. The aforementioned two evaluation measures are complementary and commonly employed in medical image segmentation.

1.2.2 Limited supervision

As mentioned before, most deep models rely upon abundant image-label pairs to learn a reasonable model. Acquiring such paired data involves pixel-wise labeling for the image segmentation task. This process is labor-intensive and prone to subject variability. In addition, the labeling task is magnified since medical images often require 3D annotations. Recently, learning techniques with limited supervision (Jiao *et al.*, 2023; Peng & Wang, 2021; Shen *et al.*, 2023) have been emerging to ease the burden of annotation in two ways: the number of annotations is reduced by leveraging unlabeled data, or the level of supervision is decreased from a stronger to a weaker form of annotations. These alternatives are commonly categorized as semi-supervised and weakly-supervised learning (Peng & Wang, 2021; Shen *et al.*, 2023).

Semi-supervised learning

Semi-supervised learning (SSL) leverages unlabeled data along with relatively few labeled samples to improve the model performance (Chapelle, Scholkopf & Zien, 2009; Jiao *et al.*, 2023; Van Engelen & Hoos, 2020). The idea behind SSL is that neighboring data is likely to have similar labels, and low-density regions separate two or more classes. These premises suggest that the data within each class should form a cluster with a smooth decision boundary (Van Engelen & Hoos, 2020). In this context, unlabeled data serve to refine the decision boundaries and help to learn the distribution of the data. Depending on how unlabeled images are employed, the recent SSL approaches in medical image segmentation are categorized into pseudo-labeling, regularization, or knowledge prior techniques. A brief overview of these strategies is provided in the following section.

Pseudo labeling: The pseudo-labeling or self-training strategy aims to generate labels for unlabeled images in order to improve the model (Lee *et al.*, 2013). It involves segmentation predictions for unlabeled images from the initial model and then assigning pseudo labels to them. These pseudo labels are added to the original labeled set so that it can be used to re-train

the model. The addition of pseudo labels to training is carried out iteratively to continuously improve the quality of new pseudo labels. Various ways of pseudo label generation have been proposed (Bai *et al.*, 2017; Du *et al.*, 2022; Seibold, Reiß, Kleesiek & Stiefelhagen, 2022). For instance, pseudo labels are generated based on thresholding of predictions (Zeng *et al.*, 2023), confidence-aware predictions under perturbation (Yao, Hu & Li, 2022), self-ensembling of predictions (Du *et al.*, 2022; Xie, Luong, Hovy & Le, 2020), post-processing (Bai *et al.*, 2017), or propagating neighboring labels (Seibold *et al.*, 2022). The challenge with pseudo-labeling approaches is that a careful addition of pseudo labels is required, as mistakes in the pseudo labels are propagated during the training process (Chapelle *et al.*, 2009).

Regularization: Regularization-based approaches are prominent in SSL due to their simplicity in leveraging unlabeled data as an unsupervised loss function during training. A wide range of regularization-based methods has been proposed for SSL. These regularization techniques are generally formulated as entropy minimization (Vu, Jain, Bucher, Cord & Pérez, 2019), adversarial learning (Chaitanya *et al.*, 2019; Nie, Gao, Wang & Shen, 2018), consistency loss (Bortsova, Dubost, Hogeweg, Katramados & Bruijne, 2019; Cui *et al.*, 2019), or co-training learning (Peng, Estrada, Pedersoli & Desrosiers, 2020). For instance, an entropy minimization strategy is a simple regularization strategy where the entropy of prediction is minimized for the unlabeled data (Vu *et al.*, 2019). In the adversarial method, the prediction of unlabeled images encourages closer to those of the labeled data (Chaitanya *et al.*, 2019; Nie *et al.*, 2018) via adversarial loss. On the other hand, the consistency or co-training methods encourage two or more segmentation predictions from the same or different networks to be consistent under different data and model perturbations (Tarvainen & Valpola, 2017) or multiple views of image (Peng *et al.*, 2020). An example of consistency regularization used in the self-ensembling framework (Cui *et al.*, 2019; Tarvainen & Valpola, 2017) is shown in Fig. 1.3. It comprises two identical models, known as student and teacher, receiving different perturbed inputs. The student model is trained with a supervised loss on labeled data, and a consistency loss encourages

models to produce similar outputs on labeled and unlabeled data. Meanwhile, the teacher model is updated using an exponential moving average (EMA) strategy.

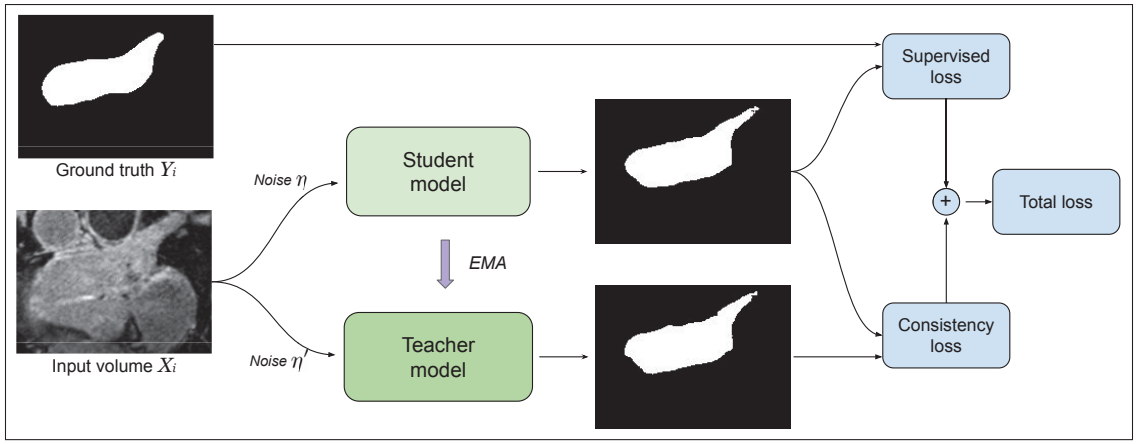

Figure 1.3 Schematic of consistency regularization in the self-ensembling approach. Adapted from Cui *et al.* (2019); Tarvainen & Valpola (2017)

Knowledge prior: To effectively leverage unlabeled data, a few approaches utilize prior information within the data. Such prior typically incorporated either from images or available labels as pre-training (He *et al.*, 2020b; Kiyasseh, Swiston, Chen & Chen, 2021), meta-learning (Li, Zhang & He, 2020a; Xue *et al.*, 2020), or unsupervised losses (Zheng *et al.*, 2019a). For instance, labeled and unlabeled images are encoded using autoencoder (He *et al.*, 2020b) or self-supervised learning (Kiyasseh *et al.*, 2021), which are subsequently utilized as priori for learning segmentation in SSL. Lately, a signed distance map (SDM) is used as shape constraints during training (Li *et al.*, 2020a; Xue *et al.*, 2020). For instance, Li *et al.* (2020a) proposes an additional task of predicting SDM to enforce similarity between labeled and unlabeled predictions. A probabilistic atlas also has been used to enforce anatomical priors on the unlabeled predictions (Huang *et al.*, 2022; Zheng *et al.*, 2019a).

Weakly-supervised learning

Weakly supervised approaches alleviate the need for dense or complete annotations. These methods utilize sparse or incomplete annotations to train the segmentation network. There are different types of weak annotation that are used for segmentation tasks, such as image-level tags (Papandreou, Chen, Murphy & Yuille, 2015), scribbles (Lin, Dai, Jia, He & Sun, 2016), points (Bearman, Russakovsky, Ferrari & Fei-Fei, 2016), or bounding boxes (Rajchl *et al.*, 2016). The image-level tags indicate whether a specific class is present or absent in the image, points or scribbles describe sparse labeling on target regions, whereas bounding boxes contain boxes around the target objects. Examples of these different weak annotations are shown in Fig. 1.4. These weak cues are easier and inexpensive to acquire compared to pixel-level annotations.

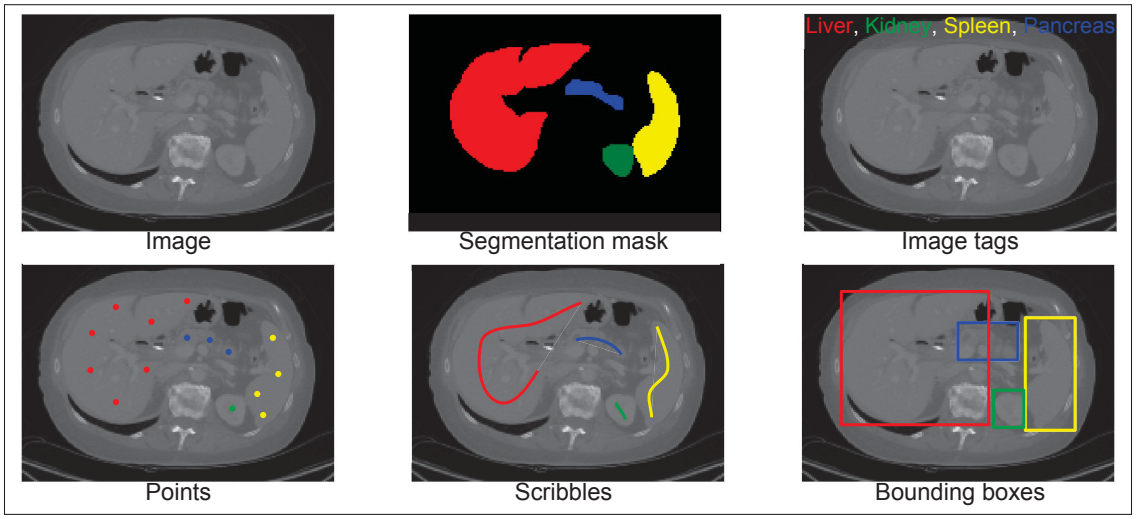

Figure 1.4 Illustrations of different weak annotations compared to dense segmentation

The literature on weakly supervised segmentation in medical imaging is growing with different weak annotations. Many methods resort to image-level labels due to the ease of obtaining such labels (Feng, Yang, Laine & Angelini, 2017; Nguyen *et al.*, 2019; Patel & Dolz, 2022). These methods derive class-specific feature maps, known as class activation maps (CAMs) (Zhou, Khosla, Lapedriza, Oliva & Torralba, 2016), from a classification network to provide a likely segmentation. However, generated CAMs are highly discriminative and result in over

or under-segmentations. Therefore, these methods focus on improving the initial CAMs using conditional random fields (CRF) (Nguyen *et al.*, 2019), segmentation proposals (Wu *et al.*, 2019), or equivariant constraints (Patel & Dolz, 2022). A schematic of a CAM-based segmentation pipeline is shown in Fig 1.5. The initial activation maps are obtained from a classification model. These maps are further refined using dense CRF to obtain proxy labels. Such labels are subsequently utilized to train a segmentation network for final predictions (Nguyen *et al.*, 2019).

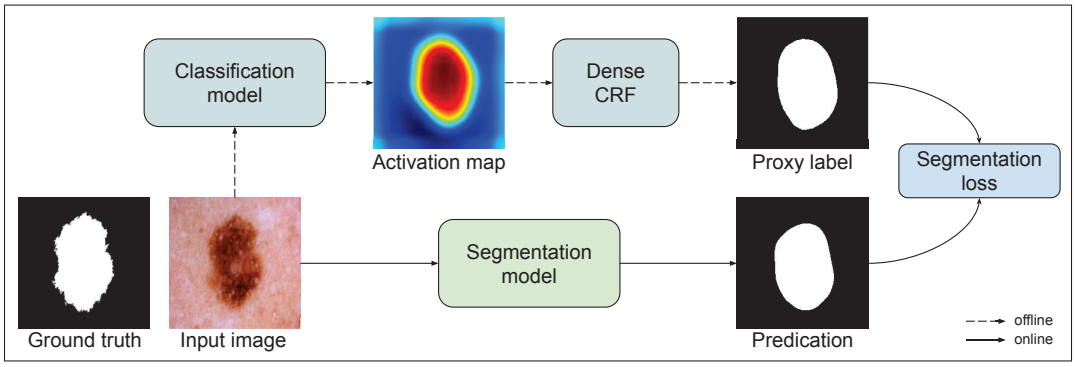

Figure 1.5 Illustration of a CAM-based weakly supervised segmentation pipeline. Adapted from Nguyen *et al.* (2019)

The points or scribble annotations significantly reduce the annotation efforts. These annotations are typically used as label propagation to generate pseudo labels using clustering (Qu *et al.*, 2020), superpixels (Chen *et al.*, 2020b), distance map (Tian *et al.*, 2020) or self-ensembling (Lee & Jeong, 2020). Conversely, a bounding box annotation is a stronger form of supervision, which describes the locations by a rectangle (2D) or cuboid (3D) that contains the region of interest, often with tight boundaries. Methods based on bounding box annotations also commonly generate pseudo labels, which are used for training the segmentation network (Rajchl *et al.*, 2016). For instance, DeepCut (Rajchl *et al.*, 2016) produces the pseudo-labeling using CRF. Alternately, a few methods incorporate additional priors into the loss function, such as constraints on the bounding box tightness or the target object size (Jia, Huang, Eric, Chang & Xu, 2017; Kervadec *et al.*, 2019; Kervadec, Dolz, Wang, Granger & Ayed, 2020).

1.3 Related Work

So far, we have discussed various machine learning techniques employed in semantic segmentation depending on the type and amount of annotations. In addition, our research objectives are interconnected with other related topics, which are discussed in this section.

1.3.1 Distance Transform

Distance transforms provide spatial information and shape-related cues that are valuable in medical image segmentation. Traditional methods often use the distance transform to model the shape of an object or to propagate labels based on distance similarities (Sabuncu, Yeo, Van Leemput, Fischl & Golland, 2010). They are generally robust to complex and irregular structures. Distance transform can be applied to an image or a mask using geodesic or Euclidean distances. Recently, deep learning-based methods incorporate these distances as an auxiliary task to regularize the segmentation network (Bui *et al.*, 2019; Dangi, Linte & Yaniv, 2019; Li *et al.*, 2020a; Xue *et al.*, 2020), an additional input to provide the contextual information to the network (Wang *et al.*, 2018; Wei *et al.*, 2022), or a post-processing operation to improve the segmentation (Bagheri, Tarokh & Ziaratban, 2021).

1.3.2 Anatomical Priors

Medical images inherently possess valuable anatomical information such as the organ size, shape, and location. These anatomical priors are incorporated explicitly during the training of the segmentation network to obtain plausible and accurate results (Nosrati & Hamarneh, 2016). Nevertheless, integrating such priors poses challenges due to the non-differentiable and complex nature of objective terms (Oktay *et al.*, 2017). Recent approaches resort to data-driven solutions to enforce such priors with global or local constraints (Painchaud *et al.*, 2020; Ravishankar, Venkataramani, Thiruvenkadam, Sudhakar & Vaidya, 2017). In (Oktay *et al.*, 2017), an autoencoder trained on segmentation masks is utilized to map predictions into an anatomically

plausible space, with the encoder serving as a global regularizer between the prediction and ground truth distributions. Alternatively, an anatomically plausible segmentation mapping is utilized to ensure the smoothness and topological correctness of the segmentation results (Gaggion, Mansilla, Mosquera, Milone & Ferrante, 2022; Larrazabal, Martínez, Glocker & Ferrante, 2020; Painchaud *et al.*, 2020; Ravishankar *et al.*, 2017). A probabilistic atlas is used as an alternative to enforce the priors for an aligned dataset (Huang *et al.*, 2021).

1.3.3 Representation Learning

Representation learning involves extracting patterns and features directly from data (Bengio, Courville & Vincent, 2013), which facilitates various downstream tasks, such as classification (Pati, Foncubierta-Rodríguez, Goksel & Gabrani, 2020), image retrieval (Sohn, 2016), clustering (Ziko, Granger & Ben Ayed, 2018), or segmentation (Liao, Gao, Oto & Shen, 2013). This representation is commonly learned with labels using supervised approaches or without explicit labels, such as with an autoencoder or self-supervised learning. The similarity between arbitrary images can also used to learn the representation. The seminal work of Siamese Networks (Bromley, Guyon, LeCun, Säckinger & Shah, 1994) learns a representation by contrasting positive and negative pairs of images such that similar images should be closer in a learned embedding space while dissimilar images should be farther apart. Likewise, metric learning presents a compelling approach to similarity learning, aiming to minimize the distance between images of the same class while maximizing the distance between images from different classes. Recently, deep metric learning emerged as a powerful approach to learning similarities, where Euclidean or cosine distances are employed to measure the similarity between pairs of images (Hadsell, Chopra & LeCun, 2006; Schroff, Kalenichenko & Philbin, 2015).

1.4 Summary

This chapter has presented a literature review on image segmentation, covering both traditional and deep learning approaches. Subsequently, we delved into various segmentation approaches and related research areas that form the groundwork for this thesis. This thesis introduces new tools that advance image segmentation under different types of supervision utilizing the interconnected topics discussed earlier. The following Chapters present the specific research objectives pursued in this thesis.

CHAPTER 2

AN INTENSITY-BASED GEODESIC SOFT LABELING FOR IMAGE SEGMENTATION

Sukesh Adiga Vasudeva[a,b], Jose Dolz[b], Herve Lombaert[a,b]

[a] Department of Software and IT Engineering, École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

[b] Department of Computer Engineering and Software Engineering, Polytechnique Montréal, 2500, chemin de Polytechnique, Montreal, Quebec, Canada H3T 1J4

Paper submitted for publication in *Journal of Machine Learning for Biomedical Imaging* (MELBA), March 2024

Presentation

This chapter presents the article *"GeoLS: an Intensity-based, Geodesic Soft Labeling for Image Segmentation"* submitted to Journal of **MELBA** (Machine Learning for Biomedical Imaging) on 18 March 2024. A preliminary version of this article was published (Adiga Vasudeva, Dolz & Lombaert, 2023) at **MIDL** (Medical Imaging with Deep Learning) 2023, presented as an oral talk in Nashville, USA. The objective of this article is to incorporate ambiguities associated with image intensity into the soft-labeling process for a segmentation task.

2.1 Introduction

Image segmentation is a highly structured and dense prediction problem where pixels in an image are grouped into a set of target regions, such as organs or tumors (Pham, Xu & Prince, 2000; Suetens, 2017). It plays a pivotal role in clinical decision systems, notably in computer-assisted prognosis and diagnosis, treatment planning, and intervention support (Duncan & Ayache, 2000; Zhou, Rueckert & Fichtinger, 2019). Recent advancements in segmentation methods are primarily due to the ability of deep learning techniques to solve such complex predictive tasks

(Hesamian, Jia, He & Kennedy, 2019; Litjens *et al.*, 2017). Training these approaches involves minimizing the deviation of the network predictions from the given ground-truth annotations using various objective functions (Lin, Goyal, Girshick, He & Dollár, 2017; Rubinstein & Kroese, 2004; Sudre *et al.*, 2017).

A common strategy to measure this deviation is to employ the cross-entropy function with the ground-truth mask represented as one-hot encoded vectors. This learning objective exhibits remarkable performance in problems needing predictions of independent classes, such as in whole-image classification (Baum & Wilczek, 1987; He *et al.*, 2016; Szegedy, Ioffe, Vanhoucke & Alemi, 2017). Nevertheless, the use of standard one-hot encoding in segmentation tasks can be sub-optimal since class predictions at each pixel are inherently conditioned with surrounding pixels. Such encoding indeed fails to capture the spatial relationships across neighborhoods as well as inter-class relationships within an image. These relationships, however, are crucial for the segmentation of medical images. For instance, labels can be similar for pixels within a homogeneous region, but vary near object boundaries due to various image ambiguities (Fig. 2.1). Such ambiguity can be attributed to partial volume effect, motion artifacts, or image acquisition, among other reasons. Moreover, the one-hot label assignments are solely based on the provided ground-truth masks, where the underlying spatial and inter-class relationships are not explicitly considered. Therefore, explicitly modeling spatial and inter-class relationships in the label assignments is sought to improve the performance of the segmentation model.

Recent attempts to incorporate the inter-class relationships in the labels (Galdran *et al.*, 2020; Szegedy, Vanhoucke, Ioffe, Shlens & Wojna, 2016) generally modify the hard one-hot encoding into a softer version. For instance, Label Smoothing (LS) (Szegedy *et al.*, 2016) uniformly redistributes a portion of the target-class probability into all non-target classes to obtain a new soft label assignment for training a deep model. In (Galdran *et al.*, 2020), a non-uniform label smoothing approach is proposed to capture the underlying structure within annotations. This method uses a Gaussian smoothing on each target class to redistribute probability over other

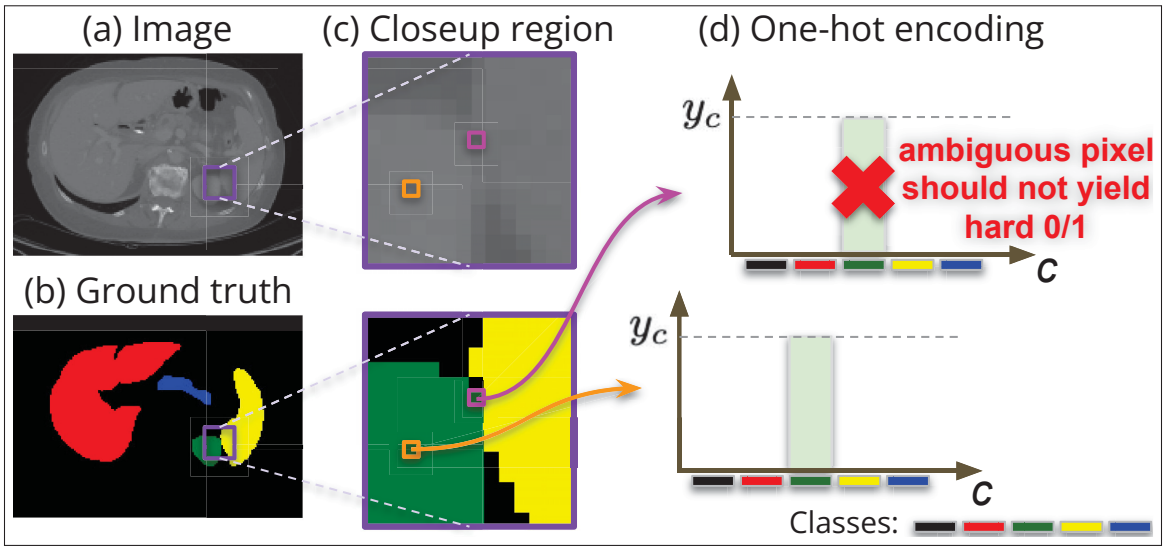

Figure 2.1 Limitation of one-hot label assignments. (a) A sample image and (b) its corresponding ground-truth mask, (c) a closeup image around the boundary region (purple), and (d) the one-hot (OH) encoding for two pixels (orange and pink in closeup images)

classes. It is particularly suitable for datasets featuring ordered class labels, such as tumor or disease grading. These label-smoothing approaches, however, disregard the spatial relationships in their soft-label assignments.

To capture the spatial relationships, a few approaches alter the target segmentation mask to obtain softer labels in the boundary regions (Gros, Lemay & Cohen-Adad, 2021; Kats, Goldberger & Greenspan, 2019). For instance, Kats *et al.* (2019) generates the soft labels in the dilated regions of the target masks by adding granularity in the object boundaries. Furthermore, a Spatially-Varying Label Smoothing (SVLS) approach models the annotation ambiguity around object boundaries in target masks (Islam & Glocker, 2021). Its soft labels capture the local structural variations by applying a Gaussian-smoothing operation on the target masks. However, the annotation ambiguities of object boundaries stem from poorly defined image intensities caused by imaging techniques or existing pathologies, which inherently leads to labeling inaccuracies (Hayward *et al.*, 2008; Joskowicz, Cohen, Caplan & Sosna, 2019). These

ambiguities are not captured in these soft-labeling methods, as they solely rely on the given ground-truth masks.

One solution is to incorporate image-based metrics in the soft-label assignments process. More specifically, a geodesic distance transform captures intensity variations and spatial distances within an image (Criminisi, Sharp & Blake, 2008; Toivanen, 1996). Our approach, therefore, leverages the geodesic distance in order to capture inter-pixel and inter-class relationships during the label smoothing process. The generated soft labels thus become intensity-aware, capturing image gradient information across object boundaries. Incorporating our geodesic soft labels in model training is found to improve the segmentation performance, as they model the underlying intensity variations across objects and labels.

2.1.1 Our contributions

This work proposes a novel Geodesic Label Smoothing (GeoLS) for image segmentation. Specifically, our originality lies in leveraging the geodesic distance transform to embed intensity variations in the soft-labeling process. In contrast to existing soft-labeling approaches, our GeoLS smooths hard labels using geodesic maps, which capture the underlying image context that is crucial for medical image segmentation. The resulting intensity-based soft labels capture class-wise relationships by considering image gradient information between two or more object categories. Furthermore, the geodesic distance between pixels captures the spatial relationships, integrating richer information than the Euclidean distance. Our approach is extensively validated on a variety of medical imaging datasets: the 2019 brain tumor segmentation (BraTS) challenge dataset (Bakas *et al.*, 2017,1), the 2021 abdominal organ segmentation dataset (Ma *et al.*, 2022), and the prostatic zone segmentation dataset (Litjens, Debats, Barentsz, Karssemeijer & Huisman, 2014). The results demonstrate the superiority of GeoLS over state-of-the-art methods that are based on soft-labeling segmentation. Moreover, our experiments include comprehensive ablation studies to highlight further the effectiveness of our geodesic soft labels for image

segmentation. In particular, we investigate the parameters influencing the generation of geodesic soft labels, such as studying the impact of intensity variation and different seeding strategies in obtaining our soft labels. Additionally, we conduct experiments focusing on the combination of our proposed loss with other losses, such as Dice, boundary, and focal loss functions, which aim to assess the synergies in combining these approaches.

2.2 Related Work

2.2.1 Soft labeling

Soft labeling has been actively investigated in the machine learning community (Müller, Kornblith & Hinton, 2019; Szegedy *et al.*, 2016; Zhang *et al.*, 2021). The early methods often leverage the nearest-neighbor points to obtain a soft label (Keller, Gray & Givens, 1985; Seo, Bode & Obermayer, 2003). Such a labeling scheme captures multiple class characteristics in the dataset, which are later used to train a classifier (El Gayar, Schwenker & Palm, 2006). More recently, Szegedy *et al.* (2016) proposed a label smoothing strategy for training deep neural networks. This smoothing strategy uniformly redistributes the portion of the one-hot label of a given class to all other classes. The model trained with these soft labels has been shown to improve the performance in classification tasks in both computer vision (Müller *et al.*, 2019; Szegedy *et al.*, 2016) and medical imaging domains (Galdran *et al.*, 2020; He, Fang, Rabbani, Chen & Liu, 2020a; Islam, Seenivasan, Ming & Ren, 2020). It is also shown to be effective in handling noisy labels (Lukasik, Bhojanapalli, Menon & Kumar, 2020; Lukov, Zhao, Lee & Lim, 2022).

In the context of image segmentation tasks, the label smoothing strategy (Szegedy *et al.*, 2016) captures inter-class relationships within an image. However, It is also essential to consider the spatial relationships within neighboring regions. Recent approaches (Gros *et al.*, 2021; Islam & Glocker, 2021; Kats *et al.*, 2019) attempt to capture such relationships with

spatially-varying smooth labels, improving segmentation performance. For instance, Kats *et al.* (2019) obtains soft labels by expanding the original binary mask using a dilation operation and subsequently assigns a soft value in the extended region. In (Gros *et al.*, 2021), non-binary pre-processing and data augmentation techniques are employed on the target mask to obtain soft labels around the boundaries. These strategies are designed for binary segmentation tasks, where they disregard the probability distribution in the label assignments. Therefore, adopting them directly to multi-class segmentation is not trivial. A SVLS approach generates the soft labels by redistributing the class probabilities based on Gaussian filtering (Islam & Glocker, 2021). Nevertheless, these soft-labeling methods are entirely based on ground-truth masks while ignoring the ambiguities arising from image intensities.

Alternately, soft labels can also be generated by averaging multi-rater annotations (Lourenço-Silva & Oliveira, 2021). Such soft labels are even more expensive to obtain in practice, as they require multiple independent annotations. Furthermore, a few methods also utilize uncertainty maps for soft segmentation (Tang *et al.*, 2022; Wang *et al.*, 2023). Nevertheless, these methods require multiple segmentation predictions to compute uncertainty maps, which are computationally expensive. Compared to these approaches, our method leverages the geodesic distance transform (Toivanen, 1996) to capture the intensity variations in the label smoothing process. The resulting intensity-based soft labels capture spatial and class-wise relationships through the geodesic maps. Moreover, the generated soft labels are computed once and incorporated into the learning objective to train a segmentation model. Also, our method generates new soft labels from a single annotation and can be seamlessly integrated into any segmentation network.

2.2.2 Geodesic Distance Transform (GDT)

The GDT is commonly used for smooth and contrast-sensitive image segmentation (Criminisi *et al.*, 2008; Protiere & Sapiro, 2007; Toivanen, 1996), as it captures the local contrast and

structural information within an image. The seminal work, GeoS (Criminisi *et al.*, 2008), proposes a generalized geodesic distance (GGD) method for segmentation tasks in an energy-based model. The effectiveness of GeoS has led to various segmentation approaches (Kontschieder, Kohli, Shotton & Criminisi, 2013; Qiu *et al.*, 2015; Wang *et al.*, 2014b). For instance, Wang *et al.* (2014b) utilizes GGDs to bring the spatial context between object boundaries in an atlas-based label propagation method. Recent approaches have leveraged GGDs in deep learning techniques to improve image segmentation (Bui *et al.*, 2019; Hammoumi, Moreaud, Ducottet & Desroziers, 2021; Wang *et al.*, 2018; Wei *et al.*, 2022). For instance, Bui *et al.* (2019) proposes a regression of the geodesic distance maps to regularize the segmentation network through an additional prediction branch. Similarly, Ying, Huang, Fu, Yang & Cheng (2023) regularizes geodesic distance maps in a dual-branch network to enhance edge details for weakly supervised segmentation. To improve initial segmentation, the geodesic distance from user interactions (Wang *et al.*, 2018) or initial network predictions (Wei *et al.*, 2022) are employed to provide the contextual information. The resulting geodesic maps are subsequently used as additional inputs to the refinement network. These existing approaches require an extra prediction branch or refinement network to integrate the geodesic maps. In contrast, our method leverages the geodesic distance to embed underlying image context information into the label smoothing process. The generated soft labels are computed once and consequently incorporated into the learning objective to train the segmentation model. Our geodesic soft-labels, therefore, can be directly combined with any segmentation network.

2.3 Method

An outline of the proposed approach comparing hard labels (OH) and existing soft labels (LS and SVLS) is shown in Fig. 2.2. Consider two closeup regions with the same masks but differing image intensities as in Fig. 2.2. The existing methods rely only on ground-truth masks to generate the soft labels. Therefore, they have the same class probability maps in both closeup regions. In contrast, our approach adds image context by leveraging geodesic distance transform in the

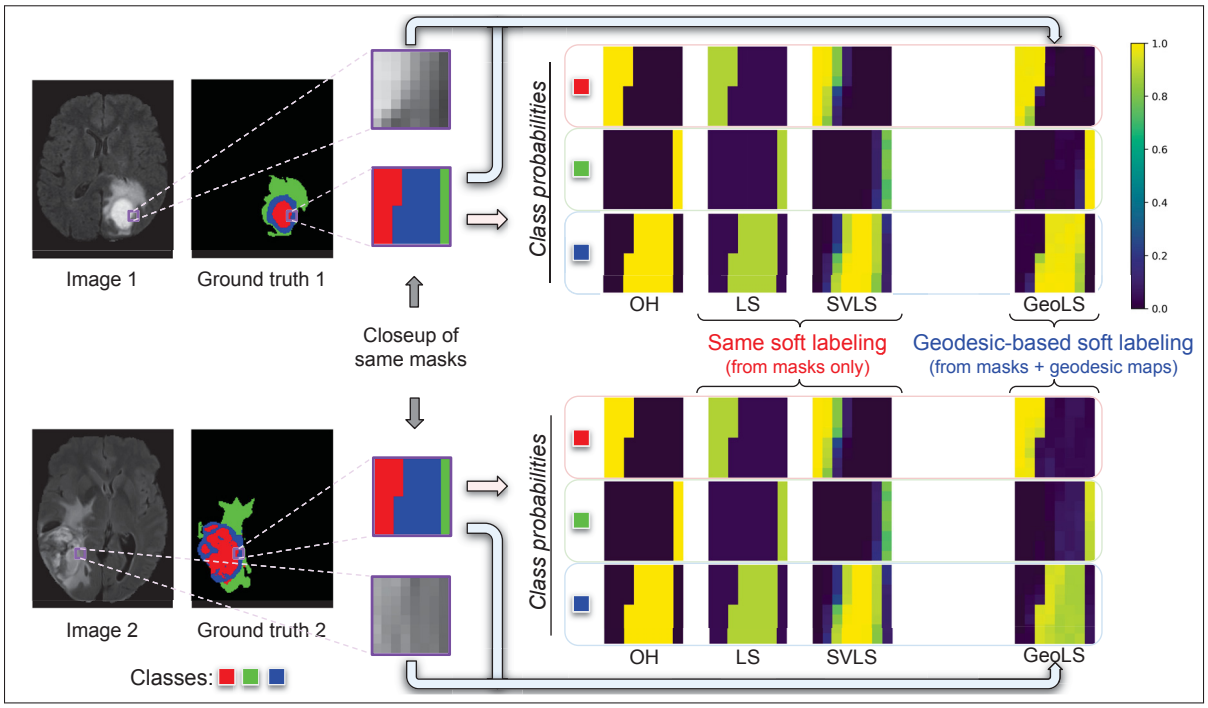

Figure 2.2 Visualization of different soft labeling. *Left side:* Two samples, their corresponding ground-truth masks, and closeup images having the same ground-truth masks around tumor regions. *Right side:* The probabilities of each class (in red, green, and blue colors) for the same closeup images from One-Hot (OH) encoding, Label Smoothing (LS), Spatially-Varying LS (SVLS), and ours (GeoLS)

soft-labeling process. The resulting intensity-based soft labels capture the underlying image ambiguities through geodesic maps. Thus, our method produces different class probability maps in the two closeups. The following subsections describe the label smoothing formulation and our proposed geodesic soft-labeling approach.

2.3.1 Preliminaries

Let $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{Y}^i)\}_{i=1}^{N_l}$ indicate the training dataset with N_l labeled samples, where $\mathbf{X}^i \in \mathbb{R}^{\Omega}$ represents an input volume with a spatial domain Ω, and $\mathbf{Y}^i \in \{1, ..., C\}^{\Omega}$ denotes the corresponding ground truth with C classes, which is provided as an OH representation, i.e.,

$[0, 1]^{C\times\Omega}$. The Cross-Entropy (CE) loss function for a given voxel is defined as:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{2.1}$$

where p_c is the predicted softmax probability from the segmentation network. For simplicity, we use i and c notations wherever necessary and assume that the cardinality of the training set normalizes the loss function.

The OH label encoding, y_c, assigns a probability of '1' for the target class and '0' for the non-target classes. Such assignments fail to provide the model with annotation ambiguity since they do not capture the underlying inter-class relationships within the image. One way to model these relationships is by softening the hard OH encoding during the training process. For instance, the LS method (Szegedy *et al.*, 2016) reduces the probability of the target class by a factor α and evenly distributes it across all classes. The resulting soft label for a given voxel is:

$$y_c^{LS} = (1 - \alpha)y_c + \frac{\alpha}{C} \tag{2.2}$$

These soft labels are subsequently used in training a segmentation network by replacing the original OH label in Eq 2.1. This strategy has been shown to improve performance in classification tasks (He *et al.*, 2020a; Islam *et al.*, 2020; Szegedy *et al.*, 2016). Nevertheless, LS ignores the intrinsic spatial structure that is essential for the segmentation tasks.

2.3.2 Geodesic Label Smoothing (GeoLS)

Existing soft-labeling approaches modify the segmentation masks to capture the spatial relationships (Gros *et al.*, 2021; Islam & Glocker, 2021; Kats *et al.*, 2019), thereby accounting for the annotation ambiguities around the object boundaries. Nevertheless, they largely overlook the annotation ambiguities coming from the image, being prone to annotation mistakes. To consider

such image ambiguities, we integrate the geodesic distance transform (Criminisi *et al.*, 2008; Toivanen, 1996) directly in the soft labeling of pixels. This addition captures the variation in intensities and spatial distance between pixels in an image. The following subsections elaborate on our geodesic label-smoothing method.

2.3.2.1 Generalized Geodesic Distance (GGD) Transform

The GGD transform (Criminisi *et al.*, 2008) computes the shortest geodesic distance between a set of reference points, known as seed points, and each pixel in an image. This transform produces a distance map derived from a spatial distance and image gradient combination. The seed points can be either a single point or a set of points selected from the object of interest. Let $\mathcal{S}_c$ represent a set of seed points upon the target class c. The generalized geodesic distance of each voxel v to the set $\mathcal{S}_c$ of a target class is described as:

$$D_c(v; \mathcal{S}_c, \mathbf{X}^i) = \min_{v' \in \mathcal{S}_c} d(v, v', \mathbf{X}^i), \tag{2.3}$$

with:

$$d(v, v', \mathbf{X}^i) = \min_{\mathbf{p} \in \mathcal{P}_{v,v'}} \int \sqrt{||\mathbf{p}'(s)||^2 + \gamma^2 (\nabla \mathbf{X}^i \cdot \mathbf{u}(s))^2} ds, \tag{2.4}$$

where $\mathcal{P}_{v,v'}$ represents the set of all paths between voxels v and v', and $\mathbf{p}(s)$ denotes one such path parameterized by $s \in [0, 1]$. We define a unit vector $\mathbf{u}(s) = \frac{\mathbf{p}'(s)}{||\mathbf{p}'(s)||}$, which is tangent in the direction of the path, and whose spatial derivative is $\mathbf{p}'(s) = \frac{\partial \mathbf{p}(s)}{\partial s}$.

In Eq. 2.4, the first term, $\mathbf{p}'(s)$, accounts for the spatial distance, while the second term captures the image gradient ($\nabla \mathbf{X}_i$). The parameter γ, termed the geodesic factor, balances the contribution of the image gradient, and the spatial distance between the seed set $\mathcal{S}_c$ and each voxel in the image. When $\gamma = 0$, Eq. 2.4 simplifies to the Euclidean Distance, whereas setting γ to 1 facilitates computation of the geodesic distance (Criminisi *et al.*, 2008). In practice, the geodesic

distance transform is optimally estimated using the raster scan algorithm (Criminisi *et al.*, 2008; Toivanen, 1996).

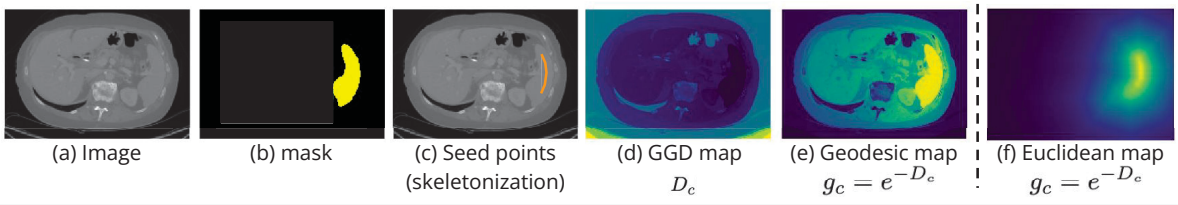

Figure 2.3 Geodesic map generation. (a) A sample image and (b) a corresponding segmentation mask of a spleen organ. (c) Seed points (in orange) are derived by skeletonization of the segmentation mask. (d) The GGD map is generated from seed sets to each pixel in the image. (e) Our final geodesic map is obtained by inverting the GGD map. (f) An Euclidean map is similarly obtained for the same seed points

An example of generating a geodesic map is shown in Fig. 2.3. The seed points are chosen by the skeletonization operation on a target mask. The GGD map is subsequently obtained using Eq. 2.4. To highlight the object of interest, we invert the GGD map to get the final geodesic map for each target class as follows:

$$g_c = e^{-D_c} \tag{2.5}$$

The resulting maps are thus in the range $[0, 1]$. The geodesic map of the background class is obtained by inverting the average of foreground geodesic maps, also in the range $[0, 1]$. In Fig. 2.3, we have also added an Euclidean distance map for comparison with a geodesic map. The Euclidean map spreads uniformly from seed points in all directions. In contrast, our geodesic map propagates based on both spatial distance and gradient information, capturing the underlying intensity similarities.

2.3.2.2 Geodesic Soft Labels

The geodesic maps encode image gradient details as a function of distance from the target objects. Such maps account for the intensity variations across object boundaries. Our approach,

therefore, avails the geodesic maps for smoothing the hard labels. In order to accomplish this, we first normalize the geodesic map of each class as $\tilde{g}_c = \frac{g_c}{\sum_c g_c}$, such that it follows a probability distribution. Subsequently, the normalized geodesic maps are integrated with the original one-hot encoding to produce the new intensity-based soft labels, as defined below:

$$y_c^{GeoLS} = (1 - \alpha) y_c + \alpha \tilde{g}_c \tag{2.6}$$

These generated soft labels are thereafter substituted in Eq. 2.1 to facilitate the training of the segmentation network. The generation of our proposed geodesic soft labels is demonstrated in Fig 2.4. As our approach incorporates intensity variations into the target label assignments through geodesic maps, it effectively guides the network toward better segmentation.

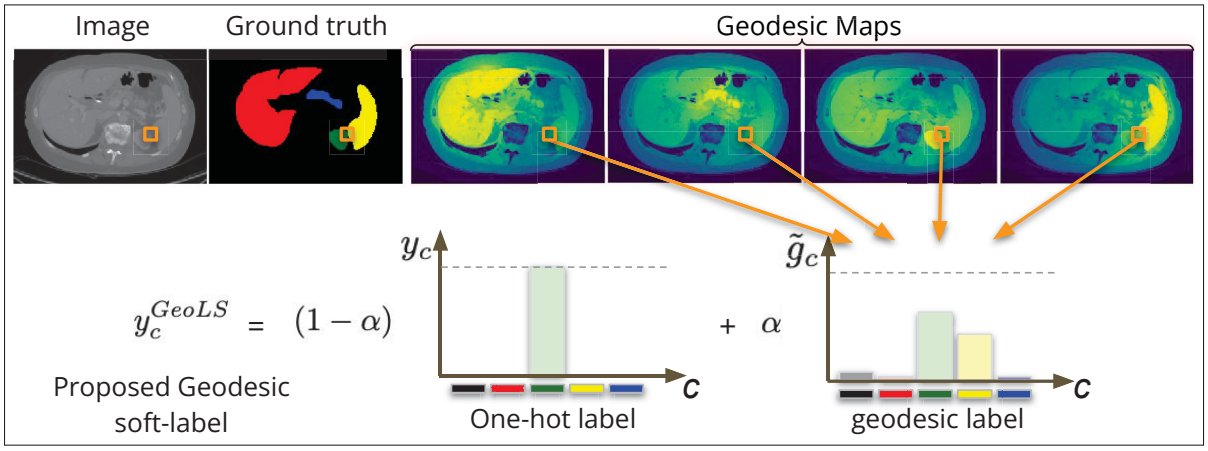

Figure 2.4 Illustration of our proposed Geodesic Label Smoothing (GeoLS). The geodesic maps for all target labels are combined to form a probability distribution. The generated geodesic label is subsequently used to modify the one-hot encoding to obtain the proposed intensity-based soft label. Our soft label captures the underlying intensity variation, thus it can better guide the segmentation network in ambiguous regions. Best viewed in color

2.4 Experiments and Results

2.4.1 Datasets

In order to validate our geodesic label-smoothing method, we utilize three publicly accessible segmentation datasets. These datasets include: a) the Brain Tumor Segmentation dataset obtained from the 2019 BraTS challenge (Bakas *et al.*, 2017,1), b) the multi-organ abdominal segmentation dataset from the 2021 FLARE challenge (Ma *et al.*, 2022), and c) the prostatic zone segmentation dataset from the ProstateX challenge (Litjens *et al.*, 2014). A detailed description of these datasets and our experimental settings are presented next.

a) BraTS:

This dataset comprises 335 multimodal MRI volumes of the brain, containing T1, T2, FLAIR, and T1ce sequences. These volumes are preprocessed with skull-striped, co-registered to a fixed template, and resampled to an isotropic resolution of 1 mm^3. The dataset contains corresponding annotations of glioma tumors, including delineations of the necrotic and non-enhancing core, edema, and enhancing tumor regions. These regions are converted into Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) for evaluation purposes. The dataset is partitioned into 235 for training, 32 for validation, and 68 for testing across all our experiments.

b) FLARE:

The dataset consists of 361 CT volumes of abdominal regions with segmentation masks of four organs: liver, kidney, spleen, and pancreas. These volumes have variable resolutions, which are standardized by resampling to a consistent resolution of $2 \times 2 \times 2.5$ mm^3. Subsequently, they are intensity normalized by retaining values within the percentile range of $[0.5, 0.95]$. We employ a predefined dataset split for all experiments, allocating 260 volumes for training, 26 for validation, and the remaining 75 for testing.

c) ProstateX:

The dataset includes 98 prostatic T2 MRI scans and corresponding segmentation labels of four anatomical zones, including the peripheral zone (PZ), transition zone (TZ), distal prostatic urethra (DPU), and anterior fibromuscular stroma (AFS). All volumes are resampled into a fixed resolution of $0.5 \times 0.5 \times 3\ mm^3$ as followed in (Islam & Glocker, 2021). For all our experiments, the dataset is split into 68 for training, 10 for validation, and the remaining 20 for testing.

2.4.2 Training and implementation details.

To assess the contribution of our geodesic soft labeling, we utilize a 3D U-net (Çiçek *et al.*, 2016) architecture for the segmentation network. This model is trained using Adam optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-4} and weight decay of 10^{-4}. The input size of $192 \times 192 \times 128$ in BraTS, $160 \times 208 \times 112$ in FLARE, and $320 \times 320 \times 24$ in ProstateX experiments are fed into the network. Data augmentations such as random flipping and rotation are utilized, as in (Islam & Glocker, 2021). The network is trained for 200 epochs with a batch size of 4. For inference, the model with the best dice score on the validation set is selected for testing. Our evaluation includes experiments with CE, Focal Loss (FL) (Lin *et al.*, 2017), LS (Szegedy *et al.*, 2016), and SVLS (Islam & Glocker, 2021) losses as training objectives. Following the literature, commonly utilized hyperparameter values are considered for each baseline approach, and the result is reported for a value with the best dice score on the validation set. In particular, the focusing parameter γ in FL is set to $\{1, 2, 3\}$. In the case of LS, $\alpha \in \{0.1, 0.2, 0.3\}$ are used, whereas $\sigma \in \{0.5, 1, 2\}$ values are employed in SVLS with a kernel size of 3. In our method, the geodesic factor γ is explored for $\{0.5, 0.75, 1\}$ values with a fixed smoothing factor of $\alpha = 0.1$.

To obtain the geodesic maps, an open-source library, *GeodisTK* [1], is employed with a skeletonization of a segmentation mask as seed points. Note that our soft labels are computed offline,

[1] https://github.com/taigw/GeodisTK

requiring virtually no additional computation during the training process. The only additional cost is loading the geodesic maps, whose computational burden is negligible. The geodesic maps are not needed during the inference step, resulting in exactly the same computation cost as existing approaches. All our experiments were executed on an NVIDIA RTX A6000 GPU with PyTorch 1.8.0. Our GeoLS implementation is available at: https://github.com/adigasu/GeoLS.

2.4.3 Evaluation

The segmentation performance is evaluated with standard and widely used evaluation measures, such as the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD). The former measure estimates the overlap between ground truth labels and predictions, whereas the latter measures the distance between ground truth and predicted segmentation boundaries. To ensure a fair comparison, we conducted all experiments three times with fixed seed sets on identical machines, presenting results with mean and standard deviation values.

2.4.4 Comparison with the state-of-the-art.

The performance of the proposed geodesic soft-labeling approach is first compared with the state-of-the-art soft-labeling methods (LS (Szegedy *et al.*, 2016) and SVLS (Islam & Glocker, 2021)), and their discriminative results are reported in Tables 2.1-2.3 for all three datasets. The table also includes the hyperparameter value corresponding to the best-performing model for each method.

The performance of various methods on multi-class brain tumor segmentation dataset is shown in Table 2.1. The results show that employing soft labels improves the segmentation performance compared to models trained with a CE loss on hard labels in both scores. Among soft-labeling baselines, FL and SVLS achieve the best DSC and HD scores, respectively. Our approach outperforms these best-performing baselines in both DSC and HD scores in all tumor categories. Notably, we observe that the proposed GeoLS indeed benefits in the enhancing tumor (ET)

Table 2.1 Segmentation results on the BraTS test set. In all tumor structures (ET, TC, WT), our method yields the best DSC and HD scores. For each tumor structure, bold and underlined indicate the best and second-best methods

	Methods	ET	TC	WT	Average
DSC (%) ↑	CE	72.05 ± 2.14	82.38 ± 0.91	90.09 ± 0.39	81.51 ± 1.03
	FL ($\gamma = 1$)	73.55 ± 0.49	82.82 ± 0.20	90.37 ± 0.16	82.25 ± 0.20
	LS ($\alpha = 0.1$)	73.28 ± 0.85	82.65 ± 0.30	90.46 ± 0.08	82.13 ± 0.35
	SVLS ($\sigma = 1.0$)	73.15 ± 2.82	82.67 ± 1.96	90.43 ± 0.78	82.08 ± 1.81
	Ours ($\gamma = 0.75$)	**74.61 ± 0.79**	**83.51 ± 0.24**	**90.88 ± 0.12**	**83.00 ± 0.31**
HD (mm) ↓	CE	14.55 ± 1.61	7.64 ± 1.15	6.28 ± 0.86	9.49 ± 1.20
	FL ($\gamma = 1$)	12.81 ± 1.11	7.31 ± 0.32	5.96 ± 0.18	8.69 ± 0.31
	LS ($\alpha = 0.1$)	13.52 ± 0.35	7.23 ± 0.16	5.95 ± 0.16	8.90 ± 0.21
	SVLS ($\sigma = 1.0$)	12.83 ± 2.70	6.93 ± 1.37	5.72 ± 1.10	8.50 ± 1.70
	Ours ($\gamma = 0.75$)	**12.36 ± 0.56**	**6.08 ± 0.61**	**5.22 ± 0.52**	**7.89 ± 0.32**

region. Such a region is often irregular and poorly defined, which leads to imprecise annotation (Menze *et al.*, 2014). Our method improves this challenging region by 1.06% in DSC score and 0.45 mm in HD, highlighting the advantage of combining the intensity information in our soft labels. These results demonstrate the merit of using our geodesic soft-labeling over hard-labeling and existing soft-labeling approaches.

Table 2.2 Segmentation results on the FLARE test set. Our method produces the best DSC and HD scores on average results as well as on a challenging pancreas organ. For each abdominal organ, bold and underlined indicate the best and second-best methods

	Methods	Liver	Kidney	Spleen	Pancreas	Average
DSC (%) ↑	CE	94.88 ± 0.31	94.70 ± 0.33	95.46 ± 0.85	72.52 ± 0.61	89.39 ± 0.14
	FL ($\gamma = 1$)	94.84 ± 1.08	94.38 ± 0.35	95.56 ± 0.72	69.66 ± 2.02	88.61 ± 0.90
	LS ($\alpha = 0.1$)	**95.96 ± 1.11**	**94.89 ± 0.35**	95.61 ± 0.63	73.07 ± 1.35	89.88 ± 0.38
	SVLS ($\sigma = 0.5$)	95.76 ± 0.34	94.28 ± 0.34	95.01 ± 0.09	73.39 ± 0.16	89.61 ± 0.10
	Ours ($\gamma = 1.0$)	95.60 ± 0.87	94.80 ± 0.37	**96.52 ± 0.30**	**73.72 ± 1.02**	**90.16 ± 0.44**
HD (mm) ↓	CE	4.15 ± 1.10	2.94 ± 0.11	2.98 ± 1.06	6.72 ± 1.18	4.20 ± 0.19
	FL ($\gamma = 1$)	3.28 ± 1.28	3.22 ± 0.32	2.80 ± 1.08	8.03 ± 0.46	4.33 ± 0.61
	LS ($\alpha = 0.1$)	2.87 ± 1.14	2.93 ± 0.37	2.60 ± 0.24	6.37 ± 1.03	3.69 ± 0.26
	SVLS ($\sigma = 0.5$)	**2.61 ± 1.06**	3.17 ± 0.78	**1.42 ± 0.18**	6.26 ± 0.48	3.36 ± 0.20
	Ours ($\gamma = 1.0$)	3.01 ± 1.05	**2.40 ± 0.50**	1.49 ± 0.55	**5.59 ± 0.20**	**3.12 ± 0.21**

Table 2.3 Segmentation results on the ProstateX test set. Our method is competitive in most cases and achieves the best DSC score on average results. At the same time, baselines are ranked differently across prostatic zones (PZ, TZ, DPU, and AFS). For each prostatic zone, bold and underlined indicate the best and second-best methods

	Methods	PZ	TZ	DPU	AFS	Average
DSC (%) ↑	CE	71.56 ± 0.55	86.34 ± 0.28	48.39 ± 2.46	<u>38.27 ± 4.46</u>	61.14 ± 1.21
	FL (γ = 1)	**72.18 ± 1.11**	<u>86.38 ± 0.20</u>	51.19 ± 2.73	35.50 ± 6.85	61.31 ± 1.96
	LS (α = 0.2)	70.52 ± 0.31	86.34 ± 0.46	**53.31 ± 2.89**	35.16 ± 6.65	<u>61.33 ± 1.29</u>
	SVLS (σ = 1.0)	<u>72.08 ± 1.89</u>	85.89 ± 0.64	51.10 ± 4.14	35.67 ± 3.08	61.19 ± 2.12
	Ours (γ = 1.0)	70.86 ± 1.11	**86.51 ± 0.36**	<u>51.50 ± 0.50</u>	**39.50 ± 2.60**	**62.09 ± 0.75**
HD (mm) ↓	CE	<u>6.51 ± 0.34</u>	**3.22 ± 0.10**	11.28 ± 0.44	**9.58 ± 1.21**	7.65 ± 0.24
	FL (γ = 1)	**5.76 ± 0.97**	3.38 ± 0.39	7.89 ± 3.34	<u>9.68 ± 0.59</u>	**6.68 ± 1.05**
	LS (α = 0.2)	6.64 ± 0.69	<u>3.33 ± 0.15</u>	<u>7.28 ± 2.20</u>	9.75 ± 1.14	<u>6.75 ± 0.70</u>
	SVLS (σ = 1.0)	7.04 ± 0.84	3.73 ± 0.24	10.94 ± 5.75	10.2 ± 1.26	7.98 ± 1.59
	Ours (γ = 1.0)	7.83 ± 2.72	**3.22 ± 0.06**	**6.50 ± 0.52**	9.78 ± 0.26	6.83 ± 0.78

Table 3.2 presents the results of the multi-organ abdominal segmentation on the FLARE test set. A similar pattern is observable in the LS, SVLS, and GeoLS results compared to those obtained from the BraTS dataset (Table 2.1). Nevertheless, there is an apparent performance gap in FL compared to CE results, which may be attributed to the over-emphasis on mislabeled pixels present in the data. Overall, our GeoLS yields the best segmentation performance corresponding to the baselines, notably enhancing the segmentation in the challenging pancreas and spleen regions.

The results of the multi-class prostatic zone segmentation on the ProstateX dataset are reported in Table 2.3. A similar trend in FL, LS, and GeoLS results is observed as in Table 2.1. However, SVLS produces a drop in performance compared to CE results (HD), possibly due to the over-suppression of original one-hot encoding in the boundaries. Moreover, existing methods are ranked differently across datasets and evaluation measures, indicating that these approaches are sensitive to datasets. In contrast, our GeoLS outperforms the state-of-the-art approaches in most cases. Based on these results, we can conclude that our method remains consistent across diverse datasets, highlighting the robustness of our intensity-based soft labels.

2.4.5 Qualitative Results

Figure 2.5 shows the visual comparison of different segmentation results on brain tumors from BraTS, abdominal organs from FLARE, and prostatic zones from ProstateX datasets. In brain tumor segmentations (top row), the results of existing approaches (OH, FL, SVLS) are predominantly over-segmenting in non-enhancing core regions (blue), whereas the LS and GeoLS reduce the segmentation errors. In the middle row of Fig. 2.5, the existing methods struggle to segment the challenging pancreas organ (yellow) organ. In contrast to these baselines, our GeoLS delivers a superior segmentation of the pancreas organ. The prostatic zone segmentations are arguably challenging due to imprecise boundaries between different zones. In the bottom row, the results of prostatic zone segmentations are poor in all approaches. Our method produces reasonable segmentation results, notably in the AFS prostatic zone (yellow).

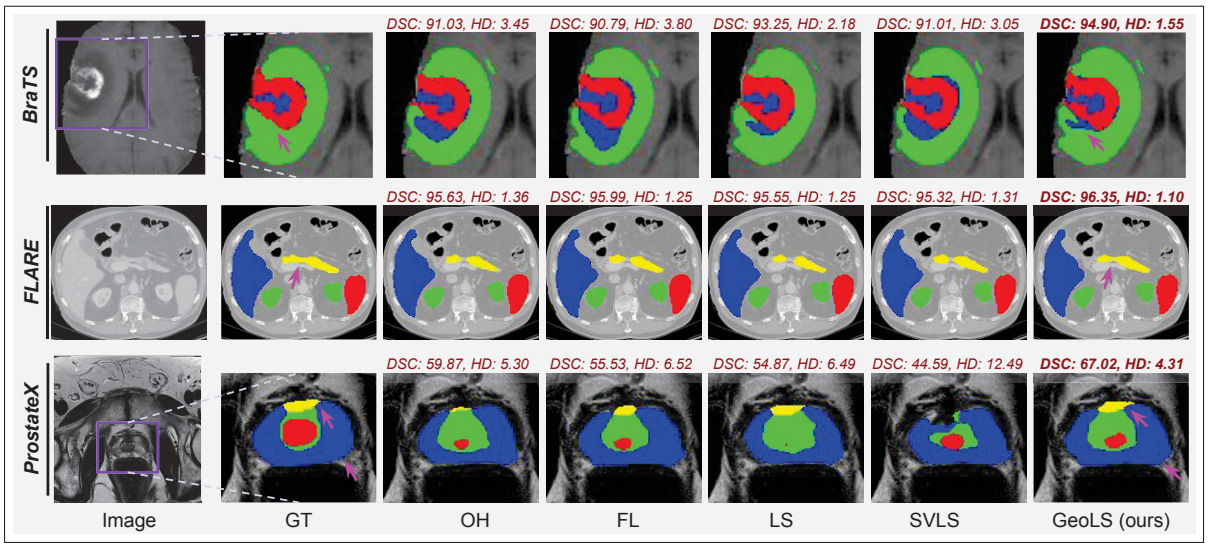

Figure 2.5 Qualitative results on BraTS, FLARE, and ProstateX datasets. Our GeoLS minimizes classification errors in ambiguous regions, such as the non-enhancing core (blue) in BraTS (top), the pancreas (yellow) in FLARE (middle), and PZ (blue) and AFS (yellow) zones in the ProstateX (bottom) examples

In addition, the prediction probability maps of baseline and our method for the same examples are shown in Fig. 2.6. Our GeoLS produces reasonably low probabilities in poorly defined image intensities and misclassified regions, ensuring segmentation accuracy even in challenging areas.

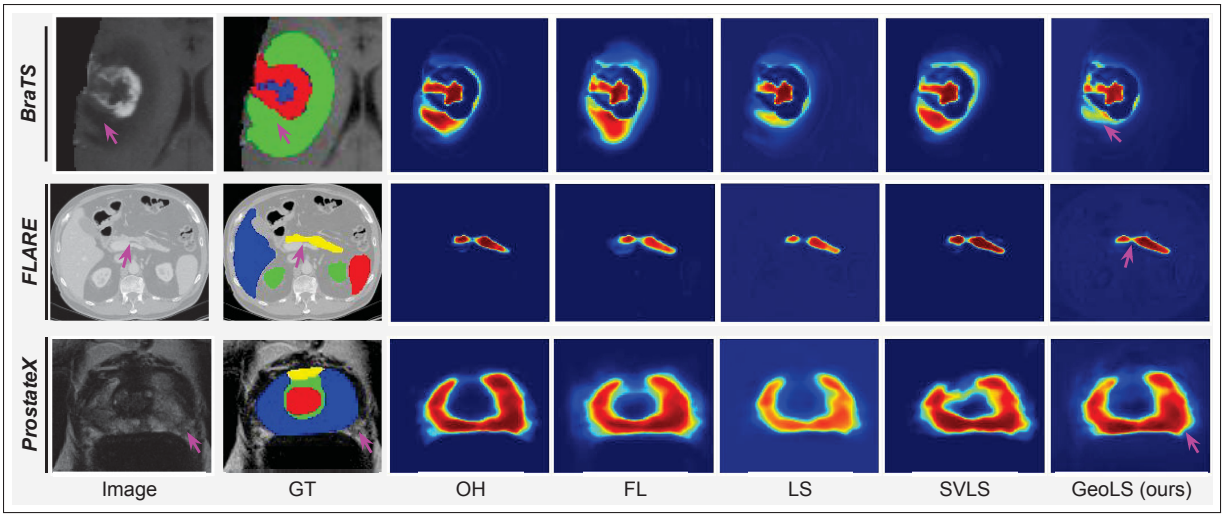

Figure 2.6 Predicted probability maps. The probability maps indicate the non-enhancing core (blue) in BraTS (top), the pancreas (yellow) in FLARE (middle), and PZ (blue) in ProstateX (bottom) examples

At the same time, it consistently maintains high probabilities in well defined image intensities regions. Furthermore, the quantitative results presented in Sec. 2.4.4 support these visual results. These results indicate that supplying image gradient information through geodesic maps in our intensity-based soft-labeling approach enhances the segmentation performance.

2.4.6 Sensitivity to γ

The hyperparameter γ in Eq. 2.4 plays a crucial role in balancing between the Geodesic Distance and the Euclidean Distance. Since the intensity variations and spatial distance can influence the geodesic distance transform, we investigate the segmentation performance by varying the γ parameter and report their results in Fig. 2.7, across all datasets. Additionally, we include the segmentation result obtained from a model trained with $\gamma = 0$, i.e., utilizing only the Euclidean Distance for soft labels. The results demonstrate that the segmentation performance is better for higher γ values compared to the models solely relying on Euclidean distance maps. This indicates that incorporating geodesic information based on image gradients in our soft labels positively impacts the performance of segmentation tasks.

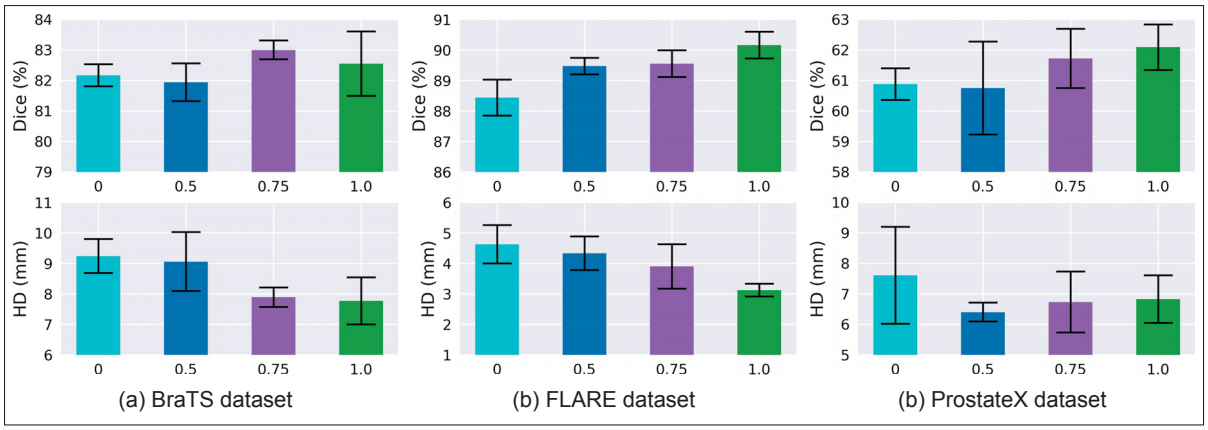

Figure 2.7 Sensitivity of geodesic factor γ on segmentation performance. Each bar indicates the average DSC ↑ (top) and HD ↓ (bottom) scores on each dataset. $\gamma = 0$ here uses only using Euclidean Distance. Segmentation accuracy improves when the γ value is increased towards 1, indicating a higher emphasis on Geodesic Distance in soft labels

2.4.7 Choice of seed set $\mathcal{S}$

Our soft label relies on the geodesic maps, which vary with the different choices of seed set $\mathcal{S}$. Therefore, to validate the effectiveness of our seeding strategy on segmentation performance, we conduct experiments with different seed-set strategies. These strategies involve obtaining a random selection of pixels within each target class. For this, our experiments include 3, 5, and 7 randomly selected pixels as seed points. Such seed points are inadequate for large regions, such as the liver, or multiple instances of a class label, such as the kidney. To address this issue, seed sets are also obtained using the remainings of the skeletonization and erosion operations applied to each target class. The results of these experiments are reported in Table 2.4. It shows that the segmentation performances are comparable for different seed-set choices, which further demonstrates the strength of our geodesic soft labels. Furthermore, the results suggest that the skeleton-based seed strategy consistently yields favorable results across all datasets, which indicates that this seeding strategy could also be viable on new datasets.

Table 2.4 Performance under different seed sets $\mathcal{S}$. Average DSC and HD scores on each dataset are reported. Segmentation accuracy is consistent across datasets for skeleton-based seed points. The bold and underlined indicate the best and second-best results

Datasets	BraTS		FLARE		ProstateX	
choice of $\mathcal{S}$	**DSC (%) ↑**	**HD (mm) ↓**	**DSC (%) ↑**	**HD (mm) ↓**	**DSC (%) ↑**	**HD (mm) ↓**
random-3	82.98 ± 0.68	8.10 ± 0.09	87.83 ± 1.02	4.79 ± 0.16	58.65 ± 3.73	7.41 ± 1.59
random-5	82.51 ± 0.80	9.00 ± 0.70	89.46 ± 1.00	4.20 ± 0.97	60.88 ± 0.85	7.07 ± 0.33
random-7	82.36 ± 0.48	8.89 ± 0.81	89.23 ± 0.21	4.41 ± 0.49	**61.76 ± 2.62**	6.84 ± 0.91
skeleton	**83.00 ± 0.31**	**7.89 ± 0.32**	**90.16 ± 0.44**	**3.12 ± 0.21**	61.72 ± 0.97	**6.73 ± 1.00**
erosion	81.93 ± 0.93	9.17 ± 0.68	89.56 ± 0.08	3.63 ± 0.27	61.72 ± 0.90	6.96 ± 0.55

2.4.8 Combining with other loss function

The main goal of this work is to provide an alternative to state-of-the-art soft labeling losses by leveraging geodesic distance transform. Nevertheless, the proposed approach is orthogonal to other types of segmentation losses, including widely used Dice loss (Sudre *et al.*, 2017). Moreover, combined CE and Dice losses are often employed to train segmentation models for medical images (Ma *et al.*, 2021a; Taghanaki *et al.*, 2019). Thus, we investigate whether the findings observed when comparing the CE loss hold when we combine the proposed GeoLS with the Dice loss. These results, depicted in Fig. 2.8, demonstrate that adding the Dice loss improves the segmentation performance of both CE and GeoLS across all datasets. Moreover, combining GeoLS and Dice losses achieves the best results in most cases, demonstrating the consistency of our geodesic label-smoothing approach.

Furthermore, we performed experiments by combining ours and CE loss with boundary loss (BL) first and then with focal loss (FL), whose results are reported in Fig. 2.9. The results show a similar trend as with a combination of Dice loss. Combining our method with BL and FL yields better segmentation results than combining CE with BL and FL across all three datasets. These results demonstrate the robustness of the proposed GeoLS when combined with other loss functions.

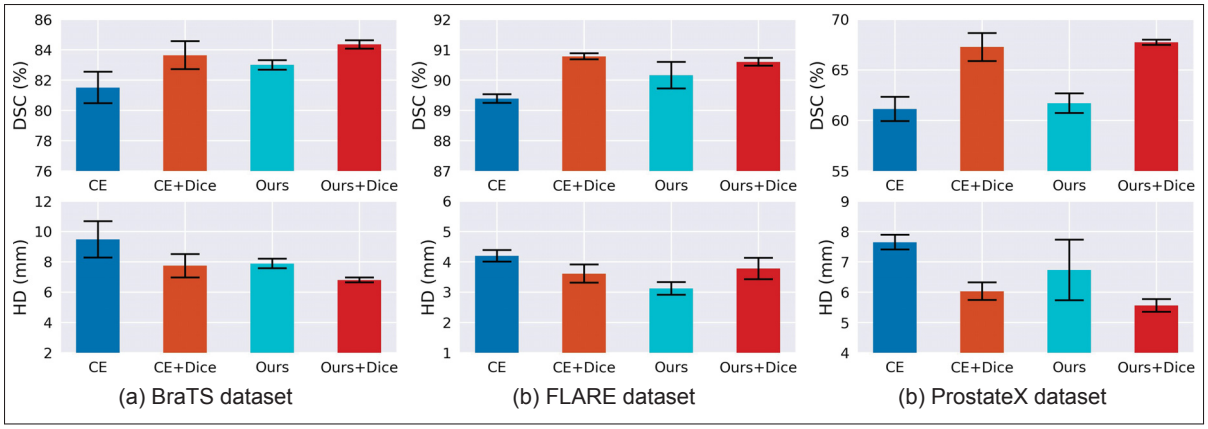

Figure 2.8 Segmentation performance with a combination of Dice loss. Each bar indicates the DSC (top) and HD (bottom) scores on all three datasets. The segmentation performance improves by adding Dice loss on both CE and our models. A combination of ours and Dice loss yield consistently best in most cases

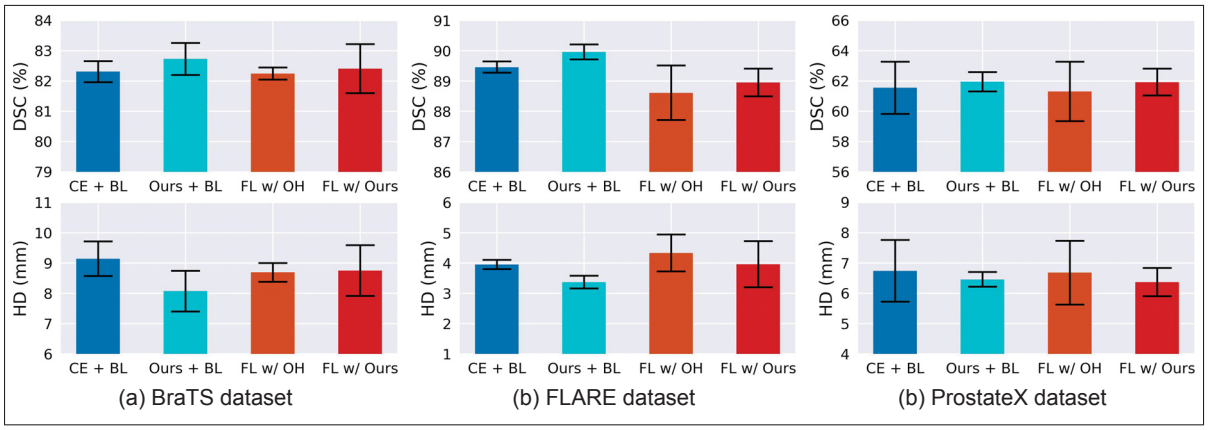

Figure 2.9 Segmentation performance with a combination of Boundary loss (BL) and Focal loss (FL). Each bar indicates the average DSC ↑ (top) and HD ↓ (bottom) scores on all three datasets. Combining our method with BL and FL consistently provides better segmentation results compared to combining CE with BL and FL in most cases

2.5 Discussion and Conclusion

Despite the growing popularity of contemporary soft-labeling approaches, the underlying image context information associated with the label is largely overlooked in the soft labels for image segmentation. This work demonstrates that incorporating such information into standard hard labels would improve the performance of deep segmentation networks. To that

effect, our contribution, a Geodesic label smoothing, incorporates intensity variation details into the soft-labeling process through geodesic distance transforms. More specifically, our proposed approach generates new intensity-based soft labels that capture ambiguity between neighboring target regions. Employing our soft labels in the training of segmentation models has consequently demonstrated an improved segmentation performance. Our results have in fact shown that our geodesic-based smoothing consistently outperforms state-of-the-art approaches in soft-labeling, across three different datasets: multi-class tumor segmentation in brain MRIs, organ segmentation in abdominal CTs, and zone segmentation in prostatic MR volumes. Both quantitative and qualitative results indicate notable improvements in the segmentation of known challenging regions, such as of enhancing tumors, as well as the pancreas.

Furthermore, the ablation study conducted on the geodesic factor parameter indicates that our geodesic maps integrate richer intensity information in the yielded soft labels, effectively producing an improved segmentation performance than utilizing only Euclidean distance maps. Our experiments have also evaluated several key seeding strategies for generating soft labels. These results show that the skeleton-based strategy remains consistent across all datasets. The design of the seeding process can be further explored in order to better capture the intrinsic structures of target objects. This work provides, therefore, a valuable alternative to hard-labeling and existing soft-labeling losses. Nonetheless, our geodesic label smoothing loss can also be combined with other segmentation losses, such as the conventional Dice loss. The use of such loss has in fact shown further improvements in the segmentation accuracy within our experiments. As future work, our approach could also be potentially applicable to segmentation tasks under noisy annotations (Karimi, Rollins, Velasco-Annis, Ouaalam & Gholipour, 2023; Lukasik *et al.*, 2020). Overall, our proposed geodesic-based soft-labeling could be virtually leveraged in broader ranges of applications where annotation remains challenging due to ambiguities in image intensities across regions.

CHAPTER 3

ANATOMICALLY-AWARE UNCERTAINTY FOR SEMI-SUPERVISED IMAGE SEGMENTATION

Sukesh Adiga Vasudeva[a], Jose Dolz[a], Herve Lombaert[a]

[a] Department of Software and IT Engineering, École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *Journal of Medical Image Analysis* (MedIA), October 2023

Presentation

This chapter presents the article *"Anatomically-aware Uncertainty for Semi-supervised Image Segmentation"* (Adiga Vasudeva, Dolz & Lombaert, 2024) submitted to Journal of **MedIA** (Medical Image Analysis) on 11 December 2022, revised on 11 August 2023, and accepted for publication on 18 October 2023. An initial article was published (Adiga Vasudeva, Dolz & Lombaert, 2022b) in the conference of **MICCAI** (Medical Image Computing and Computer Assisted Intervention), held in Singapore. This article aims to guide a segmentation model with reliable target regions through anatomically-aware uncertainty estimation within semi-supervised scenarios.

3.1 Introduction

Segmentation is a fundamental task in medical image analysis, where image pixels are associated with a target object, such as an organ, structure, or abnormal region. It is a vital pre-processing step in many clinical applications, notably in computer-assisted diagnosis, intervention assistance, treatment planning, and personalized medicine (Ayache & Duncan, 2016; Duncan & Ayache, 2000). Recent segmentation methods based on deep learning techniques are driving progress under the full-supervision regime, often outperforming traditional methods (Litjens *et al.*, 2017).

Such a regime, however, relies on a large amount of annotations, which is time-consuming. Delineating an image at a pixel-level is indeed challenging, especially in homogeneous or low-contrast regions, and often requires prohibitive clinical expertise. The burden of image annotation motivates new learning strategies with limited supervision (Cheplygina, de Bruijne & Pluim, 2019).

Semi-supervised learning is an emerging strategy that alleviates annotation scarcity by leveraging unlabeled data with a small set of labeled data. Current semi-supervised segmentation methods typically utilize unlabeled data either in the form of pseudo labels (Bai *et al.*, 2017; Zheng *et al.*, 2020) or in a regularization term (Cui *et al.*, 2019; Nie *et al.*, 2018; Peng *et al.*, 2020). The former strategies augment the original labeled dataset with unlabeled data alongside its corresponding model predictions, commonly referred to as pseudo labels. Later techniques incorporate unlabeled data into the training process by constraining predictions with a regularizer term. Training these semi-supervised approaches typically involves a supervised loss associated with labeled data and an unsupervised loss associated with unlabeled data.

Among regularization techniques, consistency-based approaches (Laine & Aila, 2017; Tarvainen & Valpola, 2017) are often used in semi-supervision due to simple ways to leverage unlabeled data. Their approach encourages two or more segmentation predictions to be consistent under different perturbations of the input data (Bortsova *et al.*, 2019; Cui *et al.*, 2019; Li *et al.*, 2020b). However, the segmentation predictions can be unreliable and noisy for unlabeled data since its annotations are unavailable. To alleviate this issue, uncertainty-aware regularization methods (Sedai *et al.*, 2019; Yu, Wang, Li, Fu & Heng, 2019) have been proposed to gradually add reliable target regions in predictions. Although these methods perform well in low-labeled data regime, their high computation and complex training techniques remain a limiting factor to broader applications. For instance, the pixel-level uncertainty approximation with Monte-Carlo Dropout (MCDO) (Gal & Ghahramani, 2016) or ensembling (Lakshminarayanan, Pritzel & Blundell, 2017) requires multiple predictions per image, thereby increasing the computation of

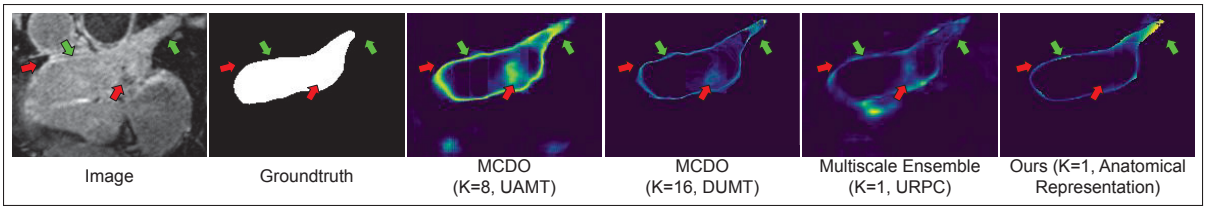

Figure 3.1 Uncertainty maps from different semi-supervision methods. K denotes the number of inferences. Green arrows in regions of probable uncertainty due to unclear boundaries or annotator cut preference (such as in pulmonary veins cut in top right). Red arrows in regions of lower uncertainty as they depict high image gradients in uninformative clear boundary or inner foreground content

each training step. Moreover, these approaches do not consider global information to estimate uncertainty. The resulting uncertainty maps capture pixel-wise disparity, most likely around boundaries (Kendall, Badrinarayanan & Cipolla, 2017). However, high gradient regions near anatomical boundaries or inner content of anatomical structures should have a certain labeling mask. For instance, Fig. 3.1 shows uncertainty captured by MCDO mostly over boundaries, while regions with high gradients (red arrows) could indicate certain boundaries or anatomical details with certainty. Probable uncertainty may lie in areas of low image gradients. For instance, anatomical boundaries may be unclear due to imaging or even non-existent in case of an arbitrary cut from an annotator (green arrows), as illustrated in the pulmonary veins in Fig. 3.1. Existing methods could benefit from capturing informative uncertainty in images beyond highlighting high image gradients or all over boundaries.

The global information of the anatomical regions is one promising direction to provide cues about informative uncertainty in images. Our approach will, therefore, exploit and capture global anatomical information by leveraging available masks to approximate segmentation uncertainty. Our main idea is to learn an anatomically-aware representation from a training set of segmentation masks. The learnt representation maps incorrect model predictions onto an anatomically-plausible segmentations. The plausible segmentation is subsequently used to estimate the uncertainty maps and further guide training of the segmentation network. We show that the proposed uncertainty estimates are more robust and computationally less expensive

than deriving them from a standard entropy variance-based method, which requires multiple inferences for each training step.

3.1.1 Our contributions

We propose a novel approach to estimate the uncertainty maps from an anatomically-aware representation of the segmentation masks, in order to guide the training of a semi-supervised segmentation model. More precisely, we innovate semi-supervised segmentation with uncertainty-based training by integrating a pre-trained denoising autoencoder (DAE) into the training of our segmentation network to: (i) map the inaccurate model predictions to plausible segmentation masks and (ii) estimate new uncertainty maps that guide the training of our segmentation model. As we approximate the uncertainty based on the difference between predicted segmentation and its DAE reconstruction learned from the segmentation mask, it can better integrate anatomical information. In contrast to most uncertainty-based approaches, estimating the uncertainty map requires a single inference from the DAE model, thereby reducing computational complexity. Our method is extensively evaluated on two medical imaging datasets: the 2018 Atrial segmentation challenge dataset (Xiong *et al.*, 2021) and the 2021 Abdominal organ segmentation dataset (Ma *et al.*, 2022). Results demonstrate the superiority of our approach over the state-of-the-art methods in semi-supervised segmentation. Moreover, we investigate the impact of various design choices made in our anatomically-aware (DAE) module and training settings to highlight the robustness of our method for image segmentation. Additionally, a qualitative comparative analysis of uncertainty for different methods and their computation time is provided, showing the merit of our anatomically-aware uncertainty estimation.

3.2 Related Work

3.2.1 Semi-Supervised Segmentation

Semi-supervised learning (SSL) is an established approach in the literature under the paradigm of learning with limited supervision (Jiao *et al.*, 2023). A wide range of SSL strategies have been explored for segmentation, such as self-training (Bai *et al.*, 2017; Zheng *et al.*, 2020), entropy minimization (Grandvalet & Bengio, 2004; Wu, Fan, Zhang, Lin & Li, 2021a), consistency regularization (Bortsova *et al.*, 2019; Cui *et al.*, 2019), co-training (Peng *et al.*, 2020; Xia *et al.*, 2020) or adversarial learning (Chaitanya *et al.*, 2019; Nie *et al.*, 2018). For instance, self-training methods (Bai *et al.*, 2017; Zheng *et al.*, 2020) typically employ pseudo-labels on unlabeled data to train models in an iterative way. However, potential labeling mistakes in the pseudo labels can quickly propagate during training, causing undesired segmentation outcomes. Entropy minimization strategies (Wu *et al.*, 2021a) circumvent such issues by enforcing high confidence in predictions but can also easily lead to trivial solutions if additional priors are not used. Co-training approaches (Peng *et al.*, 2020; Xia *et al.*, 2020) avoid iterations but at the cost of simultaneously training two or more networks with multi-view images. Adversarial methods (Chaitanya *et al.*, 2019; Nie *et al.*, 2018) encourage the predictions of unlabeled images to be closer to those of the labeled images. However, they remain challenging in terms of convergence (Salimans *et al.*, 2016). Among the existing SSL strategies, consistency regularization-based methods (Laine & Aila, 2017; Tarvainen & Valpola, 2017) are widespread due to their simple assumption that predictions should not change significantly under different realistic data perturbations. This notion is formulated as a consistency regularization term in the loss function, which encourages predictions to be consistent between data and its perturbed version (Bortsova *et al.*, 2019; Cui *et al.*, 2019; Li *et al.*, 2020b). Similarly, our method leverages unlabeled data with a consistency regularizer.

3.2.2 Uncertainty-based methods

Uncertainty estimation approaches often employ Bayesian neural networks (Neal, 2012), however, their training process poses significant computational challenges. Recent deep learning methods address this limitation by approximating uncertainty through the generation of multiple samples (Abdar *et al.*, 2021). For instance, Monte-Carlo Dropout (MCDO) (Gal & Ghahramani, 2016) performs several forward passes via the same model with dropout enabled at test time to generate multiple samples for the same input. Deep ensembles (Lakshminarayanan *et al.*, 2017) train a set of independent models to generate multiple samples. These approaches, however, tackle the problem of approximating *epistemic* uncertainty associated with the model output but not the *aleatoric* uncertainty associated with the input (Kendall & Gal, 2017). A set of recent methods models the *aleatoric* uncertainty by using intra-/inter-annotation variability as a proxy to the underlying input uncertainties (Baumgartner *et al.*, 2019; Kohl *et al.*, 2018; Monteiro *et al.*, 2020). Aforementioned methods have been shown to produce reliable uncertainty estimations in fully-supervised segmentation (Camarasa *et al.*, 2021; Mehta *et al.*, 2022).

In the context of semi-supervised segmentation, the uncertainty in the prediction is widely used within the optimization process (Wang *et al.*, 2021,2; Yu *et al.*, 2019). In particular, the uncertainty information assists the segmentation models by providing reliable target regions on unlabeled data during each training step. For instance, Yu *et al.* (2019) first approximates an uncertainty map using a predictive entropy of several predictions under data and model perturbations. The generated uncertainty map is later used to gradually add the reliable target regions in the consistency loss term. This idea was further extended to integrate uncertainty on a feature-level (Wang *et al.*, 2020) and multiple prediction branches (Wang *et al.*, 2022a). The uncertainty estimation in these approaches commonly use MCDO (Gal & Ghahramani, 2016) or ensembling (Lakshminarayanan *et al.*, 2017), which inherently relies on multiple predictions per image. In addition to being computationally expensive, estimating such entropy-based uncertainty is suboptimal in a multi-class scenario since it disregards inter-class overlaps

(Van Waerebeke, Lodygensky & Dolz, 2022). More recently, multi-scale (Luo *et al.*, 2022) or multi-decoder (Wu *et al.*, 2022) approaches have been proposed to overcome the expensive computation of uncertainty using multiple predictions in a single forward pass. Nevertheless, these methods often fail to capture the actual uncertainty regions. In contrast to existing strategies, our method leverages an anatomically-aware representation from the available annotations to estimate the uncertainty in a single inference step. This strategy leads to a lower computational complexity and an improved computational efficiency.

3.2.3 Towards anatomically-plausible segmentations

Recent approaches incorporate anatomically-aware priors in a segmentation network (Oktay *et al.*, 2017; Painchaud *et al.*, 2020; Ravishankar *et al.*, 2017) by learning the variability of structures in a medical imaging dataset. For instance, Oktay *et al.* (2017) first learn an anatomically-aware representation with an autoencoder-based architecture using segmentation masks. This representation is later utilized to map a prediction into an anatomically-plausible space. These methods use the encoder of the representation as a global shape regularizer that enforces the model predictions to follow the ground truth distribution. The anatomically-aware representation can also map an erroneous mask into an anatomically-plausible segmentation. Such mapping is subsequently used to correct the segmentation predictions as a post-processing step (Larrazabal *et al.*, 2020; Painchaud *et al.*, 2020) or improve the segmentation on unseen test images (Karani, Erdil, Chaitanya & Konukoglu, 2021). In order to encode the masks in the anatomically-aware representation, a substantial amount of annotations are used either from the given dataset (Larrazabal *et al.*, 2020; Painchaud *et al.*, 2020) or the source domain dataset (Karani *et al.*, 2021). The anatomically-aware representation is alternately substituted with a probabilistic atlas to enforce the priors (Huang *et al.*, 2022; Zheng *et al.*, 2019a), which requires an aligned dataset. For instance, Dalca, Guttag & Sabuncu (2018) learns an anatomically-aware representation on aligned labelings and subsequently uses it for unsupervised segmentation on aligned images. In contrast to these approaches, our method leverages an anatomically-aware

representation in a low-data regime with the goal of obtaining uncertainty maps in order to guide the segmentation network during the training process.

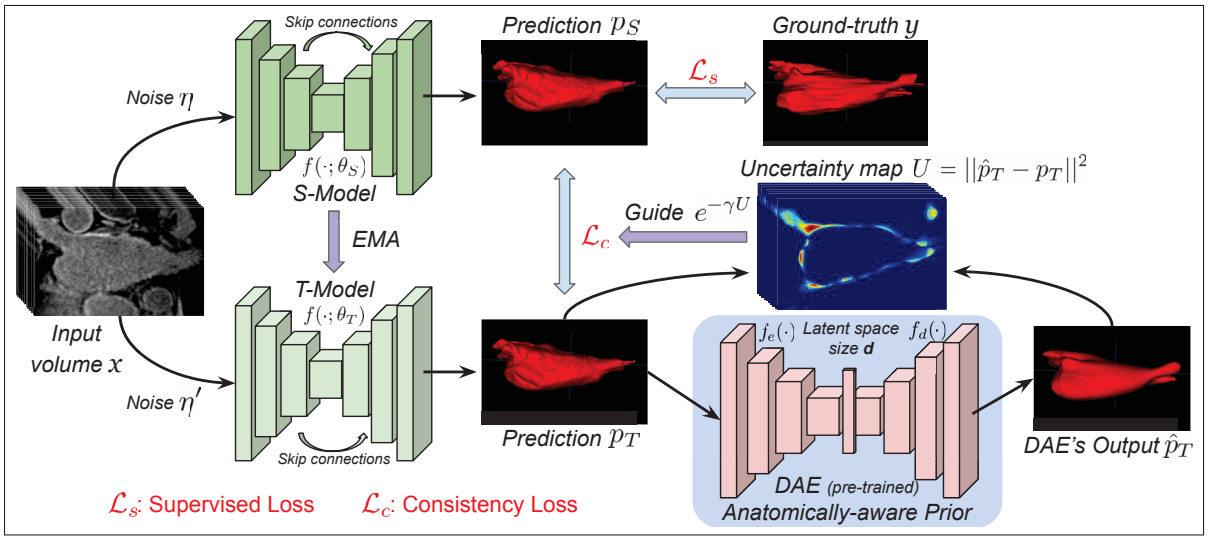

Figure 3.2 Overview of anatomically-aware uncertainty estimation for semi-supervised segmentation. A pre-trained anatomically-aware representation (DAE) module is integrated into the training of the mean teacher model, which maps the teacher prediction $\mathbf{P}_T$ into a plausible segmentation $\hat{\mathbf{P}}_T$. The uncertainty map ($\mathbf{U}$) is subsequently estimated with the output of the teacher and the DAE model in order to further guide the student model

3.3 Method

An overview of the proposed anatomically-aware uncertainty estimation for semi-supervised segmentation is shown in Fig 3.2. The main idea is to exploit an anatomically-aware representation that maps the segmentation prediction into a plausible mask. The reconstructed segmentation will be indicative in estimating an uncertainty map, which later is used to guide the segmentation training. The following subsections describe the semi-supervised setting, anatomically-aware representation, and uncertainty estimation process.

3.3.1 Preliminaries

The standard semi-supervised learning consists of N_l labeled and N_u unlabeled data in the training set, where $N_l \ll N_u$. Let $\mathcal{D}_L = \{(\mathbf{X}^i, \mathbf{Y}^i)\}_{i=1}^{N_l}$ and $\mathcal{D}_U = \{(\mathbf{X}^i)\}_{i=1}^{N_u}$ denote the labeled and unlabeled sets, where an input volume is represented as $\mathbf{X}^i \in \mathbb{R}^{\Omega}$ with a spatial domain Ω, and its corresponding segmentation mask is $\mathbf{Y}^i \in \{1, 2, ..., C\}^{\Omega}$, with C being the number of classes. The objective is to train a segmentation network with a combination of supervised loss $\mathcal{L}_s$ and unsupervised loss $\mathcal{L}_u$ using labeled and unlabeled data, i.e., $\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u$, where λ controls the weight of unsupervised loss.

3.3.2 Mean Teacher Formulation

Following current literature (Yu *et al.*, 2019), we adopt the common mean teacher approach (Tarvainen & Valpola, 2017) for training a segmentation network. It consists of a student (S) and a teacher (T) model, both having the same segmentation architecture. The overall objective function is defined as follows:

$$\mathcal{L} = \min_{\theta_S} \sum_{i=1}^{N_l} \mathcal{L}_s(f(\mathbf{X}^i; \theta_S), \mathbf{Y}^i) + \lambda_c \sum_{i=1}^{N_l+N_u} \mathcal{L}_c(f(\mathbf{X}^i, \eta; \theta_S), f(\mathbf{X}^i, \eta'; \theta_T)), \tag{3.1}$$

where $f(\cdot)$ denotes the segmentation network, and θ_S and θ_T are the learnable weights of the student and teacher models. The supervised loss $\mathcal{L}_s$ measures the segmentation quality on the labeled data, whereas the unsupervised consistency loss ($\mathcal{L}_c = \mathcal{L}_u$) measures the prediction consistency between the student and the teacher models for the same input volume $\mathbf{X}^i$ under different perturbations (η and η'). The balance between the supervised and unsupervised loss is controlled by a ramp-up weighting coefficient λ_c, which is defined as

$$\lambda_c = \beta * e^{-r(1-\frac{t}{tmax})^2}, \tag{3.2}$$

where β is a consistency weight, r controls the rate of ramp-up, t and t_{max} denote the current and maximum training steps. For training, the student model parameters (θ_S) are optimized with stochastic gradient descent (SGD), whereas the teacher model parameters (θ_T) are updated using an exponential moving average (EMA) at each training step t. The EMA is defined as

$$\theta_T^t = \alpha\theta_T^{t-1} + (1-\alpha)\theta_S^t, \tag{3.3}$$

where α is the smoothing coefficient of EMA that controls the update rate.

3.3.3 Anatomically-aware Uncertainty Approach

The reliability of the model prediction on the unlabeled dataset plays an essential role in the consistency loss. An uncertainty-aware scheme can assist this loss by providing reliable target regions. The existing approaches (Wang *et al.*, 2020; Yu *et al.*, 2019) estimate uncertainty at a pixel-level, which fails to consider global information within the dataset. To address this limitation, our approach learns an anatomically-aware representation prior in order to capture global information. The measurable deviations from this prior provide informative cues about the uncertainty of the segmentation mask. The following subsections elaborate on our anatomically-aware uncertainty method.

3.3.3.1 Anatomically-aware Representation Prior

Incorporating anatomically-aware prior in deep segmentation models is not obvious. One of the reasons is that, in order to integrate such prior knowledge during training, one needs to augment the learning objective with a differentiable term, which is not trivial. To circumvent these difficulties, a simpler solution is to resort to an autoencoder trained with segmentation masks, which maps the predictions into anatomically-plausible segmentation. This strategy has been adopted for fully-supervised learning as a global regularizer during training in (Oktay *et al.*, 2017) and as a post-processing step in (Larrazabal *et al.*, 2020) to correct the segmentation predictions.

Motivated by this concept, we encode the available segmentation masks in a non-linear latent space of a denoising autoencoder (DAE) (Vincent *et al.*, 2010) to learn an anatomically-aware representation prior. This learned representation captures the global information from the segmentation masks such that it maps an inaccurate prediction into a plausible segmentation.

The DAE model consists of an encoder $f_e(\cdot)$ and a decoder $f_d(\cdot)$ with a d-dimensional latent space as shown in the Fig. 3.2. The DAE is trained to reconstruct the clean label $\mathbf{Y}^i$ from its corrupted version $\tilde{\mathbf{Y}}^i$, which can be achieved with a mean squared error loss: $\frac{1}{|\Omega|}\sum_{v\in\Omega}||f_d(f_e(\tilde{\mathbf{Y}}^i_v)) - \mathbf{Y}^i_v||^2$, where v is a voxel. Additionally, the dice loss is added to handle the class imbalance between foreground and background in the labels.

3.3.3.2 Anatomically-aware Uncertainty

The role of the uncertainty is to gradually update the student model with reliable target regions from the teacher model predictions. Our proposed method estimates the uncertainty directly from the anatomically-aware representation network $f_d(f_e(\cdot))$, requiring only one inference step. First, we map the segmentation prediction from the teacher model ($\mathbf{P}^i_T$) with a DAE model to produce a plausible segmentation $\hat{\mathbf{P}}^i_T = f_d(f_e(\mathbf{P}^i_T))$. We subsequently estimate the uncertainty as the pixel-wise difference between the DAE output and the prediction, which is given as:

$$\mathbf{U}^i = ||\hat{\mathbf{P}}^i_T - \mathbf{P}^i_T||^2. \tag{3.4}$$

Note that the uncertainty formulation is related to the conventional sample variance-based uncertainty estimation. Specifically, for a given input, $\mathbf{X}^i$, and its corresponding multiple model predictions, $\mathbf{P}^{i_s}$, the sample variance estimation is defined as follows:

$$var(\mathbf{P}^i) = \frac{1}{S-1}\sum_{s=1}^{S}(\mathbf{P}^{i_s} - \bar{\mathbf{P}}^i)^2, \tag{3.5}$$

where $\bar{\mathbf{P}}^i$ represents the sample mean and is defined as $\bar{\mathbf{P}}^i = \frac{1}{S}\sum_{s=1}^{S}(\mathbf{P}^{i_s})$. The parameter S denotes the number of prediction samples. When S is set to 2, the sample mean $\bar{\mathbf{P}}^i$ reduces to $\frac{\mathbf{P}^{i_1}+\mathbf{P}^{i_2}}{2}$, resulting in the variance estimation taking the form of:

$$\begin{aligned} var(\mathbf{P}^i) &= (\mathbf{P}^{i_1} - \frac{\mathbf{P}^{i_1} + \mathbf{P}^{i_2}}{2})^2 + (\mathbf{P}^{i_2} - \frac{\mathbf{P}^{i_1} + \mathbf{P}^{i_2}}{2})^2, \\ &= (\frac{\mathbf{P}^{i_1} - \mathbf{P}^{i_2}}{2})^2 + (\frac{\mathbf{P}^{i_2} - \mathbf{P}^{i_1}}{2})^2, \\ var(\mathbf{P}^i) &= \frac{1}{2}(\mathbf{P}^{i_1} - \mathbf{P}^{i_2})^2. \end{aligned} \tag{3.6}$$

The above equation is equivalent to our uncertainty formulation in Eq. 3.4, where two samples are drawn from the output of the teacher model and the DAE model.

The resulting uncertainty maps from Eq. 3.4 are subsequently used to obtain the reliable target regions as follows: $e^{-\gamma \mathbf{U}^i}$, similarly to (Luo *et al.*, 2022), where γ is an uncertainty weighting factor empirically set to 1. The reliable targets are finally combined in a consistency loss as:

$$\mathcal{L}_c(\mathbf{P}_S^i, \mathbf{P}_T^i) = \frac{\sum_v e^{-\gamma \mathbf{U}^i} ||\mathbf{P}_S^i - \mathbf{P}_T^i||^2}{\sum_v e^{-\gamma \mathbf{U}^i}}, \tag{3.7}$$

where v is a voxel. Note that the consistency loss $\mathcal{L}_c$ will be equivalent to a standard mean teacher method (Tarvainen & Valpola, 2017) when $\gamma = 0$. Overall, we jointly optimize the consistency loss $\mathcal{L}_c$ and supervised loss $\mathcal{L}_s$ as learning objectives, where $\mathcal{L}_s$ is a combination of cross-entropy and dice losses.

3.4 Experiments

3.4.1 Datasets

The performance of our method is validated on two publicly available benchmarks: (a) the left atrium (LA) binary segmentation dataset from the 2018 atrial challenge (Xiong *et al.*, 2021),

and (b) the abdominal multi-organ segmentation dataset from the FLARE challenge (Ma *et al.*, 2022).

(a) LA dataset

It consists of 100 3D late gadolinium-enhanced magnetic resonance imaging (LGE-MRI) scans and corresponding LA segmentation masks. These scans have an isotropic resolution of 0.625 mm^3 and are center cropped at the heart region. The dataset is split into 80 for training and the remaining 20 for testing as in the literature (Li *et al.*, 2020a; Luo, Chen, Song & Wang, 2021; Wang *et al.*, 2020; Yu *et al.*, 2019).

(b) FLARE dataset

This dataset consists of 361 CT scans of the abdominal region and corresponding segmentation masks of four organs, namely liver, kidney, spleen, and pancreas. These scans are collected from multiple medical centers, having varying resolutions. Each image is first resampled to a uniform resolution of $2 \times 2 \times 2.5$ mm^3 and then normalized by clipping the intensity values outside $[0.5, 0.95]$ percentile range. For all our experiments, we use a fixed dataset split of 260 for training, 26 for validation, and the remaining 75 for testing.

3.4.2 Implementation and Training details

To validate our proposed method, we employ a V-Net (Milletari *et al.*, 2016) as a backbone architecture for the segmentation networks, as followed in earlier work (Luo *et al.*, 2021; Wang *et al.*, 2020; Yu *et al.*, 2019). Our anatomically-aware representation prior module (i.e., a DAE) follows a similar architecture as V-Net but without skip connections. Such design effectively makes it an autoencoder-style architecture, which is also comparable to prior work (Larrazabal *et al.*, 2020; Oktay *et al.*, 2017). To encode the segmentation mask in a latent space, a dense layer of d-dimension is added at the bottleneck layer of the DAE module as shown in Fig. 3.2.

For training, the student model uses a SGD optimizer with an initial learning rate (lr) of 0.1 and a momentum of 0.9 with a cosine annealing decaying (Loshchilov & Hutter, 2017). The teacher weights (in Eq. 3.3) are updated by an EMA with a rate of $\alpha = 0.99$ (Tarvainen & Valpola, 2017). The DAE model is also trained using a SGD optimizer with an initial $lr = 0.1$, a momentum of 0.9, and decaying the lr by a factor of 2 every 5000 iterations. Following the literature (Luo *et al.*, 2022; Yu *et al.*, 2019), the consistency weight β and ramp-up factor r in Eq. 3.2 are set to 0.1 and 5, respectively. Inputs to both segmentation and DAE networks are randomly cropped to a size of $112 \times 112 \times 80$ and $144 \times 144 \times 96$ for LA and FLARE datasets, respectively. We employ online standard data augmentation techniques such as random flipping and rotation. In addition, input labels to the DAE are corrupted with a random swapping of pixels around class boundaries, morphological operations (erosion and dilation), resizing, and adding/removing basic shapes (Van der Walt *et al.*, 2014). The latent space of the DAE is injected with a small noise drawn from a Gaussian distribution to explore different sets of plausible segmentation during training of the segmentation network. The training set is partitioned into N_l labeled and N_u unlabeled splits, which are fixed across all experiments. The batch size is set to 4 in both networks. Input batch for the segmentation network uses two labeled and unlabeled data. During the inference phase, the segmentation predictions are generated using the sliding window strategy. For the cardiac dataset (LA), following the literature (Li *et al.*, 2020a; Luo *et al.*, 2021; Yu *et al.*, 2019), the final model is evaluated at the last training iteration (i.e., 6000), whereas the best validation model is selected in the case of the abdominal dataset (FLARE). All our experiments were run on an NVIDIA RTX A6000 GPU with PyTorch 1.8.0. The implementation of our work is available at: https://github.com/adigasu/Anatomically-aware_Uncertainty_for_Semi-supervised_Segmentation.

3.4.3 Evaluation

We employ common Dice Score Coefficient (DSC) and 95% Hausdorff Distance (HD) evaluation measures to assess quantitative segmentation performance. The DSC score evaluates the degree of overlap between ground truth and prediction regions. In contrast, the HD score measures the

distance between ground truth and predicted segmentation boundaries. For a fair comparison, all experiments are run three times with a fixed set of seeds on the same machine, and their average results are reported.

3.5 Results

3.5.1 Comparison with the state-of-the-art

We first compare our method with relevant semi-supervised segmentation approaches and report the quantitative results in Tables 3.1 and 3.2. The upper and lower bound from the backbone architecture V-Net (Milletari *et al.*, 2016) are reported at the top of each section. Furthermore, non-uncertainty-based methods such as MT (Tarvainen & Valpola, 2017), DTC (Luo *et al.*, 2021), and SASSnet (Li *et al.*, 2020a) and uncertainty-based methods UAMT (Yu *et al.*, 2019), DUMT (Wang *et al.*, 2020), and URPC (Luo *et al.*, 2022) are included in our evaluation.

(a) Left Atrium segmentation

Table 3.1 shows the segmentation performance on the Left Atrium (LA) test set under the standard 10% (top) and 20% (bottom) annotation settings. From the top half of the table, we observe that leveraging unlabeled data improves the lower bound across all models. The uncertainty-based approaches typically outperform their non-uncertainty counterparts in terms of DSC, but yield inferior results in terms of HD. Among these methods, UAMT and DTC achieve the best DSC and HD scores, respectively. Nevertheless, compared to these best-performing baselines, our method brings improvements in both DSC (1.5%) and HD (0.8mm) scores. Moreover, uncertainty estimation in our method requires a single inference from an anatomically-aware representation, whereas UAMT uses K=8 inferences per training step to obtain an uncertainty map. This highlights the efficiency of the proposed approach, which yields a better segmentation performance yet requires substantially less computational time at each training step.

Furthermore, we validate our method on the 20% annotation scenario, whose results are reported in bottom half of Table 3.1. We observe a similar trend in these results, with uncertainty-based approaches outperforming non-uncertainty-based methods in DSC, whereas their performance in terms of HD is degraded. An interesting observation is that existing methods are ranked differently across the two annotation settings, indicating that they might be sensitive to the annotation scenario. For example, while UAMT achieves the best DSC score under the 10% annotation setting, URPC yields the best results in the 20% annotation case. Similarly, the best models are different for HD metric, i.e., DTC under the 10% setting and SASSNet in the 20% setting. In contrast, our method consistently outperforms each existing approach in both DSC and HD scores, highlighting its robustness against the amount of labeled data.

Table 3.1 Segmentation results on the LA test set for the 10% and 20% annotation settings. Uncertainty-based methods with K inferences per training step are grouped at the bottom of each section, while K = - indicates non-uncertainty-based methods. Ours achieves the best DSC and HD scores in both annotation scenarios. The best and second-best results are highlighted in bold and underlined, whereas the statistical significance between the top two results is denoted in *

N_l/N_u	**Methods**	**#K**	**DSC (%) ↑**	**HD (mm) ↓**
80/0	Upper bound	-	91.23 ± 0.44	6.08 ± 1.84
8/0	Lower bound	-	76.07 ± 5.02	28.75 ± 0.72
8/72 (10%)	MT (Tarvainen & Valpola, 2017)	-	78.22 ± 6.89	16.74 ± 4.80
	SASSnet (Li *et al.*, 2020a)	-	83.70 ± 1.48	16.90 ± 1.35
	DTC (Luo *et al.*, 2021)	-	83.10 ± 0.26	<u>12.62 ± 1.44</u>
	UAMT (Yu *et al.*, 2019)	8	<u>85.09 ± 1.42</u>	18.34 ± 2.80
	DUMT (Wang *et al.*, 2020)	16	82.97 ± 1.76	14.43 ± 0.67
	URPC (Luo *et al.*, 2022)	1	84.47 ± 0.31	17.11 ± 0.60
	Ours	1	**86.58 ± 1.03***	**11.82 ± 1.42**
16/0	Lower bound	-	81.46 ± 2.96	23.61 ± 4.94
16/64 (20%)	MT (Tarvainen & Valpola, 2017)	-	86.06 ± 0.81	11.63 ± 3.40
	SASSnet (Li *et al.*, 2020a)	-	87.81 ± 1.45	<u>10.18 ± 0.55</u>
	DTC (Luo *et al.*, 2021)	-	87.35 ± 1.26	10.25 ± 2.49
	UAMT (Yu *et al.*, 2019)	8	87.78 ± 1.03	11.10 ± 1.91
	DUMT (Wang *et al.*, 2020)	16	87.42 ± 0.97	10.78 ± 2.26
	URPC (Luo *et al.*, 2022)	1	<u>88.58 ± 0.10</u>	13.10 ± 0.60
	Ours	1	**88.60 ± 0.82**	**7.61 ± 0.78***

Table 3.2 Segmentation results on the FLARE test set for the 10% and 20% annotation settings. Uncertainty-based methods with K inferences per training step are grouped at the bottom of each section, while K = - indicates non-uncertainty-based methods. Our method produces the best results on average. The best and second-best results are highlighted in bold and underlined, whereas * denotes statistical significance between the top two results

	N_l/N_u	Methods	#K	Average	Liver	Kidney	Spleen	Pancreas
DSC (%) ↑	260/0	Upper bound	-	85.80 ± 1.42	94.95 ± 0.30	93.20 ± 0.81	89.65 ± 2.91	65.38 ± 2.57
	26/0	Lower bound	-	70.09 ± 2.77	88.37 ± 2.31	81.12 ± 2.49	70.74 ± 4.41	40.14 ± 3.84
		MT (Tarvainen & Valpola, 2017)	-	70.76 ± 2.79	88.77 ± 3.11	83.34 ± 1.22	72.91 ± 4.35	38.01 ± 2.62
		SASSnet (Li *et al.*, 2020a)	-	61.43 ± 14.3	86.94 ± 2.88	63.59 ± 43.0	59.83 ± 18.6	35.36 ± 5.05
	26/234	DTC (Luo *et al.*, 2021)	-	68.07 ± 1.42	87.99 ± 1.79	83.11 ± 3.93	66.04 ± 3.40	35.15 ± 1.26
	(10%)	UAMT (Yu *et al.*, 2019)	8	73.63 ± 0.65	**91.65 ± 0.49**	84.70 ± 2.39	76.16 ± 2.58	42.01 ± 2.24
		DUMT (Wang *et al.*, 2020)	16	69.04 ± 1.39	87.28 ± 0.82	80.47 ± 3.88	68.23 ± 6.79	40.18 ± 2.59
		URPC (Luo *et al.*, 2022)	1	73.31 ± 1.11	91.09 ± 0.62	85.88 ± 1.82	75.40 ± 2.64	40.89 ± 4.05
		Ours	1	**75.28 ± 1.54***	90.78 ± 1.26	**87.09 ± 1.89**	**78.13 ± 1.23**	**45.12 ± 2.20***
HD (mm) ↓	260/0	Upper bound	-	6.37 ± 1.15	5.50 ± 2.86	3.31 ± 1.10	7.49 ± 1.94	9.17 ± 0.66
	26/0	Lower bound	-	18.51 ± 4.01	15.26 ± 0.90	9.89 ± 2.13	30.51 ± 11.9	18.40 ± 3.53
		MT (Tarvainen & Valpola, 2017)	-	18.58 ± 1.66	12.09 ± 3.72	8.70 ± 0.85	35.89 ± 7.47	17.64 ± 1.53
		SASSnet (Li *et al.*, 2020a)	-	27.76 ± 8.51	24.59 ± 23.0	15.1 ± 11.1	51.86 ± 21.3	19.53 ± 0.89
	26/234	DTC (Luo *et al.*, 2021)	-	23.11 ± 6.01	21.63 ± 16.7	18.8 ± 11.3	32.64 ± 16.8	19.31 ± 2.07
	(10%)	UAMT (Yu *et al.*, 2019)	8	14.30 ± 1.94	**10.44 ± 1.45**	8.08 ± 1.41	20.44 ± 6.18	18.24 ± 3.04
		DUMT (Wang *et al.*, 2020)	16	22.35 ± 3.82	13.23 ± 2.28	19.21 ± 13.9	36.17 ± 15.5	20.77 ± 3.58
		URPC (Luo *et al.*, 2022)	1	14.23 ± 1.97	11.71 ± 2.37	**7.41 ± 1.16**	20.82 ± 5.02	16.96 ± 3.00
		Ours	1	**13.69 ± 0.68**	10.85 ± 1.69	9.48 ± 2.10	**18.45 ± 4.17**	**15.98 ± 1.33**
DSC (%) ↑	52/0	Lower bound	-	70.15 ± 1.58	88.40 ± 1.24	81.91 ± 2.07	68.40 ± 5.68	41.88 ± 7.44
		MT (Tarvainen & Valpola, 2017)	-	72.10 ± 1.84	89.82 ± 2.30	85.15 ± 1.66	71.87 ± 4.28	41.55 ± 2.99
		SASSnet (Li *et al.*, 2020a)	-	69.74 ± 4.43	88.41 ± 1.10	86.19 ± 3.13	64.11 ± 12.1	40.25 ± 3.07
	52/208	DTC (Luo *et al.*, 2021)	-	68.49 ± 1.30	89.61 ± 0.71	83.31 ± 4.39	62.76 ± 5.64	38.29 ± 3.38
	(20%)	UAMT (Yu *et al.*, 2019)	8	74.72 ± 1.15	89.54 ± 3.10	87.92 ± 1.52	73.07 ± 3.91	**48.34 ± 1.41**
		DUMT (Wang *et al.*, 2020)	16	72.08 ± 2.77	90.11 ± 1.66	85.43 ± 4.82	71.83 ± 0.92	40.94 ± 4.17
		URPC (Luo *et al.*, 2022)	1	74.26 ± 1.02	91.02 ± 0.54	87.91 ± 2.47	72.06 ± 1.82	46.03 ± 0.40
		Ours	1	**76.69 ± 0.81***	**91.84 ± 1.00***	**88.72 ± 0.74**	**78.07 ± 0.69***	48.14 ± 1.73
HD (mm) ↓	52/0	Lower bound	-	15.63 ± 0.33	15.18 ± 4.46	11.93 ± 4.64	20.50 ± 2.56	**14.91 ± 2.78**
		MT (Tarvainen & Valpola, 2017)	-	16.39 ± 3.34	**11.04 ± 0.58**	10.89 ± 0.91	25.70 ± 9.08	17.94 ± 4.50
		SASSnet (Li *et al.*, 2020a)	-	23.84 ± 0.79	34.01 ± 14.3	11.89 ± 8.66	32.28 ± 1.53	17.16 ± 1.69
	52/208	DTC (Luo *et al.*, 2021)	-	22.46 ± 2.12	25.23 ± 20.1	18.09 ± 8.14	29.05 ± 4.84	17.46 ± 1.02
	(20%)	UAMT (Yu *et al.*, 2019)	8	14.50 ± 2.46	16.60 ± 4.11	7.83 ± 0.76	17.91 ± 8.34	15.66 ± 0.76
		DUMT (Wang *et al.*, 2020)	16	15.53 ± 2.75	11.74 ± 2.27	8.64 ± 0.95	25.43 ± 8.42	16.31 ± 0.89
		URPC (Luo *et al.*, 2022)	1	14.16 ± 0.68	11.16 ± 2.09	8.47 ± 2.79	20.66 ± 0.80	16.33 ± 1.70
		Ours	1	**13.11 ± 0.45**	11.32 ± 2.29	**7.79 ± 2.69**	**17.38 ± 4.19**	15.94 ± 0.28

(b) Abdominal multi-organ segmentations

Table 3.2 presents the performance of the abdominal multi-organ segmentations on the FLARE test set. The results of 10% and 20% annotation experiments are grouped in the top and bottom half of the table, respectively. We report individual organs as well as average results. From the top half of the table, we first notice that the performance of most existing methods is improved when compared to the lower bound in both DSC and HD scores, except SASSNet, DTC, and DUMT. The gap in the segmentation performance of SASSNet and DTC is due to the use of

signed distance maps (SDM), which are designed for binary segmentation. Adopting these methods for multi-class segmentation is challenging since it requires careful hyperparameter tuning of per-class SDM predictions, which is beyond the scope of this work. Note that DUMT did not outperform the simple baseline under a multi-class setting, which is consistent with the observations in (Van Waerebeke *et al.*, 2022). Among the existing methods, the uncertainty-based methods (UAMT and URPC) perform well in both segmentation measures. These methods improve the segmentation of liver and spleen regions, achieving the best average DSC and HD scores in UAMT and URPC, respectively. Compared to these best-performing baselines, our method predominantly improves the segmentation of challenging regions, notably the pancreas organ. Overall, our anatomically-aware method consistently performs well in all regions and improves average DSC (1.65%) and HD (0.6mm) scores.

The results of the 20% annotation scenario are reported in the bottom half of Table 3.2. We notice a similar trend in the results when compared to the 10% annotation setting. All existing methods, except SASSNet and DTC, improve the segmentation performance over the lower bound in both DSC and HD scores. Our method outperforms the best-performing baselines (UAMT and URPC) in most cases and improves the average DSC (1.95%) and average HD (1mm) scores. These results show that our method consistently outperforms the existing approaches across different datasets and labeling scenarios. We can, therefore, argue that including our novel anatomically-aware module is a valuable alternative to existing semi-supervised segmentation approaches.

3.5.2 Qualitative Analysis

Visual results of the left atrium (LA) segmentation obtained by different methods are depicted in Fig. 3.3. In the top row (10% annotation setting), the existing approaches produce segmentation output with holes (SASSnet, UAMT) and noisy boundaries (SASSnet, DTC, UAMT, DUMT). In contrast, URPC and our methods produce smoother segmentations, but URPC generates

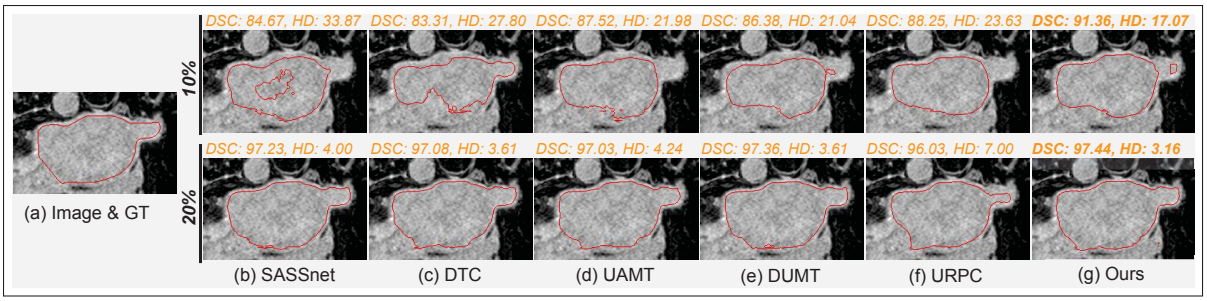

Figure 3.3 Qualitative comparison under the 10% and 20% annotation settings on LA dataset. DSC (%) and HD (mm) scores are mentioned at the top of each image. Each image is overlaid with a contour of segmentation prediction or ground truth (red)

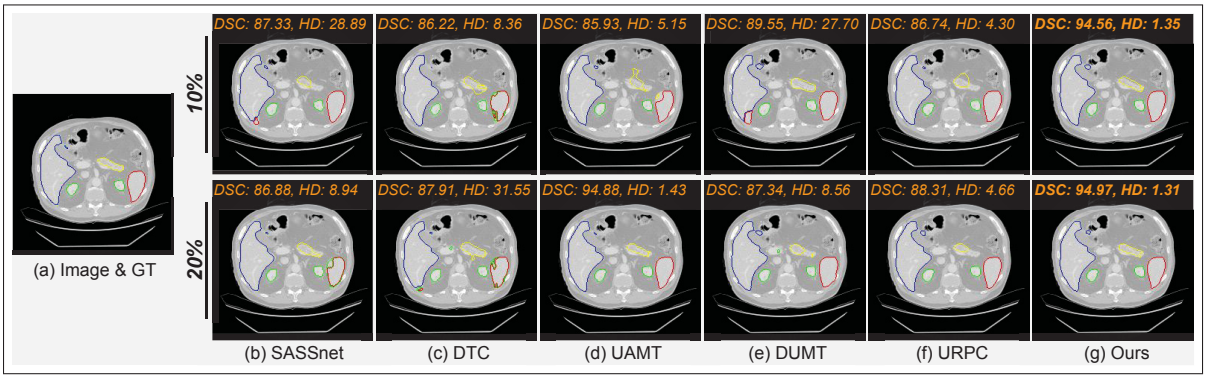

Figure 3.4 Qualitative comparison under the 10% and 20% annotation settings on FLARE dataset. Average DSC (%) and average HD (mm) scores are mentioned at the top of each image. The colorings are liver (blue), kidney (green), spleen (red), and pancreas (yellow)

under-segmented output compared to our method. Note that a post-processing tool is commonly employed in SASSNet to improve the segmentation performance. However, this is avoided in our experiments for a fair comparison. In the 20% annotation setting (bottom row), with access to more labeled data, all methods reduce segmentation errors. Even in this case, our method produces promising and smoother segmentations when compared to existing approaches.

To highlight the deficiencies of these approaches in multi-class segmentation, we now show qualitative results on abdominal organs in Fig. 3.4. In the 10% annotation setting (top row), we first observe that misclassification between different organs is a common problem across existing approaches, notably in SASSnet, DTC, UAMT, and DUMT. For instance, part of the liver is

segmented as a spleen in SASSnet and DUMT, whereas the parts of the spleen are misclassified as kidneys in DTC and as pancreas in UAMT. This misclassification could be due to either similar intensity characteristics across different organs (Durieux, Gevenois, Muylem, Howarth & Keyzer, 2018) or the inefficiency of networks in discriminating multi-class distributions (Van Waerebeke *et al.*, 2022). Furthermore, most methods (SASSnet, DTC, UAMT, URPC) have failed to capture the challenging pancreas region. In contrast, our method provides an improved segmentation in this challenging region and minimizes classification errors. In the bottom row of Fig. 3.4, adding more labeled images to the training (20% annotation setting) also reduces classification errors (UAMT, URPC). Our method similarly improves the segmentation performance in all observed regions. The quantitative results from the previous section further support the superiority of our approach. Overall, we argue that the observed improvements in both datasets could be attributed to the knowledge derived from the anatomically-aware representation.

3.5.3 Choice of Latent Space in DAE

Our anatomically-aware prior (DAE) plays a vital role in guiding the segmentation model. Therefore, we investigate the impact of the design choices made in the DAE on the final segmentation performance. The latent space (LS) of our DAE is first studied under varying sizes (d) across two datasets in Fig. 3.5. The results show that the segmentation performance varies with LS sizes. The best results are achieved for d=128 in binary left atrium segmentations and d=512 in abdominal multi-organ segmentations. It indicates that the choice of the latent space size, d, depends on the complexity of the dataset.

Furthermore, the LS of the DAE is perturbed with an addition of a Gaussian noise. This facilitates a different set of reconstructions from the DAE when training the segmentation model. The different reconstructions aid in better guiding the segmentation model. To validate this notion, we conduct experiments with and without adding a noise in the LS across both datasets in Fig.3.6. The results demonstrate that the final segmentation performance improves up to 1.79%

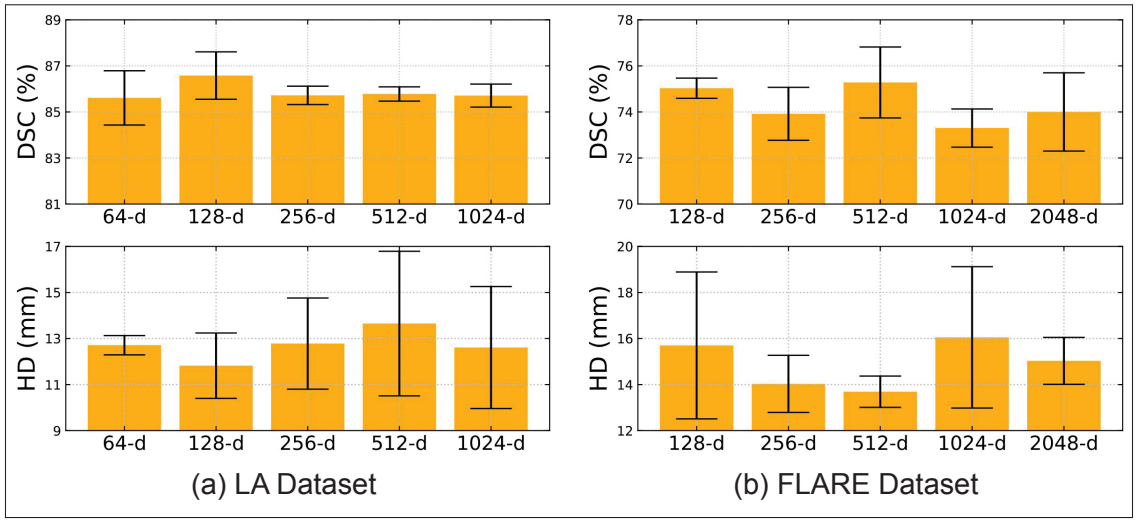

Figure 3.5 Segmentation performance with different latent space sizes of DAE. Each bar indicates the DSC (top) and HD (bottom) scores under the 10% annotation setting. The best results are obtained for the latent space size d=128 in binary LA segmentations (a), whereas d=512 is needed for abdominal multi-organ segmentations (b)

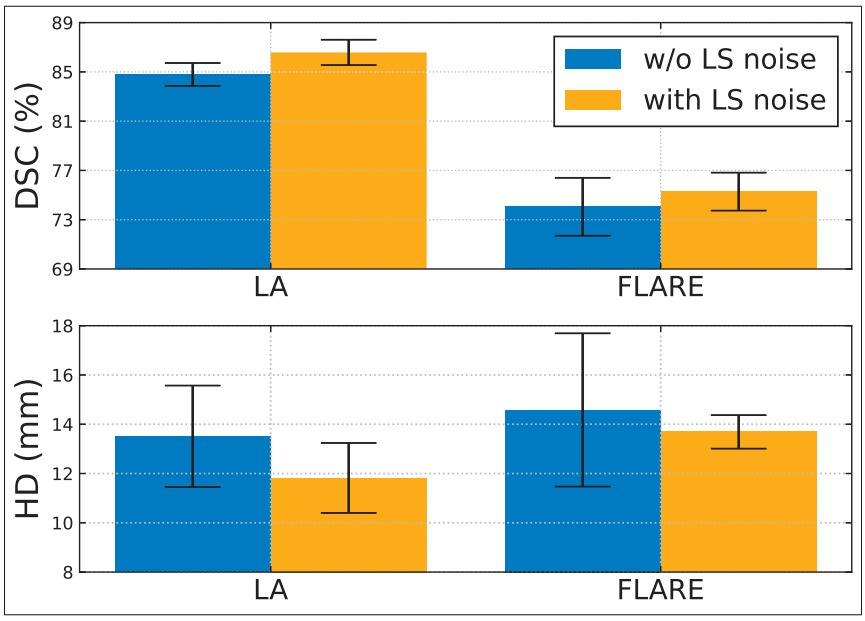

Figure 3.6 Impact of noise in the latent space of DAE on segmentation performance. Each bar indicates the DSC (top) and HD (bottom) scores under the 10% annotation setting. The addition of a noise (orange) in latent space improves DSC and HD scores

Table 3.3 Effectiveness of our proposed uncertainty estimation on segmentation results using different strategies. N_l and N_u indicate the number of labeled and unlabeled data

		LA Dataset		FLARE Dataset	
N_l/N_u	**Methods**	**DSC (%) ↑**	**HD (mm) ↓**	**DSC (%) ↑**	**HD (mm) ↓**
8/72 (10%)	UAMT (Yu *et al.*, 2019)	85.09 ± 1.42	18.34 ± 2.80	73.63 ± 0.65	14.30 ± 1.94
	Ours (Threshold)	85.39 ± 0.91	12.96 ± 3.05	74.25 ± 1.76	14.47 ± 1.63
	Ours (Entropy)	85.92 ± 1.52	**11.16 ± 0.82**	74.01 ± 0.62	15.03 ± 2.00
	Ours	**86.58 ± 1.03**	11.82 ± 1.42	**75.28 ± 1.54**	**13.69 ± 0.68**
16/64 (20%)	UAMT (Yu *et al.*, 2019)	87.78 ± 1.03	11.10 ± 1.91	74.72 ± 1.15	14.50 ± 2.46
	Ours (Threshold)	88.12 ± 1.16	8.44 ± 1.96	74.80 ± 0.80	14.09 ± 1.83
	Ours (Entropy)	87.76 ± 0.36	8.90 ± 0.48	74.57 ± 0.53	15.38 ± 2.57
	Ours	**88.60 ± 0.82**	**7.61 ± 0.78**	**76.69 ± 0.81**	**13.11 ± 0.45**

in DSC and 1.69mm in HD by adding a noise in the LS of the DAE module. These analyses show the impact of our design choices in the anatomically-aware prior on the segmentation performance.

3.5.4 Ablation Study on uncertainty

To validate the effectiveness of our uncertainty estimation on the segmentation performance, we conducted two experiments by adopting a threshold strategy and a predictive entropy scheme used in UAMT. Specifically, a threshold strategy filters out the most unreliable region from the uncertainty map ($\mathbf{U}^i$), defined as $H > \mathbf{U}^i$ with a threshold, H, set with a ramp-up function, as in UAMT (Yu *et al.*, 2019). In the entropy experiments, we estimate the uncertainty ($\mathbf{U}^i$) using the entropy of the DAE prediction ($\hat{\mathbf{P}}^i_T$) and then combining it in a consistency loss as in Eq 3.7. The results of these ablation experiments on the LA and FLARE datasets under the 10% and 20% annotation settings are reported in Table 3.3. Compared to UAMT, our threshold and entropy experiments improve the segmentation performance in both DSC and HD scores in most cases. At the same time, our proposed uncertainty method (Sec. 3.3.3.2) achieves the best performance in all the settings. These results show the merit of our anatomically-aware uncertainty estimation for guiding the segmentation model.

3.5.5 Impact of γ and β hyperparameters

The sensitivity of the uncertainty weight γ (in Eq.3.7) and the consistency weight β on the segmentation performance is shown in Fig. 3.7. In particular, we evaluate the segmentation performance using DSC and HD scores by varying the γ and β values across the LA and FLARE datasets. In Fig. 3.7(a)-(b), increasing the gamma value leads to an improvement in the segmentation performance in both DSC and HD scores across both datasets. The best results are usually observed for $\gamma = 1$. Beyond that, performance generally decreases, possibly due to an exponential decrease in the weight (Eq.3.7) of the reliable target regions.

Figure 3.7(c)-(d) shows the segmentation performance for varying the β values. The results show that increasing the beta value improves the segmentation performance. The best result is achieved for β=0.1 except in the LA dataset (in Fig. 3.7(c)), where β=1 produces the best scores. Nevertheless, we chose to set β=0.1 across all our experimental scenarios, as this value is widely adopted in the literature on consistency-based approaches (Tarvainen & Valpola, 2017; Wang *et al.*, 2021) and for a fair comparison with our baselines (Luo *et al.*, 2022; Wang *et al.*, 2020; Yu *et al.*, 2019).

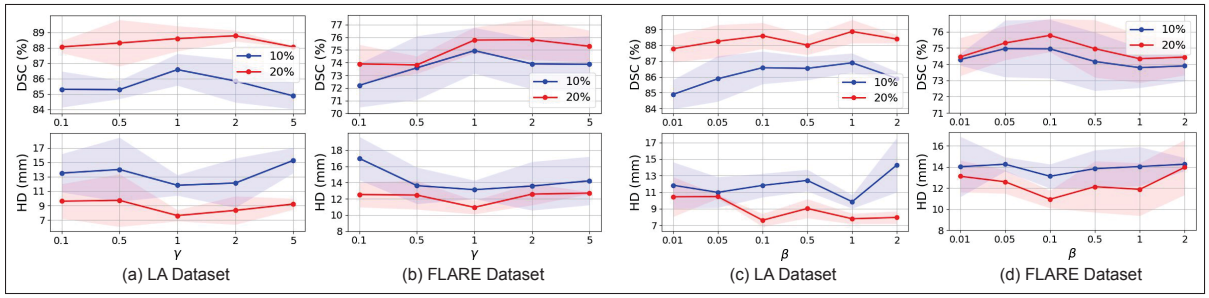

Figure 3.7 Sensitivity of the consistency weight β (a, b) and the uncertainty weight γ (c, d). Each point in a line indicates the DSC (top) and HD (bottom) scores on LA and FLARE datasets under 10% (blue) and 20% (red) annotation settings

3.5.6 Training time

To evaluate the speed of our uncertainty estimation, we compare the computation time required for each training iteration by the proposed and the baseline methods in Table 3.4. From the table, we observe that the non-uncertainty-based methods (SASSnet, DTC) are slower when compared to uncertainty-based methods across both datasets, LA and FLARE. The relative slow speed of SASSnet and DTC is attributed to the additional computational overhead required for predicting the signed distance maps (SASSnet, DTC) and the inclusion of a discriminator module (SASSnet). On the other hand, ours and the URPC method are faster than the MCDO-based methods (UAMT and DUMT) due to the need of only one inference when estimating the uncertainty (#K=1). Overall, our approach adds a minimal overhead on top of the mean teacher (MT) approach for estimating uncertainty while producing superior segmentation results on both datasets.

Table 3.4 Comparison of average training times in seconds per iteration. Our method adds a minimal overhead on top of the mean teacher (MT) approach for uncertainty estimation

Methods	#K	LA	FLARE
MT (Tarvainen & Valpola, 2017)	-	0.612	1.108
SASSnet (Li *et al.*, 2020a)	-	1.442	5.856
DTC (Luo *et al.*, 2021)	-	0.989	4.874
UAMT (Yu *et al.*, 2019)	8	1.207	2.429
DUMT (Wang *et al.*, 2020)	16	3.804	7.678
URPC (Luo *et al.*, 2022)	1	0.779	1.504
Ours	1	0.745	1.266

3.5.7 Uncertainty Analysis

The predicted segmentation and uncertainty map from different uncertainty-based methods are shown in Fig. 3.8. The top row shows the 10% annotation setting, where uncertainties are all over the predicted regions for UAMT. These uncertainties inside the prediction regions are reduced in DUMT, possibly due to more inferences and the addition of feature uncertainty. However, the

uncertainties are highly focused on the prediction boundaries. The uncertainty is produced at arbitrary regions in URPC due to their multi-scale discrepancy-based uncertainty estimation. Our method produces uncertainty in challenging regions, such as unclear anatomical boundaries or annotator cuts (as in pulmonary veins), which are estimated using anatomically-aware representation. In the below row of Fig. 3.8, increasing labeled samples (i.e., 20% setting) improves the predictions and uncertainty in most cases. Nevertheless, uncertainties are all over the boundaries, or arbitrary regions remain in the existing methods. Our method further improves the uncertainties due to the improvement of anatomically-aware representation using more access to labels. Moreover, our method requires a single inference when compared to entropy-based methods.

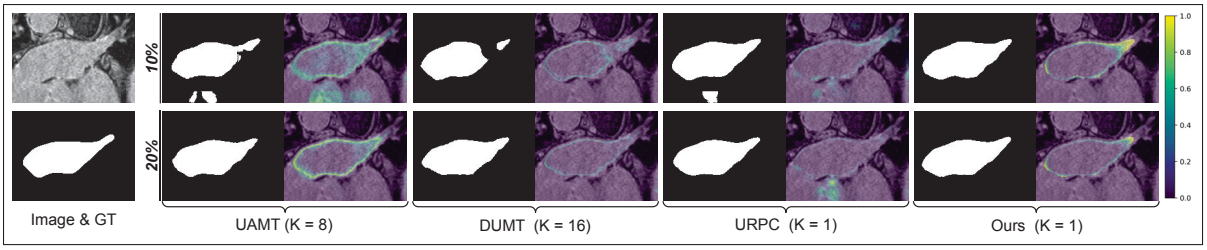

Figure 3.8 Uncertainty analysis on the left atrium dataset. Prediction and uncertainty map (overlaid on its image) are shown for each uncertainty-based method. The number of inferences for generating the uncertainty map is denoted as K

3.6 Discussion and Conclusion

This work proposes a novel anatomically-aware uncertainty estimation method for semi-supervised image segmentation. Our approach consists of leveraging an anatomically-aware representation of labeling masks to estimate the segmentation uncertainty. The obtained uncertainty maps guide the training of the segmentation model within reliable regions of the predicted masks. Our experimental results demonstrate that the proposed method yields improved segmentation results when compared to state-of-the-art baselines on two publicly available benchmarks using left atria and abdominal organs. The qualitative results also show how our anatomically-aware approach improves segmentation in challenging image areas. The

ablation studies demonstrate the effectiveness and robustness of our uncertainty estimation when compared to entropy-based methods. Adding noise in the latent space of our representation helps to map the predictions into a better set of plausible segmentations, which improves the segmentation accuracy. Unlike most uncertainty-based approaches, our anatomically-aware uncertainty requires a single inference, thereby reducing computational complexity. Moreover, as our anatomically-aware representation is independent of any image information, it can be further enhanced with existing segmentation masks from different datasets or imaging modalities (Karani *et al.*, 2021), potentially further improving the modeling capacity of our representation. The learning representation with an additional constraint can also be explored separately as a post-processing tool that maps the erroneous prediction into anatomically-plausible segmentation (Larrazabal *et al.*, 2020; Painchaud *et al.*, 2020). Additionally, our anatomically-aware representation prior could also benefit from the image intensity information to learn a joint representation (Judge *et al.*, 2022; Oktay *et al.*, 2017) for uncertainty estimation in a limited supervision problem. Overall, our proposed approach could be leveraged to a broader range of applications where uncertainties could be related to anatomical information.

CHAPTER 4

ATTENTION-BASED DYNAMIC SUBSPACE LEARNERS FOR MEDICAL IMAGE ANALYSIS

Sukesh Adiga Vasudeva[a], Jose Dolz[a], Herve Lombaert[a]

[a] Department of Software and IT Engineering, École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *IEEE Journal of Biomedical And Health Informatics* (JBHI), September 2022

Presentation

This chapter presents the article *"Attention-based Dynamic Subspace Learners for Medical Image Analysis"* (Adiga Vasudeva, Dolz & Lombaert, 2022a) submitted to the IEEE **JBHI** (Journal of Biomedical And Health Informatics), sent on 11 April 2021, revised 13 October 2021 and 21 January 2022, and accepted for publication on 10 June 2022. The journal article was also presented as a short paper (Adiga Vasudeva, Dolz & Lombaert, 2022c) at the conference **MIDL** (Medical Imaging with Deep Learning) in Zurich, Switzerland. The objective of this article is to develop attention-based dynamic representation learners for various medical image analysis applications, including segmentation, clustering, and retrieval tasks.

4.1 Introduction

Learning the similarity between arbitrary images is a fundamental problem in many key areas of computer vision, such as image retrieval (He, Zhou, Zhou, Bai & Bai, 2018; Movshovitz, Toshev, Leung, Ioffe & Singh, 2017; Sohn, 2016), recommender system (Ma, Zhou, Cui, Yang & Zhu, 2019), duplicate detection (Zheng, Song, Leung & Goodfellow, 2016), clustering (Ziko *et al.*, 2018), or zero-shot learning (Zhang & Saligrama, 2016). In this context, metric learning is commonly used for measuring similarities by learning a distance function over objects (Kulis

et al., 2012; Weinberger, Blitzer & Saul, 2006). Recently, deep metric learning (DML) has been raised as a powerful approach to learn these similarities (Kaya & Bilge, 2019). More specifically, the goal of DML is to learn an embedding space where images from the same classes are encouraged to be close to one another. In contrast, images belonging to different classes are pushed away in the embedding space. In recent DML approaches, the loss function can be typically expressed in Euclidean distances or cosine similarities between pairs or tuples of images in the embedding space. Well-known losses employed in DML include: contrastive loss (Hadsell *et al.*, 2006), triplet loss (Wang *et al.*, 2014a), lifted structure loss (Oh Song, Xiang, Jegelka & Savarese, 2016), N-pairs loss (Sohn, 2016), margin loss (Wu, Manmatha, Smola & Krahenbuhl, 2017), angular loss (Wang, Zhou, Wen, Liu & Lin, 2017b), or ProxyNCA loss (Movshovitz *et al.*, 2017). In addition to novel learning objectives, recent efforts are also devoted to designing efficient sample-mining (Wu *et al.*, 2017), or sample weighting (Wang, Han, Huang, Dong & Scott, 2019b) strategies.

Most of these methods use a single-metric learner to learn the embedding mapping function. However, medical images have complex distributions consisting of different object attributes such as color, shape, size, or artifacts. Thus, learning the complex similarity associated with these different object attributes may be inadequate with only one single learner. A few attempts have been made towards leveraging multiple metric learners to address this complexity (Kim, Goyal, Chawla, Lee & Kwon, 2018; Lombaert, Zikic, Criminisi & Nicholas, 2014; Sanakoyeu, Tschernezki, Buchler & Ommer, 2019). For example, Kim *et al.* (2018) ensemble multiple learners, whereas a *divide-and-conquer* strategy is used in (Sanakoyeu *et al.*, 2019) by splitting the manifold into several embedding subspaces. One main limitation of these approaches is a need to empirically find the optimal number of learners, which requires a new validation for every new setting, including every use of a new dataset. Furthermore, the sizes of the embedding subspaces associated with each learner might differ since learning the various sets of object attributes requires varying degrees of modeling complexity.

Despite the popularity of DML, surprisingly few works attempt to visually explain which regions contribute to the similarity between images in embedding networks (Hu, Vasu & Hoogs, 2022). These visualizations are of pivotal importance since they provide an efficient mechanism to understand the predictions of the model. Recent efforts have been devoted to the interpretability of deep neural networks, resulting in a variety of different approaches (Belharbi *et al.*, 2021; Chen, Chen, Ren, Huang & Zhang, 2019; Koh & Liang, 2017; Selvaraju *et al.*, 2017; Zeiler & Fergus, 2014). Among these methods, GradCAM (Selvaraju *et al.*, 2017) has been widely employed to explain deep classification models. This method uses gradients to highlight the discriminative regions of an image. Nevertheless, since the gradients are not available during testing, directly applying this strategy in embedding networks is not feasible (Chen, Chen, Hajimirsadeghi & Mori, 2020a). Integrating interpretability in embedding networks requires either attaching an additional classification branch (Zheng, Karanam, Wu & Radke, 2019b) or employing multiple images simultaneously (Stylianou, Souvenir & Pless, 2019; Zhu, Yang & Chen, 2021). Needless to say, interpretability is of particular interest in medical imaging, as visual explanations of predictions directly impact the diagnosis, therapy planning, and follow-up of many diseases. Thus, existing DML approaches may be inadequate to visually uncover what constitutes similarities among a complex set of medical images.

Motivated by these gaps and the scarcity of the DML literature in medical imaging, we propose a novel attention-based dynamic subspace learners approach. The underlying metric learning method is inspired by the idea of a *divide-and-conquer* strategy. More specifically, we propose to follow the approach of (Sanakoyeu *et al.*, 2019) in order to capture different object attributes, each of them processed with an independent subspace learner. These subspace learners having variable sizes are learned dynamically as and when the network accuracy is plateauing during training. Thereby avoids the need to find *apriori* the number of subspace learners while retaining the state-of-the-art performance. Furthermore, the visual interpretation of the embedding is addressed by integrating an attention module after feature extraction layers, encouraging the learners to focus on the discriminative areas of target objects. Consequently, the learning process

provides a visual insight of which image region considerably contributes to the clustering of image sets in the form of pixel-wise interpretable predictions.

4.1.1 Our Contribution

We contribute a novel approach to the state-of-the-art method in deep metric learning and illustrate its application in medical image analysis. More precisely, we propose a training strategy that (i) explores the dynamic learning of an embedding, (ii) overcomes the empirical search of an optimal number of subspaces in approaches based on multiple metric learners, and (iii) produces compact subspaces of variable size to attend different object attributes. Furthermore, the integration of an attention module in our dynamic learner approach focuses the attention of each independent learner on the discriminative regions of an object of interest. This attention mechanism provides the added benefit of visually interpreting relevant embedded features. The evaluation of our proposed method is conducted by extensive experiments on three publicly available benchmarks: ISIC19 (Codella *et al.*, 2019; Combalia *et al.*, 2019), MURA (Rajpurkar *et al.*, 2018), and HyperKvasir (Borgli *et al.*, 2020). The performance is evaluated on clustering and image retrieval tasks, showing that the proposed method achieves competitive results with the state-of-the-art without requiring the grid searches over optimal numbers of learners. We also demonstrate that the attention maps produced by our method can be used as proxy labels to train deep segmentation models. In particular, we evaluate our approach on ISIC18 (Codella *et al.*, 2019; Tschandl, Rosendahl & Kittler, 2018) in a weakly supervised segmentation task and show improvements to the visual attention and class activation maps obtained from recent state-of-the-art methods, including the method specifically designed for skin lesion detection (Zhang, Xie, Xia & Shen, 2019).

4.2 Related Work

4.2.1 Deep Metric Learning

Metric learning is a widely explored research field in the learning community (Bromley *et al.*, 1994; Weinberger *et al.*, 2006). The seminal work of Siamese Networks (Bromley *et al.*, 1994) represents the first attempt to use neural networks for feature embedding. Its concept is to employ two identical neural networks that learn a contrastive embedding from a pair of images. With the advent of deep learning, deep metric learning (DML) has gained popularity, becoming a mainstay in many modern computer vision problems, such as image retrieval (Opitz, Waltner, Possegger & Bischof, 2017), person re-identification (Liao, Hu, Zhu & Li, 2015), or few-shot learning (Snell, Swersky & Zemel, 2017). In DML, the images are mapped into a manifold space via deep neural networks. Euclidean or cosine distances can then be directly used as a metric distance between two images in this mapped space. Typical losses employed in DML include contrastive (Hadsell *et al.*, 2006) or triplet loss (Schroff *et al.*, 2015). The contrastive loss (Hadsell *et al.*, 2006) encourages images from the same class to stay closer –in the learned manifold– while pushing away samples from different classes, which should be separated by a given fixed distance. Nevertheless, forcing the same distance for all pairs of images can discourage any potential distortion in the embedded space. In contrast, this assumption is relaxed in triplet loss (Hadsell *et al.*, 2006), which only imposes that negative pairs of images should be further away than positive pairs.

In the same direction as our work, (Kim *et al.*, 2018) and (Sanakoyeu *et al.*, 2019) have leveraged the use of multiple learners to diversify the learning space towards different object attributes. While Kim *et al.* (2018) proposes an ensemble of multiple learners driven by attention, a *divide and conquer* strategy is employed in (Sanakoyeu *et al.*, 2019), which promotes the discovery of multiple subspaces. For example, Sanakoyeu *et al.* (2019) explicitly splits the embedded space into a predefined number of learners with fixed-size subspaces. Then, each learner independently

learns a part of an embedding space, i.e., a subspace, from a portion of clustered data, and the final embedding is later refined by multiple learners. Even though this strategy leads to improvements over its single-learner counterpart, a grid search is needed to find an optimal number of learners with each new dataset. Furthermore, the size of the embedding space is uniform across the learners, whereas some attributes, such as color, might require smaller embeddings to encode the information than other attributes, such as shape.

4.2.2 Metric Learning in Medical Image Analysis

Despite the interest in other domains, metric learning, and more particularly DML, remains almost unexplored in medical imaging. In the pre-deep learning era, related work includes (Yang *et al.*, 2008), which employed distance metric learning in a traditional boosting framework in a medical image retrieval scenario. More recently, Yan *et al.* (2018) investigates the use of DML to model the similarity relationship between lesions in the context of radiology images, where a triplet loss is employed to learn the lesion embeddings. Gupta *et al.* (Gupta, Thapar, Bhavsar & Sao, 2019) also resorts to the triplet loss to learn the underlying manifold space for the task of Mitotic classification, whose embedded features are subsequently used as input for a Support Vector Machine classifier. Recently, a combination of cross-entropy loss and a contrastive loss or triplet loss has been used to classify whole slide images in digital pathology (Pati *et al.*, 2020; Teh & Taylor, 2019). In (Sikaroudi *et al.*, 2020), a triplet loss is used to learn a representation of source domain images, which is later used for target domain classification under the few-shot learning paradigm. In (Teh & Taylor, 2020), DML is used to pre-train a model in the application of digital pathology classification, where authors use a ProxyNCA loss for learning transferable features. To enhance the embedding, (Yang *et al.*, 2019; Zhong *et al.*, 2021) has integrated a multi-similarity loss to DML in the context of chest radiography and liver histopathology images, respectively. Nevertheless, most of these methods are developed with the goal of classification tasks and do not effectively leverage the geometrical information of the underlying embedding space.

4.2.3 Weakly Supervised Segmentation

Weakly supervised segmentation (WSS) has emerged as an alternative to alleviate the need for large amounts of *pixel-level* labeled data. These labels can come in the form of *image-level* labels (Papandreou *et al.*, 2015), scribbles (Lin *et al.*, 2016), points (Bearman *et al.*, 2016), bounding boxes (Rajchl *et al.*, 2016) or direct losses (Kervadec *et al.*, 2019). Among them, image-level labels are easier and inexpensive to obtain (Bearman *et al.*, 2016). Particularly, class activation maps (CAM) (Zhou *et al.*, 2016) have gained popularity in identifying saliency regions based on image labels. It is achieved by associating feature maps of the last layers and weighting their activation using a global average pooling (GAP) layer. However, generated saliency maps are typically spread around the target object, only focusing on the most discriminant areas. This limits its usability as pixel-level supervision for semantic segmentation. To enhance the generated saliency regions, some alternatives based on back-propagation (GradCAM (Selvaraju *et al.*, 2017)) or super-pixels (SP-CAM (Kwak, Hong, Han *et al.*, 2017)) have been proposed. Nevertheless, these methods demand additional gradient computations (Selvaraju *et al.*, 2017) or supervisions (Kwak *et al.*, 2017).

The literature on WSS in medical imaging with deep learning remains scarce. While few methods resort to direct losses, hence requiring additional priors, such as the target size (Jia *et al.*, 2017; Kervadec *et al.*, 2019), other approaches rely on stronger forms of supervision, for instance, using bounding boxes (Rajchl *et al.*, 2016) or scribbles (Can *et al.*, 2018). Tackling WSS from a perspective of *image-level* labels typically involves visual features, which has not been thoroughly investigated (Dubost *et al.*, 2020; Feng *et al.*, 2017; Meng *et al.*, 2019; Nguyen *et al.*, 2019). For example, Nguyen *et al.* (2019) has proposed a CAM-based approach for the segmentation of uveal melanoma. In their method, the CAMs generated by the classification network are further refined using an active shape model and conditional random fields (Krähenbühl & Koltun, 2011). More recently, CAMs derived from image-level labels have been combined with attention scores to refine lesion segmentation in brain images (Wu *et al.*, 2019). By doing so, they

have demonstrated a performance improvement compared to the vanilla version of CAMs. Nevertheless, these methods typically integrate CAM or GradCAM with complex models to enhance the performance of a final segmentation.

4.3 Methodology

4.3.1 Overview

An overview of the proposed approach is depicted in Fig. 4.1. The main idea is to split the embedding space into multiple subspaces (K) such that the original embedding space can be learned by refining its subspaces. Contrary to (Sanakoyeu *et al.*, 2019), the embedding space is split dynamically, which removes the need to search for the optimal number of learners K in each scenario. The whole process is divided into two iterative steps. First, input images are mapped into the lower dimension embedding space using the entire embedding layer $\mathbf{e}$ of d-dimension, where they are clustered into different groups. Second, the clustered data is consequently assigned to an individual subspace learner, where their corresponding images are used to train each subspace. These two steps are repeated at regular intervals, as well as each time a new learner is added. The key idea is that each subspace learner learns a part of the embedding space from a subgroup of images instead of learning a whole embedded representation vector. Finally, all subspaces are combined to generate a full embedding space. Furthermore, an attention module is integrated into the learning process to guide the learning of distance metrics. The following sections describe the deep metric learning formulation and present the proposed dynamic subspace metric learning and attention module.

4.3.2 Deep Metric learning Formulation

Let the training dataset be defined as $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{Y}^i)\}_{i=1}^{N_I}$, where i-th image is denoted as $\mathbf{X}^i \in \mathbb{R}^{\Omega}$ with spatial domain Ω, and $\mathbf{Y}^i \in \{1, 2, ..., C\}$ is its corresponding class label. C defines the total number of classes. The goal of deep metric learning is to learn an embedding function

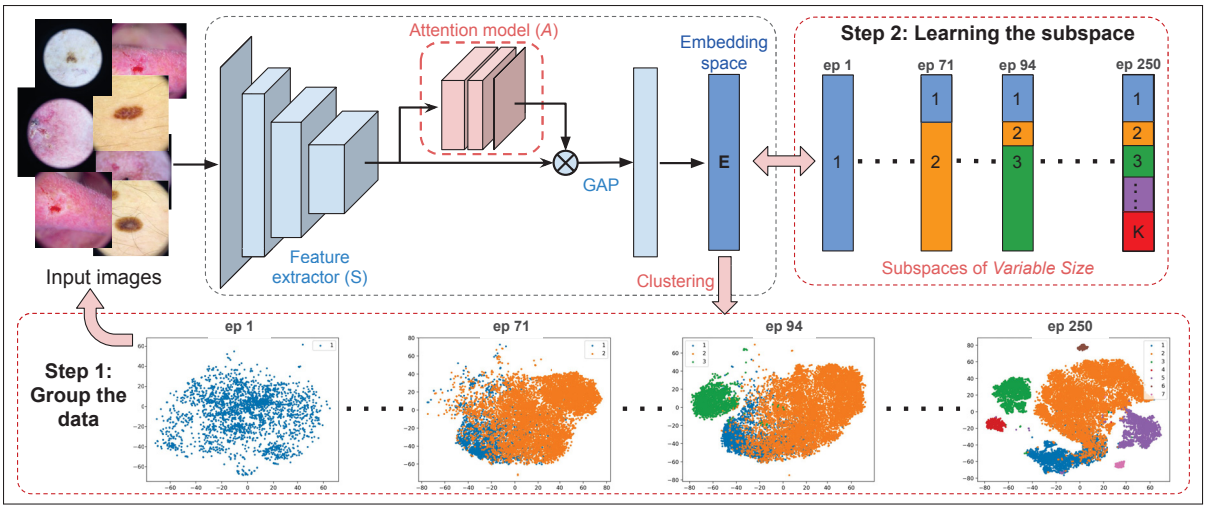

Figure 4.1 Overview of our proposed attention-based dynamic subspace learners. The embedding space is dynamically divided into multiple subspaces (K) of varying sizes during training. *Step 1:* the dataset is first split into K groups (e.g., 3 groups for epoch 94) using full embedding space (**e**) and assigns each data subgroup to an individual subspace learner. *Step 2:* each learner then only attends the data from its subgroup in the learning stage

$f_\theta(\cdot) : \mathbb{R}^\Omega \rightarrow \mathbb{R}^d$, which discriminatively maps semantically similar images (same class) in the input space $\mathbb{R}^\Omega$ onto metrically close points in the learned manifold $\mathbb{R}^d$. Similarly, semantically dissimilar images (different class) in $\mathbb{R}^\Omega$ should be mapped metrically far in $\mathbb{R}^d$. The parameters θ of the mapping function are learned by a convolutional neural network. Formally, the distance metric $d(\mathbf{X}^i, \mathbf{X}^j) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, between two images in the embedding space $\mathbb{R}^d$ can be defined as:

$$d(\mathbf{X}^i, \mathbf{X}^j) = ||f_\theta(\mathbf{X}^i) - f_\theta(\mathbf{X}^j)||, \tag{4.1}$$

where $||\cdot||$ denotes the Euclidean norm. This distance can be minimized in different ways, depending on the loss function employed. In this work, we resort to the Margin loss (Wu *et al.*, 2017):

$$\mathcal{L}_{margin} = \sum_{(\mathbf{X}^i, \mathbf{X}^j) \sim B} [\alpha + \mu_{ij}(d(\mathbf{X}^i, \mathbf{X}^j) - \beta)]_+, \tag{4.2}$$

where β is the boundary between the similar and dissimilar pairs, α is a separation margin, and $\mu_{ij} \in \{-1, 1\}$ indicates whether the images in the pair are similar ($\mu_{ij} = 1$) or different ($\mu_{ij} = -1$). Note that any other metric learning loss function can be employed with our approach.

4.3.3 Dynamic Subspace Learners

The complexity of the original problem can be solved by dividing the problem into smaller sub-problems, which are easier to solve. We follow the approach in (Sanakoyeu *et al.*, 2019), where an embedding space $\mathbb{R}^d$ and dataset is split into multiple groups. Specifically, splitting of the embedding space is conducted by slicing the $\mathbb{R}^d$ space, i.e., the last dense layer of the network, into K sub-vectors of the same size, d/K. Furthermore, data is clustered into K groups based on their pairwise distance in the embedding space $\mathbb{R}^d$, for instance, using K-means. Then, a set of K independent learners is used to learn over each subspace by using a fraction of the input data, thereby reducing the complexity of the original problem. Nevertheless, a major bottleneck is finding an optimal number of subspaces K to learn an effective embedding, which must be found empirically for every new dataset. Moreover, the subspace is divided equally, which is ineffective as not all the object attributes require the same size to encode the information.

Contrary to (Sanakoyeu *et al.*, 2019), our proposed learning strategy finds an optimal embedding by dynamically splitting the embedding space and associating it with a metric learner during training. To construct each subspace, we group highly contributing neurons of the embedding layer **e**, which is repeated until network convergence. Initially, the entire embedding space is learned with all the data, with an initial single learner $K = 1$. As the learning progresses, the accuracy of the model starts to reach an initial plateau. At this stage, we compute the score of each neuron (e_i) in the embedding layer, similarly to the pruning strategy as in (Molchanov, Tyree, Karras, Aila & Kautz, 2017). In particular, the low-scoring neurons are pruned such that the performance drop of the model is minimal, i.e., $|\Delta f_\theta(e_i)| = |f_\theta(\mathcal{D}, e_i = 0) - f_\theta(\mathcal{D}, e_i)|$. By using Taylor expansion, as in (Molchanov *et al.*, 2017), the scoring of each neuron e_i can be

reduced to:

$$s(e_i) = |\Delta f_\theta(e_i)| = \left| f_\theta(\mathcal{D}, e_i) - \frac{\partial f_\theta}{\partial e_i} e_i - f_\theta(\mathcal{D}, e_i) \right| = \left| \frac{\partial f_\theta}{\partial e_i} e_i \right| \tag{4.3}$$

Thus, the scoring of neurons is simplified to multiplying the activation and the gradient output in the embedding layer. This score $s(e_i)$ is computed for each training example separately, and is consequently averaged across all training data and normalized to $[0, 1]$. The neurons having high normalized scores are subsequently grouped to form a new subspace. Particularly, the neurons having more than 50% of the confidence score, i.e., $s(e_i) > 0.5$, are grouped as a new subspace. The current metric learner ($\mathcal{L}_k$) is later assigned to this group of neurons. The remaining neurons of the embedding layer, e_r, are eventually reset, similar to the pruning technique (Molchanov *et al.*, 2017) and assign a new metric learner as in Eq. 4.4. After adding this new learner, the training data is clustered by mapping into the entire embedding space using K-means with the updated K ($K = 2$ for the second iteration). Note that the entire embedding space here is a combination of all the subspaces. Each learner is eventually assigned a subgroup of data from the clustering, resulting in each learner being trained with a fraction of the input data. The addition of a new learner is repeated with the remaining neurons e_r when the network performance reaches a new plateau, until convergence. In the end, it results in K mapping functions, $\mathbf{f} = [f^1, f^2, ..., f^K]$, where each mapping function f^k will project the images $\mathbb{R}^\Omega$ into the corresponding subspace of $\mathbb{R}^{d_k}$, each with a variable size.

All learners are trained jointly by resorting to the margin loss (Wu *et al.*, 2017), which for each learner can be defined as:

$$\mathcal{L}_k^{f_{\theta_k}^k}(\mathbf{X}^i, \mathbf{X}^j) = \sum_{(\mathbf{X}^i, \mathbf{X}^j) \sim B} [\alpha + \mu_{ij}(d_{f_{\theta_k}^k}(\mathbf{X}^i, \mathbf{X}^j) - \beta)]_+, \tag{4.4}$$

where $(\mathbf{X}^i, \mathbf{X}^j) \sim B$ is the current mini-batch (uniformly sampled from each data group) having both positive and negative classes, and $d_{f_{\theta_k}^k}$ is the distance metric (similar to Eq.4.1) for the k-th learner. Once individual learners are trained, these are merged to compose the entire

Algorithm 4.1 Dynamic Subspace Learner Pseudocode

Inputs : $\mathcal{D}$, $\mathcal{D}_{test}$: Training and test data
θ : backbone network parameters
e : Embedding space
T_c, T_p : clustering and network plateau threshold

Initialize : K $\leftarrow$ 1, number of learners
$B \leftarrow 0$, Best epoch
ep $\leftarrow$ 1, current epoch
$e_r \leftarrow$ **e**, remaining embedding space
RC $\leftarrow$ True, re-clustering flag

while *Not converged* **do**
  if *RC* **then** ▷ Re-cluster the data
    e $\leftarrow$ ConcatEmbedding($\{e_1, e_2,...e_{K-1}, e_r\}$)
    emb $\leftarrow$ ComputeEmbedding($\mathcal{D}$, θ, **e**)
    $\{C_1, C_2,...,C_K\} \leftarrow$ ClusterData(emb, K)
    $\{e_1, e_2,...e_{K-1}, e_r\} \leftarrow$ SplitEmbedding(**e**, K)
    RC $\leftarrow$ False
  end
  repeat ▷ Train all learners
    $C_k \sim \{C_1, C_2,...,C_K\}$
    b $\leftarrow$ GetBatch(C_k)
    $L_k \leftarrow$ FPass(b, θ, f^k)
    θ, $f^k \leftarrow$ BPass(L_k, θ, f^k)
  until *epoch completed*
  ep $\leftarrow$ ep + 1
  e $\leftarrow$ ConcatEmbedding($\{e_1, e_2,...e_{K-1}, e_r\}$)
  RC $\leftarrow$ (ep mod T_c == 0)
  if *Evaluate($\mathcal{D}_{test}$, θ,* **e***, ep)* $> B$ **then** ▷ Is better than best
    B $\leftarrow$ ep
  end
  else if $ep \geq (B + T_p)$ **then** ▷ Is network plateaued
    K $\leftarrow$ K + 1 ▷ Update new learner
    $\{e_{K-1}, e_r\} \leftarrow$ splitLearner($\{e_r\}$) ▷ using Eq.4.3
    $\{e_1,..e_{K-1}, e_r\} \leftarrow$ SplitEmbedding(**e**, K, e_{K-1})
    reset(e_r)
    RC $\leftarrow$ True
  end
end
e $\leftarrow$ ConcatEmbedding($\{e_1, e_2,...e_{K-1}, e_r\}$)
θ, **e** $\leftarrow$ FineTune($\mathcal{D}$, θ, **e**)
Output: θ, **e**

embedding space, which is refined with the entire training set. Furthermore, assuming that the learned embedding space is improving over time, we re-cluster the images at every T_c epoch by mapping all the images using the entire embedding space $\mathbf{e}$. An outline of the proposed method is presented in Algorithm 4.1.

4.3.4 Attentive Dynamic Subspace Learners

Deep attention is raising as an efficient mechanism to focus the learning on the objects of interest in a wide range of applications, such as person re-identification (Li, Zhu & Gong, 2018a), object classification (Wang *et al.*, 2017a), or medical image segmentation (Schlemper *et al.*, 2019; Sinha & Dolz, 2020). Inspired by these advances, we introduce an attention module to learn attentive features, with the goal of enhancing the learning of the embedding space. For a given input image $\mathbf{X}^i$, feature extractor $S(\cdot)$ produces a set of feature maps $\mathbf{S}^i = S(\mathbf{X}^i) \in \mathbb{R}^{c \times m \times n}$, where m, n denote the spatial dimension of the feature map and c the number of channels. The attention map produced by the attention module $A(\cdot)$ can be then defined as $\mathbf{A}^i = A(\mathbf{S}^i) \in \mathbb{R}^{m \times n}$. The generated attention map is multiplied with each feature map, $\mathbf{A}^i \odot \mathbf{S}^i$, where $\odot$ is the element-wise product, resulting in the set of attentive features. Last, the attentive features are combined to produce a $c-$dimensional vector by using global average pooling (GAP), which is mapped into the manifold space using a dense layer (Fig. 4.1).

4.3.5 Attention maps for Weakly Supervised Segmentation

The attention maps obtained by our proposed method can serve as proxy *pixel-level* labels to train a segmentation network in a fully-supervised manner. Specifically, the input image $\mathbf{X}^i$ and corresponding attention map $\mathbf{A}^i$ are used as a training pair. To differentiate foreground pixels from the background pixels in $\mathbf{A}^i$, we threshold the attention maps with T_s (i.e., pixels in $\mathbf{A}^i$ greater than T_s are set to 1, 0 otherwise) before training the segmentation network. The network is trained with binary cross-entropy as a loss function, which is computed over pixel-wise

softmax probabilities, defined as:

$$\mathcal{L}_{BCE}(\mathbf{X}, \mathbf{A}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2}\mathbf{A}_c^i \cdot log(F_{\theta_s}(\mathbf{X}_c^i)), \tag{4.5}$$

where F_{θ_s} is a segmentation network parameterized by θ_s. Note that the learning objective that trains a segmentation network is same in both the fully and weakly supervised scenario. However, the main difference lies in the labels employed in the cross-entropy term. In particular, while the former resorts to given segmentation masks, e.g., $\mathbf{Y}^i$, the latter leverages the obtained attention masks as pseudo-labels, i.e., $\mathbf{A}^i$.

4.4 Experiments

4.4.1 Experimental Setting

The performance of the proposed attention-based dynamic subspace learners (ADSL) is compared to other deep metric learning methods applied in medical imaging (Gupta *et al.*, 2019; Pati *et al.*, 2020; Sikaroudi *et al.*, 2020; Teh & Taylor, 2019; Yan *et al.*, 2018), which resort to contrastive or triplet loss. To assess the effectiveness of the dynamic learner training strategy, we compare it with the divide and conquer approach (DCML) (Sanakoyeu *et al.*, 2019). Since we use class labels information, we compare with the classification network trained using a cross-entropy loss. For a fair evaluation, the backbone architecture and hyper-parameters are fixed across the different methods. In addition, experiments across all the models and datasets are run three times, and their average performances are reported. Note that the baselines based on triplet and contrastive loss rely on single-learner, whereas models based on the *divide-and-conquer* strategy and our method employ multiple learners.

To assess the performance of our approach in terms of segmentation, we benchmark the resulting attention maps against the popular GradCAM (Selvaraju *et al.*, 2017) from the classification networks. We include a recent Attention Residual Learning (ARL) approach in (Zhang *et al.*,

2019) since it has been similarly proposed in the context of skin lesion analysis. We also include a recently proposed weakly supervised segmentation method, Embedded Discriminative Attention Mechanism (EDAM) (Wu *et al.*, 2021b), applied for the natural image. Lastly, we include as an upper bound the results obtained by U-Net (Ronneberger *et al.*, 2015) that was trained on the provided pixel-level masks. Note that the model architecture and hyperparameters are fixed across the different methods. Nevertheless, the ARL model employs a carefully modified ResNet50 backbone with soft-attention blocks in each layer. It is noteworthy to mention that it also uses an offline multi-scale patch extraction strategy, resulting in extra images during training. At the same time, the EDAM model employs a collaborative multi-head attention module after the feature extraction layer to directly generate the discriminative activation masks.

4.4.1.1 Datasets

The performance of the proposed method, in terms of clustering and image retrieval, is evaluated on three diverse medical imaging datasets: skin lesion images from the ISIC 2019 Challenge (Codella *et al.*, 2019; Combalia *et al.*, 2019), musculoskeletal radiographs from the MURA dataset (Rajpurkar *et al.*, 2018), and gastrointestinal tract images from the HyperKvasir dataset (Borgli *et al.*, 2020). To assess the segmentation performance, we resort to the skin lesion dataset from the ISIC 2018 Challenge (Codella *et al.*, 2019; Tschandl *et al.*, 2018).

ISIC19

This dataset consists of 25,331 images across eight different categories. In our experiments, following the standard procedure in DML, we split our dataset into independent training and testing sets. Specifically, 20,000 images were used for training and the remaining 5,331 for testing.

MURA

It consists of 40,561 images from 9,045 normal and 5,818 abnormal musculoskeletal radiography studies across seven standard upper extremity types. We configure this as 14 categories (7 normal and 7 abnormal) to represent the data in a manifold. We use the provided split of 36,808 images for training and 3,197 images for testing.

HyperKvasir

This dataset consists of 110,079 images, of which 10,662 images are labeled across 23 different classes of findings. We randomly split the data into 8,567 images for training and the remaining 2,095 images for testing.

ISIC18

This dataset is composed of 2,594 images and their corresponding pixel-level masks. The segmentation dataset is randomly split into three sets: training (1,042), validation (520), and testing (1,038). We leverage the attention maps and GradCAMs generated on the ISIC19 dataset (25,331 images) as proxy labels to train the segmentation networks. In contrast, the training set is used to train the upper-bound model, i.e., fully-supervised.

4.4.1.2 Evaluation

We follow the evaluation protocol typically employed in deep metric learning (Oh Song *et al.*, 2016; Sanakoyeu *et al.*, 2019). In particular, we employ the normalized mutual information (NMI) to assess the clustering performance using K-means and the Recall score (with k = 1 and 4) to evaluate the image retrieval quality. To assess the segmentation performance, we employ the common Dice score coefficient.

4.4.1.3 Implementation details

As in (Sanakoyeu *et al.*, 2019), we use ResNet50 (He *et al.*, 2016) as the backbone architecture. The feature extractor layers consist of the first three residual blocks of ResNet50, used as input to the attention module. The attention module consists of three convolution layers with 3×3 kernel and filters size of $\{128, 32, 1\}$, with a ReLU activation between each convolutional layer. Last, a sigmoid activation is integrated into the final layer to produce the activation map. An input image size of 224×224 is used for all our experiments. All models are trained using the Adam optimizer (Kingma & Ba, 2015) with a batch size of $B = 32$. In each mini-batch, 8 images per class are sampled to ensure a class-balanced scenario, and experiments are trained for 300 epochs. The last 50 epochs are fine-tuned with full embedding. The re-clustering parameter is set to $T_c = 2$ as in (Sanakoyeu *et al.*, 2019), and the network plateau threshold is empirically set to $T_p = 10$. The margin loss parameters are set to $\alpha = 0.2$, $\beta = 1.2$, as in (Wu *et al.*, 2017). Last, since most DML approaches (Sanakoyeu *et al.*, 2019; Wu *et al.*, 2017) employ an embedding space of size $d = 128$, we use the same latent dimension in all our experiments. The PyTorch implementation of our work is publicly available here: https://github.com/adigasu/Dynamic_subspace_learners.

Regarding the segmentation task, we use U-Net (Ronneberger *et al.*, 2015) architecture with an initial kernel size of 32 with two convolution layers and a depth of 3. It is trained with Adam optimizer with batch sizes of 16. For each method, the threshold parameter T_s is set to maximize the Dice score on the initial maps of the validation set (Fig. 4.6).

4.4.2 Clustering and image retrieval results

4.4.2.1 Impact of number of learners K

One of the motivations of this work is to remove the need to empirically search for the optimal number of learners. To validate this hypothesis, we first study the performance of DCML (Sanakoyeu *et al.*, 2019) by varying the number of subspace learners (K). Figure 4.2 depicts

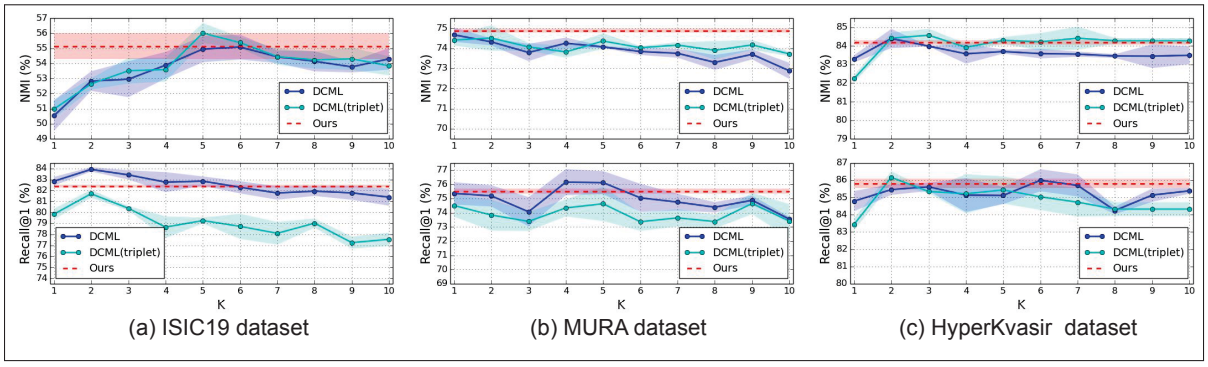

Figure 4.2 Impact of number of learners K in DCML (Sanakoyeu *et al.*, 2019). Each line indicates the NMI (top) and Recall@1 (bottom) scores across the three datasets. The default loss function employed is margin loss, whereas models with a triplet loss are explicitly mentioned. Best seen in color

the results of this experiment across the three datasets and under two different loss functions: margin and triplet loss. In these plots, it can be observed that the optimal K value significantly differs across datasets and metrics. Thus, this limitation of the DCML approach results in extra time-consuming steps to fine-tune the model in each dataset. In contrast, the proposed method (dotted line) eliminates the need to manually define K by dynamically exploring the manifold yet achieves on-par results with the best-performing DCML setting.

We also report the average K values obtained from our method over three runs, as well as the K value of DCML that achieves the best result in Table 4.1. The table shows that the K value has no relation to the number of ground-truth classes. The dynamically obtained K in our method is driven by image content, not by the number of ground-truth classes, which explains their uncorrelated values.

Table 4.1 Comparison of the obtained K value from our method and the DCML best K value with respect to the number of ground-truth classes

Dataset	#classes	ADSL - Avg. K	DCML - Best K
ISIC19	8	7	6
MURA	14	4.67	1
HyperKvasir	23	4.33	2

Table 4.2 Quantitative evaluation on ISIC19 test set. The NMI, Recall, and average scores from the different methods. Our method is emphasized with light gray, whereas the best and second-best results are highlighted in bold and underlined

Method	**NMI (↑)**	**R@1 (↑)**	**R@4 (↑)**	**Avg. of NMI + R@1 (↑)**
Classification network	45.41 ± 1.95	77.85 ± 0.86	90.54 ± 0.51	61.63 ± 1.40
Contrastive loss	31.47 ± 0.39	78.13 ± 0.59	91.13 ± 0.08	54.80 ± 0.49
Triplet loss	50.97 ± 0.61	79.84 ± 0.49	91.70 ± 0.26	65.41 ± 0.55
DCML (worst NMI, K = 1)	50.53 ± 1.01	**82.84 ± 0.39**	91.51 ± 0.43	66.69 ± 0.70
DCML (best NMI, K = 6)	55.08 ± 0.83	82.29 ± 0.56	91.73 ± 0.36	68.69 ± 0.70
ADSL (free from K, ours)	**55.14 ± 0.87**	82.39 ± 0.11	**92.11 ± 0.27**	**68.77 ± 0.49**

Table 4.3 Quantitative evaluation on MURA test set. The NMI, Recall, and average scores from the different methods. Our method is emphasized with light gray, whereas the best and second-best results are highlighted in bold and underlined

Method	**NMI (↑)**	**R@1 (↑)**	**R@4 (↑)**	**Avg. of NMI + R@1 (↑)**
Classification network	71.09 ± 1.25	74.21 ± 0.27	92.59 ± 0.40	72.65 ± 0.76
Contrastive loss	74.28 ± 0.53	71.65 ± 0.53	92.07 ± 0.36	72.97 ± 0.53
Triplet loss	74.41 ± 0.27	74.51 ± 0.78	**92.95 ± 0.33**	74.46 ± 0.53
DCML (worst NMI, K = 10)	72.88 ± 0.40	73.55 ± 0.16	91.17 ± 0.19	73.22 ± 0.28
DCML (best NMI, K = 1)	74.67 ± 0.35	75.36 ± 0.79	92.89 ± 0.18	75.02 ± 0.57
ADSL (free from K, ours)	**74.88 ± 0.09**	**75.52 ± 0.18**	92.25 ± 0.42	**75.20 ± 0.15**

Table 4.4 Quantitative evaluation on HyperKvasir test set. The NMI, Recall, and average scores from the different methods. Our method is emphasized with light gray, whereas the best and second-best results are highlighted in bold and underlined

Method	**NMI (↑)**	**R@1 (↑)**	**R@4 (↑)**	**Avg. of NMI + R@1 (↑)**
Classification network	80.13 ± 2.34	85.66 ± 0.39	**94.42 ± 0.39**	82.90 ± 1.87
Contrastive loss	83.89 ± 0.15	78.52 ± 0.86	93.44 ± 0.48	81.21 ± 0.51
Triplet loss	82.24 ± 0.19	83.44 ± 0.34	93.92 ± 0.22	82.84 ± 0.27
DCML (worst NMI, K = 1)	83.31 ± 0.19	84.79 ± 0.59	94.05 ± 0.26	84.05 ± 0.39
DCML (best NMI, K = 2)	**84.40 ± 0.52**	85.46 ± 0.31	94.19 ± 0.28	84.93 ± 0.42
ADSL (free from K, ours)	84.18 ± 0.12	**85.82 ± 0.27**	94.24 ± 0.41	**85.00 ± 0.20**

4.4.2.2 Comparison to prior literature

We now compare our method with recent prior work as baselines, whose results are reported in Tables 4.2-4.4. As the performance of DCML varies with K, we report only the best and worst models. Note that the DCML with a single-learner, i.e., $K = 1$, is equivalent to a margin

loss method (Wu *et al.*, 2017). We also report the performance of the embedding space learned by the classification network. From the Tables 4.2-4.4, we observe that the proposed method consistently achieves the best results in terms of NMI across the three datasets while performing on par with the best setting of the DCML approach on image retrieval metrics. As shown previously, it is important to note that the performance of DCML heavily depends on the value of K. For instance, the difference between the worst and best DCML configuration in the NMI score can be up to 5% on the ISIC19 dataset. Compared to single-learner approaches, our method brings up to 5 and 2% improvements in NMI and Recall scores on the ISIC19 dataset and up to a 1% improvement in both scores on the MURA and HyperKvasir datasets. This highlights the potential of exploring embeddings via multiple subspaces.

Furthermore, the comparison with the conventional classification network shows that our method consistently outperforms its accuracy by up to 10% in terms of NMI score on ISIC19, and up to 4% NMI score on MURA and HyperKvasir datasets, and up to 4% and 1.5% in terms of Recall scores on the ISIC19 and MURA datasets. The averaged NMI and R@1 results of the proposed method slightly outperform the best DCML configuration, which is consistent across all the datasets. The standard deviation of our method is smaller in all cases for all metrics compared to the DCML. Overall, our method shows a better robustness with compared to the state-of-the-art methods in the learning manifold space. The performance of our method is also in line with the recent literature (Allegretti *et al.*, 2021; Barata & Santiago, 2021).

Ablation study on the use of attention

Adding an attention module brings additional value to our model in terms of interpretability. Nevertheless, to assess whether this improvement is also reflected in the model performance, we compare our model to its non-attention counterpart, denoted as Dynamic Subspace Learners (DSL). Results from this study are reported in Table 4.5, which shows that adding attention typically leads to a boost in the model performance. In particular, the attentive model brings

Table 4.5 Impact of attention module. Per-dataset and average results of the proposed model with (ADSL) and without (DSL) the attention module. Best result is highlighted in bold for each dataset as well as for the average results

Dataset	Method	**NMI (↑)**	**R@1 (↑)**	**R@4 (↑)**
ISIC19	DSL	54.11	**82.74**	91.95
	ADSL	**55.14**	82.39	**92.11**
MURA	DSL	74.21	**75.85**	**92.26**
	ADSL	**74.88**	75.52	92.25
HyperKvasir	DSL	**84.44**	85.36	93.54
	ADSL	84.18	**85.82**	**94.24**
Average	DSL	70.92	**81.32**	92.58
	ADSL	**71.40**	81.24	**92.87**

0.5 and 0.3% improvement on average over the three datasets for the NMI and R@4 metrics, respectively, while achieving on par results for R@1. Additionally, the attention module minimally increases the model memory by 5 MB (includes parameters, forward and backward pass size) when compared to the non-attention counterpart, which is arguably negligible with respect to the overall model size (607 MB) in the case of deployment.

Impact of the embedding size

We also evaluate the effect of representing the embedding space with different sizes. In particular, we assess the clustering and image retrieval performance on the ISIC19 dataset by fixing the embedding dimension size to 64, 128, 256, and 512. Figure 4.3 shows that increasing the embedding size results in a performance improvement, which is reflected in both NMI and recall metrics. Nevertheless, beyond a 256-dimension embedding, the performance of both models typically decreases.

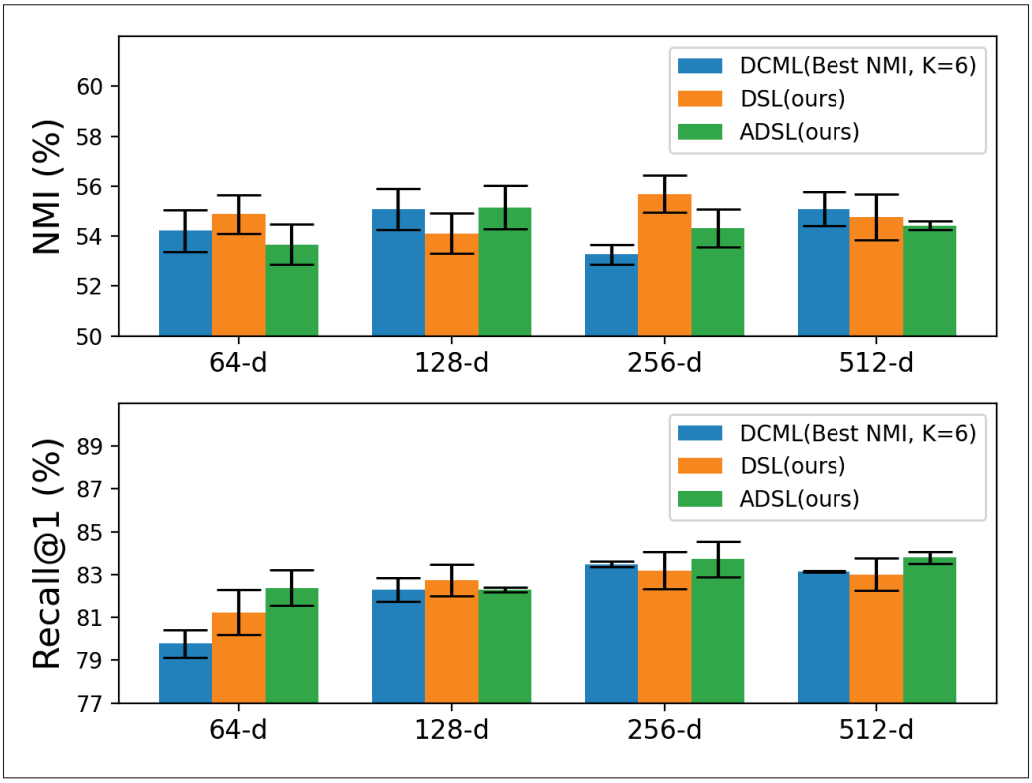

Figure 4.3 Impact of the embedding size. Each bar indicates the NMI (top) and Recall@1 (bottom) scores on the ISIC19 dataset. Compared to the best model of DCML, our method produces better NMI and Recall scores for most cases

Qualitative Analysis

To show the inter and intra-class representation power in the embedding space across different models, we visualize a t-SNE mapping (Maaten & Hinton, 2008) on the ISIC19 test set (Fig. 4.4). The classification network fails to discover clear boundaries across classes in the embedding space (Fig. 4.4b). This could be because the cross-entropy loss when coupled with softmax, does not explicitly guarantee the minimization of intra-class variance or maximization of inter-class variance, which results in suboptimal discriminative features (Liu, Wen, Yu & Yang, 2016). The single metric learner, i.e., DCML $K = 1$ (Fig. 4.4c), improves the class boundaries when compared to the classification network, yet they fail to possess compact clusters. On the other hand, inter-class discrimination is visually enhanced when resorting to multiple learners, i.e.,

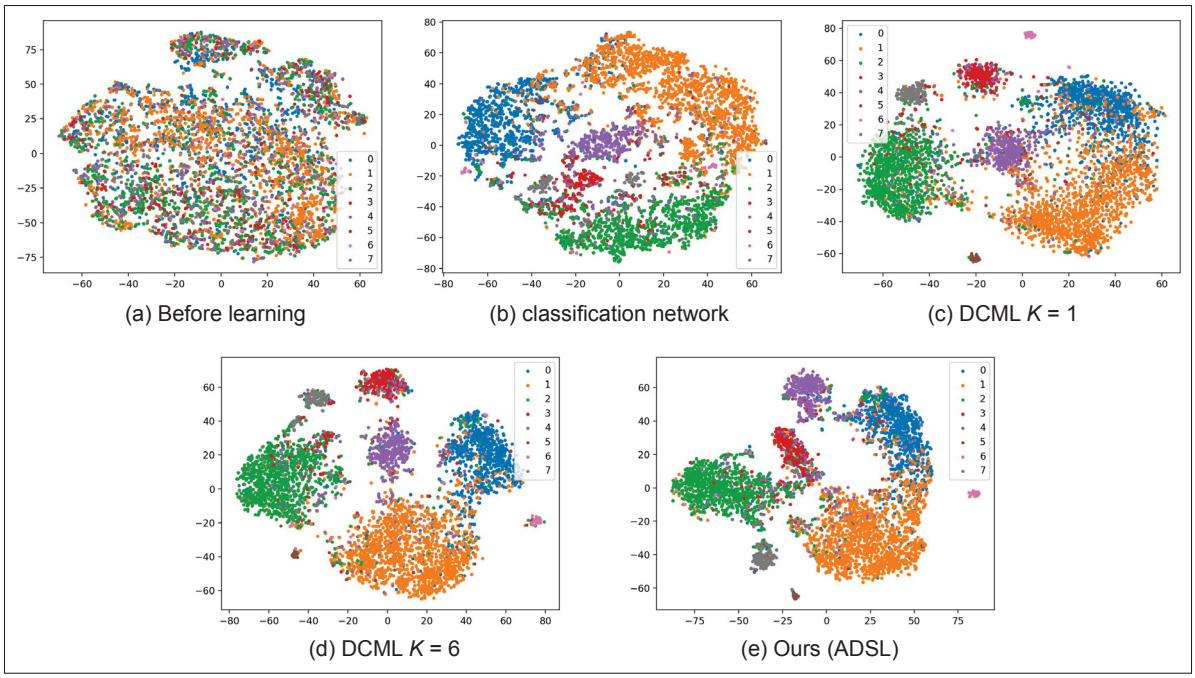

Figure 4.4 Visualization of ISIC19 test set in embedding space using t-SNE. Each class is indicated by its individual color. When compared to a standard classification network, DCML $K = 1$ (a single-learner) improves the separation between classes. The multi-learner methods, DCML $K = 6$, and our method further improve the separation between classes, while our method has the advantage of being free from the number of learners K

DCML $K = 6$ (Fig. 4.4d) and our approach (Fig. 4.4e). Further, we can also observe that the proposed model yields more compact clusters than the DCML approach, which might be due to the freedom of our model to explore the manifold.

Qualitative evaluation in terms of image retrieval is assessed in Fig. 4.5, where a given random query with its five nearest neighbors, found using both DCML and our method, are shown. Additionally, we overlay the contour of our attention maps (having a probability above 0.5) from the proposed method over their respective retrieved image. First, our method indeed retrieves images having similar lesions and colors from the ISIC19 dataset. In radiography wrist images, both DCML and our method have similar retrieval errors. Finally, retrieval images from the HyperKvasir dataset have similar image semantics in terms of texture and probe length using our method when compared to DCML. The coherence of image retrievals indicates that the

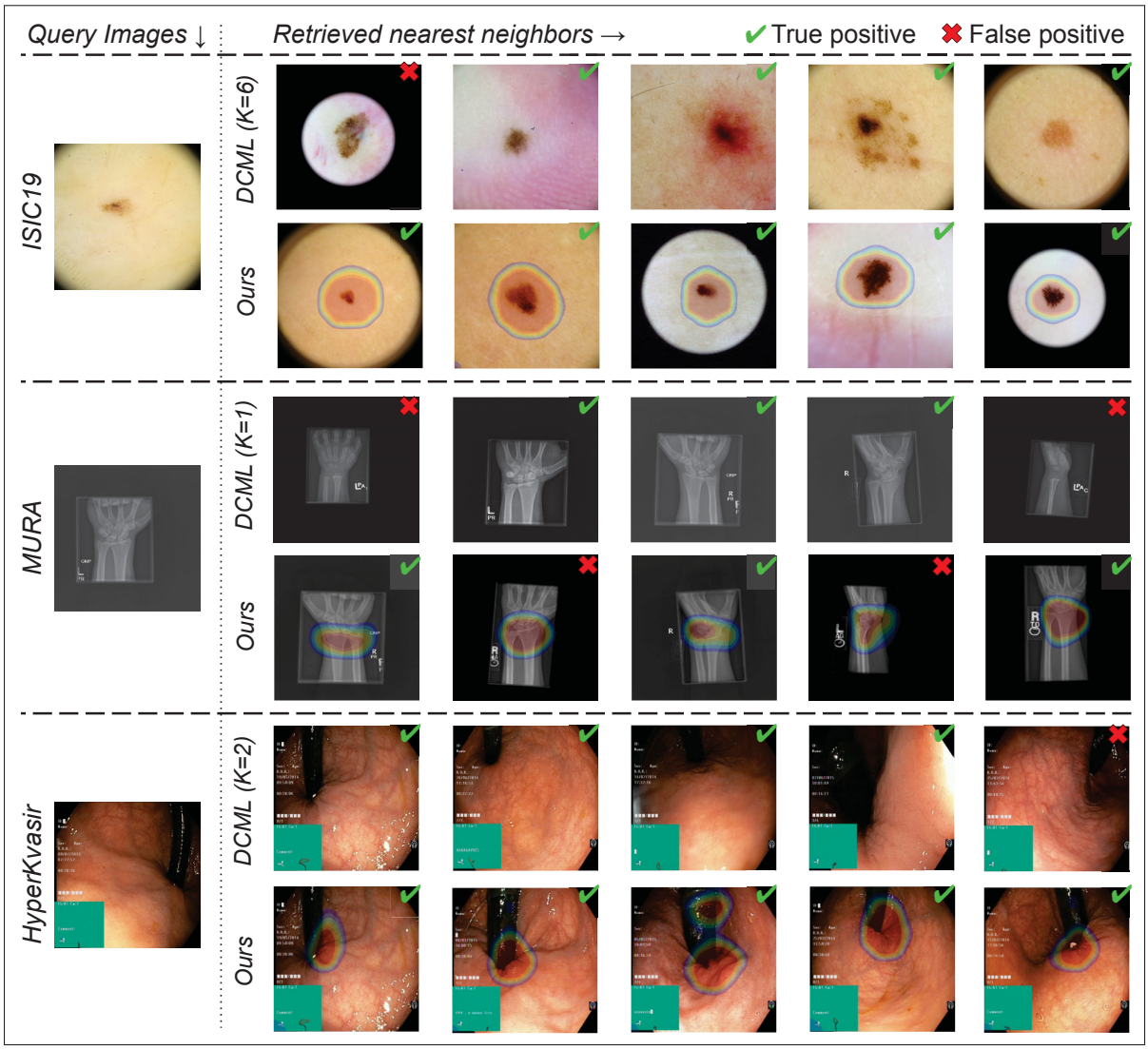

Figure 4.5 Performance of image retrieval on test sets. Each query image and its five nearest neighbors in ascending order of distance are shown (left to right) from the DCML (best K) and our method with an overlay of our attention maps (probability above 0.5)

intra- and inter-class similarities have been captured by our method and thereby demonstrate the robustness of our learned embedding. Moreover, our attention maps mainly concentrate on the lesion in the skin images, the wrist in the radiography images, and the probe contact region in the endoscopic images, demonstrating that our model decisions are consistent over all retrievals.

4.4.3 Weakly Supervised Segmentation results

Table 4.6 reports the results of the segmentation experiments. In this table, *Init maps* are used to denote the raw visual salient regions from either GradCAM or attention maps. *Refined* refers to the performance of the segmentation network trained on the *Init maps*. First, we can observe that segmentation results obtained by raw attention maps and GradCAMs are considerably low, with Dice values around 40%. This is likely due to the well-known fact that both are highly discriminative, resulting in over-segmented regions. The Attention Residual Learning (ARL) significantly outperforms these baselines, whose improvement could be due to the use of attentive residual blocks and additional multiscale data augmentation. The attention maps from the recent Embedded Discriminative Attention Mechanism (EDAM) method perform at a similar level when compared to ARL. Last, the attention maps from the proposed approach bring a significant boost compared to all the other methods. In particular, our model outperforms the baselines by nearly 30% and the recent ARL model by 13%. These results are typically consistent if we employ the initial maps as proxy labels to train a segmentation network. In this case, raw attention maps or GradCAMs barely improve or even decrease the initial segmentation performance. In contrast, ARL, EDAM, and the proposed method reach higher Dice values, with about 1%, 3.5%, and 3% of increase, respectively. This represents a difference of 15% in Dice with respect to ARL. On the other hand, by only using image-level information, the proposed model bridges the gap with a fully-supervised network, with only a 14% difference. This suggests that the proposed model generates reliable segmentations.

4.4.3.1 Ablation study of threshold T_s on the raw visual maps

We evaluate the effect of threshold values T_s on the Dice score for raw visual maps from attention maps and GradCAMs, as shown in Fig 4.6. First, the attention maps and GradCAMs from the classification network have an almost flat Dice score of around 40% until $T_s = 0.4$, succeeded by a gradual decrease. The ARL and EDAM have a gradually increasing Dice score until $T_s = 0.4$

Table 4.6 Performance of weakly supervised segmentation. "Initial maps" and "Refined" are Dice scores (in %) on the ISIC18 test set for different methods. Our method yields the best results. *, †, and $^{\diamond}$ are from ResNet50, ResNet101, and modified ResNet50, respectively, indicating the used architecture in visual map

Method	Init maps	Refined
Attention *	38.45	33.43
Attention †	38.52	38.38
GradCAM *	41.55	40.76
GradCAM †	39.80	41.27
ARL (Zhang *et al.*, 2019) $^{\diamond}$	56.78	57.60
EDAM (Wu *et al.*, 2021b) $^{\diamond}$	51.99	55.50
ADSL (ours) $^{\diamond}$	**69.23**	**72.42**
Full-supervision (upperbound)	-	86.15

and $T_s = 0.6$ with a maximum score of 57.33% and 50.89%, respectively, followed by a gradual decrease. Our method outperforms the baselines for all threshold values in Dice scores with a maximum dice score of 69.0%, showing the robustness of the attention maps derived from our method. This study assists in setting a threshold value T_s for each method before training the segmentation network.

4.4.3.2 Qualitative Performance Evaluation

Visual results of the different methods are shown in Fig. 4.7. In this figure, *Init maps* (rows 1 and 3) are raw visual salient regions from either GradCAM or attention maps shown as heatmaps, whereas *Refined* (rows 2 and 4) refers to the performance of the segmentation network trained using *Init maps* as proxy labels. The attention maps (rows 1 and 3) produced by the classification network spread all over the image, capturing some discriminative regions on the target lesion. GradCAMs spread around the target, highlighting discriminative regions of the lesion but failing to capture the whole context. The saliency map produced by the ARL method is focused on the target lesion. The attention maps obtained by the recent EDAM method spread around the target lesion, including the artifact regions, and fail to capture the target object context. In contrast,

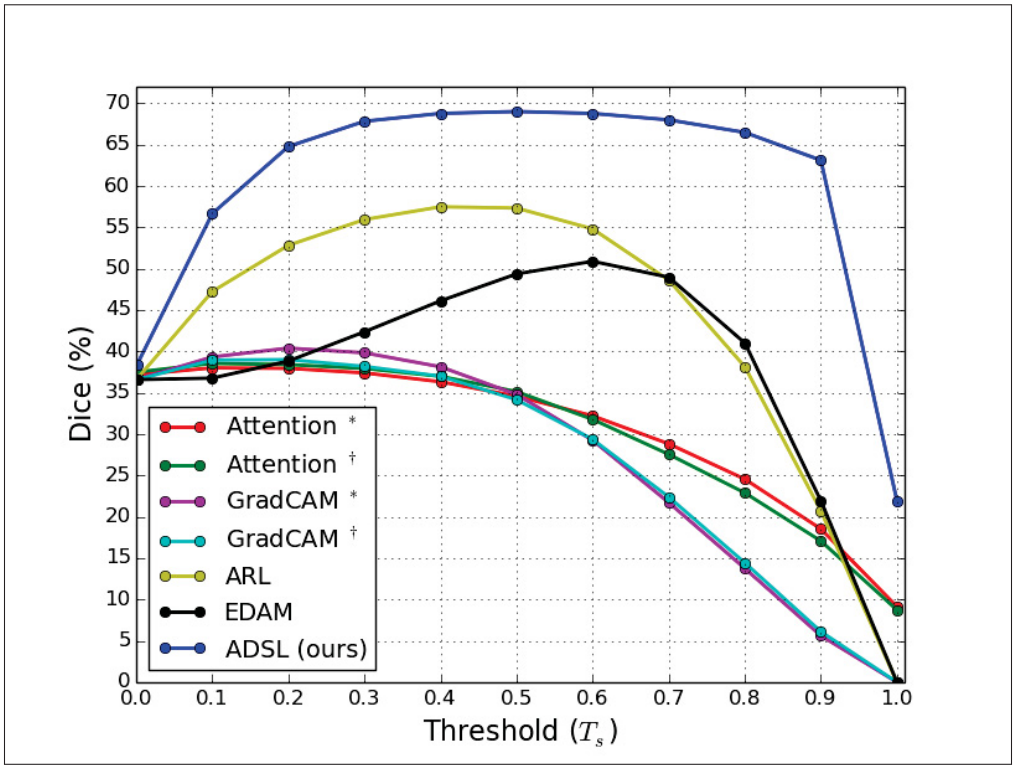

Figure 4.6 Threshold T_s selection. Each line indicates the Dice scores of initial maps on the ISIC18 validation set for different methods. Our method outperforms the baselines for all T_s values. * and † are obtained by classification networks using ResNet50 and ResNet101, respectively

the attention maps derived from our approach better capture the attentive region, which mostly covers the lesion regions. The results show that our proposed approach generates superior attention maps compared to attention maps or GradCAMs from classification networks.

The results obtained by training a segmentation network on the initial salient regions (rows 1 and 3) are depicted in rows 3 and 4. These images demonstrate the feasibility of our method to weakly generate pixel-level labels that are usable for training segmentation networks.

4.5 Discussion and Conclusion

This paper presents a novel attention-based dynamic subspace metric learning approach for medical image analysis. The proposed algorithm leverages recent advances in deep metric

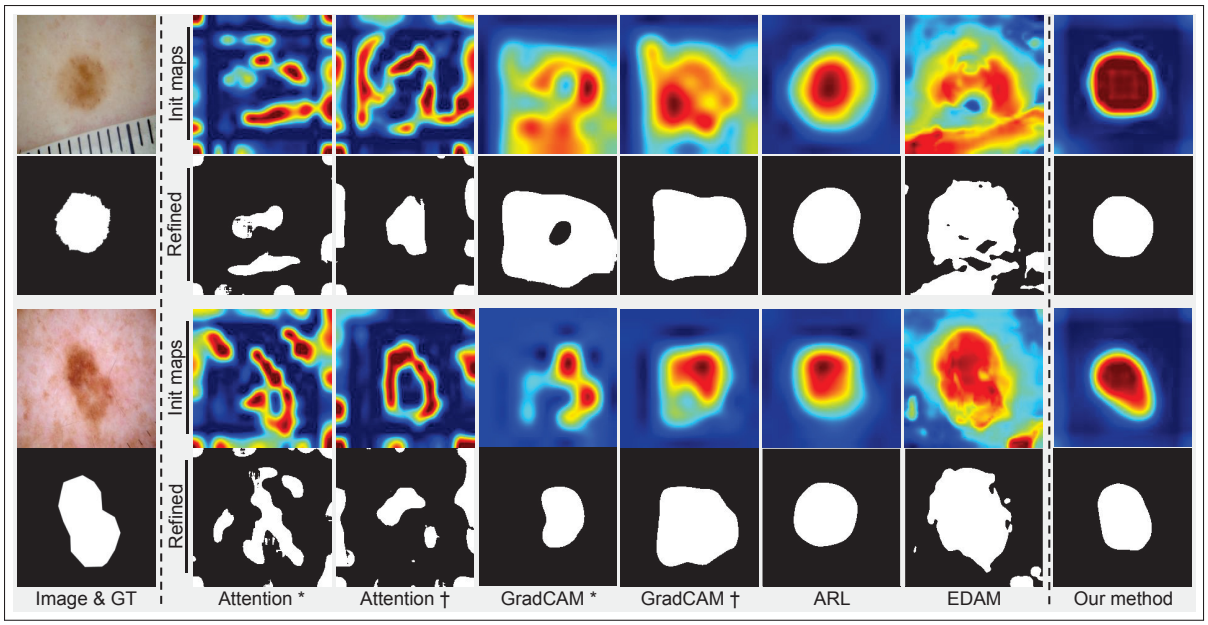

Figure 4.7 Visual results of weakly supervised segmentation. Saliency maps *(Init maps)* obtained by different methods and their segmentation results *(Refined)*. * and † are obtained by GradCAM on classification networks using ResNet50 and ResNet101, respectively

learning using multiple metric learners. Our contribution improves the state-of-the-art method (Sanakoyeu *et al.*, 2019) with dynamic exploitation of subspace learners to learn the embedding space. Specifically, our novel training strategy overcomes the empirical search of the optimal number of subspace learners parameters while achieving competitive results in clustering and image retrieval tasks. Performance is extensively evaluated on three publicly available benchmark datasets: skin lesions, musculoskeletal radiography, and endoscopic images. Results demonstrate that our dynamic learner approach achieves the best results in clustering performance across all three datasets. Compared to the single-learner method, our method brings a maximum of 5 and 2% improvements in clustering and image retrieval scores on the ISIC19 dataset. Furthermore, our method significantly outperforms the classification network in all the datasets with a maximum of 10% and 4% improvements in clustering and retrieval scores on the ISIC19 dataset. Overall, the proposed method slightly outperforms in averaged results and has a smaller standard deviation when compared to the state-of-the-art methods in multiple metric learning. Our experiments have shown consistency across all the datasets, demonstrating the robustness

of our method. Qualitative results show that the proposed method produces compact clustering and coherent image retrievals.

The addition of the attention module to our subspace learners provides the visual interpretability of the learned embedding space in terms of attention maps and improves the clustering metrics. Our method offers new tools in multiple metric learners approaches, notably dynamically learning the number of learners and providing attention maps to hint at salient information caught by the learners. Studying the clinical usability of these tools remains to be explored. Nevertheless, A recent study (Barata & Santiago, 2021) shows that the use of a retrieval network, in a single learner, yields an improvement of 9.2% in the decision accuracy of dermatologists. Our method indeed suggests that multiple learners capture a data embedding that yields a higher accuracy in clustering and retrieval tasks over single-learner methods, while additionally offering visual saliency from our attention mechanism.

The attention maps produced by our proposed method can serve as proxy pixel-level labels to train a segmentation network. The segmentation results outperform a state-of-the-art method, Attention Residual Learning (ARL) (Zhang *et al.*, 2019), as well as the recent Embedded Discriminative Attention Mechanism (EDAM) (Wu *et al.*, 2021b) by a margin of 15% and 17% in Dice scores, respectively, on the skin lesion dataset. The qualitative results demonstrate that the produced attention maps and their segmentation masks focus on the target lesion, demonstrating the effectiveness and robustness of our method. These attention maps produced in our subspace learning approach could therefore be potentially beneficial to a broader range of weakly supervised tasks, where the feature space remains challenging to represent using a single metric model within a specific task.

CONCLUSION AND RECOMMENDATIONS

The introductory chapter of this thesis outlines the general challenge encountered in learning the deep segmentation networks under various levels of supervision and availability of annotated data. These challenges are tackled by exploring a set of cues associated with labels to enhance medical image segmentation. Notably, we have presented three research objectives that leverage uncertainty cues across different levels of annotations, such as with fully, semi, and weakly supervised approaches. This chapter summarizes the contributions of three research objectives, discussing their limitations and potential future recommendations.

5.1 Summary of contributions

1) Intensity-based soft labeling for image segmentation

As a first contribution, **Chapter 2** presented a Geodesic label smoothing (GeoLS) technique that incorporates intensity variation into the label smoothing process. Mainly, it leverages the geodesic distance maps to generate intensity-based soft labels. The resulting soft labels capture the underlying image ambiguities associated with labels. Training a network with our soft labels has shown improved segmentation performance, especially in challenging regions. Results also demonstrated that the proposed approach consistently outperforms the existing soft-labeling approach across three diverse sets of segmentation datasets, including tumors in brain MRIs, multi-organs in abdominal CTs, and multiple zones in prostatic MRIs. The ablation experiments also highlight the effectiveness of integrating the intensity information in our geodesic soft labels rather than solely utilizing the Euclidean distance maps. This work introduces novel soft labels that can be incorporated into any network to enhance the segmentation, particularly impacting applications where labels prove challenging due to ambiguities in image intensities.

2) Anatomically-aware uncertainty for semi-supervision

In **Chapter 3**, a novel anatomically-aware uncertainty estimation approach is proposed for image segmentation under semi-supervised settings. This work utilizes the anatomically-aware representation of segmentation masks to estimate the uncertainty maps. These maps emphasize reliable regions of the predicted masks while regularizing the model. The experiments on left atria and multi-organ abdominal datasets reveal that the proposed approach outperforms current semi-supervised segmentation methods. Our studies also thoroughly evaluate various design choices made in our anatomically-aware representation module. In contrast to existing entropy-based methods, our uncertainty estimation needs a single inference, which minimizes the computation time. The proposed anatomically-aware approach effectively utilizes the limited annotated data with improved segmentation performance, thereby minimizing the need for extensive annotation efforts.

3) Attention-based representation for weak-supervision:

Chapter 4 presents an attention-based dynamic representation learners for image segmentation, retrieval, and clustering tasks. The proposed approach integrates the attention module in the embedding network in order to obtain direct visual maps and dynamically explores the subspace to learn the overarching embedding space. The generated visual maps serve as proxy labels for a segmentation network. Experiments on the skin lesion dataset show that our attention maps yield significantly superior results compared to existing methods relying on attention and class activation maps of classification networks. Moreover, our dynamic subspace learners advance the recent multiple-learner method in clustering and retrieval tasks on skin lesions, musculoskeletal radiography, and endoscopic benchmarks. In addition, our novel dynamic training approach eliminates the need for empirical search in determining an optimal subspace learner parameter. Our representation can be beneficial for the structural organization of the

data while providing interpretability through attention maps. These attention maps are valuable to a wider range of weakly supervised segmentation tasks.

5.2 Limitations and recommendations for future work

This section discusses the main limitations of our work and offers possible avenues for future research.

Limited and interactive supervision: The proposed geodesic soft labeling requires seed points for the geodesic distance transform. These seed points are chosen within segmentation masks so that distance maps capture the similarity intensity values. In our study, different seeding strategies are examined to validate their effectiveness. One can extend our idea of soft labeling to accommodate limited supervision by generating soft labels based on pseudo or proxy labels. To elaborate, we can derive soft labels by selecting seed points within these pseudo or proxy labels. In addition, a recent library supports a PyTorch-based geodesic distance transform generation (Asad, Dorent & Vercauteren, 2022), enabling end-to-end learning with limited supervision. Furthermore, these strategies work with the assumption that the target region is homogeneous. In this case, we can extend our soft labeling strategy to interactive supervision, where the user can provide seed points or scribbles to capture the entire target region (Wang *et al.*, 2018).

Domain adaptation: In this thesis, an anatomically-aware representation is developed for estimating uncertainty maps. Our representation is learned on available segmentation masks in limited data settings. The learned representation subsequently provides reliable target regions through uncertainty estimates in order to regularize the model predictions. This representation is helpful, especially for regularizing the prediction of unlabeled images. One can extend the benefits of our approach to domain adaptation scenarios (Bateson, Dolz, Kervadec, Lombaert & Ayed, 2021; Karani *et al.*, 2021; Perone, Ballester, Barros & Cohen-Adad, 2019). Particularly, we can enhance our anatomically-aware representation by utilizing the existing

labels from the source domain datasets. This enhanced representation is likewise utilized to identify reliable regions of target images. This idea can help during the domain adaptation process, as anatomically-aware representation is independent of imaging data.

Anatomically-aware foundation model: The anatomically-aware representation contributes to uncertainty maps. These uncertainty maps require a single inference, thereby decreasing the computational burden compared to most uncertainty-based approaches. Our representation, however, is learned exclusively on available labels, which can be suboptimal for the clinical use case of uncertainty estimation. We can use the existing labels from the other datasets or modalities to enhance this representation. Another potential direction could be developing a foundation model for anatomical labels. One can train this model on a broad spectrum of publicly available annotations using self-supervised learning strategies. Such a model could better map incorrect predictions into anatomically plausible segmentation, which could be subsequently used as uncertainty estimation. Furthermore, the anatomically-aware foundation model could be used as a post-processing tool with additional constraints to improve the segmentation prediction (Larrazabal *et al.*, 2020; Painchaud *et al.*, 2020). This model could also be combined with image information to learn a joint representation (Judge *et al.*, 2022) to further improve the uncertainty estimation and post-processing of segmentation.

Mixed-supervision: The visual maps using our attention-based representation serve as proxy labels for a weakly supervised segmentation. These proxy labels are assessed for a binary segmentation task, which substantially yields segmentation performance compared to existing weakly supervised methods. Nevertheless, a performance gap prevails compared to fully supervised segmentation methods. One can combine our attention maps with other weak annotations or limited supervision to improve the segmentation. Such a hybrid method is referred to as mixed supervised segmentation (Liu, Desrosiers, Ayed & Dolz, 2023; Papandreou *et al.*, 2015; Wang *et al.*, 2019a). For instance, combining our proxy labels with a small amount of

labeled data could significantly improve the refinement of the segmentation network. Also, it could be combined with other weak supervision, such as bounding boxes, scribbles, or points, to improve the downstream segmentation network.

Model calibration: This thesis presents a set of uncertainty-aware tools in deep segmentation models under various levels of supervision. These tools tackle annotation ambiguities and noisy pseudo labels while training segmentation models. Incorporating these tools has improved the segmentation predictions during the inference. One can also evaluate the calibration (Mehrtash, Wells, Tempany, Abolmaesumi & Kapur, 2020; Murugesan *et al.*, 2023) and uncertainty quantification (Camarasa *et al.*, 2021; Mehta *et al.*, 2022) of model predictions. Such validation can be crucial for the trustworthiness of our uncertainty-aware models. Furthermore, integrating such validated tools in clinical workflows and ensuring compliance with ethical standards will be vital for successful deployment in healthcare settings.

Robustness of foundation models: A recent foundation model has shown significant promise in generalizing across various tasks, including medical image segmentation (Ma *et al.*, 2024; Silva-Rodríguez, Dolz & Ayed, 2023). These models, trained on extensive and diverse datasets, perform reasonably well on out-of-distribution datasets. However, ensuring their robustness across diverse clinical scenarios is essential for their reliable application. Integrating our uncertainty-aware tools into these models could enhance their robustness. Such reliable and generalizable models ultimately improve clinical outcomes and foster trust in artificial intelligence-driven medical imaging solutions.

In conclusion, this thesis introduces new uncertainty-aware tools that improve medical image segmentation under full, semi, and weak supervision. The first objective led to new intensity-based soft labels that enhanced segmentation algorithms in challenging regions. The following research objective led to anatomically-aware uncertainty estimation for effectively utilizing the limited data, thereby minimizing the annotation cost. The final research objective led to

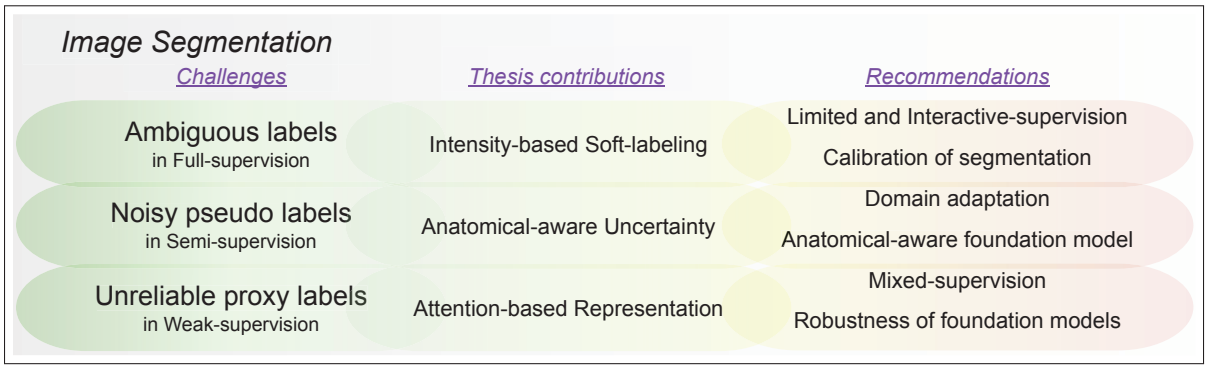

Figure 5.1 Summary of the key contributions and recommendations for future work. **Intensity-based soft labeling** extending to segmentation under limited and interactive supervision. **Anatomically-aware uncertainty estimation** expanding to domain adaptation application and uncertainty quantification from anatomically-aware foundation model. **Attention-based representation** leads to mixed-supervised segmentation. Overall, **uncertainty-aware tools** aid in calibration and foundation model robustness

attention-based dynamic representation that improved segmentation using solely image-level labels. It also enabled the structured organization of the dataset with direct visual interpretability. These new sets of tools, along with future recommendations in this thesis, could assist clinicians in precisely identifying target areas through an automated system under low-data and uncertainty regimes. Such a system may be helpful for the future of artificial intelligence-based interventions, treatments, screening, computer-aided diagnosis, and prognosis.

BIBLIOGRAPHY

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R. et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2274–2282.

Adams, R. & Bischof, L. (1994). Seeded region growing. *Transactions on Pattern Analysis and Machine Intelligence*, 16(6), 641–647.

Adiga, S. (2019). Retinal Image Quality Improvement via Learning. International Institute of Information Technology Hyderabad. [Online; accessed 25-Mar-2024], Retrieved from: https://cvit.iiit.ac.in/research/thesis/thesis-students/retinal-image-quality-improvement-via-learning.

Adiga Vasudeva, S., Dolz, J. & Lombaert, H. (2022a). Attention-based dynamic subspace learners for medical image analysis. *Journal of Biomedical and Health Informatics*, 26(9), 4599–4610.

Adiga Vasudeva, S., Dolz, J. & Lombaert, H. (2022b). Leveraging Labeling Representations in Uncertainty-Based Semi-supervised Segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 265–275.

Adiga Vasudeva, S., Dolz, J. & Lombaert, H. (2022c). Attention-based Dynamic Subspace Learners. *Medical Imaging with Deep Learning*.

Adiga Vasudeva, S., Dolz, J. & Lombaert, H. (2023). GeoLS: Geodesic Label Smoothing for Image Segmentation. *Medical Imaging with Deep Learning*.

Adiga Vasudeva, S., Dolz, J. & Lombaert, H. (2024). Anatomically-aware uncertainty for semi-supervised image segmentation. *Medical Image Analysis*, 91, 103011.

Alexander, A., McGill, M., Tarasova, A., Ferreira, C. & Zurkiya, D. (2019). Scanning the future of medical imaging. *Journal of the American College of Radiology*, 16(4), 501–507.

Allegretti, S., Bolelli, F., Pollastri, F., Longhitano, S., Pellacani, G. & Grana, C. (2021). Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval. *International Conference on Pattern Recognition*, pp. 8053–8060.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M. et al. (2019). BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56, 122–139. Retrieved from: https://iciar2018-challenge.grand-challenge.org/. [Online; accessed 20-Jan-2023].

Asad, M., Dorent, R. & Vercauteren, T. (2022). FastGeodis: Fast Generalised Geodesic Distance Transform. *Journal of Open Source Software*, 7(79), 4532.

Ayache, N. & Duncan, J. (2016). 20th anniversary of the medical image analysis journal (MedIA). *Medical Image Analysis*, 33, 1–3.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.

Bagheri, F., Tarokh, M. J. & Ziaratban, M. (2021). Skin lesion segmentation based on mask RCNN, Multi Atrous Full-CNN, and a geodesic method. *International Journal of Imaging Systems and Technology*, 31(3), 1609–1624.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P. M. & Rueckert, D. (2017). Semi-supervised learning for network-based cardiac MR image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 253–260.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K. & Davatzikos, C. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1), 1–13.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M. et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. *arXiv preprint arXiv:1811.02629*.

Barata, C. & Santiago, C. (2021). Improving the Explainability of Skin Cancer Diagnosis Using CBIR. *Medical Image Computing and Computer-Assisted Intervention*, pp. 550–559.

Bateson, M., Dolz, J., Kervadec, H., Lombaert, H. & Ayed, I. B. (2021). Constrained domain adaptation for image segmentation. *Transactions on Medical Imaging*, 40(7), 1875–1887.

Baum, E. & Wilczek, F. (1987). Supervised learning of probability distributions by neural networks. *Advances in Neural Information Processing Systems*.

Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O. & Konukoglu, E. (2019). PHiSeg: Capturing uncertainty in medical image segmentation. *Medical Image Computing and Computer Assisted Intervention*, pp. 119–127.

Bearman, A., Russakovsky, O., Ferrari, V. & Fei-Fei, L. (2016). What's the point: Semantic segmentation with point supervision. *European Conference on Computer Vision*, pp. 549–565.

Becker, A. S., Chaitanya, K., Schawkat, K., Muehlematter, U. J., Hötker, A. M., Konukoglu, E. & Donati, O. F. (2019). Variability of manual segmentation of the prostate in axial T2-weighted MRI: a multi-reader study. *European Journal of Radiology*, 121, 108716.

Belharbi, S., Rony, J., Dolz, J., Ayed, I. B., McCaffrey, L. & Granger, E. (2021). Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. *Transactions on Medical Imaging*, 41(3), 702–714.

Bengio, Y., Courville, A. & Vincent, P. (2013). Representation learning: A review and new perspectives. *Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828.

Bezdek, J. C., Hall, L. & Clarke, L. (1993). Review of MR image segmentation techniques using pattern recognition. *Medical Physics*, 20(4), 1033–1048.

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D. et al. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1), 1–14.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I. & Bruijne, M. d. (2019). Semi-supervised medical image segmentation via learning consistency under transformations. *Medical Image Computing and Computer-Assisted Intervention*, pp. 810–818.

Boykov, Y., Veksler, O. & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *Transactions on Pattern Analysis and Machine Intelligence*, 23(11), 1222–1239.

Bradley, W. G. (2008). History of medical imaging. *Proceedings of the American Philosophical Society*, 152(3), 349–361.

Braune, W. (1872). Topographisch-anatomischer Atlas: nach Durchschnitten an gefrornen Cadavern. Leipzig: Verlag von Veit & Comp. [Online; accessed 25-Jan-2024], Retrieved from: https://www.nlm.nih.gov/exhibition/historicalanatomies/braune_home.html.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1994). Signature verification using a "Siamese" time delay neural network. *Advances in Neural Information Processing Systems*, pp. 737–744.

Bui, T. D., Wang, L., Chen, J., Lin, W., Li, G. & Shen, D. (2019). Multi-task learning for neonatal brain segmentation using 3D Dense-UNet with dense attention guided by geodesic distance. *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data - Medical Image Computing and Computer Assisted Intervention Workshop*, pp. 243–251.

Burton, R. J., Albur, M., Eberl, M. & Cuff, S. M. (2019). Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC medical informatics and decision making*, 19(1), 1–11.

Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P. J., Kooi, E., van der Lugt, A. & de Bruijne, M. (2021). A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation. *Journal of Machine Learning for Biomedical Imaging*, 13, 1–39.

Can, Y. B., Chaitanya, K., Mustafa, B., Koch, L. M., Konukoglu, E. & Baumgartner, C. F. (2018). Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 236–244). Springer.

Canny, J. (1986). A computational approach to edge detection. *Transactions on Pattern Analysis and Machine Intelligence*, (6), 679–698.

Chaitanya, K., Karani, N., Baumgartner, C. F., Becker, A., Donati, O. & Konukoglu, E. (2019). Semi-supervised and task-driven data augmentation. *International Conference on Information Processing in Medical Imaging*, pp. 29–41.

Chan, T. & Vese, L. (1999). An active contour model without edges. *International Conference on Scale-space Theories in Computer Vision*, pp. 141–151.

Chapelle, O., Scholkopf, B. & Zien, A. (2009). Semi-supervised learning. *Transactions on Neural Networks*, 20(3), 542–542.

Chen, L., Chen, J., Hajimirsadeghi, H. & Mori, G. (2020a). Adapting Grad-CAM for Embedding Networks. *Winter Conference on Applications of Computer Vision*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.

Chen, R., Chen, H., Ren, J., Huang, G. & Zhang, Q. (2019). Explaining neural networks semantically and quantitatively. *International Conference on Computer Vision*, pp. 9187–9196.

Chen, Z., Chen, Z., Liu, J., Zheng, Q., Zhu, Y., Zuo, Y., Wang, Z., Guan, X., Wang, Y. & Li, Y. (2020b). Weakly supervised histopathology image segmentation with sparse point annotations. *Journal of Biomedical and Health Informatics*, 25(5), 1673–1685.

Cheplygina, V., de Bruijne, M. & Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54, 280–296.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*.

Coleman, G. B. & Andrews, H. C. (1979). Image segmentation by clustering. *Proceedings of the IEEE*, 67(5), 773–785.

Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Halpern, A. C., Puig, S. & Malvehy, J. (2019). BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*. Retrieved from: https://api.isic-archive.com/collections/249/. [Online; accessed 23-Jan-2024].

Criminisi, A., Sharp, T. & Blake, A. (2008). GeoS: Geodesic image Segmentation. *European Conference on Computer Vision*, pp. 99–112.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X. & Ye, C. (2019). Semi-supervised brain lesion segmentation with an adapted mean teacher model. *Information Processing in Medical Imaging*, 554–565. Retrieved from: https://link.springer.com/chapter/10.1007/978-3-030-20351-1_43. [Online; accessed 25-Mar-2024].

Dalca, A. V., Guttag, J. & Sabuncu, M. R. (2018). Anatomical priors in convolutional networks for unsupervised biomedical segmentation. *Computer Vision and Pattern Recognition*, pp. 9290–9299.

Dangi, S., Linte, C. A. & Yaniv, Z. (2019). A distance map regularized CNN for cardiac cine MR image segmentation. *Medical Physics*, 46(12), 5637–5651.

Du, Y., Shen, Y., Wang, H., Fei, J., Li, W., Wu, L., Zhao, R., Fu, Z. & Liu, Q. (2022). Learning from future: A novel self-training framework for semantic segmentation. *Advances in Neural Information Processing Systems*, 35, 4749–4761.

Dubost, F., Adams, H., Yilmaz, P., Bortsova, G., van Tulder, G., Ikram, M. A., Niessen, W., Vernooij, M. W. & de Bruijne, M. (2020). Weakly supervised object detection with 2D and 3D regression neural networks. *Medical Image Analysis*, 65, 101767.

Duncan, J. S. & Ayache, N. (2000). Medical image analysis: Progress over two decades and the challenges ahead. *Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 85–106.

Durieux, P., Gevenois, P. A., Muylem, A. V., Howarth, N. & Keyzer, C. (2018). Abdominal attenuation values on virtual and true unenhanced images obtained with third-generation dual-source dual-energy CT. *American Journal of Roentgenology*, 210(5), 1042–1058.

El Gayar, N., Schwenker, F. & Palm, G. (2006). A study of the robustness of KNN classifiers trained using soft labels. *Artificial Neural Networks in Pattern Recognition IAPR Workshop*, pp. 67–80.

Feng, X., Yang, J., Laine, A. F. & Angelini, E. D. (2017). Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. *Medical Image Computing and Computer-Assisted Intervention*, pp. 568–576.

Gaggion, N., Mansilla, L., Mosquera, C., Milone, D. H. & Ferrante, E. (2022). Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest X-ray analysis. *Transactions on Medical Imaging*, 42(2), 546–556.

Gal, Y. & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, pp. 1050–1059.

Galdran, A., Chelbi, J., Kobi, R., Dolz, J., Lombaert, H., Ben Ayed, I. & Chakor, H. (2020). Non-uniform label smoothing for diabetic retinopathy grading from retinal fundus images with deep neural networks. *Translational Vision Science and Technology*, 9(2), 34–34.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT press.

Gould, S., Gao, T. & Koller, D. (2009). Region-based segmentation and object detection. *Advances in Neural Information Processing Systems*, 22.

Grady, L. (2006). Random walks for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 28(11), 1768–1783.

Grandvalet, Y. & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems*, 17.

Gros, C., Lemay, A. & Cohen-Adad, J. (2021). SoftSeg: Advantages of soft versus binary training for image segmentation. *Medical Image Analysis*, 71, 102038.

Gupta, K., Thapar, D., Bhavsar, A. & Sao, A. K. (2019). Deep Metric Learning for Identification of Mitotic Patterns of HEp-2 Cell Images. *Computer Vision and Pattern Recognition Workshops*.

Hadsell, R., Chopra, S. & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *Computer Vision and Pattern Recognition*, 2, 1735–1742.

Hall, L. O., Bensaid, A. M., Clarke, L. P., Velthuizen, R. P., Silbiger, M. S. & Bezdek, J. C. (1992). A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. *Transactions on Neural Networks*, 3(5), 672–682.

Hammoumi, A., Moreaud, M., Ducottet, C. & Desroziers, S. (2021). Adding geodesic information and stochastic patch-wise image prediction for small dataset learning. *Neurocomputing*, 456, 481–491.

Hardy, M. & Harvey, H. (2020). Artificial intelligence in diagnostic imaging: impact on the radiography profession. *The British Journal of Radiology*, 93(1108), 20190840.

Hayward, R. M., Patronas, N., Baker, E. H., Vézina, G., Albert, P. S. & Warren, K. E. (2008). Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. *Journal of Neuro-oncology*, 90, 57–61.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Computer Vision and Pattern Recognition*, pp. 770–778.

He, X., Fang, L., Rabbani, H., Chen, X. & Liu, Z. (2020a). Retinal Optical Coherence Tomography image classification with label smoothing generative adversarial network. *Neurocomputing*, 405, 37–47.

He, X., Zhou, Y., Zhou, Z., Bai, S. & Bai, X. (2018). Triplet-center loss for multi-view 3D object retrieval. *Computer Vision and Pattern Recognition*, pp. 1945–1954.

He, Y., Yang, G., Yang, J., Chen, Y., Kong, Y., Wu, J., Tang, L., Zhu, X., Dillenseger, J.-L., Shao, P. et al. (2020b). Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Medical Image Analysis*, 63, 101722.

Heimann, T. & Meinzer, H.-P. (2009). Statistical shape models for 3D medical image segmentation: a review. *Medical Image Analysis*, 13(4), 543–563.

Hesamian, M. H., Jia, W., He, X. & Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, 32(4), 582–596.

Hsieh, S. S., Cook, D. A., Inoue, A., Gong, H., Sudhir Pillai, P., Johnson, M. P., Leng, S., Yu, L., Fidler, J. L., Holmes III, D. R. et al. (2022). Understanding Reader Variability: A 25-Radiologist Study on Liver Metastasis Detection at CT. *Radiology*, 306(2), e220266.

Hu, B., Vasu, B. & Hoogs, A. (2022). X-MIR: EXplainable Medical Image Retrieval. *Winter Conference on Applications of Computer Vision*, pp. 440–450.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Computer Vision and Pattern Recognition*, pp. 4700–4708.

Huang, H., Zheng, H., Lin, L., Cai, M., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.-W. et al. (2021). Medical image segmentation with deep atlas prior. *Transactions on Medical Imaging*, 40(12), 3519–3530.

Huang, H., Chen, Q., Lin, L., Cai, M., Zhang, Q. W., Iwamoto, Y., Han, X., Furukawa, A., Kanasaki, S., Chen, Y. W. et al. (2022). MTL-ABS3Net: Atlas-Based Semi-Supervised Organ Segmentation Network with Multi-Task Learning for Medical Images. *Journal of Biomedical and Health Informatics*, 26(8), 3988–3998.

Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. (1993). Comparing images using the Hausdorff Distance. *Transactions on Pattern Analysis and Machine Intelligence*, 15(9), 850–863.

Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, pp. 448–456.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211.

Islam, M. & Glocker, B. (2021). Spatially varying label smoothing: Capturing uncertainty from expert annotations. *Information Processing in Medical Imaging*, pp. 677–688.

Islam, M., Seenivasan, L., Ming, L. C. & Ren, H. (2020). Learning and reasoning with the graph structure representation in robotic surgery. *Medical Image Computing and Computer-Assisted Intervention*, pp. 627–636.

Jia, Z., Huang, X., Eric, I., Chang, C. & Xu, Y. (2017). Constrained deep weak supervision for histopathology image segmentation. *Transactions on Medical Imaging*, 36(11), 2376–2388.

Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R. & Jin, C. (2023). Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 107840.

Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. (2019). Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3), 1391–1399.

Judge, T., Bernard, O., Porumb, M., Chartsias, A., Beqiri, A. & Jodoin, P.-M. (2022). CRISP-reliable uncertainty estimation for medical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 492–502.

Karani, N., Erdil, E., Chaitanya, K. & Konukoglu, E. (2021). Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68, 101907.

Karimi, D., Rollins, C. K., Velasco-Annis, C., Ouaalam, A. & Gholipour, A. (2023). Learning to segment fetal brain tissue from noisy annotations. *Medical Image Analysis*, 85, 102731.

Kass, M., Witkin, A. & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1(4), 321–331.

Kats, E., Goldberger, J. & Greenspan, H. (2019). Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation. *International Symposium on Biomedical Imaging*, pp. 1563–1566.

Kaya, M. & Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry*, 11(9), 1066.

Keele, K. D. (1964). Leonardo da Vinci's influence on Renaissance anatomy. *Medical history*, 8(4), 360–370.

Keller, J. M., Gray, M. R. & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *Transactions on Systems, Man, and Cybernetics*, (4), 580–585.

Kendall, A. & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Kendall, A., Badrinarayanan, V. & Cipolla, R. (2017). Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *British Machine Vision Conference*.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ayed, I. B. (2019). Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, 54, 88–99.

Kervadec, H., Dolz, J., Wang, S., Granger, E. & Ayed, I. B. (2020). Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *Medical Imaging with Deep Learning*, pp. 365–381.

Kim, W., Goyal, B., Chawla, K., Lee, J. & Kwon, K. (2018). Attention-based ensemble for deep metric learning. *European Conference on Computer Vision*, pp. 736–751.

Kingma, D. P. & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 5.

Kiyasseh, D., Swiston, A., Chen, R. & Chen, A. (2021). Segmentation of left atrial MR images via self-supervised semi-supervised meta-learning. *Medical Image Computing and Computer Assisted Intervention*, pp. 13–24.

Koh, P. W. & Liang, P. (2017). Understanding black-box predictions via influence functions. *International Conference on Machine Learning*, 70, 1885–1894.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D. & Ronneberger, O. (2018). A probabilistic U-net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems*, 31.

Konstantinidis, K. (2023). The shortage of radiographers: A global crisis in healthcare. *Journal of Medical Imaging and Radiation Sciences*.

Kontschieder, P., Kohli, P., Shotton, J. & Criminisi, A. (2013). GeoF: Geodesic Forests for learning coupled predictors. *Computer Vision and Pattern Recognition*, pp. 65–72.

Krähenbühl, P. & Koltun, V. (2011). Efficient inference in fully connected CRFs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, pp. 109–117.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

Kulis, B. et al. (2012). Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4), 287–364.

Kwak, S., Hong, S., Han, B. et al. (2017). Weakly supervised semantic segmentation using superpixel pooling network. *Association for the Advancement of Artificial Intelligence*, pp. 4111–4117.

Laine, S. & Aila, T. (2017). Temporal Ensembling for Semi-Supervised Learning. *International Conference on Learning Representations*.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Langlotz, C. P. (2019). Will artificial intelligence replace radiologists? Radiological Society of North America.

Larrazabal, A. J., Martínez, C., Glocker, B. & Ferrante, E. (2020). Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders. *Transactions on Medical Imaging*, 39(12), 3813–3820.

Learned-Miller, E. G. (2005). Data driven image models through continuous joint alignment. *Transactions on Pattern Analysis and Machine Intelligence*, 28(2), 236–250.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, 2.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *International Conference on Machine Learning - Workshop on challenges in representation learning*, 3(2), 896.

Lee, H. & Jeong, W.-K. (2020). Scribble2Label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. *Medical Image Computing and Computer Assisted Intervention*, pp. 14–23.

Li, S., Zhang, C. & He, X. (2020a). Shape-aware semi-supervised 3D semantic segmentation for medical images. *Medical Image Computing and Computer-Assisted Intervention*, pp. 552–561.

Li, W., Zhu, X. & Gong, S. (2018a). Harmonious attention network for person re-identification. *Computer Vision and Pattern Recognition*, pp. 2285–2294.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L. & Heng, P.-A. (2020b). Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *Transactions on Neural Networks and Learning Systems*, 32(2), 523–534.

Li, Z., Zhang, X., Müller, H. & Zhang, S. (2018b). Large-scale retrieval for medical image analytics: A comprehensive review. *Medical Image Analysis*, 43, 66–84.

Liao, S., Hu, Y., Zhu, X. & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. *Computer Vision and Pattern Recognition*, pp. 2197–2206.

Liao, S., Gao, Y., Oto, A. & Shen, D. (2013). Representation learning: a unified deep learning framework for automatic prostate MR segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 254–261.

Lin, D., Dai, J., Jia, J., He, K. & Sun, J. (2016). ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. *Computer Vision and Pattern Recognition*, pp. 3159–3167.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017). Focal loss for dense object detection. *International Conference on Computer Vision*, pp. 2980–2988.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N. & Huisman, H. (2014). Computer-aided detection of prostate cancer in MRI. *Transactions on Medical Imaging*, 33(5), 1083–1092.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

Liu, B., Desrosiers, C., Ayed, I. B. & Dolz, J. (2023). Segmentation with mixed supervision: Confidence maximization helps knowledge distillation. *Medical Image Analysis*, 83, 102670.

Liu, W., Wen, Y., Yu, Z. & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. *International Conference on Machine Learning*, 2(3), 7.

Lombaert, H., Zikic, D., Criminisi, A. & Nicholas, A. (2014). Laplacian forests: semantic image segmentation by guided bagging. *Medical Image Computing and Computer Assisted Intervention*, pp. 496–504.

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Computer Vision and Pattern Recognition*, pp. 3431–3440.

Loshchilov, I. & Hutter, F. (2017). SGDR: Stochastic Gradient Descent with warm Restarts. *International Conference on Learning Representations*.

Lourenço-Silva, J. & Oliveira, A. L. (2021). Using soft labels to model uncertainty in medical image segmentation. *Medical Image Computing and Computer Assisted Intervention - Brainlesion Workshop*, pp. 585–596.

Lukasik, M., Bhojanapalli, S., Menon, A. & Kumar, S. (2020). Does label smoothing mitigate label noise? *International Conference on Machine Learning*, pp. 6448–6458.

Lukov, T., Zhao, N., Lee, G. H. & Lim, S.-N. (2022). Teaching with soft label smoothing for mitigating noisy labels in facial expressions. *European Conference on Computer Vision*, pp. 648–665.

Luo, X., Chen, J., Song, T. & Wang, G. (2021). Semi-supervised Medical Image Segmentation through Dual-task Consistency. *AAAI Conference on Artificial Intelligence*, 35(10), 8801–8809.

Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D. N. & Zhang, S. (2022). Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80, 102517.

Ma, J., Zhou, C., Cui, P., Yang, H. & Zhu, W. (2019). Learning disentangled representations for recommendation. *Advances in Neural Information Processing Systems*.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X. & Martel, A. L. (2021a). Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71, 102035.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X. et al. (2021b). AbdomenCT-1K: Is abdominal organ segmentation a solved problem. *Transactions on Pattern Analysis and Machine Intelligence*. Retrieved from: https://flare.grand-challenge.org/. [Online; accessed 18-Jan-2024].

Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y. et al. (2022). Fast and Low-GPU-memory Abdomen CT organ sEgmentation: The FLARE challenge. *Medical Image Analysis*, 82, 102616.

Ma, J., He, Y., Li, F., Han, L., You, C. & Wang, B. (2024). Segment Anything in Medical Images. *Nature Communications*, 15, 1–9.

Maaten, L. v. d. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.

Mahesh, M., Ansari, A. J. & Mettler Jr, F. A. (2022). Patient Exposure from Radiologic and Nuclear Medicine Procedures in the United States and Worldwide: 2009–2018. *Radiology*, 307(1), e221263.

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P. & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *Transactions on Medical Imaging*, 39(12), 3868–3878.

Mehta, R. & Arbel, T. (2018). 3D U-Net for brain tumour segmentation. *Medical Image Computing and Computer Assisted Intervention - Brainlesion Workshop*, pp. 254–266.

Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Murugesan, G. K. et al. (2022). QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation-Analysis of Ranking Scores and Benchmarking Results. *Journal of Machine Learning for Biomedical Imaging*, 1.

Meng, Q., Sinclair, M., Zimmer, V., Hou, B., Rajchl, M., Toussaint, N., Oktay, O., Schlemper, J., Gomez, A., Housden, J. et al. (2019). Weakly supervised estimation of shadow confidence maps in fetal ultrasound imaging. *Transactions on Medical Imaging*, 38(12), 2755–2767.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *Transactions on Medical Imaging*, 34(10), 1993–2024.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *International Conference on 3D Vision*, pp. 565–571.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3523–3542.

Molchanov, P., Tyree, S., Karras, T., Aila, T. & Kautz, J. (2017). Pruning convolutional neural networks for resource efficient inference. *International Conference on Learning Representations*.

Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M. & Glocker, B. (2020). Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33, 12756–12767.

Movshovitz, A. Y., Toshev, A., Leung, T. K., Ioffe, S. & Singh, S. (2017). No fuss distance metric learning using proxies. *International Conference on Computer Vision*, pp. 360–368.

Müller, R., Kornblith, S. & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, 32.

Murugesan, B., Adiga Vasudeva, S., Liu, B., Lombaert, H., Ben Ayed, I. & Dolz, J. (2023). Trust your neighbours: Penalty-based constraints for model calibration. *Medical Image Computing and Computer-Assisted Intervention*, pp. 572–581.

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. *International Conference on Machine Learning*, pp. 807–814.

Neal, R. M. (2012). *Bayesian learning for neural networks*. Springer Science and Business Media.

Nguyen, H.-G., Pica, A., Hrbacek, J., Weber, D. C., La Rosa, F., Schalenbourg, A., Sznitman, R. & Cuadra, M. B. (2019). A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps. *Medical Imaging with Deep Learning*, 370–379. Retrieved from: https://proceedings.mlr.press/v102/nguyen19a.html. [Online; accessed 25-Mar-2024].

Nie, D., Gao, Y., Wang, L. & Shen, D. (2018). ASDNet: attention based semi-supervised deep networks for medical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 370–378.

Nock, R. & Nielsen, F. (2004). Statistical region merging. *Transactions on Pattern Analysis and Machine Intelligence*, 26(11), 1452–1458.

Nosrati, M. S. & Hamarneh, G. (2016). Incorporating prior knowledge in medical image segmentation: a survey. *arXiv preprint arXiv:1607.01092*.

Oakden-Rayner, L. (2020). Exploring Large-scale Public Medical Image Datasets. *Academic radiology*, 27(1), 106–112.

Oh Song, H., Xiang, Y., Jegelka, S. & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. *Computer Vision and Pattern Recognition*, pp. 4004–4012.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S. A., De Marvao, A., Dawes, T., O'Regan, D. P. et al. (2017). Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *Transactions on Medical Imaging*, 37(2), 384–395.

Opitz, M., Waltner, G., Possegger, H. & Bischof, H. (2017). BIER-boosting independent embeddings robustly. *International Conference on Computer Vision*, pp. 5189–5198.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A. & Jodoin, P.-M. (2020). Cardiac segmentation with strong anatomical guarantees. *Transactions on Medical Imaging*, 39(11), 3703–3713.

Papandreou, G., Chen, L.-C., Murphy, K. & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. *International Conference on Computer Vision*.

Paszke, A., Gross, S., Massa, F., Lerer, Chintala, S. et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Patel, G. & Dolz, J. (2022). Weakly supervised segmentation with cross-modality equivariant constraints. *Medical Image Analysis*, 77, 102374.

Pati, P., Foncubierta-Rodríguez, A., Goksel, O. & Gabrani, M. (2020). Reducing annotation effort in digital pathology: A Co-Representation learning framework for classification tasks. *Medical Image Analysis*, 101859.

Peng, B., Zhang, L. & Zhang, D. (2013). A survey of graph theoretical approaches to image segmentation. *Pattern recognition*, 46(3), 1020–1038.

Peng, J. & Wang, Y. (2021). Medical image segmentation with limited supervision: a review of deep network models. *IEEE Access*, 9, 36827–36851.

Peng, J., Estrada, G., Pedersoli, M. & Desrosiers, C. (2020). Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107, 107269.

Perone, C. S., Ballester, P., Barros, R. C. & Cohen-Adad, J. (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194, 1–11.

Pham, D. L., Xu, C. & Prince, J. L. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1), 315–337.

Prince, S. J. (2023). *Understanding Deep Learning*. MIT press.

Protiere, A. & Sapiro, G. (2007). Interactive image segmentation via adaptive weighted distances. *Transactions on Image Processing*, 16(4), 1046–1057.

Qiu, W., Yuan, J., Rajchl, M., Kishimoto, J., Chen, Y., de Ribaupierre, S., Chiu, B. & Fenster, A. (2015). 3D MR ventricle segmentation in pre-term infants with post-hemorrhagic ventricle dilatation (PHVD) using multi-phase geodesic level-sets. *NeuroImage*, 118, 13–25.

Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G. M., De, S., Zhang, S. & Metaxas, D. N. (2020). Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *Transactions on Medical Imaging*, 39(11), 3655–3666.

Rajchl, M., Lee, M. C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M. A., Hajnal, J. V., Kainz, B. et al. (2016). DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. *Transactions on Medical Imaging*, 36(2), 674–683.

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R. L. et al. (2018). MURA: Large dataset for abnormality detection in musculoskeletal radiographs. *Medical Imaging with Deep Learning*.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P. & Vaidya, V. (2017). Learning and incorporating shape models for semantic segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 203–211.

Richards, M., Maskell, G., Halliday, K. & Allen, M. (2022). Diagnostics: a major priority for the NHS. *Future Healthcare Journal*, 9(2), 133.

Roentgen, W. C. (1931). On a new kind of rays. *The British Journal of Radiology*, 4(37), 32–33.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 234–241. Retrieved from: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28. [Online; accessed 19-Feb-2024].

Rubinstein, R. Y. & Kroese, D. P. (2004). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*. Springer.

Sabuncu, M. R., Yeo, B. T., Van Leemput, K., Fischl, B. & Golland, P. (2010). A generative model for image segmentation based on label fusion. *Transactions on Medical Imaging*, 29(10), 1714–1729.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. & Chen, X. (2016). Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29.

Sanakoyeu, A., Tschernezki, V., Buchler, U. & Ommer, B. (2019). Divide and conquer the embedding space for metric learning. *Computer Vision and Pattern Recognition*, pp. 471–480.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B. & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53, 197–207.

Schroff, F., Criminisi, A. & Zisserman, A. (2008). Object Class Segmentation using Random Forests. *British Machine Vision Conference*, pp. 1–10.

Schroff, F., Kalenichenko, D. & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *Computer Vision and Pattern Recognition*, pp. 815–823.

Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W. & Garnavi, R. (2019). Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. *Medical Image Computing and Computer-Assisted Intervention*, pp. 282–290.

Seibold, C. M., Reiß, S., Kleesiek, J. & Stiefelhagen, R. (2022). Reference-guided pseudo-label generation for medical semantic segmentation. *Association for the Advancement of Artificial Intelligence Conference*, 36(2), 2171–2179.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Conference on Computer Vision*, pp. 618–626.

Seo, S., Bode, M. & Obermayer, K. (2003). Soft nearest prototype classification. *Transactions on Neural Networks*, 14(2), 390–398.

Sezgin, M. & Sankur, B. l. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1), 146–168.

Shen, W., Peng, Z., Wang, X., Wang, H., Cen, J., Jiang, D., Xie, L., Yang, X. & Tian, Q. (2023). A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *Transactions on Pattern Analysis and Machine Intelligence*.

Shen, X., Spann, M. & Nacken, P. (1998). Segmentation of 2D and 3D images through a hierarchical clustering based on region modelling. *Pattern Recognition*, 31(9), 1295–1309.

Shi, F., Hu, W., Wu, J., Han, M., Wang, J., Zhang, W., Zhou, Q., Zhou, J., Wei, Y., Shao, Y. et al. (2022). Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. *Nature Communications*, 13(1), 6566.

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905.

Shotton, J., Johnson, M. & Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. *Computer Vision and Pattern Recognition*, pp. 1–8.

Sikaroudi, M., Safarpoor, A., Ghojogh, B., Shafiei, S., Crowley, M. & Tizhoosh, H. R. (2020). Supervision and source domain impact on representation learning: A histopathology case study. *Engineering in Medicine and Biology Society*, pp. 1400–1403.

Silva-Rodríguez, J., Dolz, J. & Ayed, I. B. (2023). Towards foundation models and few-shot parameter-efficient fine-tuning for volumetric organ segmentation. *Medical Image Computing and Computer-Assisted Intervention Workshop on Foundation Models*, pp. 213–224.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Sinha, A. & Dolz, J. (2020). Multi-scale self-guided attention for medical image segmentation. *Journal of Biomedical and Health Informatics*, 25(1), 121–130.

Smith-Bindman, R., Kwan, M. L., Marlow, E. C., Theis, M. K., Bolch, W., Cheng, S. Y., Bowles, E. J., Duncan, J. R., Greenlee, R. T., Kushi, L. H. et al. (2019). Trends in use of medical imaging in US health care systems and in Ontario, Canada, 2000-2016. *Jama*, 322(9), 843–856.

Snell, J., Swersky, K. & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, pp. 4077–4087.

Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. *Advances in Neural Information Processing Systems*, pp. 1857–1865.

Sokolovskaya, E., Shinde, T., Ruchman, R. B., Kwak, A. J., Lu, S., Shariff, Y. K., Wiggins, E. F. & Talangbayan, L. (2015). The effect of faster reporting speed for imaging studies on the number of misses and interpretation errors: a pilot study. *Journal of the American College of Radiology*, 12(7), 683–688.

Strauss, H. W., Zaret, B. L., Hurley, P. J., Natarajan, T. & Pitt, B. (1971). A scintiphotographic method for measuring left ventricular ejection fraction in man without cardiac catheterization. *The American Journal of Cardiology*, 28(5), 575–580.

Stylianou, A., Souvenir, R. & Pless, R. (2019). Visualizing deep similarity networks. *Winter Conference on Applications of Computer Vision*, pp. 2029–2037.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 240–248). Springer.

Suetens, P. (2017). Fundamentals of medical imaging. Cambridge University Press. [Online; accessed 12-Jan-2024], Retrieved from: https://www.cambridge.org/core/books/fundamentals-of-medical-imaging/E9D727DBE7EB6150768A74F655C07BAC.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Computer Vision and Pattern Recognition*, pp. 2818–2826.

Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Association for the Advancement of Artificial Intelligence*, 31(1).

Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D. & Hamarneh, G. (2019). Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75, 24–33.

Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J. & Wang, Y. (2022). Unified medical image segmentation by learning from uncertainty in an end-to-end manner. *Knowledge-Based Systems*, 241, 108215.

Tarvainen, A. & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30. Retrieved from: https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf. [Online; accessed 25-Mar-2024].

Taylor, R. H. (1996). *Computer-integrated surgery: technology and clinical applications*. MIT Press.

Teh, E. W. & Taylor, G. W. (2019). Metric learning for patch classification in digital pathology.

Teh, E. W. & Taylor, G. W. (2020). Learning with less data via weakly labeled patch classification in digital pathology. *International Symposium on Biomedical Imaging*, pp. 471–475.

Tian, K., Zhang, J., Shen, H., Yan, K., Dong, P., Yao, J., Che, S., Luo, P. & Han, X. (2020). Weakly-supervised nucleus segmentation based on point annotations: A coarse-to-fine self-stimulated learning strategy. *Medical Image Computing and Computer Assisted Intervention*, pp. 299–308.

Toivanen, P. J. (1996). New geodesic distance transforms for gray-scale images. *Pattern Recognition Letters*, 17(5), 437–450.

Tschandl, P., Rosendahl, C. & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E. & Yu, T. (2014). Scikit-image: image processing in Python. *PeerJ*, 2, e453.

Van Engelen, J. E. & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2), 373–440.

Van Ginneken, B., Schaefer-Prokop, C. M. & Prokop, M. (2011). Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3), 719–732.

Van Leemput, K., Maes, F., Vandermeulen, D. & Suetens, P. (1999). Automated model-based tissue classification of MR images of the brain. *Transactions on Medical Imaging*, 18(10), 897–908.

Van Waerebeke, M., Lodygensky, G. & Dolz, J. (2022). On the pitfalls of entropy-based uncertainty for multi-class semi-supervised segmentation. *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, 36–46.

Vesalius, A. (1543). *De humani corporis fabrica libri septem*.

Vincent, L. & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *Transactions on Pattern Analysis and Machine Intelligence*, 13(06), 583–598.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A. & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

Vu, T.-H., Jain, H., Bucher, M., Cord, M. & Pérez, P. (2019). ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. *Computer Vision and Pattern Recognition*, pp. 2517–2526.

Wang, D., Li, M., Ben-Shlomo, N., Corrales, C. E., Cheng, Y., Zhang, T. & Jayender, J. (2019a). Mixed-supervised dual-network for medical image segmentation. *Medical Image Computing and Computer Assisted Intervention*, pp. 192–200.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. & Tang, X. (2017a). Residual attention network for image classification. *Computer Vision and Pattern Recognition*, pp. 3156–3164.

Wang, G., Zuluaga, M. A., Li, W., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S. et al. (2018). DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 41(7), 1559–1572.

Wang, J., Zhou, F., Wen, S., Liu, X. & Lin, Y. (2017b). Deep metric learning with angular loss. *International Conference on Computer Vision*, pp. 2593–2601.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. & Wu, Y. (2014a). Learning fine-grained image similarity with deep ranking. *Computer Vision and Pattern Recognition*, pp. 1386–1393.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L. & Wang, Y. (2021). Tripled-Uncertainty Guided Mean Teacher Model for Semi-supervised Medical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 450–460.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L. & Wang, Y. (2022a). Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79, 102447.

Wang, L., Ye, X., Ju, L., He, W., Zhang, D., Wang, X., Huang, Y., Feng, W., Song, K. & Ge, Z. (2023). Medical matting: Medical image segmentation with uncertainty from the matting perspective. *Computers in Biology and Medicine*, 158, 106714.

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H. & Nandi, A. K. (2022b). Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5), 1243–1267.

Wang, X.-Y., Wang, T. & Bu, J. (2011). Color image segmentation using pixel wise support vector machine classification. *Pattern Recognition*, 44(4), 777–787.

Wang, X., Han, X., Huang, W., Dong, D. & Scott, M. R. (2019b). Multi-similarity loss with general pair weighting for deep metric learning. *Computer Vision and Pattern Recognition*, pp. 5022–5030.

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y. & He, Z. (2020). Double-uncertainty weighted method for semi-supervised learning. *Medical Image Computing and Computer-Assisted Intervention*, pp. 542–551.

Wang, Z., Bhatia, K. K., Glocker, B., Marvao, A., Dawes, T., Misawa, K., Mori, K. & Rueckert, D. (2014b). Geodesic patch-based segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 666–673.

Wei, J., Wu, Z., Wang, L., Bui, T. D., Qu, L., Yap, P.-T., Xia, Y., Li, G. & Shen, D. (2022). A cascaded nested network for 3T brain MR image segmentation guided by 7T labeling. *Pattern Recognition*, 124, 108420.

Weinberger, K. Q., Blitzer, J. & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, pp. 1473–1480.

Wikipedia contributors. (2024a). Leonardo da Vinci — Wikipedia, The Free Encyclopedia. [Online; accessed 25-Jan-2024], Retrieved from: https://en.wikipedia.org/wiki/Leonardo_da_Vinci.

Wikipedia contributors. (2024b). X-ray — Wikipedia, The Free Encyclopedia. [Online; accessed 25-Jan-2024], Retrieved from: https://en.wikipedia.org/wiki/X-ray.

Winder, M., Owczarek, A. J., Chudek, J., Pilch-Kowalczyk, J. & Baron, J. (2021). Are we overdoing it? Changes in diagnostic imaging workload during the years 2010–2020 including the impact of the SARS-CoV-2 pandemic. *Healthcare*, 9(11), 1557.

Wolbarst, A. B., Capasso, P. & Wyant, A. R. (2013). *Medical Imaging: Essentials for Physicians*. John Wiley and Sons.

Wu, C.-Y., Manmatha, R., Smola, A. J. & Krahenbuhl, P. (2017). Sampling matters in deep embedding learning. *International Conference on Computer Vision*, pp. 2840–2848.

Wu, J., Fan, H., Zhang, X., Lin, S. & Li, Z. (2021a). Semi-Supervised Semantic Segmentation via Entropy Minimization. *International Conference on Multimedia and Expo*, pp. 1–6.

Wu, K., Du, B., Luo, M., Wen, H., Shen, Y. & Feng, J. (2019). Weakly Supervised Brain Lesion Segmentation via Attentional Representation Learning. *Medical Image Computing and Computer-Assisted Intervention*, pp. 211–219.

Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X. & Liu, C. H. (2021b). Embedded Discriminative Attention Mechanism for Weakly Supervised Semantic Segmentation. *Computer Vision and Pattern Recognition*, pp. 16765–16774.

Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y. & Cai, J. (2022). Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81, 102530.

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A. & Roth, H. (2020). 3D semi-supervised learning with uncertainty-aware multi-view co-training. *Winter Conference on Applications of Computer Vision*, pp. 3646–3655.

Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. (2020). Self-training with noisy student improves Imagenet classification. *Computer Vision and Pattern Recognition*, pp. 10687–10698.

Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X. et al. (2021). A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67, 101832.

Xu, C., Pham, D. L. & Prince, J. L. (2000). Image segmentation using deformable models. *Handbook of Medical Imaging*, 2(20), 0.

Xue, Y., Tang, H., Qiao, Z., Gong, G., Yin, Y., Qian, Z., Huang, C., Fan, W. & Huang, X. (2020). Shape-aware organ segmentation by predicting signed distance maps. *AAAI Conference on Artificial Intelligence*, 34(07), 12565–12572.

Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A. P., Bagheri, M. & Summers, R. M. (2018). Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. *Computer Vision and Pattern Recognition*, pp. 9261–9270.

Yang, L., Jin, R., Mummert, L., Sukthankar, R., Goode, A., Zheng, B., Hoi, S. C. & Satyanarayanan, M. (2008). A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. *Transactions on Pattern Analysis and Machine Intelligence*, 32(1), 30–44.

Yang, P., Zhai, Y., Li, L., Lv, H., Wang, J., Zhu, C. & Jiang, R. (2019). Liver Histopathological Image Retrieval Based on Deep Metric Learning. *Bioinformatics and Biomedicine*, pp. 914–919.

Yao, H., Hu, X. & Li, X. (2022). Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. *Association for the Advancement of Artificial Intelligence Conference*, 36(3), 3099–3107.

Ying, J., Huang, W., Fu, L., Yang, H. & Cheng, J. (2023). Weakly supervised segmentation of uterus by scribble labeling on endometrial cancer MR images. *Computers in Biology and Medicine*, 167, 107582.

Yu, L., Wang, S., Li, X., Fu, C.-W. & Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, pp. 818–833.

Zeng, L.-L., Gao, K., Hu, D., Feng, Z., Hou, C., Rong, P. & Wang, W. (2023). SS-TBN: A Semi-Supervised Tri-Branch Network for COVID-19 Screening and Lesion Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, A., Lipton, Z. C., Li, M. & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.

Zhang, C.-B., Jiang, P.-T., Hou, Q., Wei, Y., Han, Q., Li, Z. & Cheng, M.-M. (2021). Delving deep into label smoothing. *Transactions on Image Processing*, 30, 5984–5996.

Zhang, J., Xie, Y., Xia, Y. & Shen, C. (2019). Attention residual learning for skin lesion classification. *Transactions on Medical Imaging*, 38(9), 2092–2103.

Zhang, S. & Metaxas, D. (2016). Large-Scale medical image analytics: Recent methodologies, applications and Future directions. Elsevier.

Zhang, Z. & Saligrama, V. (2016). Zero-shot learning via joint latent similarity embedding. *Computer Vision and Pattern Recognition*, pp. 6034–6042.

Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Tong, R. & Wu, J. (2019a). Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. *Medical Image Computing and Computer-Assisted Intervention*, pp. 148–156.

Zheng, H., Motch Perrine, S. M., Pitirri, M. K., Kawasaki, K., Wang, C., Richtsmeier, J. T. & Chen, D. Z. (2020). Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation. *Medical Image Computing and Computer-Assisted Intervention*, pp. 802–812.

Zheng, M., Karanam, S., Wu, Z. & Radke, R. J. (2019b). Re-identification with consistent attentive siamese networks. *Computer Vision and Pattern Recognition*, pp. 5735–5744.

Zheng, S., Song, Y., Leung, T. & Goodfellow, I. (2016). Improving the robustness of deep neural networks via stability training. *Computer Vision and Pattern Recognition*, pp. 4480–4488.

Zhong, A., Li, X., Wu, D., Ren, H., Kim, K., Kim, Y., Buch, V., Neumark, N., Bizzo, B., Tak, W. Y. et al. (2021). Deep Metric Learning-based Image Retrieval System for Chest Radiograph and its Clinical Applications in COVID-19. *Medical Image Analysis*, 70, 101993.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016). Learning deep features for discriminative localization. *Computer Vision and Pattern Recognition*, pp. 2921–2929.

Zhou, S. K., Rueckert, D. & Fichtinger, G. (2019). *Handbook of medical image computing and computer assisted intervention*. Academic Press.

Zhou, S. K., Greenspan, H. & Shen, D. (2023). *Deep learning for medical image analysis*. Academic Press.

Zhu, S., Yang, T. & Chen, C. (2021). Visual explanation for deep metric learning. *Transactions on Image Processing*, 30, 7593–7607.

Ziko, I., Granger, E. & Ben Ayed, I. (2018). Scalable laplacian K-modes. *Advances in Neural Information Processing Systems*, pp. 10041–10051.