Annotation-Efficient and Reliable Medical Image Segmentation

by

Mélanie GAILLOCHET

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

MONTREAL, JUNE 2nd, 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Hervé Lombaert, Thesis supervisor
Department of Computer Engineering, Polytechnique Montréal

Mr. Christian Desrosiers, Thesis Co-Supervisor
Department of Software and IT Engineering, École de Technologie Supérieure

Mr. Marco Pedersoli, Chair, Board of Examiners
Department of Systems Engineering, École de Technologie Supérieure

Mr. Jose Dolz, Member of the Jury
Department of Software and IT Engineering, École de Technologie Supérieure

Mr. Sylvain Bouix, Member of the Jury
Department of Software and IT Engineering, École de Technologie Supérieure

Mr. Bernhard Kainz, External Independent Examiner
Department of Computing, Imperial College London

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MAY 11^{th}, 2026

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

To my family

and

To the little girl who always doubted herself

ACKNOWLEDGEMENTS

"The journey of a thousand miles begins with a single step."
— Lao Tzu

It is often said that a PhD thesis is like a marathon. Looking back, this journey will probably be the longest endurance race I will ever run, and the one I will cherish the most. I am profoundly grateful to everyone who has been by my side along the way, whether cheering from the sidelines or pacing me through the most difficult miles. Each of you has helped me grow not only as a researcher, but also as a person, a friend, and a sportswoman. This achievement is as much yours as it is mine.

First and foremost, I would like to thank my supervisors, Hervé and Christian. Throughout the years, you have guided me with unwavering support, understanding, and encouragement. Hervé, thank you for pushing me to become a better writer and communicator, and for teaching me how to grow into an independent researcher. Christian, your insight and the technical depth you brought to our weekly discussions have been invaluable. I feel incredibly fortunate to have benefited from your combined wisdom and kindness.

I extend my gratitude to my colleagues at ShapeÉTS, Livia, Neuro-IX, and PolyShape. Karthik and Sukesh, thank you for being awesome colleagues and for introducing me to your Indian culture. Fereshteh, Boshra, Gustavo, Sadaf and Ghazal—seeing you at the lab always brightened my day. Malik and Florent, I am so glad our friendship has extended to running adventures. Sylvain, thank you for our discussions and for inviting me to your lab events and projects. Nairouz, Florence, Pranav, Matéi and Dexter, although our paths crossed only recently, I am grateful to have shared my final months in the lab with you. Although I may not name everyone individually, I am deeply thankful to all of you, fellow researchers and friends, for making this journey far richer and more meaningful.

To all the people I met and collaborated with during my involvement in the ÉTS Student Association and the Graduate Student Committee, thank you for your trust. Nathalie, Geneviève, Martin, Alison, thank you for your ongoing dedication to supporting and guiding students. Prasun, your unfailing good mood and your ability to seamlessly switch between English and French mid-sentence never cease to impress me. Inoussa, you were always there to help bring our events to life. My involvement would not have been the same without you.

A special thanks to my friends at the Club de Ski de Fond de l'UdeM for being such incredible companions in adventure. Lauriane, Vincent, Lesly, Koldo, Raph, Aurel, Justine and Fab, thank you for the shared trails on snow and in the mud, the endless road trips, the energy and the laughter. To my book club companions, Sam, Caroline, Alex, Alessia, Daneese and Sarah, thank you for the great reads, the shared meals and the conversations that only got around to the book after several hours of life updates and laughter. And to Amytis and all the friends who made the trip to Montréal these past years, I am glad our friendship has stayed strong despite the years and the distance. All these moments we shared have made this journey so enjoyable and have been a welcome reminder that a good life is made of many, many things.

Enfin, à ma famille pour son amour et son soutien inconditionnel tout au long de cette aventure: Maman, Papa, Julien, Regina et ma toute nouvelle nièce Aurélie, je vous aime du fond du coeur.

Brice, je te garde pour la fin, car tu as été mon plus indéfectible soutien. Merci d'avoir toujours été là pour moi. J'ai hâte de commencer notre prochain chapitre ensemble.

Segmentation Médicale Fiable et Peu Coûteuse en Annotations

Mélanie GAILLOCHET

RÉSUMÉ

La segmentation d'images médicales, délimitation précise des structures anatomiques ou des lésions, est une étape clef pour le diagnostic médical, la planification chirurgicale et le suivi des pathologies. Les algorithmes d'apprentissage profond ont permis d'atteindre de hautes performances en segmentation automatique. Cependant, leur intégration en milieu clinique reste contrainte par deux obstacles majeurs : le coût prohibitif de l'annotation manuelle des données et l'absence de garanties statistiques sur les prédictions des modèles.

L'objectif de cette thèse est de **développer des algorithmes de segmentation plus fiables et moins coûteux en annotations** afin de faciliter leur adoption clinique. En particulier, nous abordons chaque obstacle à leur déploiement en ciblant des étapes distinctes du cycle de développement des modèles. **Premièrement**, nous proposons une méthode d'apprentissage actif basée sur des lots stochastiques pour sélectionner stratégiquement, *avant* l'entraînement du modèle, les données les plus informatives pour annotation. Cela permet d'atteindre des performances de segmentation élevées avec une fraction du budget d'annotation traditionnel. **Deuxièmement**, nous utilisons uniquement, *durant* l'entraînement, des annotations sous forme de boîtes englobantes, moins coûteuses à obtenir, pour spécialiser et automatiser les modèles de fondation visuels. Nous introduisons une stratégie d'apprentissage de prompts faiblement supervisée basée sur des contraintes et sur la régularisation. **Troisièmement**, nous développons un cadre de prédiction conforme anatomiquement cohérent pour fournir des garanties statistiques sur les masques de segmentation prédits *après* déploiement. Grâce à l'intégration au processus conforme d'une étape de diffusion par marche aléatoire basée sur les représentations des modèles de fondation, les ensembles de prédiction générés deviennent à la fois géométriquement cohérents et statistiquement valides.

Évaluées sur plusieurs modalités d'imagerie médicale et tissus anatomiques, ces contributions ouvrent la voie à une segmentation d'images médicales à la fois plus sûre et moins coûteuse en annotations, pour un meilleur déploiement clinique.

Mots-clés: Analyse d'images médicales, Annotation, Apprentissage actif, Apprentissage faiblement supervisé, Boîte englobante, Contraintes, Incertitude, Modèle fondateur, Prédiction conforme, Segmentation

Annotation-Efficient and Reliable Medical Image Segmentation

Mélanie GAILLOCHET

ABSTRACT

Medical image segmentation, which automatically delineates anatomical structures or lesions, is a key step for diagnosis, surgical planning, and disease monitoring. While deep learning-based algorithms have delivered state-of-the-art segmentation performance, their integration into the clinical workflow is hindered by two major challenges: the high cost of acquiring large annotated datasets and the lack of guarantees on model reliability.

The objective of this thesis is to **develop more reliable and annotation-efficient segmentation algorithms** in order to facilitate their clinical adoption. In particular, we address each limitation by targeting a distinct stage of the model development life-cycle. **First**, we propose a stochastic batch active learning framework that computes uncertainty at the batch level to strategically select, *before* training, the most informative samples to annotate from large unlabeled datasets. By implicitly enforcing diversity in the selection without additional computational cost, our strategy achieves better segmentation performance than purely uncertainty-based and random sampling, given a fixed data sampling budget. **Second**, we introduce a weakly-supervised prompt learning framework that adapts large promptable vision foundation models to medical tasks using only bounding box annotations. By applying box-based spatial constraints and consistency regularization to compensate for the reduced label information, our approach avoids costly pixel-level supervision *during* training and enables resource-efficient segmentation. Results show that weakly-supervised prompt learning is a scalable alternative to fully-supervised specialization of both foundation and non-foundation models. **Third**, we develop an anatomically-aware conformal prediction framework to provide statistical guarantees on the segmentation outputs *after* deployment. Specifically, we draw on the dense feature representations of vision foundation models to integrate anatomical context into the conformal process and construct geometrically consistent and statistically valid prediction sets. Our framework can be appended to any trained segmentation model without retraining, making it broadly applicable across architectures and clinical tasks.

Evaluated across multiple medical imaging modalities and anatomical targets, these contributions bring medical image segmentation closer to clinical deployment by reducing the annotation burden and providing reliability guarantees.

Keywords: Active learning, Annotation, Bounding box, Computer-aided diagnosis, Conformal prediction, Constraints, Foundation model, Medical image analysis, Prompt, Segmentation, Uncertainty, Weakly-supervised learning

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

ACDC	Automated Cardiac Diagnosis Challenge dataset
AL	Active learning
CAMUS	Cardiac Acquisitions for Multi-structure Ultrasound Segmentation dataset
CNN	Convolution neural network
CP	Conformal prediction
CRC	Conformal risk control
CT	Computer tomography
DL	Deep learning
ÉTS	École de Technologie Supérieure
FNR	False Negative Rate
GT	Ground truth
HC18	Head Circumference dataset
HU	Hounsfield unit
i.i.d.	Independent and identically distributed
kNN	k-nearest neighbour
LV	Left ventricle
MSD	Medical Segmentation Decathlon dataset
MRI	Magnetic resonance imaging
PROMISE12	Prostate MR Image Segmentation dataset

RV	Right ventricle
RW-CP	Random Walk Conformal Prediction
SAM	Segment Anything Model
SB	Stochastic batches
SSL	Semi-supervised learning
TTA	Test-time augmentation
US	Ultrasound
ViT	Vision transformer

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

ASSD	Average Symmetric Surface distance
DSC	Dice Similarity coefficient
HD	Hausdorff distance
HD95	95% Hausdorff distance
IoU	Intersection-over-Union
JSD	Jensen–Shannon divergence
$C_\lambda(.)$	Conformal set parametrized by λ
$\mathcal{D}_L$	Labeled dataset
$\mathcal{D}_U$	Unlabeled dataset
$f_\theta(\cdot)$	Model parameterized by θ
$\mathcal{L}$	Loss
Ω	Spatial image domain
$\mathbb{R}$	Real
$(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$	Image/ground-truth mask pair sample
$(\mathbf{x}^{(i)}, \mathbf{m}^{(i)})$	Image/bounding box label pair sample
$\mathbf{Z}_i$	Image embedding
$\mathbf{Z}_p$	Prompt embedding

INTRODUCTION

Medical image analysis and computer-aided diagnosis

Since the discovery of X-rays by Wilhelm Röntgen in 1895, medical images have served as non-invasive windows into the human body (Scatliff & Morris, 2014). They are now a standard diagnostic tool, allowing clinicians to detect anomalies, plan treatment or monitor disease progression. Modern medical imaging now covers a wide range of modalities, including computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), ultrasonography, and optical techniques such as dermoscopy and microscopy (see Fig. 0.1). Each modality exploits different regions of the electromagnetic spectrum or distinct physical phenomena to visualize anatomical structures, tissues, and pathologies.

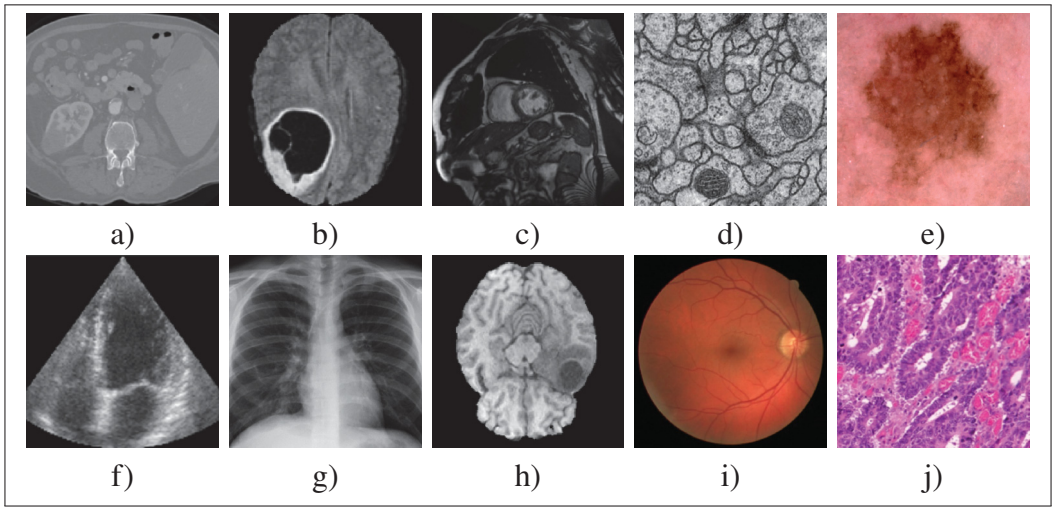


Figure 0.1 **Examples of medical imaging modalities depicting various regions of the human body.** a) abdominal CT, b) T2-FLAIR MRI of the brain, c) cine-MRI, d) electron microscopy, e) dermoscopy of skin lesion, f) cardiac ultrasound, g) chest X-ray, h) T1-weighted MRI of the brain, i) eye fundus, and j) histopathology.
Adapted from Huang et al. (2024b)

Interpreting medical images is a complex and time-consuming task that requires extensive expertise and training. The growing availability of imaging devices and the rapid increase in

the number of examinations have further intensified clinician workload (Winder, Owczarek, Chudek, Pilch-Kowalczyk & Baron, 2021). However, the limited availability of specialists, high consultation costs, and the inherent subjectivity of visual assessment constrain effective image interpretation (Sarvamangala & Kulkarni, 2022; Shi *et al.*, 2022). This situation places considerable pressure on healthcare systems, contributing to diagnostic delays and medical errors, even among experienced physicians (Winder *et al.*, 2021).

Medical image analysis addresses these limitations by providing computational tools that automatically process and interpret the content of imaging data, in order to support clinical decision-making (Lloyd, Monaco & Bui, 2016). Compared to manual interpretation, automated algorithms offer faster and more consistent inference, reducing analysis time and the risk of errors from clinician fatigue or workload (Sarvamangala & Kulkarni, 2022; Shi *et al.*, 2022). They also scale naturally to the demands of large screening programs and population studies.

Deep learning (DL) (LeCun, Bengio & Hinton, 2015) has become the dominant approach for computer-aided diagnosis, as models can learn task-relevant representations directly from data without hand-crafted features (Xia *et al.*, 2025). In medical imaging, DL is applied across six main tasks (Fig. 0.2): *regression* predicts a continuous variable (e.g., patient age or disease severity); *reconstruction* recovers high-quality images from degraded acquisitions; *registration* aligns images across time points, patients or modalities; *classification* assigns a diagnostic label to an image (e.g., malignant vs. benign); *detection* localizes anatomical structures or lesions; and *segmentation* assigns a class label to every pixel or voxel (Sistaninejhad, Rasi & Nayeri, 2023). The last three applications are part of a domain called image recognition, which aims to identify objects in images.

Segmentation is the most demanding task in medical image analysis, because it requires every pixel to be assigned a label. It is an active area of research (see Fig. 0.3) and forms the **primary application of this thesis**.

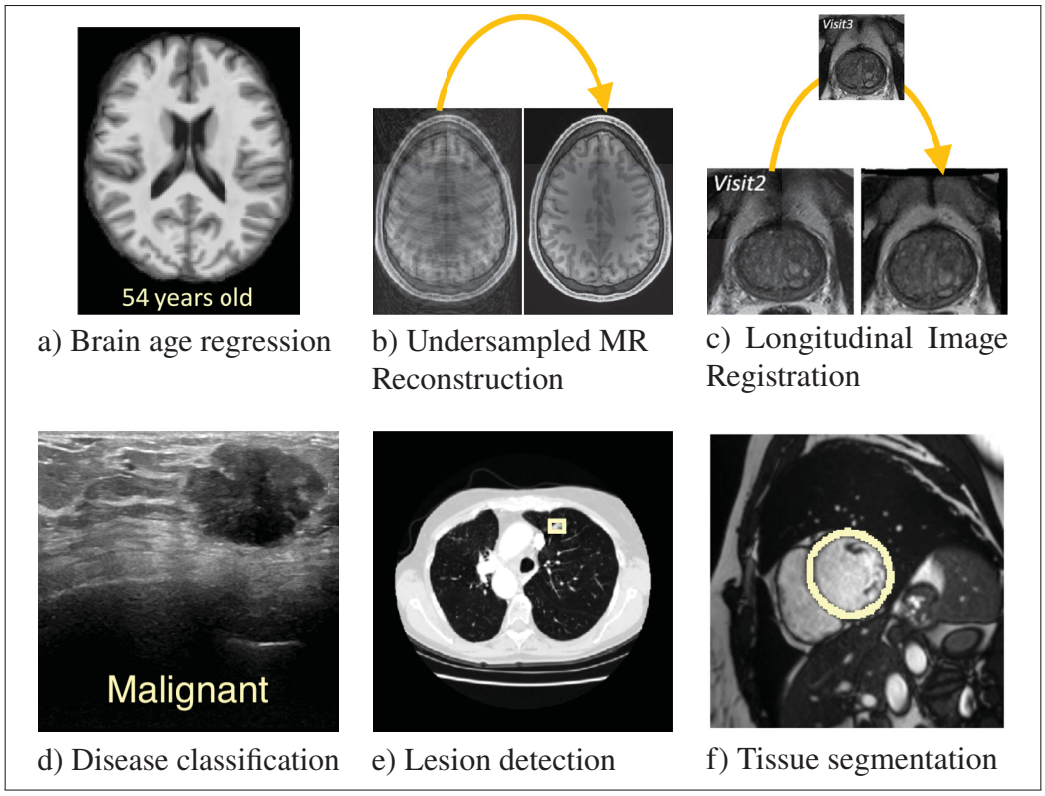


Figure 0.2 **Different applications of medical image analysis.** a,d-f) Image is overlayed with annotation corresponding to task.
Taken and adapted from Jonsson et al. *(2019); Gaillochet* et al. *(2020); Yang* et al. *(2020); Al-Dhabyani* et al. *(2020); Armato* et al. *(2011) and Bernard* et al. *(2018)*

Medical image segmentation

The precise delineation of anatomical structures or lesions in medical images is traditionally performed manually by trained radiologists. This process is very time-consuming, as annotating a single patient's volumetric scan can require several hours of expert effort (Shi *et al.*, 2022). It is also poorly reproducible, as manual delineation is subject to annotator expertise and level of focus. These limitations make manual segmentation impractical at clinical scale and motivate the development of reliable automatic alternatives.

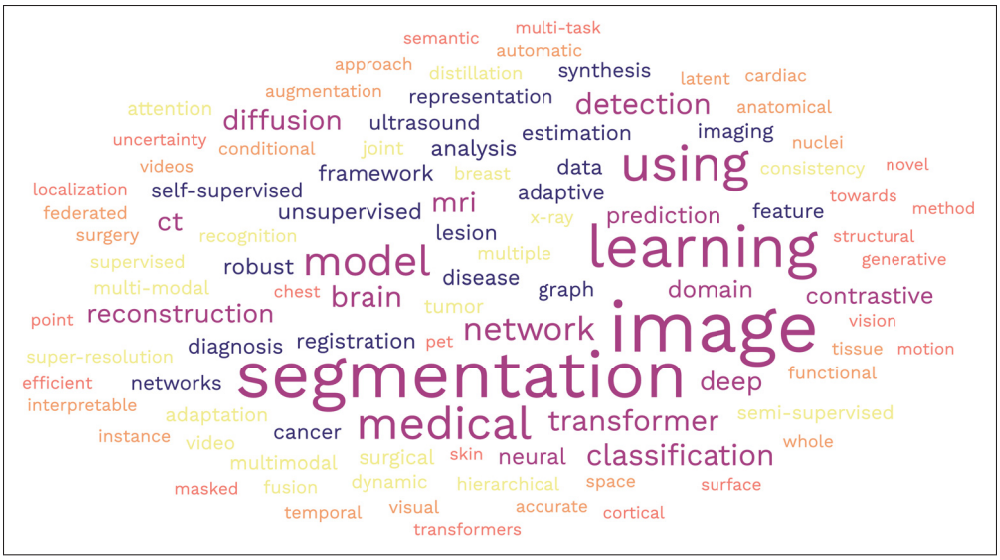


Figure 0.3 **Word cloud of paper titles** from the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2023 conference. Segmentation remains an active area of research

Medical image segmentation automatically partitions an image into meaningful regions by assigning a label to each pixel (2D) or voxel (3D). As a result, it can rapidly provide quantitative information on the location, size and morphology of structures in the body.

Segmentation supports a wide range of clinical workflows. In diagnosis, it enables quantitative measurements such as cortical thickness, whose reduction is associated with neuronal loss and neurodegenerative disease, or ventricular ejection fraction, whose decline reflects impaired cardiac function and may indicate heart failure. In surgical planning, accurate delineation of target structures and surrounding anatomy guides resection boundaries and reduces intra-operative risk. In post-operative settings, it supports assessment of tissue recovery. For chronic conditions, longitudinal segmentation allows objective monitoring of disease progression (e.g. tracking tumor growth).

Regardless of the clinical context, the accuracy of segmentation directly affects clinical decisions and patient outcomes. For instance, the stakes are particularly high in radiotherapy planning,

where oncologists must precisely delineate tumor volumes and organs-at-risk prior to dose calculation. Spatial errors of even a few millimeters can lead to tumor under-treatment or unnecessary irradiation of healthy tissue (Eber *et al.*, 2025). Reliable segmentation is therefore a prerequisite for safe and effective patient care (Xia *et al.*, 2025).

Challenges and motivation: between annotation cost and prediction reliability

The promising results obtained in classification and image segmentation tasks have placed deep learning (DL)-based techniques at the forefront of medical image analysis and computer-aided diagnosis. However, despite their state-of-the-art performance, their integration into the clinical workflow is hindered by two main limitations: the high cost of acquiring large annotated datasets, and the lack of statistical guarantees on model outputs (Xia *et al.*, 2025; Kumar *et al.*, 2025; Budd, Robinson & Kainz, 2021).

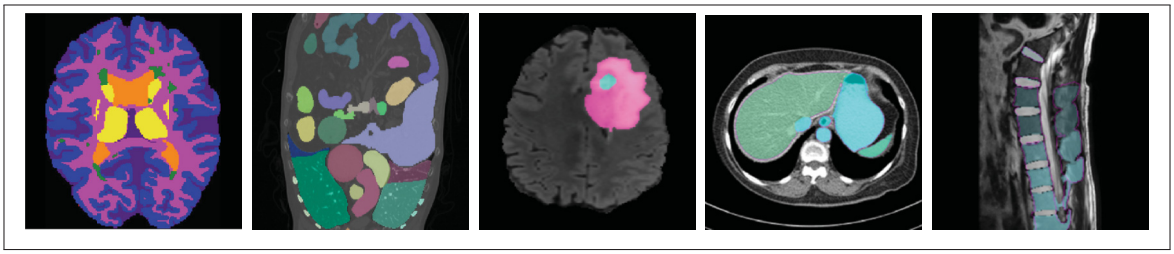


Figure 0.4 **Examples of segmented medical images.** Segmentation involves assigning a class to every pixel in the image. The segmented structures (in colour) often exhibit complex shapes and ambiguous boundaries, making segmentation a challenging task and the acquisition of accurate ground-truth annotations labor-intensive.
Taken from Li et al. *(2020), Cheng* et al. *(2023) and Ma* et al. *(2024)*

First, although automatic segmentation algorithms are designed to alleviate the burden of manual annotation, they typically rely on vast amounts of dense, pixel-level labels (see Fig. 0.4) to achieve high performance and generalize effectively. In practice, producing such annotations is extremely time-consuming: trained radiologists may spend several hours or even days annotating a single patient's data (Shi *et al.*, 2022). As a result, constructing large, densely labeled datasets

is prohibitively expensive and limits the scalability of these methods. Reducing the dependence on pixel-wise manual annotation is therefore a central challenge in medical image segmentation.

Second, in domains such as medical diagnosis, the ability to quantify and control the risk of model failure is critical. Clinicians should be able to assess when and where predictions are likely to be incorrect and with what frequency. Yet, standard deep learning models inherently lack the mechanisms to quantify uncertainty and to provide insight on the model's trustworthiness (Abdar *et al.*, 2021). Providing segmentation models with performance guarantees would significantly improve their reliability and clinical usefulness.

Objectives and contributions

The challenges outlined above prevent segmentation models from being deployed at clinical scale. The main objective of this thesis is therefore to develop more reliable and annotation-efficient segmentation algorithms in order to facilitate their clinical adoption.

Research questions

To investigate whether accurate and reliable automatic segmentation of medical images can be achieved with reduced human annotation effort, this thesis addresses the following questions:

1. Can randomness serve as an effective criterion for selecting diverse samples for annotation?
2. Can weak supervision reach near state-of-the-art performance when combined with promptable foundation models?
3. Can foundation model features enrich predicted probability maps with spatial context and improve uncertainty quantification?

Specific objectives

We tackle this main objective by elaborating on three specific objectives, each targeting a different phase of model development (see Fig. 0.5).

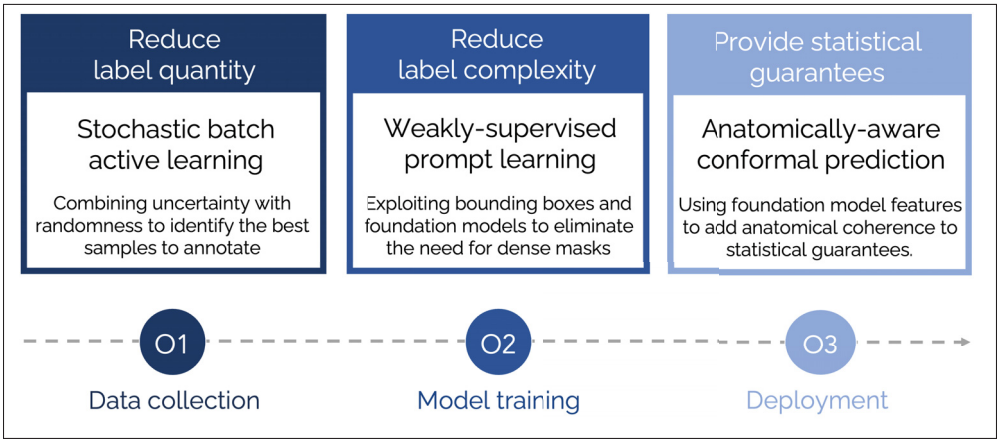


Figure 0.5 **Outline of thesis contributions.** This thesis explores different strategies across the model development life-cycle to make segmentation algorithms more annotation-efficient and reliable. Each research objective, highlighted in blue, is addressed by a contribution presented in the same frame

The first objective is to develop an active learning strategy to identify the most valuable samples to annotate *before* training. The second objective is to adapt vision foundation models to clinical segmentation tasks using only weak bounding box supervision *during* training. The third objective is to provide guaranteed coverage control *after* deployment through an anatomically-aware conformal prediction framework. The specific details of these three objectives of this thesis are as follows:

1. **Prioritize annotation via stochastic batch active learning.** Acquiring large, annotated medical datasets is extremely time-consuming and costly, but remains a prerequisite for high segmentation performance. Active learning addresses annotation cost by identifying the most informative samples to annotate and use during training, instead of indiscriminately annotating all samples or sampling randomly from an unlabeled pool. In practice,

however, existing strategies struggle to deliver consistent gains over a random selection, and their performance tends to be sensitive to architecture and hyperparameter choice. Moreover, purely uncertainty-based selection tends to query redundant or outlier samples, and approaches that enforce diversity through latent space coverage are computationally expensive and do not scale well to high-dimensional segmentation tasks. In Chapter 2, we tackle these issues by developing a stochastic batch querying strategy that can be used on top of any uncertainty-based AL method. Rather than selecting samples purely by individual uncertainty scores, we compute uncertainty over batches of samples that are generated randomly. Our strategy naturally promotes diversity without requiring pairwise distance computations or distribution estimation. Furthermore, it outperforms standard AL strategies on medical image segmentation while remaining robust to changes in architecture, hyperparameters, and annotation budget.

2. **Improve weakly-supervised segmentation via prompt learning with constraints.** Dense pixel-wise annotations remain the dominant form of supervision for training segmentation models, yet their collection is labour-intensive and increasingly difficult to scale given the growing volume of medical imaging data. Vision foundation models offer a promising solution through their zero-shot abilities (Kirillov *et al.*, 2023; Ma *et al.*, 2024), but fully exploiting them still typically requires either manual prompting at test time, or dense annotations for specialization via fine-tuning or prompt-learning. Bounding boxes offer a practical alternative: they are easier to obtain and carry spatial information on the object to segment. Building on this observation, Chapter 3 proposes a prompt learning framework that adapts promptable vision foundation models to clinical tasks using only sparse bounding box annotations. More specifically, keeping the foundation model frozen, we train an auxiliary module to automatically generate prompt embeddings from the input image, thus bypassing the need for manual prompting at inference. To compensate for the reduced supervision signal, training combines multiple box-based constraints with consistency regularization,

enabling the model to produce accurate segmentation masks without requiring ground truth annotations during training. The resulting framework achieves near fully-supervised segmentation performance while relying only on coarse, cheap-to-obtain labels.

3. **Guarantee coverage of predicted segmentations via anatomically-aware conformal prediction.** Reliable clinical deployment requires uncertainty estimates that provide rigorous, user-controlled guarantees on prediction errors. Heuristic uncertainty measures such as Monte Carlo dropout (Gal & Ghahramani, 2016b) or deep ensembles (Lakshminarayanan, Pritzel & Blundell, 2017b) produce intuitively reasonable estimates, but carry no formal guarantee. In addition, they are sensitive to model design choices and distribution shifts. Conformal prediction offers an attractive solution by constructing statistically valid prediction sets, guaranteed to contain the true segmentation within a pre-specified error rate, under minimal assumptions. However, directly applying conformal prediction to segmentation outputs and enforcing the coverage constraint often leads to severe over-segmentation, a trade-off that is rarely quantified in the literature. In Chapter 4, we introduce the Random-Walk Conformal Prediction framework that tackles this issue by integrating conformal prediction with a random walk diffusion guided by the rich feature embeddings of vision foundation models. By propagating predicted probabilities through semantically similar neighborhoods, the framework regularizes and denoises uncertainty estimates. The result is a conformal prediction set that retains its statistical validity while respecting the anatomical geometry of the structure being segmented.

Impact

Our contributions pave the way for fast and efficient clinical deployment of segmentation models. Specifically, they:

- **Reduce resource requirements**: Active learning and weak supervision reduce how much expert labeling is needed. Instead of annotating large datasets upfront, our methods identify

which samples are worth labeling and accept coarser annotations where possible. This makes model development viable even in settings where radiologist time is limited.

- **Accelerate clinical workflow**: Once trained, segmentation models reduce clinician workload by automating the delineation of structures that would otherwise take hours per patient. In addition, the ability to adapt foundation models with minimal supervision means that institutions can apply these models rapidly to new modalities or anatomical targets.
- **Improve patient safety**: In high-stakes applications such as surgical planning, radiotherapy or tumor delineation, segmentation errors can lead to severe clinical consequences. Our conformal prediction framework improves reliability by producing prediction sets that satisfy user-defined coverage guarantees, allowing clinicians to quantify and control the risk of model failure. This supports safer clinical deployment by enabling more informed decision-making.

Published work

The research presented in this thesis has resulted in the following publications. Chapters 2–4 are each based on one or more of these publications, with additional details provided in the respective chapter introductions.

- Journal articles -

1. **M. Gaillochet**, C. Desrosiers and H. Lombaert. "Anatomically-Aware Conformal Prediction for Medical Image Segmentation with Random Walks". Submitted to *IEEE Journal of Biomedical and Health Informatics (IEEE JBHI)*, 2026.
2. **M. Gaillochet**, M. Noori, S. Dastani, C. Desrosiers and H. Lombaert. "Prompt Learning with Bounding Box Constraints for Medical Image Segmentation". *IEEE Transactions on Biomedical Engineering (IEEE TBME)*, vol. 73, no. 1, pp. 359-368, 2025.

3. **M. Gaillochet**, C. Desrosiers and H. Lombaert. "Active Learning for Medical Image Segmentation with Stochastic Batches". *Medical Image Analysis (MedIA)*, vol. 90, p. 102958, 2023.

- Conference articles (peer-reviewed and archived) -

1. **M. Gaillochet**, C. Desrosiers and H. Lombaert. "Automating MedSAM by Learning Prompts with Weak Few-shot Supervision." *MICCAI Workshop on Foundation Models for General Medical AI (MICCAI-MedAGI)*, p. 61-70, 2024.
2. **M. Gaillochet**, C. Desrosiers and H. Lombaert. "TAAL: Test-time Augmentation for Active Learning in Medical Image Segmentation". *MICCAI Workshop on Data Augmentation, Labeling, and Imperfections (MICCAI-DALI)*, p. 43-53, 2022.

- Conference short articles -

1. **M. Gaillochet**, C. Desrosiers and H. Lombaert. "Active Learning for Medical Image Segmentation with Stochastic Batches". *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2023.

- Co-authored publications -

In addition to the publications cited above, I also contributed to the following work during my doctoral studies:

1. G. Danaee, **M. Gaillochet**, C. Desrosiers, H. Lombaert and S. Bouix. "Exploring Entropy-based Active Learning for Fair Brain Segmentation." Accepted to *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2025.

- Seminar presentations -

Beyond peer-reviewed publications, I presented at several seminars and workshops for the local medical imaging community.

1. "Weakly Supervised Prompt Learning with Box-based Constraints for Medical Image Segmentation." *Montreal Medical Imaging Workshop*, May 2025.
2. "Automating MedSAM via Prompt Learning with Weak Few-Shot Supervision." *Montreal Medical Imaging Seminar*, May 2024.
3. "Active Learning for Medical Image Segmentation with Stochastic Batches". *Workshop in Medical Imaging with Deep Learning*, October 2023.
4. "Active Learning Methods for Medical Image Segmentation". *ÉTS Medical Imaging Seminar*, April 2023.

- Scientific Communication Competitions -

Finally, I also engaged in science communication competitions aimed at making my research accessible to a non-specialist audience.

1. *Ma Thèse en 180 Secondes (Three Minute Thesis).* National competition, 2026. **Finalist.**
2. *Ma Thèse en 180 Secondes (Three Minute Thesis).* École de technologie supérieure, Montréal, 2026. **Winner.**
3. *Speed Science Competition.* Mila - Quebec AI Institute, Montréal, 2025. **Runner-up.**
4. *Poster Competition for Science Popularization.* École de technologie supérieure, Montréal, 2021. **Finalist.**

Code availability

All implementations presented in this thesis were developed using the Python programming language[1] and the PyTorch library[2]. The source code and relevant scripts are publicly accessible at:

1. Test-time augmentation for active learning
   https://github.com/Minimel/TAAL
2. Active learning with stochastic batches
   https://github.com/Minimel/StochasticBatchAL
3. Automating MedSAM by learning prompts with weak few-shot supervision
   https://github.com/Minimel/MedSAMWeakFewShotPromptAutomation
4. Prompt learning with bounding box constraints
   https://github.com/Minimel/box-prompt-learning-VFM
5. Anatomically-aware conformal prediction with random walks
   https://github.com/Minimel/RW_ConformalPrediction

[1] https://www.python.org

[2] https://pytorch.org

CHAPTER 1

BACKGROUND

1.1 Overview of medical imaging modalities

Medical imaging modalities are non-invasive methods to visualize the internal anatomy and biological functions of the human body. They have become essential to the modern clinical workflow, enabling diagnosis, treatment planning and follow-up monitoring without surgical intervention. Unlike natural images, medical images span multiple modalities (i.e. acquisition technologies), such as Magnetic Resonance Imaging, X-ray, Computed Tomography, Positron Emission Tomography or ultrasound. Each modality exploits different physical phenomena and produces images with distinct contrast properties, noise characteristics, resolution limits and boundary ambiguities. The work presented in this thesis focuses on the three modalities most represented in public benchmarks for segmentation: CT, MRI, and ultrasound.

1.1.1 Computed tomography

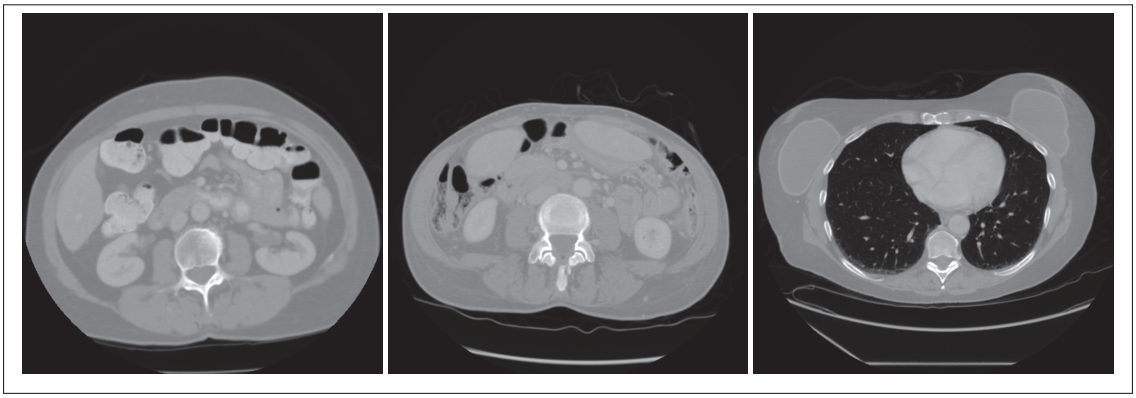


Figure 1.1 **Examples of axial slices from CT volumes** illustrating the high-contrast differentiation between bone, soft tissue and air.
Taken from Antonelli et al. (2022)

Computed Tomography (CT) is a radiographic modality that produces detailed 3D volumes by combining multiple 2D X-ray projections acquired from different angles. A rotating X-ray source and detector array measure the attenuation of radiation through the body, and reconstruction

algorithms, typically filtered back-projection or iterative methods, assemble these projections into cross-sectional slices.

CT intensity values are standardized using the Hounsfield unit (HU), which expresses the linear attenuation coefficient of a tissue relative to that of water (0 HU) and air (-1000 HU). Typically, bone tissue appears near +400 HU while soft tissue falls between -100 and +100 HU. The fact that voxel intensities hold the same physical meaning across different scanners and acquisition protocols makes CT a convenient modality for automated analysis. It is, for example, commonly used for lung cancer detection.

CT offers standardized intensity values and high spatial resolution. However, its reliance on ionizing radiation motivates the use of alternative modalities for soft-tissue characterization and longitudinal monitoring. Magnetic Resonance Imaging addresses both of these limitations.

1.1.2 Magnetic resonance imaging

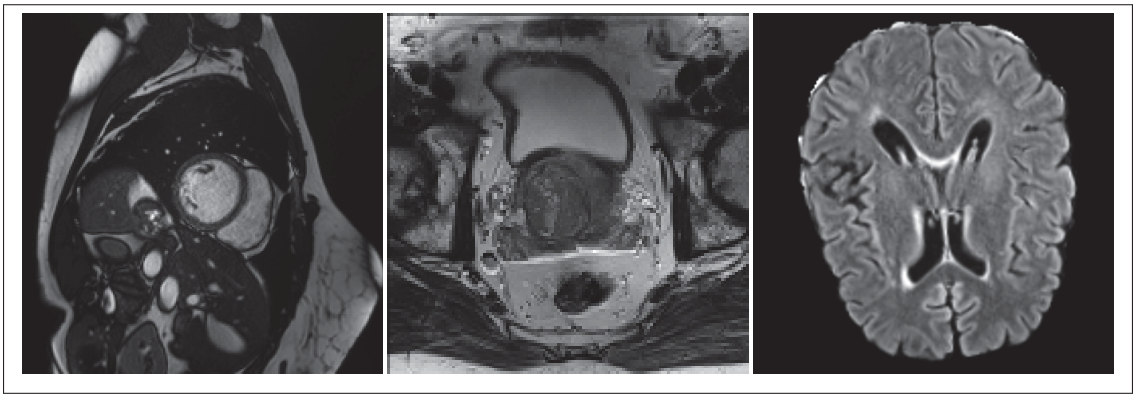


Figure 1.2 **Examples of slices from MR volumes** of the heart, prostate and brain, illustrating the variability in tissue appearance.
Taken from Bernard et al. (2018), Litjens et al. (2014) and Antonelli et al. (2022)

Magnetic Resonance Imaging (MRI) is a non-ionizing modality that creates high-resolution images by exploiting the magnetic properties of hydrogen nuclei, present in large quantity in the human body. A strong static magnetic field aligns the proton spins of the body, and a radiofrequency pulse perturbs this alignment. The time it takes for protons to return to

equilibrium, characterized by the longitudinal relaxation time T_1 and the transverse relaxation time T_2, varies across tissue types and is measured to construct the MR image.

The primary advantage of MRI lies in its exceptional soft-tissue contrast. By tuning acquisition parameters such as echo time or repetition time, different tissue contrasts can be emphasized, enabling clinicians to distinguish between healthy and pathological tissue that might appear identical on a CT scan. This flexibility is particularly valuable for spine or brain tumor imaging. Furthermore, unlike CT, MRI does not involve ionizing radiation, making it safer for repeated examinations.

However, MRI acquisition remains costly, time-consuming (typical scans range from 15 to 90 minutes) and susceptible to motion artifacts (Petralia *et al.*, 2021). In addition, both CT and MRI require specialized, fixed infrastructure and are unsuitable for real-time or bedside use. Ultrasound imaging fills this gap as a portable, low-cost alternative.

1.1.3 Ultrasound

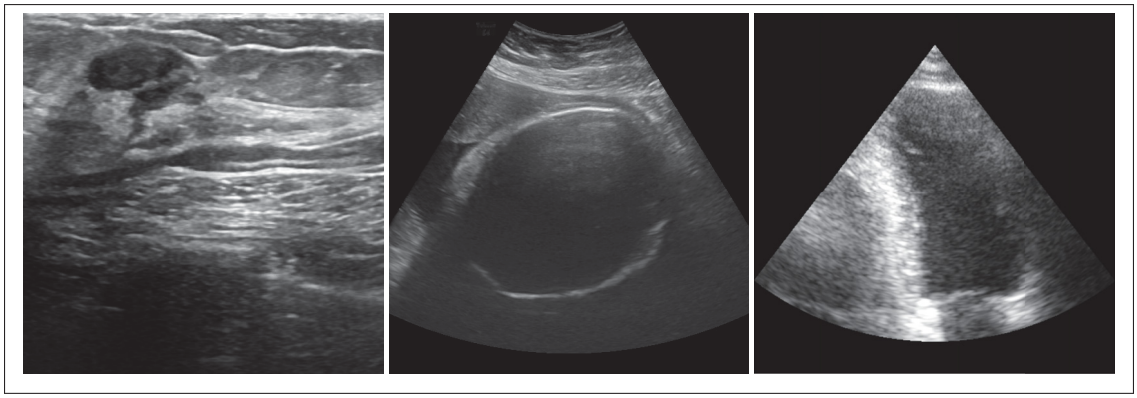


Figure 1.3 **Examples of B-mode ultrasound images.**
Taken from Al-Dhabyani et al. *(2020), van den Heuvel* et al. *(2018) and Leclerc* et al. *(2019)*

Ultrasound imaging is a real-time modality based on the transmission and reflection of high-frequency sound waves (typically 2–15 MHz). A transducer probe converts electrical pulses into mechanical vibrations via the piezoelectric effect. The sound waves propagate through the

body, and the time-of-flight of reflected echoes is used to reconstruct a 2D (or 3D) image of internal structures. The resulting image is referred to as a B-mode (i.e. brightness mode) scan.

Because of its portability, low cost and absence of ionizing radiation, ultrasound has become the primary tool for real-time monitoring across a range of clinical domains, including obstetrics, cardiac imaging and musculoskeletal assessment.

From a computer vision perspective, however, ultrasound is widely considered the most challenging modality. The images are inherently noisy due to speckle, a granular interference pattern. In addition, ultrasound images suffer from acoustic shadowing, which occurs when a dense structure such as bone reflects all incident sound energy, creating a dark cone of missing information below the structure. These artifacts, combined with the relatively low spatial resolution compared to CT and MRI, make the boundaries between anatomical structures particularly ambiguous in ultrasound images.

1.2 Deep learning for medical image segmentation

Medical images were historically analyzed using conventional computer vision techniques based on hand-crafted features, such as thresholding, region growing, edge detection or Markov random fields (Sistaninejhad *et al.*, 2023; Held *et al.*, 1997). While effective in pre-defined settings, these heuristic approaches require significant domain expertise to design and tune, and often fail to generalize to the complex variability and high dimensionality of real clinical data (Xia *et al.*, 2025).

Medical image segmentation underwent a major transformation with the widespread adoption of neural networks and deep learning (Goodfellow, Bengio & Courville, 2016), which replaced manual feature engineering with end-to-end representation learning (Kumar *et al.*, 2025). Deep learning-based models automatically extract a hierarchical sequence of increasingly abstract features directly from raw pixel data, and are trained by minimizing a task-specific loss function through stochastic gradient descent. As a result, segmentation accuracy has improved considerably, especially for organs and pathologies with ill-defined boundaries or high

inter-patient variability. Initially dominated by Convolutional Neural Networks (CNNs), deep learning models for segmentation have more recently been influenced by Vision Transformers (ViTs) and large-scale Foundation Models.

1.2.1 CNN-based approaches

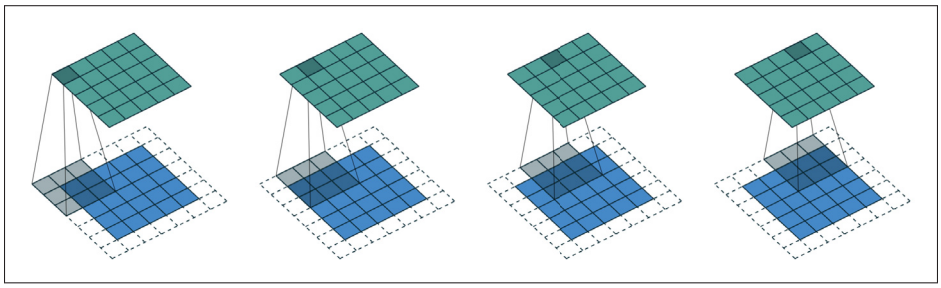


Figure 1.4 **Convolution operation applied to a 5×5 image**. The 3×3 kernel is applied with stride 1 and zero-padding, to keep the output the same size as the input.
Taken from Dumoulin & Visin (2018)

Convolutional Neural Networks (Lecun, Bottou, Bengio & Haffner, 1998; LeCun *et al.*, 2010) have been widely adopted for learning-based segmentation owing to their ability to automatically extract relevant spatial features from raw pixel data through shared, learnable filters (see Fig. 1.4).

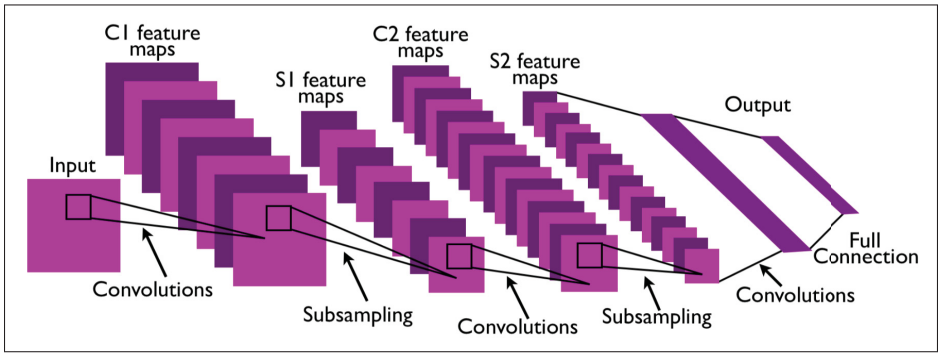


Figure 1.5 **Fundamental components of a CNN**. Are depicted the interleaving of convolutional filters, spatial downsampling via subsampling, and non-linear activation.
Taken from LeCun et al. (2010)

A typical CNN architecture consists of a sequence of functional blocks (see Fig. 1.5). First, convolutional layers apply learnable filters to capture local spatial patterns. Then, subsampling operations (e.g. max pooling or strided convolutions) reduce the spatial resolution of feature maps to increase the effective receptive field. Finally, non-linear activation functions such as ReLU introduce the expressive capacity necessary to approximate complex input-output mappings. The inductive biases inherent to CNNs, namely local receptive fields and translation equivariance, make them particularly well-suited for image-based tasks and enable robust generalization even when training data is limited.

Among CNN-based approaches, the U-Net (Ronneberger *et al.*, 2015) is arguably the most influential architecture in medical image segmentation. It adopts a symmetric encoder-decoder structure: the encoder gradually reduces spatial dimensions while capturing semantic context, whereas the decoder upsamples feature maps to recover fine-grained spatial resolution (see Fig. 1.6). In addition, skip connections concatenate high-resolution feature maps from each encoder stage directly to their symmetric decoder counterpart, effectively preserving spatial detail that would otherwise be lost during downsampling. This design enables both high-level semantic understanding and precise localization.

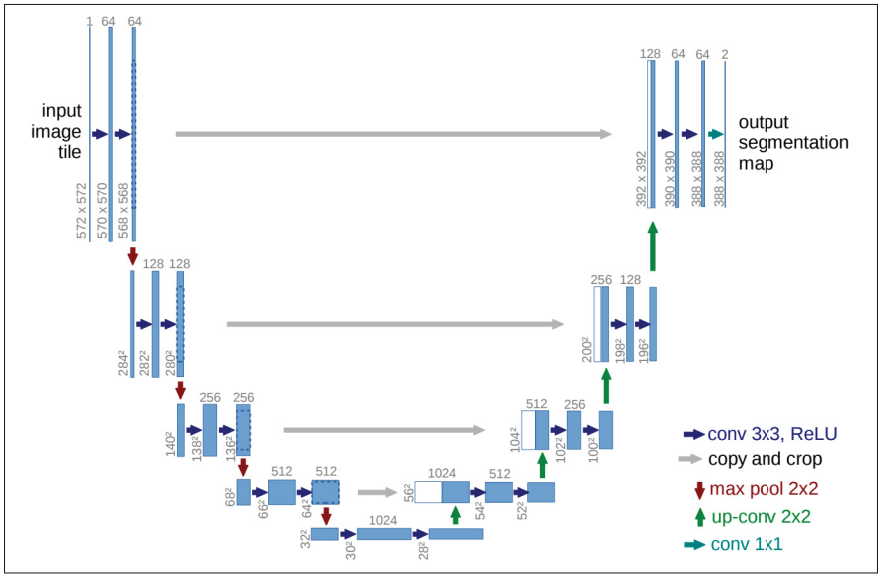


Figure 1.6 **Overview of the U-Net architecture.**
Taken from Ronneberger et al. *(2015)*

Because of its success, subsequent work has extended the U-Net in numerous directions. ResUNet (Yang *et al.*, 2019) incorporates residual connections into the encoder path to ease gradient flow, Attention U-Net (Oktay *et al.*, 2018) introduces soft spatial attention gates to focus on relevant regions, and nnU-Net (Isensee, Jaeger, Kohl, Petersen & Maier-Hein, 2021) provides a self-configuring framework that automatically adapts the architecture and training pipeline to a given dataset.

1.2.2 Vision transformers and foundation models

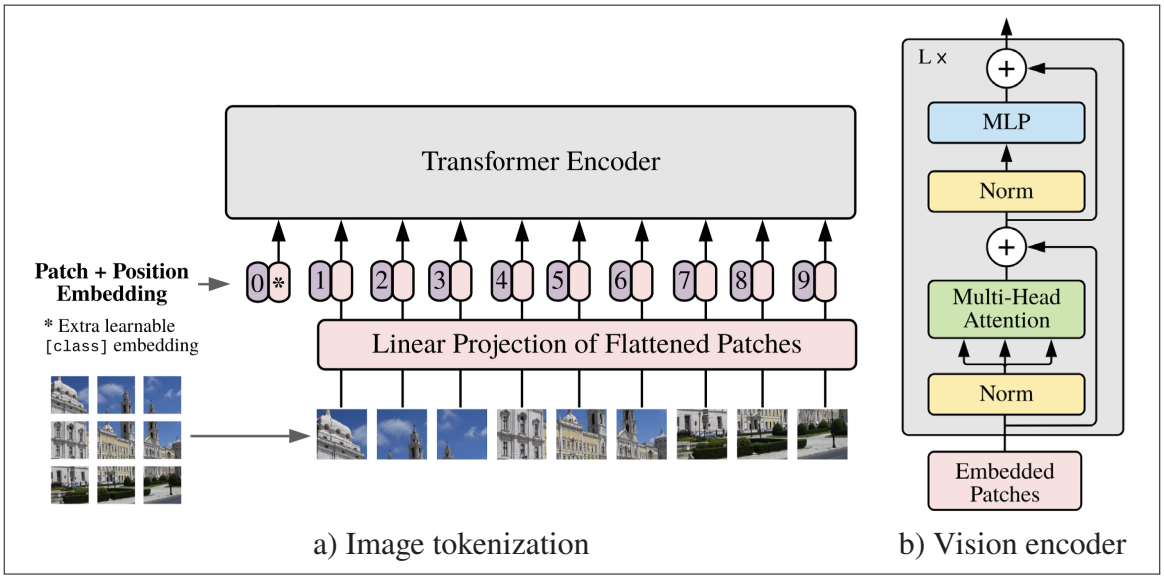


Figure 1.7 **Overview of a vision transformer encoder.** a) The input image is split into patches which are converted into embeddings and fed into a b) Standard transformer encoder with multi-head attention.
Adapted from Dosovitskiy et al. *(2021)*

Vision Transformers (Dosovitskiy *et al.*, 2021) have challenged the dominance of CNNs by adapting the transformer architecture, originally introduced for natural language processing (Vaswani *et al.*, 2017), to 2D image data. A ViT first partitions an image into a sequence of fixed-size, non-overlapping patches, which are linearly projected into token embeddings (see Fig. 1.7). These tokens, augmented with positional encodings, are then processed by successive self-attention layers. Unlike CNNs, whose local convolutional filters inherently limit the receptive

field to spatial neighbourhoods, self-attention computes pairwise interactions between all tokens simultaneously. This enables the model to capture long-range spatial dependencies at every layer.

Vision foundation models

To compensate for the absence of the translation equivariance and local connectivity biases built into CNNs, ViTs typically require substantially larger datasets for pre-training. This limitation has motivated the development of foundation models, large-scale models pre-trained on massive and diverse datasets and designed to generalize across a wide range of downstream tasks.

A first family of vision foundation models learns joint image-text representations through contrastive objectives. For instance, CLIP (Radford *et al.*, 2021) and its variants (Wang, Wu, Agarwal & Sun, 2022; Zhang *et al.*, 2024b) align image and text embeddings in a shared latent space from image-caption pairs, enabling strong zero-shot recognition. A second family exploits purely visual self-supervised objectives. Masked Autoencoders (MAE) (He *et al.*, 2022) train the network to reconstruct randomly masked image patches, yielding scalable representations without labels. The DINO series (Caron *et al.*, 2021; Oquab *et al.*, 2024; Siméoni *et al.*, 2025) instead adopts a teacher-student self-distillation scheme on millions of curated images, producing rich, transferable features without manual annotation. A third family targets general-purpose task execution. A prominent example is the Segment Anything Model (SAM) (Kirillov *et al.*, 2023), a promptable segmentation model trained on over one billion masks from natural images (see Fig. 1.8). At its release in 2023, SAM set a new standard in zero-shot image segmentation and prompt-based interaction through clicks, bounding boxes, or scribbles.

Together, these models establish a new groundwork in computer vision: general-purpose representations that can be transferred, fine-tuned, or prompted for specialized domains.

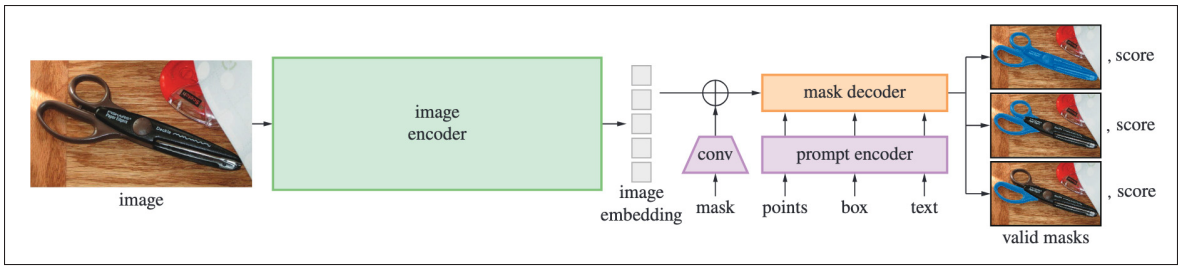


Figure 1.8 **Overview of the Segment Anything Model**.
Taken from Kirillov et al. (2023)

Prompt learning versus fine-tuning

Adapting a pre-trained model to a new domain such as medical imaging can be approached in several ways. Fine-tuning updates all or some of the backbone model parameters using the target dataset. However, effectively updating model parameters increases the risk of catastrophic forgetting of general representations and requires a large amount of labeled data. Parameter-efficient alternatives such as low-rank adaptation (LoRA) mitigate this by inserting small trainable adapter modules into the frozen backbone, considerably reducing the number of updated parameters. However, these approaches still require gradient flow through the backbone at training time, which can be prohibitively expensive for very large foundation models. Prompt learning, by contrast, keeps the pre-trained model frozen and instead learns a small set of parameters, either prompt embeddings or a dedicated auxiliary module that generates them, to adapt the model's behavior to the target task. Because the backbone is never updated, training is faster, memory requirements are lower, and the backbone representations are left intact. These properties make prompt learning particularly well-suited to the low-data, resource-constrained settings common in medical imaging. It is therefore the adaptation strategy adopted in this thesis.

1.3 The annotation bottleneck in medical AI and label-efficient strategies

The aim of supervised learning is to update the weights of deep neural networks to minimize the difference between the prediction and the ground-truth (i.e. target). Supervised learning

has reached state-of-the-art performance in many medical image segmentation tasks when large annotated datasets are available. However, collecting such large datasets remains a challenge, both in terms of cost, time and expertise required. This annotation bottleneck has spurred the development of a range of label-efficient methods, which seek to maximize segmentation performance while minimizing the burden placed on human annotators.

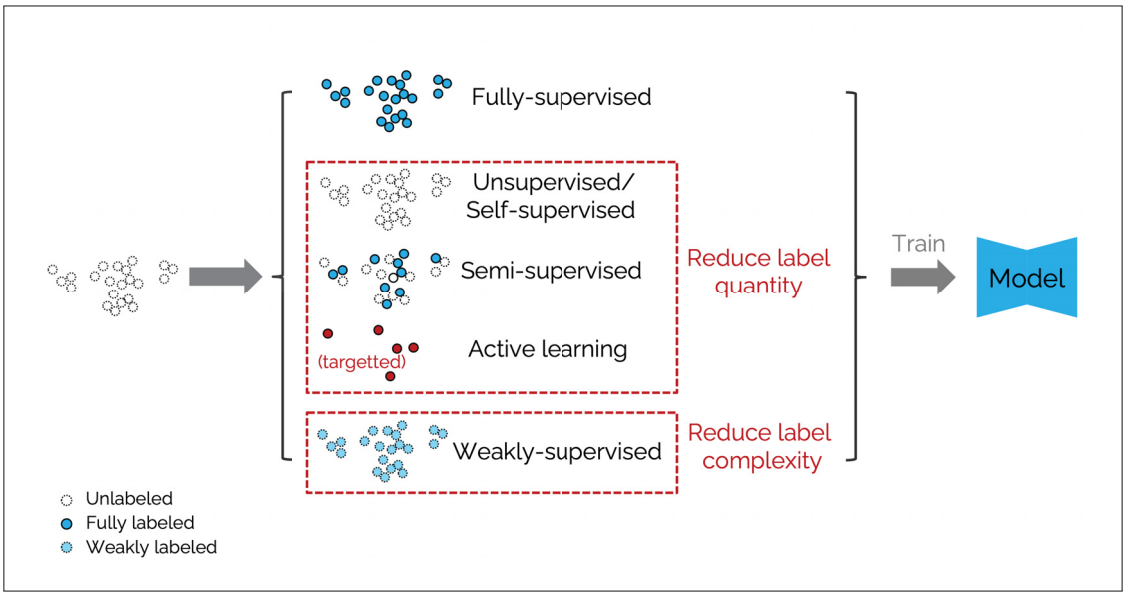


Figure 1.9 **Label-efficient alternatives to fully-supervised learning.** *Self-supervised learning* learns the spatial and semantic structure of unlabeled data. *Semi-supervised learning* uses both some labeled and many unlabeled samples. *Active learning* iteratively selects few informative samples to annotate and add to the training set. *Weakly-supervised learning* uses labels that are less informative and precise but easier to obtain than dense annotations

Strategies that seek to minimize annotation cost can be divided into two categories: those reducing the *quantity* of examples that require annotation and those reducing the annotation *complexity* required per example (see Fig. 1.9). However, as depicted in Fig. 1.10, both these approaches directly impact the output model's performance.

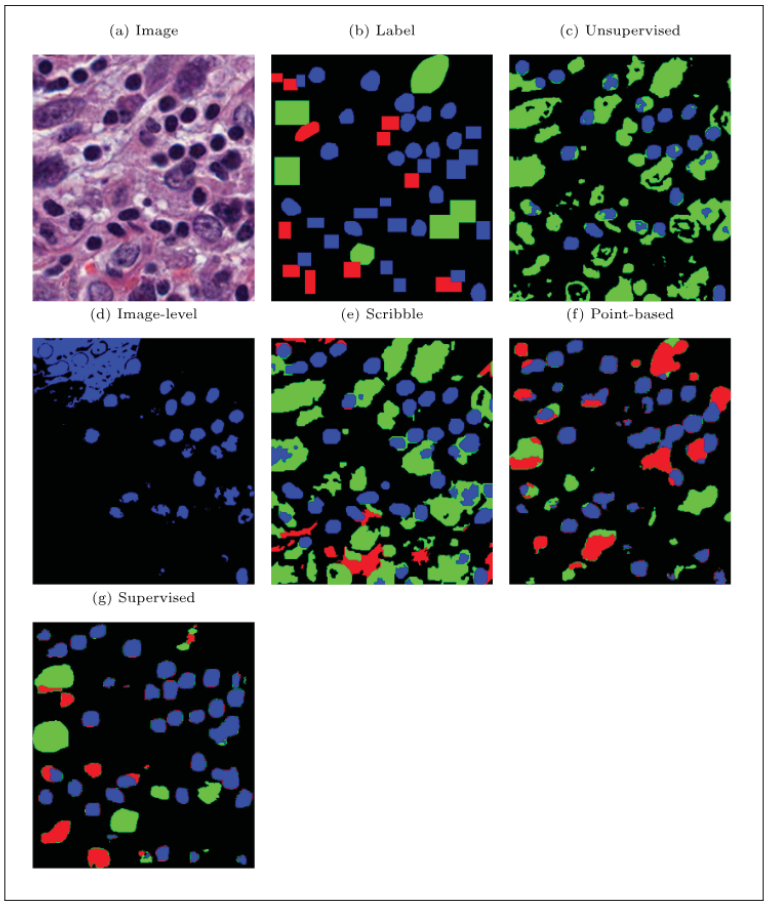


Figure 1.10 **Examples of segmentation prediction for different learning approaches.** First row: a) image and b) ground-truth mask and c) prediction from unsupervised training. Second row: prediction from weak supervision given d) image labels, e) scribbles and f) point-based labels. Third row: prediction from supervised training. Training a model when training samples are not all fully annotated is a challenging task.
Taken from Fernández-Moreno et al. (2023)

1.3.1 Unsupervised and self-supervised learning

Classical unsupervised methods, such as clustering and autoencoders, extract structure from unlabeled data but typically yield representations that are too generic for segmentation tasks (Doersch, Gupta & Efros, 2015). Self-supervised learning addresses this by constructing supervision signals directly from the unlabeled data, forcing the model to solve tasks that require spatial and

semantic understanding. Early attempts relied on context reconstruction (Doersch *et al.*, 2015; Pathak, Krähenbühl, Donahue, Darrell & Efros, 2016), image colorization (Zhang, Isola & Efros, 2016) or rotation prediction (Gidaris, Singh & Komodakis, 2018). These were quickly overtaken by contrastive learning approaches, which train a model to produce similar representations for augmented views of the same image while pushing apart views of different images (He, Fan, Wu, Xie & Girshick, 2020; Chen, Kornblith, Norouzi & Hinton, 2020). More recently, masked image modeling, which predicts randomly masked input patches, has emerged as the dominant self-supervised approach for vision transformers, with masked autoencoders (He *et al.*, 2022) achieving strong transfer performance across downstream tasks. In parallel, self-distillation methods such as DINO (Caron *et al.*, 2021) and its successors (Oquab *et al.*, 2024; Siméoni *et al.*, 2025) have demonstrated that ViT-based encoders trained on large unlabeled datasets produce semantically rich patch-level features that generalize across domains.

1.3.2 Semi-supervised learning

Semi-supervised learning (SSL) (van Engelen & Hoos, 2020) makes use of a large pool of unlabeled data in addition to few labeled samples in order to improve the representation learned from data. SSL methods exploit the unlabeled data through a variety of mechanisms: consistency regularization enforces the stability of model predictions under data augmentation or model perturbations (Laine & Aila, 2017; Tarvainen & Valpola, 2017); pseudo-labeling assigns predicted soft labels to unlabeled examples and incorporates them into training (Lee, 2013); and contrastive learning (Chaitanya, Erdil, Karani & Konukoglu, 2020) encourages the model to learn representations that cluster semantically similar examples together in feature space. While SSL can significantly reduce the annotation requirement, it implicitly assumes that the unlabeled data distribution matches the labeled one, an assumption that may not hold in clinical settings. Moreover, semi-supervised learning frameworks do not address the question of how to select the labeled subset, a task which active learning undertakes.

1.3.3 Active learning

Training samples do not contribute equally to the performance of learning-based algorithms (Settles, 2009). Hence, rather than passively use all available data like in conventional supervised learning, active learning (AL) selects the most informative subset of unlabeled examples to present to the annotator and use for training (Budd *et al.*, 2021; Ren *et al.*, 2021). By prioritizing examples where annotation would most improve generalization, AL aims to achieve comparable performance with a fraction of the annotation budget. Active learning is hence particularly relevant in medical imaging, where manual annotation cost is high and expert time scarce.

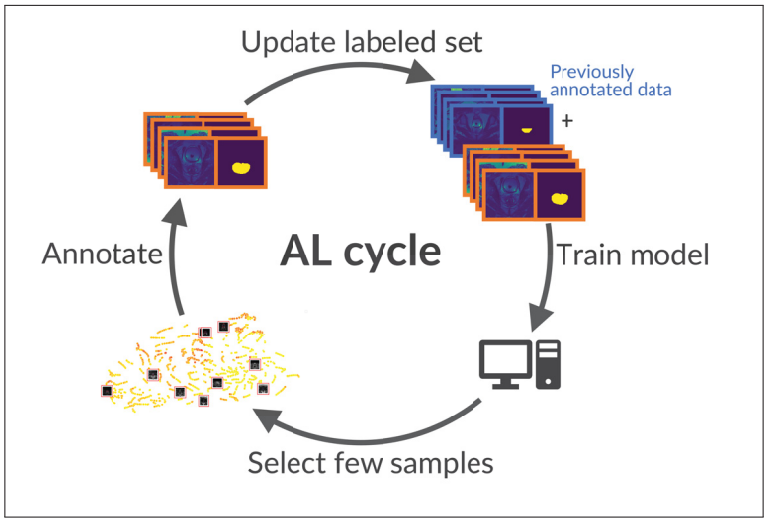


Figure 1.11 **Typical active learning cycle.** A small subset of samples is selected for annotation, added to the training set and used to train the model. The model is then used to identify the next batch of samples to annotate. The cycle is repeated until the labeling budget is exhausted.
Adapted from Gaillochet et al. *(2023)*

AL approaches can typically be divided into three groups: uncertainty-based methods, diversity-based methods and hybrid strategies which combine both. *Uncertainty-based sampling* strategies target the samples for which the model is least confident. Initial attempts typically used heuristics such as prediction entropy or margin sampling (Beluch, Genewein, Nurnberger & Kohler, 2018; Ren *et al.*, 2021). Subsequent work combined the traditional uncertainty measures

with geometric measures (Konyushkova, Sznitman & Fua, 2019), pseudo-labeling (Wang, Zhang, Li, Zhang & Lin, 2017) or Bayesian approximations via Monte Carlo dropout (Gal, Islam & Ghahramani, 2017; Kirsch, van Amersfoort & Gal, 2019). Recent work also used the loss predicted from an auxiliary model as proxy for uncertainty (Yoo & Kweon, 2019). Although widely adopted, purely uncertainty-based selection tends to query redundant, correlated samples, wasting annotation resources and potentially introducing bias into the model. *Diversity-based methods* address this issue by selecting samples that best represent the overall dataset distribution. This requires computing latent embeddings for each sample, either using the current model (Sener & Savarese, 2018) or an auxiliary network (Sinha, Ebrahimi & Darrell, 2019). Diversity is then enforced by minimizing the distance between labeled and unlabeled embeddings (Sener & Savarese, 2018), or by training a discriminator to assess whether unlabeled embeddings resemble labeled ones (Sinha *et al.*, 2019). However, the computational hurdle of these methods hinders their applicability to high-dimensional data. *Hybrid approaches* consider uncertainty and diversity as complementary and have combined them sequentially through a two-step procedure (Yang, Zhang, Chen, Zhang & Chen, 2017; Ash, Zhang, Krishnamurthy, Langford & Agarwal, 2020; Kim, Park, Kim & Chun, 2021; Nath, Yang, Landman, Xu & Roth, 2021), or jointly within a single acquisition function (Sourati *et al.*, 2019). However, these hybrid methods typically inherit the computational burden of their individual components, limiting scalability (Ash *et al.*, 2020; Nath *et al.*, 2021) or introducing additional tuning (Kim *et al.*, 2021).

Active learning has been extensively explored in lower-dimensional tasks such as classification (Gal *et al.*, 2017; Wang *et al.*, 2017; Sener & Savarese, 2018; Beluch *et al.*, 2018; Ash *et al.*, 2020). It remains a relatively new field for segmentation, and applications to medical image segmentation are still scarce. Existing approaches either rely on non-deep-learning models (Top, Hamarneh & Abugharbieh, 2011; Konyushkova, Sznitman & Fua, 2015; Konyushkova *et al.*, 2019), are computationally prohibitive (Yang *et al.*, 2017; Nath *et al.*, 2021), or require sub-sampling of the unlabeled pool to remain tractable (Sourati *et al.*, 2019). This leaves a largely

unaddressed gap between active learning methodology and its application to large-scale medical image segmentation.

1.3.4 Weakly supervised learning

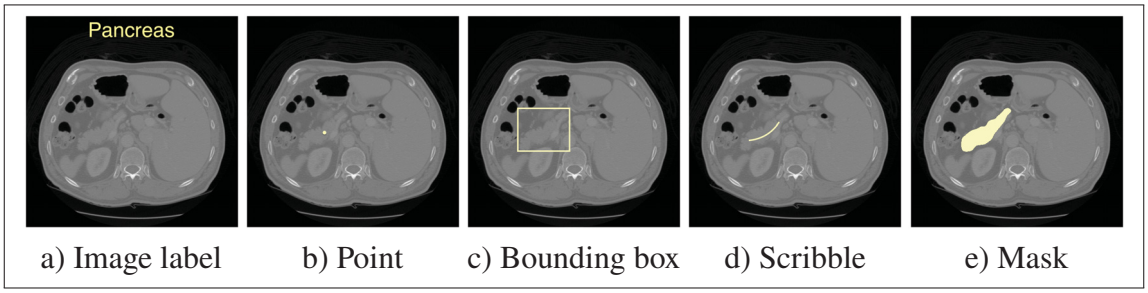


Figure 1.12 **Annotation types ordered by complexity.** From left to right: image-level label, point, bounding box, scribble and dense mask. Annotations are overlaid in yellow. *Adapted from Antonelli* et al. *(2022)*

Dense pixel-wise masks, while the most informative, are also the most expensive to obtain. Weakly supervised learning (WSL) relaxes the requirement for full, ground-truth annotation by training with annotations that are cheaper and faster to acquire, albeit less spatially informative. As illustrated in Fig. 1.12, weak annotation types have varying complexities: image-level labels indicate only the presence or absence of a structure; points provide a single approximate localization; scribbles offer sparse strokes tracing through a structure; and bounding boxes provide axis-aligned rectangles enclosing the target region.

Translating these coarse annotations into precise segmentation masks requires compensating for the missing spatial detail. Under image-level supervision, class activation maps (Chen *et al.*, 2022b) and multiple instance learning (Ilse, Tomczak & Welling, 2018) have been widely adopted to localize structures from classification labels alone. Under bounding-box supervision, early approaches such as GrabCut (Rother, Kolmogorov & Blake, 2004) and its deep learning extension DeepCut (Rajchl *et al.*, 2017) iteratively refine pseudo-labels generated from the box region. More recent methods instead impose explicit constraints on the predicted mask, such as tightness or size penalties derived directly from the box geometry (Kervadec, Dolz, Wang,

Granger & Ayed, 2020; Wang & Xia, 2021), or combine pseudo-labeling with constrained optimization (Pathak, Krahenbuhl & Darrell, 2015; Khoreva, Benenson, Hosang, Hein & Schiele, 2017). These strategies are particularly relevant to the medical domain, where bounding boxes are substantially faster to acquire than full pixel-level masks.

1.4 Uncertainty and reliability in segmentation models

Successful clinical integration of automatic segmentation models requires mechanisms to detect and respond to system failures. Uncertainty measures are a natural choice because they give an insight on how confident the model is when performing a task for a given input image. The information can then be used to quantify the segmentation performance and to assist clinicians in performing targeted manual corrections, in a post-processing step.

Uncertainty in medical image segmentation is relevant at three levels (Jungo & Reyes, 2019). First, pixel- (or voxel-)level uncertainty, embodied by uncertainty maps for an input image, indicates *where* within a predicted segmentation the model is uncertain. Second, pixel-level uncertainty can be derived to obtain an instance-level or image-level uncertainty, which gives insight on *how confident* the model is about its specific segmentation. The resulting value can then be used to filter out unreliable detections. Finally, comparing that score against a predefined threshold produces a binary task-level evaluation of whether the segmentation was *successful or not*.

Uncertainty in deep learning is typically decomposed into two components: data-related aleatoric uncertainty and model-related epistemic uncertainty. On the one hand, *aleatoric uncertainty* arises from inherent noise or ambiguity in the data itself (e.g. low tissue contrast, acoustic shadowing or other acquisition artefacts) and cannot be reduced by additional training data (Kendall & Gal, 2017). On the other hand, *epistemic uncertainty* captures the uncertainty in the model parameters caused by limited training data. It is in principle reducible as more data become available. While many methods exist for modeling these two uncertainty components,

how they interact when combined remains poorly understood, and recent work suggests they are substantially entangled (Judge *et al.*, 2022).

Let $X \in \mathbb{R}^{H \times W}$ be an input image with label $Y \in \{0,1\}^{|C| \times H \times W}$ (with H and W the image height and width, and C the set of classes), and let $\hat{Y}$ be the predicted mask from a model parametrized by θ. The following subsections describe the main families of approaches, organized by how they estimate uncertainty, from direct statistical measures to sampling-based approximations, auxiliary-based measures or conformal guarantees.

1.4.1 Statistical uncertainty

The simplest uncertainty measure is the entropy Shannon (1948) of the predicted probabilities $P^c = \mathrm{p}(\hat{Y} = c | X, \theta)$ for class c:

$$\mathrm{H}(P) = -\sum_{c \in C} P^c \log P^c \tag{1.1}$$

Entropy is high when the model assigns similar probability across classes and low when one class dominates, making it a direct measure of prediction confidence for a single forward pass. Because deep learning-based models tend to be poorly calibrated (i.e. confidence scores don't accurately reflect the true probability of being correct), entropy is often unreliable in segmentation tasks. However, its computational simplicity makes it a widely-used uncertainty measure.

1.4.2 Sampling-based approaches

Sampling-based methods generate multiple predictions for the same input and quantify disagreement across them. Let $P_t^c = \mathrm{p}(\hat{Y} = c | X, \theta_t)$ denote the predicted probability for class c under sample t. Three strategies are widely used to produce these samples:

- **MC Dropout.** Dropout (Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov, 2014), originally introduced as a regularization technique, can be interpreted as approximate Bayesian inference (Gal & Ghahramani, 2016b). A network with dropout implicitly places a

distribution over its weights. Keeping dropout active at test time yields Monte Carlo samples from the approximate posterior distribution $q(\theta \mid \mathcal{D})$, where $\mathcal{D}$ is the training data.

- **Ensembling.** Multiple model instances are run independently, generating different outputs. Instances may come from separate training runs with different random initializations or hyperparameters (Lakshminarayanan *et al.*, 2017b; Wenzel, Snoek, Tran & Jenatton, 2020) or from snapshots saved at different stages of a single training (Huang *et al.*, 2017).
- **Test-time Augmentation (TTA)**. Multiple transformed versions of the input are passed through a fixed model, resulting in multiple predictions (Wang *et al.*, 2019).

Regardless of the sampling strategy, the same set of uncertainty measures can be computed from the T collected predictions. Letting $\tilde{P}^c = \frac{1}{T}\sum_{t=1}^{T} P_t^c$ be the mean prediction for class c:

- *Variance* quantifies per-class spread around the mean: $\text{Var} = \frac{1}{C}\sum_{c\in C}\frac{1}{T}\sum_{t=1}^{T}\left(P_t^c - \tilde{P}^c\right)^2$
- *Predictive entropy* measures uncertainty in the mean prediction: $\text{H}(\tilde{P}) = -\sum_{c\in C}\tilde{P}^c \log \tilde{P}^c$
- *Mutual information* isolates the epistemic component by subtracting the average per-sample entropy from the predictive entropy: $\text{I}(\hat{Y};\, \theta) = \text{H}(\tilde{P}) - \frac{1}{T}\sum_{t=1}^{T} H(P_t)$

The main disadvantage of these sampling-based approaches is that they are often time-consuming and resource-intensive, requiring multiple forward passes through the model at inference time, for a single input.

1.4.3 Network prediction

An alternative to sampling is to have the network predict uncertainty directly as part of its output. For example, the model can be augmented with an additional output head that produces an uncertainty estimate along with the prediction. This uncertainty estimate is typically the variance of logits perturbed with Gaussian noise (Kendall & Gal, 2017) or a scalar confidence score (DeVries & Taylor, 2018). However, a model may be a poor judge of its own reliability (Jungo, Balsiger & Reyes, 2020). Instead, a dedicated auxiliary network can be trained to assess the task model's predictions post-hoc, without modifying the original training procedure. This separate network can be supervised to output the predicted probability

of the true class (Corbière, Thome, Bar-Hen, Cord & Pérez, 2019) or the segmentation error probabilities (e.g. false positives or false negatives) (Jungo *et al.*, 2020). One limitation of these methods is that the predicted uncertainty is only as reliable as the proxy supervision signal.

1.4.4 Conformal prediction

All methods described above produce uncertainty estimates that lack formal probabilistic guarantees. Conformal prediction (CP) addresses this directly, by providing statistically valid prediction sets without assumptions on the model or data distribution (Angelopoulos & Bates, 2023). Given a calibration set drawn from the same distribution as the test data, conformal methods construct post-processed prediction sets guaranteed to contain the true label with a user-specified probability (see Fig. 1.13). In practice, the size of the conformal set reflects the level of uncertainty in the prediction: bigger sets imply greater uncertainty and vice-versa.

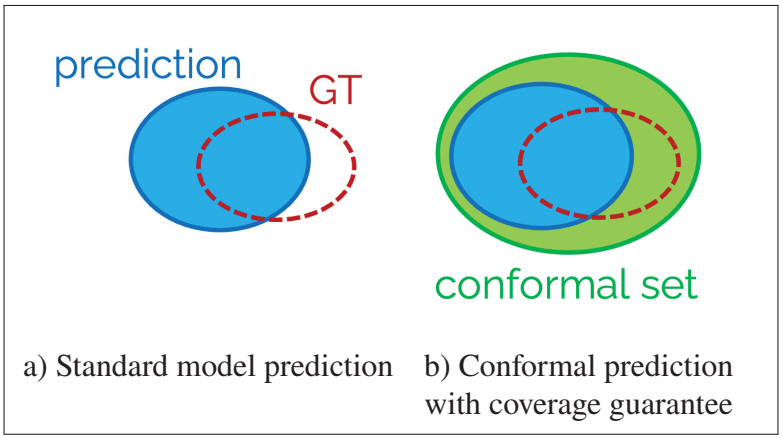


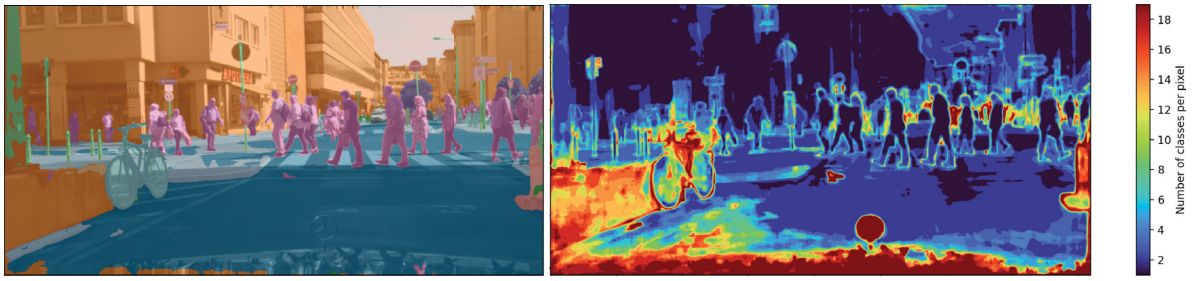
Figure 1.13 **Impact of conformal prediction sets.** a) The standard model prediction (blue) may fail to overlap with the ground truth (red). b) The conformal prediction set (green) is guaranteed to contain the ground truth with high probability. As the task difficulty increases, the conformal prediction set grows to account for greater model uncertainty

More formally, given a user-defined error rate $\alpha \in (0, 1)$ and a new test point (X_{n+1}, Y_{n+1}), the conformal set $C(.)$ constructed from calibration data $\{(X_i, Y_i)\}_{i=1}^{n}$ satisfies:

$$\mathbb{P}\left[Y_{n+1} \notin C(X_{n+1})\right] \leq \alpha. \tag{1.2}$$

This conformal set is *marginally valid*, meaning that it holds on average over the joint distribution of all $n + 1$ samples (not conditionally for any individual input). The only assumption required is exchangeability of the calibration and test samples, a weaker condition than independent and identically distributed (i.i.d.).

These properties make CP particularly attractive for clinical deployment. Unlike sampling-based or auxiliary methods, its coverage guarantee holds regardless of model miscalibration. However, conformal sets can be overly conservative when the underlying model is poorly calibrated, and estimates can be unstable when the calibration set is small.



a) Image with predicted segmentation b) Uncertainty heatmap

Figure 1.14 **Example of conformal semantic image segmentation with corresponding uncertainty map**, given a user-defined error rate. The uncertainty reflects the number of classes in the conformal set for each pixel.
Adapted from Mossina et al. (2024) ©2020, IEEE

Moreover, adapting the conformal framework to image segmentation, where predictions are pixel-wise binary masks rather than scalar labels, is not trivial and requires the careful design of both the uncertainty quantification mechanism and the coverage guarantee (see Fig. 1.14). As a result, conformal prediction for medical image segmentation, and binary segmentation in particular, remains largely unexplored and represents a gap this work aims to fill.

CHAPTER 2

ACTIVE LEARNING FOR MEDICAL IMAGE SEGMENTATION WITH STOCHASTIC BATCHES

Mélanie Gaillochet[1,2], Christian Desrosiers[1], Hervé Lombaert[2,3]

[1] Department of Software and IT Engineering , École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Mila - Quebec AI Institute,
6666 Rue Saint-Urbain, Montréal, Québec, Canada H2S 3H1
[3] Department of Computer Engineering, Polytechnique Montréal,
2500 Chem. de Polytechnique, Montréal, Québec, Canada H3T 0A3

Article published in *Medical Image Analysis (MedIA)*, December 2023

Presentation

The previous chapters presented two major limitations to the adoption of automatic image segmentation in the clinical workflow. This chapter focuses on the annotation burden. Specifically, we address the challenge of selecting which images to label in order to maximize segmentation performance with as few labeled samples as possible.

We propose a novel active learning strategy based on stochastic batch querying, which improves upon existing uncertainty-based sample selection methods by promoting diversity among candidate samples without additional computational cost.

This article, entitled *"Active learning for medical image segmentation with stochastic batches"*, was published in 2023 in **Medical Image Analysis** and presented as a short paper in 2022 at the conference **Medical Imaging with Deep Learning**. A prior work on uncertainty-based active learning, *"TAAL: Test-time Augmentation for Active Learning in Medical Image Segmentation"* (Appendix II), was presented in 2022 at the **MICCAI Workshop** on Data Augmentation, Labeling, and Imperfections.

Abstract

The performance of learning-based algorithms improves with the amount of labelled data used for training. Yet, manually annotating data is particularly difficult for medical image segmentation tasks because of the limited expert availability and intensive manual effort required. To reduce manual labelling, active learning (AL) targets the most informative samples from the unlabelled set to annotate and add to the labelled training set. On the one hand, most active learning works have focused on the classification or limited segmentation of natural images, despite active learning being highly desirable in the difficult task of medical image segmentation. On the other hand, uncertainty-based AL approaches notoriously offer sub-optimal batch-query strategies, while diversity-based methods tend to be computationally expensive. Over and above methodological hurdles, random sampling has proven an extremely difficult baseline to outperform when varying learning and sampling conditions. This work aims to take advantage of the diversity and speed offered by random sampling to improve the selection of uncertainty-based AL methods for segmenting medical images. More specifically, we propose to compute uncertainty at the level of batches instead of samples through an original use of stochastic batches (SB) during sampling in AL. Stochastic batch querying is a simple and effective add-on that can be used on top of any uncertainty-based metric. Extensive experiments on two medical image segmentation datasets show that our strategy consistently improves conventional uncertainty-based sampling methods. Our method can hence act as a strong baseline for medical image segmentation. The code is available on: https://github.com/Minimel/StochasticBatchAL.git.

2.1 Introduction

Data annotation is fundamental to medical imaging. Notably, the performance of segmentation algorithms depends on the amount of annotated training data. The manual annotation of pixel-level ground truth is therefore highly sought but remains difficult to obtain due to two challenging problems. First, the pixel-wise annotation of entire biological structures is a laborious and expensive task that requires highly trained clinicians. Second, image acquisition grows faster than the experts' ability to manually process the data, leaving large datasets mostly

unlabelled. Clinicians can realistically annotate only small sets of images with a limited capacity to scale up. This constraint creates a need for strategies that reduce the crucial but arduous annotation efforts in medical imaging.

To maximize the performance of a model with reduced annotated data during training, two types of approaches can unleash the potential of unlabelled data: active learning and semi-supervised learning. Active learning (AL) aims to identify the best samples to annotate and use during training. Meanwhile, semi-supervised learning seeks to improve the representation learned from data by exploiting unlabelled samples in addition to the few labelled ones. However, this approach still leaves the question of choosing which samples to use for the labelled set, underlining the importance of active learning.

Images in the training set do not contribute equally to the performance of learning-based algorithms (Settles, 2009). Given a large unlabelled dataset, active learning overcomes labelled data scarcity by incrementally identifying the most valuable samples to be annotated and added to a training set (Budd *et al.*, 2021; Ren *et al.*, 2021). Actively selecting which data to label conceivably maximizes the performance of machine learning models with a minimum amount of labelled data. AL strategies also have the potential of accelerating training convergence and improving robustness by targeting specific types of data points (Nath *et al.*, 2021).

Active learning methods can be divided into three broad categories: uncertainty-based sampling strategies, representative-based sampling strategies and hybrid approaches (Settles, 2009; Budd *et al.*, 2021). Uncertainty-based methods assume that the most valuable samples to annotate are the ones for which the current model is least confident. These methods, which differ in ways of calculating uncertainty, are however susceptible to target outlier samples or redundant information, particularly when querying batches of samples. To avoid bias towards narrow locals in distributions, representative-based and hybrid approaches try to diversify the set of candidate samples. Ensuring such diversity generally relies on learning a latent data representation, which requires estimating pairwise distances between all samples or computing their marginal distribution. These strategies consequently hardly scale satisfyingly to high

dimensions. Consequently, the majority of active learning approaches applied to computer vision focus on lower-dimensional tasks such as classification, while AL approaches for segmentation tend to focus on natural images with several thousands of annotated images (Sinha *et al.*, 2019; Huang, Wang, Xiong, Huan & Dou, 2021a; Kim *et al.*, 2021; Xie, Yuan, Li, Liu & Cheng, 2022). Due to its high-dimensional nature, medical image segmentation remains an ongoing challenge in active learning, despite the substantial need to minimize the high cost of manual annotation from clinical expertise.

A limited yet increasing number of works acknowledges that random sampling is, in practice, a painstakingly difficult baseline to outperform in active learning (Kirsch *et al.*, 2019; Mittal, Tatarchenko, Çiçek & Brox, 2019; Nath *et al.*, 2021; Munjal, Hayat, Hayat, Sourati & Khan, 2022; Burmeister *et al.*, 2022). Indeed, the gains of AL strategies over random sampling are often inconsistent across different experimental setups. For example, varying the sampling budget can cancel the improvements originally observed for such strategies (Bengar, van de Weijer, Twardowski & Raducanu, 2021; Munjal *et al.*, 2022). Similarly, existing methods for AL tend to be sensitive to the model architecture, hyperparameters and regularization used during training (Mittal *et al.*, 2019; Munjal *et al.*, 2022). These hurdles hinder AL advances in medical image segmentation.

This paper intends to address the limitations of current AL methods, notably their drawback of selecting batches solely based on per-sample uncertainty, the computational cost of ensuring diversity, and the significantly varying amounts of robustness in performance across experimental setups. Our work proposes to leverage the power of randomness during uncertainty-based batch sampling to improve the overall segmentation performance of AL models.

2.1.1 Contributions

We introduce the use of stochastic batch (SB) querying, a simple and effective add-on to uncertainty-based AL strategies, compatible with any uncertainty metric (see Fig.2.1). Our stochastic batch sampling strategy proves advantageous by:

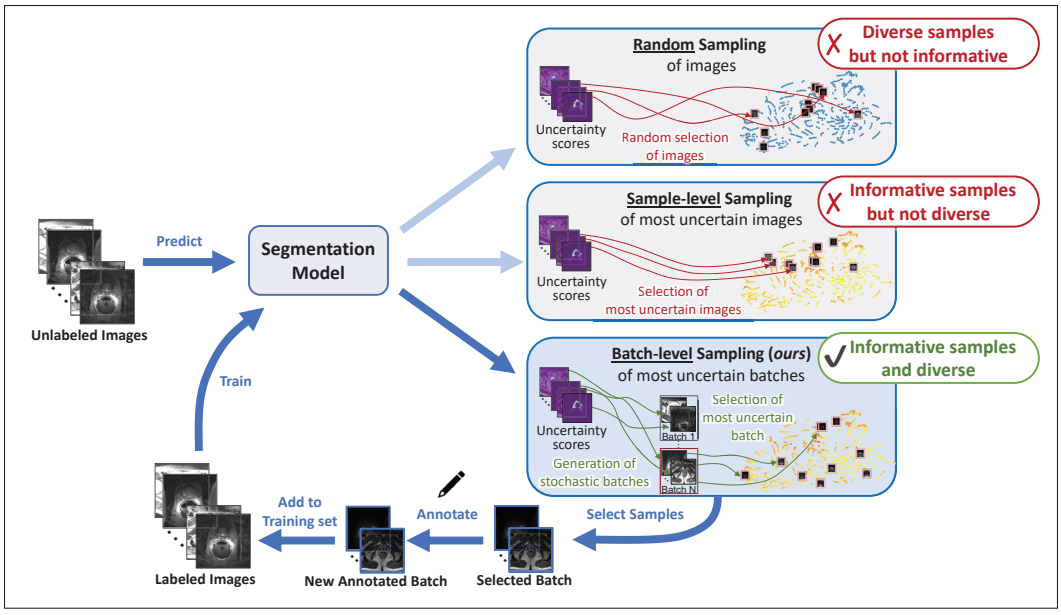


Figure 2.1 **Stochastic batch AL for uncertainty-based sampling**. Our sampling method combines the diversity of random sampling with the informativeness of uncertainty-based sampling. Adding our stochastic batch paradigm enables the data uncertainty to be estimated in a broader *batch-level* selection rather than a *sample-level* selection. After selecting a candidate set of unlabelled samples, the set is annotated and added to the existing labelled set. Finally, the segmentation model is retrained

1. Minimizing the problem of uncertainty-based strategies, often susceptible to query samples with redundant information;
2. Allowing uncertainty-based AL strategies to benefit from a larger diversity of samples in a simple and computationally-efficient way; and
3. Providing noticeably consistent gains across different experimental settings, as shown by our extensive ablation studies.

2.2 Literature review

Active learning methods maximize the future model performance by augmenting the current labelled training set with the most informative unlabelled samples. AL approaches mainly fall into uncertainty-based, representative-based or hybrid strategies, each described next.

2.2.1 Uncertainty-based AL methods

Uncertainty is one of the most prevalent criteria for sampling in active learning. Uncertainty-based methods query samples for which the current model is least confident (Settles, 2009). AL strategies for deep learning-based models have initially applied traditional AL methods that identify difficult examples using simple heuristics. However, in practice, they still hardly scale to high-dimensional data (Beluch *et al.*, 2018) or are not consistently effective for deep learning models that rely on batch selection (Sener & Savarese, 2018; Ren *et al.*, 2021). Hence, subsequent work has combined traditional uncertainty measures, such as the entropy of the output probabilities, with measures of geometric uncertainty (Konyushkova *et al.*, 2019) or with the pseudo-labelling of samples with confident predictions (Wang *et al.*, 2017). Similarly, Gal *et al.* (2017) and Kirsch *et al.* (2019) adapt existing heuristics to a Bayesian framework through Monte Carlo dropout. More recently, Yoo & Kweon (2019) developed a new uncertainty measure based on the predicted loss from the intermediate representations of the model. Although widely popular, purely uncertainty-based strategies relying on batch selection are susceptible to query samples with redundant information. However, manually annotating similar samples is a waste of annotation resources. Moreover, incorporating a set of similar samples to the labelled training set could bias the model towards an area outside the true data distribution. These samples could hence hamper rather than improve model generalization.

2.2.2 Representative-based AL methods

As opposed to uncertainty-based approaches, representative-based AL methods aim at diversifying the batch of candidate samples to improve the future performance of the model (Settles, 2009). One of the main representative-based approaches, Core-set (Sener & Savarese, 2018), identifies the most diverse and representative samples by minimizing the distance between the latent representations of labelled and unlabelled images, as given by the task model. Core-set aims for the model to perform as well with the candidate set as it would with the entire dataset. While specifically designed to be applied to complex models such as Convolutional Neural Networks (CNNs), core-set selection does not scale well to high-dimensional data since it

requires computing the Euclidean distance between all pairs of data samples. A later work, VAAL (Sinha *et al.*, 2019), learns a smooth latent-state representation of the input data via a variational auto-encoder (VAE). VAAL then selects samples different from the ones already labelled based on the learnt latent representation. Since the VAE is task-agnostic, VAAL can, however, easily query outlier data. In addition, it provides no mechanism to avoid choosing overlapping samples and requires careful tuning of its added modules.

2.2.3 Hybrid AL strategies

Against the limitations of uncertainty-based methods, hybrid strategies try to find a balance between uncertainty and diversity measures to identify the most informative samples (Settles, 2009). They usually combine existing approaches. An early study proposed to adaptively choose the best AL strategies from a candidate set of methods (Hsu & Lin, 2015). However, most hybrid methods first compute model uncertainty before ensuring sample diversity through a similarity metric. For instance, Suggestive Annotation (Yang *et al.*, 2017) applies core-set selection on a subgroup of the most uncertain samples obtained through bootstrapping. BADGE (Ash *et al.*, 2020) uses gradient embeddings to account for uncertainty (uncertain samples will have a gradient embedding with higher norm) and employs Kmeans++ initialization on top of these embeddings to ensure the diversity of selected samples. Nath *et al.* (2021) combine prevailing mutual information and entropy measures to ensure diversity and optimize training by duplicating difficult samples. Observing that uncertainty-based approaches fail to exploit the data distribution and representative-based approaches are task-agnostic, Task-aware VAAL (Kim *et al.*, 2021) incorporates the uncertainty measure proposed by the method Learning Loss (Yoo & Kweon, 2019) to VAAL's (Sinha *et al.*, 2019) latent representation. While these studies rely on a two-step approach, Sourati *et al.* (2019) directly solve an optimization problem for batch-mode sampling, yielding a distribution of candidate samples rather than specific examples. However, just like representative-based AL strategies, most of these works are difficult to scale due to their computational complexity (Ash *et al.*, 2020; Nath *et al.*, 2021; Sourati *et al.*, 2019;

Yang *et al.*, 2017). Alternatively, they may require external modules, which increase the range of parameters to tune and learn (Kim *et al.*, 2021).

2.2.4 AL for medical image segmentation

High-dimensional data remains a particularly challenging problem in AL (Ren *et al.*, 2021). Therefore, most studies on AL applied to computer vision primarily focus on low-dimensional annotation tasks such as image classification (Gal *et al.*, 2017; Wang *et al.*, 2017; Sener & Savarese, 2018; Beluch *et al.*, 2018; Sourati *et al.*, 2019; Gao *et al.*, 2020; Ash *et al.*, 2020; Zhang *et al.*, 2022). Moreover, approaches tackling pixel-wise annotations predominantly address the segmentation of natural images (Sinha *et al.*, 2019; Huang *et al.*, 2021a; Kim *et al.*, 2021; Xie *et al.*, 2022).

Earlier work applying AL to medical image segmentation has relied on geometric priors to query planes or supervoxels of maximum uncertainty, without adopting deep learning-based models (Top *et al.*, 2011; Konyushkova *et al.*, 2015, 2019). One of the initial deep AL frameworks for this task, Suggestive Annotation (Yang *et al.*, 2017), uses bootstrapping to estimate sample uncertainty and a greedy cosine similarity measure to evaluate the similarity between the candidate set and the unlabelled pool. Similarly, Li & Yin (2020) propose to select a candidate set with a high disagreement among the predictions of K models and a minimal discrepancy between the labelled and unlabelled sets. Instead of relying on multiple models, Ozdemir, Peng, Tanner, Fuernstahl & Goksel (2018) employ a Bayesian network with Monte Carlo dropout to compute prediction variance, and adopt a Borda-count-based sampling strategy to find the best-ranked candidates in terms of uncertainty and representativeness. An extension of this approach instead computes the representativeness with an infoVAE (Zhao, Song & Ermon, 2019) for a maximum-likelihood sampling in the latent space (Ozdemir, Peng, Fuernstahl, Tanner & Goksel, 2021). Nath *et al.* (2021) build a mutual information-based metric, computed between the labelled and unlabelled pools, to ensure the diversity of the candidate set. However, these approaches tend to be computationally expensive and challenging to scale to large datasets. Instead of relying on a 2-step approach, Sourati, Gholipour, Dy, Kurugol & Warfield (2018)

propose a method based on the Fisher information to directly solve an optimization problem that outputs a distribution to sample from. Alternative approaches have opted for membership query synthesis as an AL strategy, producing synthetic samples for annotation. For instance, Mahapatra, Bozorgtabar, Thiran & Reyes (2018) employ a conditional generative adversarial network (cGAN) to generate realistic-looking chest X-ray images conditioned on real images, and a Bayesian neural network to select which ones would be most informative when used as training data. Other approaches propose a sample selection strategy which also covers the initial labelled set (Smailagic *et al.*, 2018; Nath, Yang, Roth & Xu, 2022; Li *et al.*, 2023). Recently, a comparative study of existing strategies for 3D medical image segmentation found that random sampling and strided sampling served as particularly strong baselines for this type of task (Burmeister *et al.*, 2022). The study also observed that representative-based strategies did not perform well in early stages, which the authors attribute to poor feature vectors generated by the model trained on very few labelled samples.

2.3 Methods

Given a labelled set $\mathcal{D}_L = \{(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\}_{j=1}^{N}$, with data $\mathbf{x} \in \mathbb{R}^{H \times W}$ and segmentation mask $\mathbf{y} \in \mathbb{R}^{C \times H \times W}$ (H and W are respectively the image height and width, and C is the number of classes), we train a fully-supervised segmentation model $f_\theta(\cdot)$ parameterized by θ with labelled samples from $\mathcal{D}_L$.

After training the model f_θ with $\mathcal{D}_L$ (corresponding to one training cycle), we select B samples from the unlabelled set $\mathcal{D}_U = \{\mathbf{x}_u^{(j)}\}_{j=1}^{M}$. These samples are annotated by an oracle before being added to the labelled training set $\mathcal{D}_L$. The new labelled and unlabelled sets are updated such that $|\mathcal{D}_L| = N + B$ and $|\mathcal{D}_U| = M - B$. This iterative process is repeated until the total annotation budget is exhausted.

Our AL method addresses the problem of uncertainty-based strategies, generally prone to query samples with redundant information, in a simple and computationally-efficient way. It builds upon our use of stochastic batches and operates in two stages to ensure a guided sampling

diversity, summarized in Fig. 2.1. First, we generate a pool of Q batches, each containing B samples chosen uniformly at random from $\mathcal{D}_U$:

$$Batch^{(i)} = \{\mathbf{x}_u^{(i_1)}, \mathbf{x}_u^{(i_2)}, ..., \mathbf{x}_u^{(i_B)}\} \sim Uniform(\mathcal{D}_u, B) \tag{2.1}$$

For each generated batch, an uncertainty score is assigned to each unlabelled sample it contains, according to the current model $f_{\hat{\theta}}$ and the chosen uncertainty metric ($Uncert$):

$$\forall k = 1, ..., B: \quad u_{score}^{x_u^{(i_k)}} = Uncert\ (f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}). \tag{2.2}$$

The mean u_{score} is computed across each generated batch:

$$u_{score}^{Batch^{(i)}} = \frac{1}{B} \sum_{k=1}^{B} u_{score}^{x_u^{(i_k)}}. \tag{2.3}$$

The batch with the highest mean score yields the set of annotation candidates $X_{candidate}$, such that:

$$X_{candidate} \leftarrow \underset{Batch^{(i)}}{\text{argmax}} \left(u_{score}^{Batch^{(i)}} \right). \tag{2.4}$$

The algorithm for our stochastic batch selection strategy is presented in Alg. 2.1.

Algorithm 2.1 Uncertainty-based sampling with Stochastic Batches

Input: $\mathcal{D}_u,\ Q,\ B$

for $\mathbf{x}_u \in \mathcal{D}_u$ **do**
  $u_{score} \leftarrow Uncert(f_{\hat{\theta}}, \mathbf{x}_u)$;
end for
for $i \leftarrow 1$ *to* Q **do**
  $Batch^{(i)} = \{\mathbf{x}_u^{(1)}, ..., \mathbf{x}_u^{(B)}\} \leftarrow Uniform(\mathcal{D}_u, B)$;
  $u_{score}^{Batch^{(i)}} \leftarrow$ Mean u_{score} over all samples in $Batch^{(i)}$;
end for
$X_{candidate} \leftarrow \text{argmax}_{Batch^{(i)}} \left(u_{score}^{Batch^{(i)}}\right)$;

Output: $X_{candidate}$

2.4 Experiments

We assess the benefits of our proposed stochastic batches on a medical image segmentation task. Our evaluation compares the performance with and without stochastic batches of models trained with different uncertainty-based AL strategies. These strategies include Entropy-based sampling (Shannon, 1948), Dropout-based sampling (Gal & Ghahramani, 2016b), Test-time augmentation (TTA)-based sampling (Gaillochet, Desrosiers & Lombaert, 2022) and sampling based on Learning Loss (Yoo & Kweon, 2019), defined in more details in Sec. 2.4.3.2.1. We start by evaluating the gains of our stochastic batch sampling on two medical image datasets. We then assess the robustness of our method to the training and sampling procedure through a series of ablation studies on the initial labelled set size, training hyperparameters, sampling budget and stochastic pool size.

2.4.1 Datasets

We validate our method on two complementary datasets with different types of challenges: 1) the Prostate MR Image Segmentation (PROMISE) 2012 challenge (Litjens *et al.*, 2014), for prostate segmentation, with varying degrees of pixel intensity distributions (as pictured in Fig. 2.6), and

2) the Medical Segmentation Decathlon (Antonelli *et al.*, 2022) for the segmentation of anterior and posterior hippocampus, with varying degrees of anatomical shapes.

The PROMISE12 dataset contains MRI data from 50 patients, both healthy (or with benign diseases) and pathological (with prostate cancer). Each volume is converted to 2D images by slicing along the short axis. Images are then resampled to 1.0 mm isotropic resolution and resized to 128×128 pixels.

Similarly, the Medical Segmentation Decathlon contains hippocampus data from 260 patients. The MRI volumes are converted to 2D images, which are resized to 50×50 pixels while kept to the original 1.0 mm isotropic resolution. The pixel intensity of both datasets is normalized based on the 1% and 99% percentiles for each scan.

We test our model on 10 patient volumes from the prostate dataset and 50 from the hippocampus dataset, all selected uniformly at random. This yields 248 and 1757 test images, respectively. Our validation uses 109 prostate images composing 5 volumes, and 350 hippocampus images composing 10 volumes. Since active learning aims to minimize the amount of labelled data, we only use this validation set for hyperparameter search purposes. Our ablation studies show that our method remains advantageous under different hyperparameter settings. Our training set, labelled and unlabelled, comprises 1020 prostate images from 35 patients and 7163 hippocampus images from 200 patients.

2.4.2 Evaluation metrics

We evaluate our method on test volumes (3D) and individual images from these volumes (2D). We use both pixel overlap-based metrics and distance-based metrics.

In terms of overlap-based metrics, we use the well-known Dice similarity coefficient (DSC), which ranges from 0% (zero overlap) to 100% (perfect overlap):

$$\mathrm{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| \cup |Y|} \tag{2.5}$$

In our results, we report the DSC averaged over all non-background channels.

The Hausdorff distance (HD) measures the quality of the segmentation by computing the maximum shortest distance between a point from the prediction contour and a point from the target contour. Since the Hausdorff distance tends to be sensitive to outliers, we use a more robust variant which considers the 95^{th} percentile instead of the true maximum (HD95). Given $\mathrm{d}(x, Y)$ the minimum distance from the boundary pixel x to the region Y, we get:

$$\mathrm{HD95}(X, Y) = \max \left\{ 95^{th}_{x \in X}\, \mathrm{d}(x, Y),\ 95^{th}_{y \in Y}\, \mathrm{d}(X, y) \right\} \tag{2.6}$$

2.4.3 Implementation details

Medical annotations for image segmentation are typically performed on all slices of a given image volume (Ozdemir *et al.*, 2021). However, to optimize the limited annotation resources, we conduct slice-based active learning and select individual images for annotation after every cycle. We start each experiment by training our model with 10 labelled images, randomly sampled from the unlabelled set before annotation. Setting the budget to $B = 10$, we use our AL strategy to select 10 new samples from the unlabelled set, annotate them and add them to the existing labelled set. This process corresponds to the first AL cycle, which we repeat for a fixed number of cycles. Similarly to the experimental setting of previous studies, we retrain the model from scratch after each AL cycle to evaluate model performance in a consistent way (Budd *et al.*, 2021).

Random processes such as model initialization or data shuffling are seeded. We repeat each experimental setup with 5 different seeds and report the mean and standard deviation of these runs as our result. Experiments were run on NVIDIA V100 and A6000 GPUs, with CUDA 10.2

and CUDA 12.0, respectively. We implement the methods using Python 3.8.10 with the PyTorch framework.

2.4.3.1 Training

State-of-the-art methods in medical image segmentation have often adopted UNet-based architectures (Ronneberger *et al.*, 2015). Accordingly, we use a standard 4-layer UNet as a proxy for widely used architectures in our segmentation model, with dropout ($p = 0.5$), batch normalization and a leaky ReLU activation function. Employing such a model also focuses the evaluation on the improvement due to our stochastic batch strategy instead of measuring the performance of a backbone. However, without loss of generality, the use of alternative segmentation models could also be envisioned for our AL approach.

The model is trained for 75 epochs in all experiments, each iterating over 250 batches (training samples can appear in several batches), with a batch size of 4. Training is hence carried out for a fixed $75 \times 250 = 18,750$ steps in all experiments, ensuring a fairer comparison of model performance between AL cycles.

We optimize a supervised CE loss with the Adam optimizer (Kingma & Ba, 2015). We apply a gradual warmup with a cosine annealing scheduler (Loshchilov & Hutter, 2017; Goyal *et al.*, 2018) to control the learning rate. During training, we use data augmentations on the input, with parameters d and ϵ, where d is the degree of rotation in 2D, and ϵ models Gaussian noise.

When not testing for their impact, we keep the training hyperparameters fixed. We fix the initial learning rate $LR = 10^{-6}$ with optimizer weight decay set to 10^{-4}. The scheduler increments the learning rate by a factor 200 during the first 10 epochs. For augmentations, we set $d \sim \mathcal{U}(-10, 10)$ and $\epsilon \sim \mathcal{N}(0, 0.01)$.

Since active learning aims to minimize the amount of labelled data needed to train the model, we minimize the use of the validation set and avoid its use to select the final model. Our final model is instead the model obtained after the last training epoch.

2.4.3.2 Active learning sampling

2.4.3.2.1 Baselines

We compare our stochastic batches strategy with random sampling (RS), Core-set (Sener & Savarese, 2018), and four purely uncertainty-based methods:

- Entropy-based uncertainty (Shannon, 1948), which computes the entropy on the predicted output probabilities:

$$Uncert\ (f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}) = -\sum_i p(y_i|\mathbf{x}_u^{(i_k)}, \hat{\theta})\ \log p(y_i|\mathbf{x}_u^{(i_k)}, \hat{\theta}); \tag{2.7}$$

- Dropout-based uncertainty (Gal & Ghahramani, 2016b), using the divergence of K predictions obtained by multiple inferences with dropout d:

$$Uncert\ (f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}) = Div\big(f_{\hat{\theta},d_1}(\mathbf{x}_u^{(i_k)}),\ ...,\ f_{\hat{\theta},d_K}(\mathbf{x}_u^{(i_k)})\big); \tag{2.8}$$

- Test-time Augmentation (TTA)-based uncertainty (Gaillochet *et al.*, 2022), which measures the divergence of predictions obtained for K transformations Γ to the input:

$$Uncert\ (f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}) = Div\ \big(\Gamma_1^{-1}[f_{\hat{\theta}}(\Gamma_1(\mathbf{x}_u^{(i_k)}))],\ ...,\ \Gamma_K^{-1}[f_{\hat{\theta}}(\Gamma_K(\mathbf{x}_u^{(i_k)}))]\big); \tag{2.9}$$

- Learning Loss uncertainty (Yoo & Kweon, 2019), which trains an external module $L_{\tilde{\theta}}$ to predict the target losses from a feature set h extracted from the hidden layers of $f_{\hat{\theta}}$:

$$Uncert\ (f_{\hat{\theta}}, L_{\tilde{\theta}}, \mathbf{x}_u^{(i_k)}) = L_{\tilde{\theta}}(h(\mathbf{x}_u^{(i_k)})). \tag{2.10}$$

These purely uncertainty-based methods query batches made of the most uncertain samples according to a sample-level uncertainty measure.

Similarly to Gaillochet *et al.* (2022), as our divergence measure for Dropout-based and TTA-based sampling, we use a standard Jensen–Shannon divergence (JSD) on the output probability

maps obtained from $K = 8$ inferences. For TTA, augmentations Γ include Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.01)$ and rotation. To simulate more realistic transformations in medical data, we replace the 90 degrees rotations in Gaillochet *et al.* (2022) with rotations of angle $d \sim \mathcal{U}(-10, 10)$ degrees. The training parameters used for the approach based on Learning Loss (Yoo & Kweon, 2019) were obtained by grid search on 10 labelled samples. We kept these parameters fixed in all our experiments.

2.4.3.2.2 Stochastic batches

We generate the pool of stochastic batches by iteratively sampling B unlabelled images uniformly at random and without replacement. In other words, we divide the unlabelled samples into Q pools of B samples. Hence the stochastic pool has size $Q = \text{floor}\left(|\mathcal{D}_U|/B\right)$, and it reduces in size with the number of AL cycles.

2.5 Results

2.5.1 AL performance on the Prostate and Hippocampus datasets

We validate our proposed stochastic batch sampling strategy by looking at the AL performance over 5 different initial labelled sets chosen uniformly at random from the training set. Tab. 2.1 shows the average results over all AL cycles for both Prostate and Hippocampus data. Note that the standard deviations given in the table tend to be large because they are averaged over multiple initial labelled sets, initialization seeds and AL cycles. For all methods and metrics except for TTA on Prostate with the 95% Hausdorff distance metric, stochastic batch sampling constantly provides improved performance over its purely uncertainty-based counterpart, both in terms of overlap-based and distance-based metric.

We also observe that stochastic batch sampling outperforms both random sampling and Core-set (Sener & Savarese, 2018), a diversity-based AL approach. This is corroborated by Fig. 2.2 and Fig. 2.3, which show that Dropout with our stochastic batches outperforms all other baseline

Table 2.1 **Overall improvements with Stochastic Batches over varying initial labelled samples**. Mean model performance over all AL cycles. We show the mean (std) Dice score (DSC, higher is better) and 95% Hausdorff distance (HD95, lower is better) over 3D test volumes and 2D test images. The results are averaged over 5 initial labelled sets chosen uniformly at random and 6 AL cycles (we omit results with the initial labelled set as they are similar across all methods). A * indicates the statistical significance of the result with a p-value < 0.05 given a paired permutation test

		Prostate			Anterior Hippocampus			Posterior Hippocampus		
		3D DSC	**2D DSC**	**3D HD95**	**3D DSC**	**2D DSC**	**3D HD95**	**3D DSC**	**2D DSC**	**3D HD95**
RS		68.83 ±15.99	67.94 ±8.28	7.032 ±3.734	77.42 ±1.67	75.45 ±1.13	4.09 ±0.47	76.43 ±0.80	**70.02** ±1.62	4.51 ±0.70
Core-set (Sener & Savarese, 2018)		68.84 ±17.37	65.87 ±7.31	7.64 ±2.73	78.83 ±3.25	73.14 ±1.20	4.45 ±0.46	75.32 ±5.46	66.45 ±1.29	4.52 ±0.53
Entropy (Shannon, 1948)	w/o SB	67.01 ±16.68	66.88 ±8.62	7.026 ±4.271	78.22 ±1.90	75.03 ±0.97	3.79 ±0.23	74.68 ±1.60	65.70 ±1.66	5.10 ±1.10
	Ours	71.27* ±17.39	68.99* ±9.03	6.689* ±3.143	79.25* ±0.86	75.84 ±0.86	**3.72** ±0.15	76.23* ±0.87	69.01* ±1.97	**3.85*** ±0.31
Dropout (Gal & Ghahramani, 2016b)	w/o SB	67.69 ±17.16	67.07 ±9.51	6.964 ±4.952	78.22 ±1.28	74.29 ±1.10	4.04 ±0.33	74.45 ±1.20	66.78 ±2.06	4.77 ±1.19
	Ours	**72.59*** ±14.96	**69.64*** ±8.05	**6.583*** ±3.177	**79.28*** ±0.83	**76.36*** ±0.69	3.73 ±0.10	76.27* ±0.85	**68.94*** ±1.15	**3.88*** ±0.39
TTA (Gaillochet *et al.*, 2022)	w/o SB	64.07 ±21.13	65.85 ±10.25	6.918* ±4.794	77.31 ±3.24	73.66 ±1.08	4.10 ±0.54	73.84 ±2.01	64.94 ±0.59	5.07 ±0.93
	Ours	69.71* ±17.59	68.00* ±9.02	7.188 ±3.173	78.86* ±0.94	75.25 ±1.41	4.07 ±0.31	**76.44*** ±0.90	67.08* ±1.01	4.43* ±0.40
Learning Loss (Yoo & Kweon, 2019)	w/o SB	53.88 ±21.51	60.22 ±10.36	9.139 ±6.439	62.54 ±1.38	69.70 ±0.92	5.94 ±0.59	61.57 ±2.83	62.82 ±1.14	5.87 ±0.12
	Ours	65.29* ±17.72	65.72* ±8.94	7.816* ±4.384	72.09* ±2.73	74.32* ±0.85	4.51* ±0.60	71.23* ±1.29	67.75* ±1.21	5.16 ±0.63

methods in terms of 3D dice score, in almost all AL cycles. In addition, Tab. 2.2 gives the average time required by each strategy to provide a candidate set for annotation from the Hippocampus dataset. We see that using stochastic batches does not increase the sampling time of uncertainty-based methods. Furthermore, sampling with our proposed method is always much faster than with Core-set.

When looking at each dataset in more detail, the pairwise results on the Prostate dataset, shown in Fig. 2.4, validate the effectiveness of our method against different initial labelled sets. Averaged over 25 experiments with varying initial labelled sets and initialization seeds, our stochastic

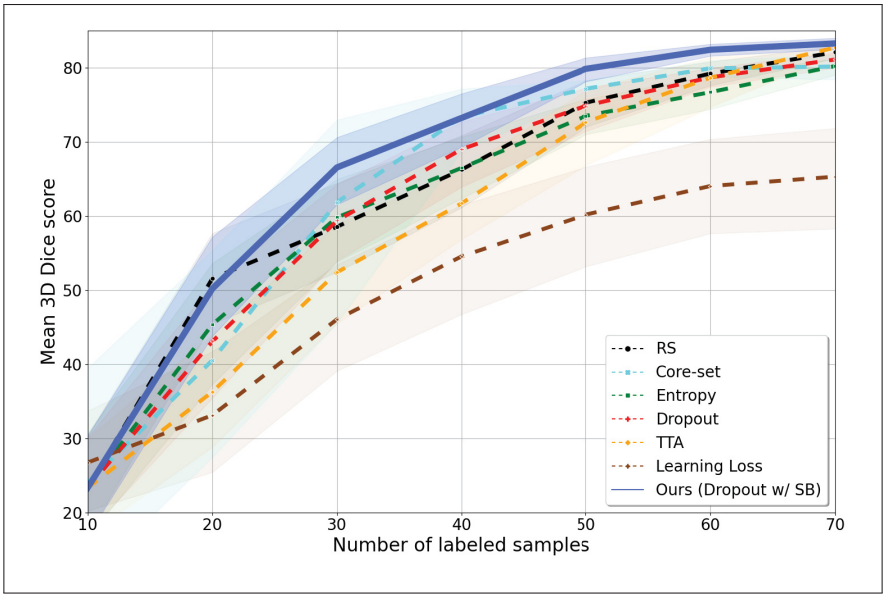


Figure 2.2 **Overall AL performance on the Prostate dataset**. Our best stochastic batch sampling method (full-blue) outperforms all other methods, including Core-set and random sampling (RS)

Table 2.2 **Sampling time.** Mean sampling time computed over all AL cycles, for the Hippocampus dataset

	RS	Core-set	Entropy		Dropout		TTA		Learning Loss	
			w/o SB	Ours	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours
Time (min.)	0.00	0.71	0.12	0.11	0.58	0.58	0.37	0.37	0.16	0.18

batch querying (blue, full lines) improves the model's performance of purely uncertainty-based strategies (orange, dashed lines). For all considered AL strategies, selecting the most uncertain batch of samples rather than the most uncertain individual samples improves the model's overall performance. The 3D dice score is always boosted, either over the score obtained by random sampling (grey, dotted) or to a level similar to that of a random sampling if the score were originally much lower, such as in the case of the Learning Loss. Indeed, Learning Loss has noticeably lower performance compared to Entropy, Dropout and TTA-based sampling. The Learning Loss approach involves backpropagating the gradient through both the task model and

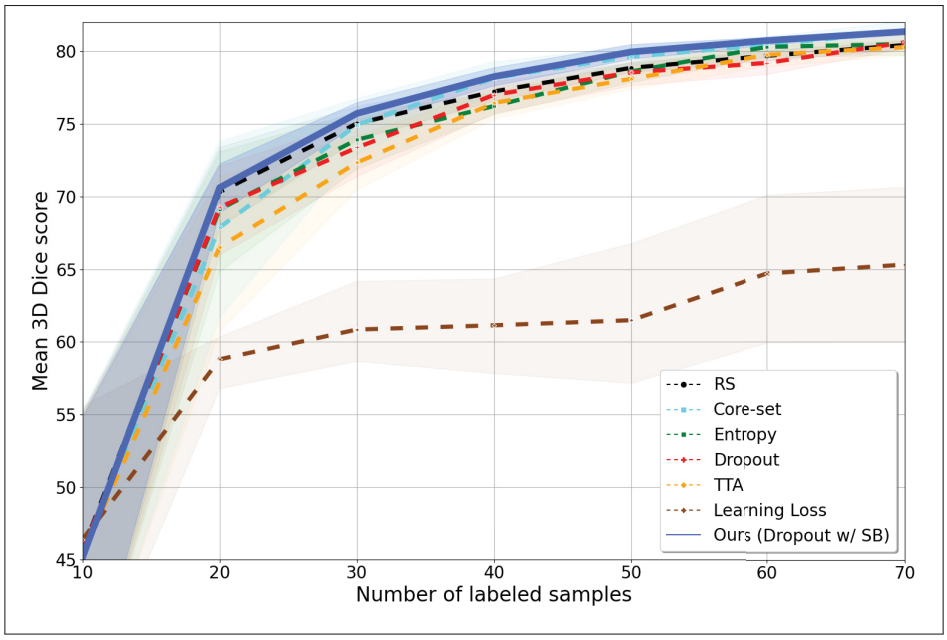


Figure 2.3 **Overall AL performance on the Hippocampus dataset**. Our best stochastic batch sampling method (full-blue) outperforms all other methods, including Core-set and random sampling (RS)

loss module during training. Both are updated simultaneously, which means that training the loss module affects the training of the task model and vice versa. For comparability reasons and following most works in AL, we tuned the hyperparameters such that the best validation performance was obtained on the first AL cycle (with the initial labelled set). We believe this could explain the poorer performance of Learning Loss with an increasing number of labelled samples. Similar observations can be made from Hippocampus data, as shown in Fig. 2.5.

We also visually investigate the benefits of using our stochastic batches with an uncertainty-based sample selection. In Fig. 2.6, we show two sets of candidate samples from the Prostate dataset identified by Entropy-based sampling, with and without our stochastic batches. The first two columns show samples selected by identifying the most uncertain randomly generated batch. The last two columns depict the most certain queried samples based on the individual entropy of their predicted output probabilities. While the samples from the first two columns seem more diverse, with more variety in the candidate set, the third column contains nearly identical samples. Indeed, tracking the first four images of the column to their corresponding 3D volume

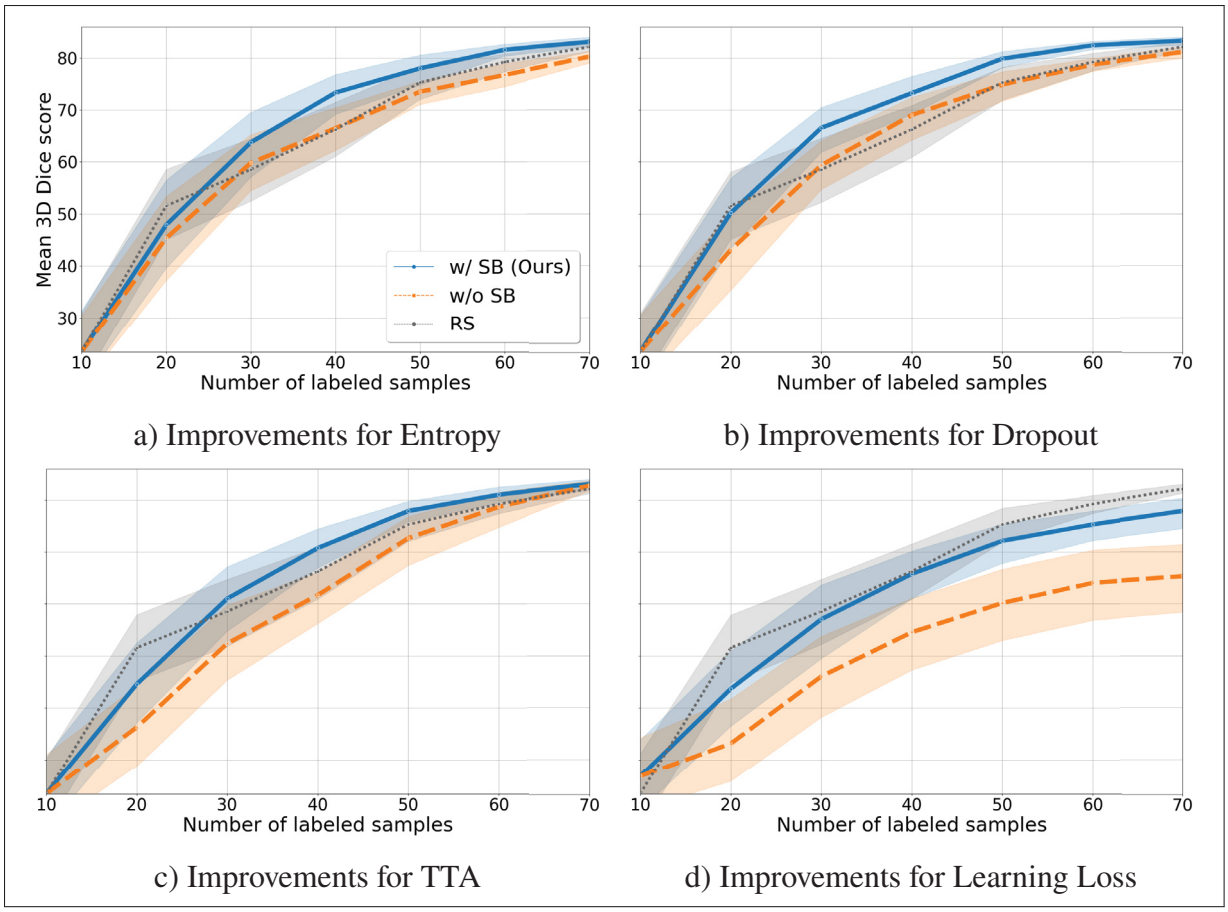


Figure 2.4 **Individual improvements with Stochastic Batches on the Prostate dataset**. Active learning results in terms of 3D test dice score and corresponding 95% confidence interval. The results are averaged over 5 different initial labelled sets and 5 initialization seeds. Depicted are the results for sampling based on a) Entropy, b) Dropout, c) Test-time augmentation and d) Learning Loss. The active learning selection is shown with (blue, full) and without (orange, dashed) stochastic batches, and random sampling is plotted in dotted grey. Stochastic batches improve the model performance of purely uncertainty-based AL strategies, regardless of the initial labelled set, repeatedly outperforming random sampling

shows that the slices were taken from the MRI volume of the same patient. This confirms our claim that purely uncertainty-based strategies are likely to select very similar samples and that our stochastic batch sampling reduces the probability of querying samples with highly overlapping information.

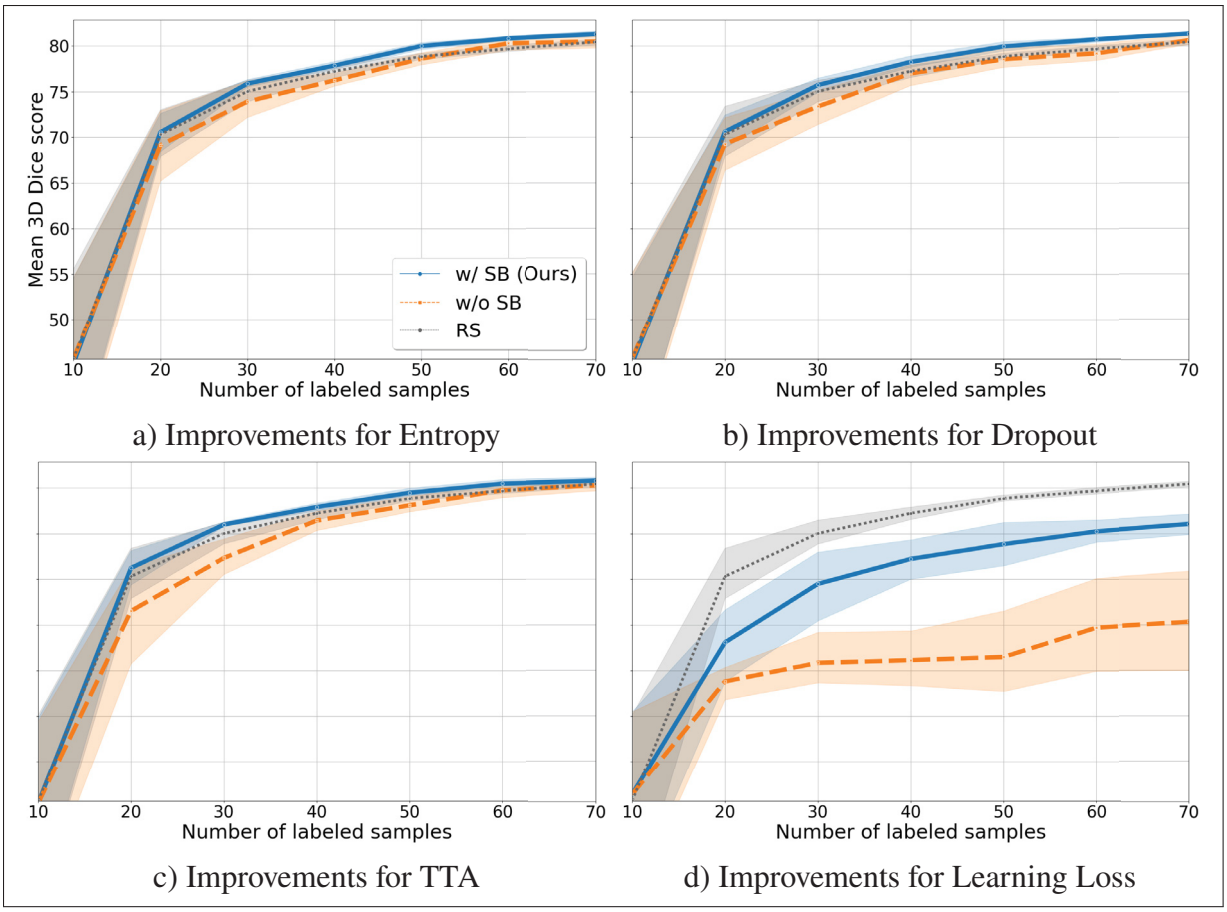


Figure 2.5 **Individual improvements with Stochastic Batches on the Hippocampus dataset**. Active learning results on the Hippocampus dataset in terms of 3D test dice score and corresponding 95% confidence interval. The results are averaged over 5 different initial labelled sets. Depicted are the results for sampling based on a) Entropy, b) Dropout, c) Test-time augmentation and d) Learning Loss. Sampling with Stochastic batches (blue, full) improves the model performance of purely uncertainty-based AL strategies (orange, dashed), regardless of the initial labelled set, boosting it above random sampling (grey, dotted) in the majority of cases

Finally, we examine the impact of our selection strategy on the segmentation of test data. In Fig. 2.7, we see that the model trained on images selected via our stochastic batch sampling method outputs better anterior and posterior hippocampus segmentations. By the fourth cycle, the segmentation reaches a mean DSC (over both classes) of 81.15%, compared to the 68.03% obtained via a purely Entropy-based sampling.

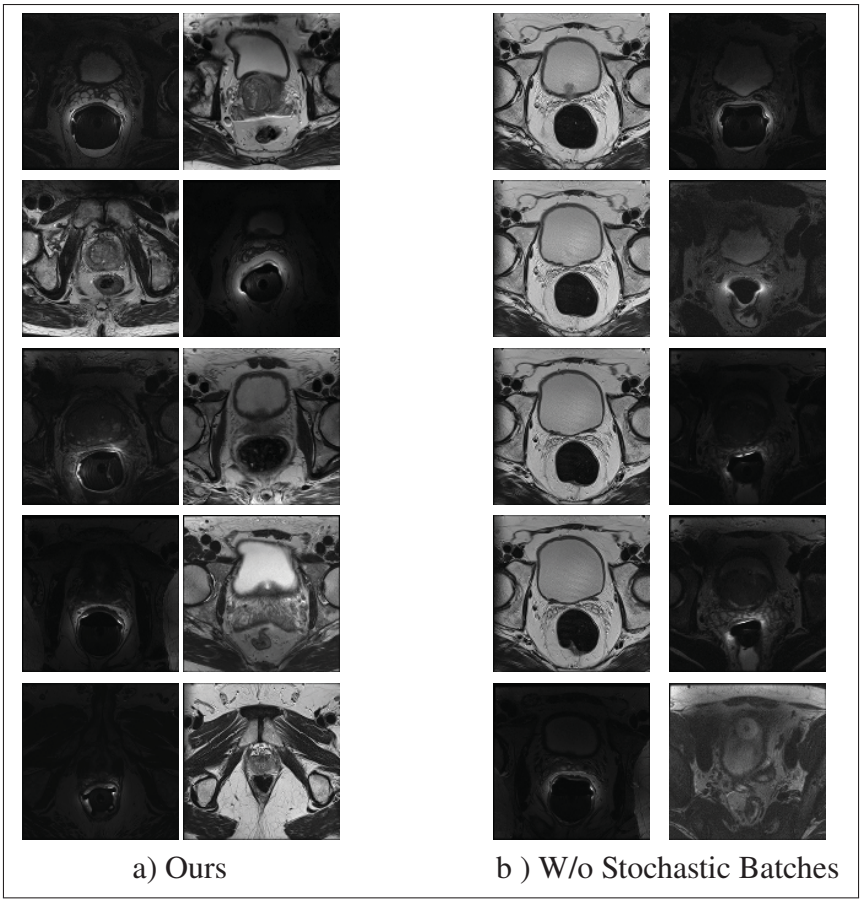


Figure 2.6 **Examples of candidate batches**. Candidate samples were selected by Entropy-based AL sampling a) with and b) without stochastic batches. While the candidate batch obtained via purely uncertainty-based sampling contains similar samples, selection with stochastic batches reduces the number of redundancies

2.5.2 Ablation experiments on the Prostate dataset

To evaluate the robustness of our method to different experimental settings, we perform a series of ablation studies on the Prostate dataset, evaluating the impact of the initial labelled set size, training hyperparameters, sampling budget and stochastic pool size.

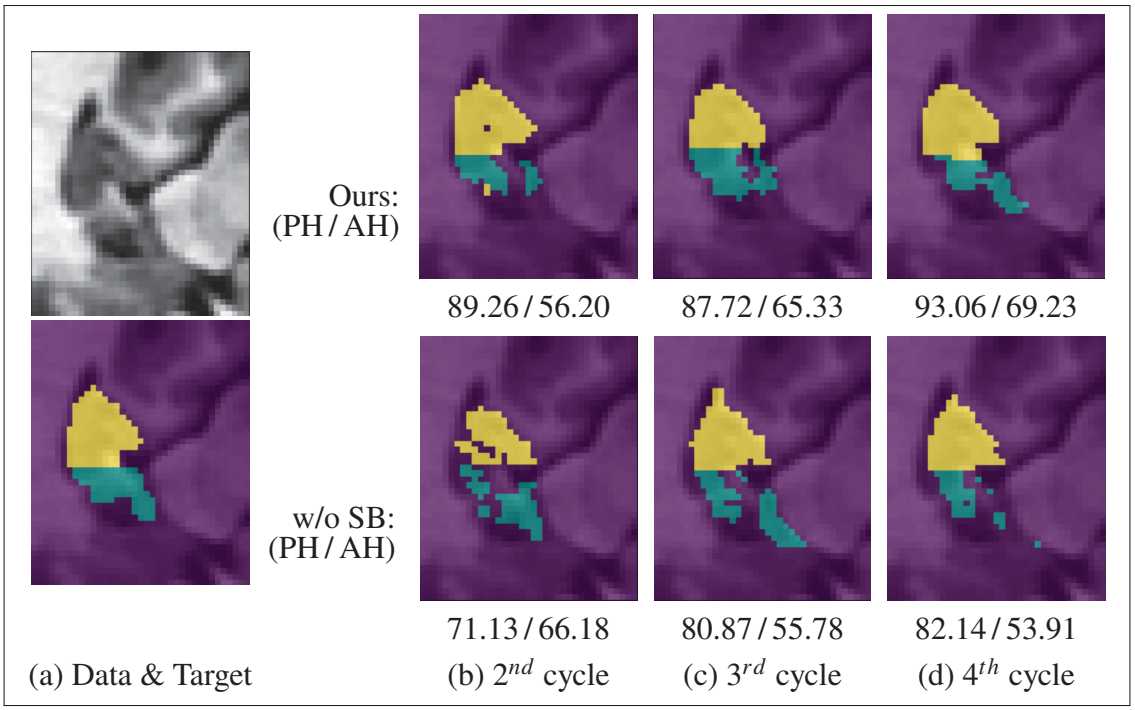


Figure 2.7 **Segmentation of a Hippocampus test sample across AL cycles**. The 2D dice score (DSC) is given for each predicted segmentation, both for the posterior hippocampus (PH, yellow) and the anterior hippocampus (AH, blue). At every AL cycle, the model trained on labelled samples selected with our stochastic batches (top row) predicts segmentations closer to the target mask (leftmost) compared to its purely Entropy-based counterpart (bottom row)

2.5.2.1 Impact of initial labelled set size

For our first ablation study, we validate the performance of models trained on initial labelled sets of varying sizes. For each given initial labelled set size, the experiment is repeated with 5 initialization seeds controlling the initial labelled samples used, the model initialization and the training updates. Table 2.3 gives the average model performance over 6 AL cycles. We observe that our stochastic batch selection strategy improves upon purely uncertainty-based selection also when we vary the initial number of labelled samples.

Table 2.3 **Overall improvements with Stochastic Batches for initial labelled sets of different sizes**. Mean model performance on the Prostate data over all AL cycles for initial sets of different sizes. We show the mean (std) Dice score (higher is better) over 3D test volumes (3D DSC). The results are averaged over 6 AL cycles (we omit results for the first AL cycle since all strategies share the same initial set). A * indicates the statistical significance of the result with a p-value < 0.05 given a paired permutation test

		Entropy (Shannon, 1948)		Dropout (Gal & Ghahramani, 2016b)		TTA (Gaillochet *et al.*, 2022)		Learning Loss (Yoo & Kweon, 2019)	
	RS	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours
5 initial samples	71.22 ±15.09	65.18 ±14.43	**72.36*** ±16.19	61.69 ±18.15	**71.45*** ±15.41	64.16 ±16.43	**67.67** ±16.70	55.06 ±21.90	**66.85*** ±19.30
10 initial samples	71.08 ±11.70	66.21 ±13.79	**73.78*** ±10.73	65.09 ±15.63	**73.30*** ±14.95	70.23 ±12.35	**73.01** ±12.24	48.51 ±9.83	**60.73*** ±9.10
15 initial samples	75.21 ±7.27	0.7319 ±7.54	**74.21** ±7.85	72.90 ±6.96	**76.81*** ±8.34	**72.00** ±10.86	71.53 ±8.84	58.19 ±12.09	**72.84*** ±10.32
20 initial samples	76.00 ±7.19	76.13 ±5.55	**80.24*** ±4.19	77.47 ±6.38	**79.81*** ±05.54	74.59 ±10.15	**78.04** ±7.89	69.81 ±6.74	**75.51*** ±6.14
25 initial samples	77.07 ±4.39	77.73 ±3.79	**79.71*** ±4.37	77.44 ±4.31	**81.08*** ±5.20	76.81 ±9.03	**78.32** ±5.88	73.61 ±5.27	**77.65*** ±5.52

2.5.2.2 Impact of training hyperparameters

Active Learning methods typically tune hyperparameters using an initial labelled set, maintaining these settings throughout all AL cycles. However, these parameters might be sub-optimal for subsequent training cycles as more labelled data becomes available. We hence explore the robustness of stochastic batches to different yet realistic training hyperparameters. We select five hyperparameter sets, each optimized for labelled set sizes of 10, 50, 100, 150, and 200. These sets included diverse augmentation parameters, scheduling parameters and loss function weights.

Results in Fig. 2.8 reveal that our stochastic batch sampling noticeably improves the performance of purely uncertainty-based sampling, particularly in the first 3 or 4 AL cycles. In addition, the spread of 3D dice scores tends to be narrower with our method than with a purely uncertainty-based sampling, showing that our strategy tends to be more stable.

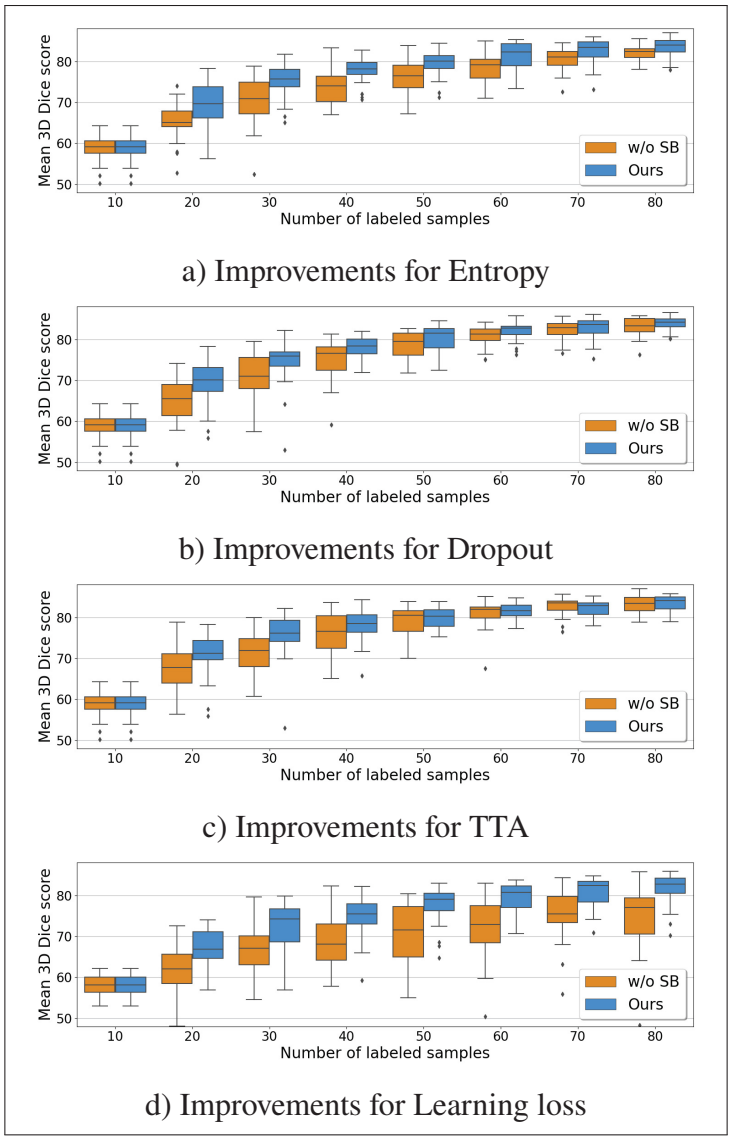


Figure 2.8 **Improvements with Stochastic Batches over varying hyperparameters**. Box plot of active learning results on Prostate data in terms of 3D test dice score, given over 5 training hyperparameters sets and 5 initialization seeds. Depicted are the results for sampling based on a) Entropy, b) Dropout, c) Test-time augmentation and d) Learning loss. The AL selection is shown with (blue) and without (orange) stochastic batches. Our stochastic batches improve the model performance of purely uncertainty-based AL strategies and boost performance, even with variations in hyperparameters

Table 2.4 **Overall improvements with Stochastic Batches over varying training hyperparameters**. Mean model performance on Prostate data over all AL cycles (omitting training with the initial labelled set). We show the mean (std) Dice score (DSC, higher is better) and 95% Hausdorff (HD95, lower is better) distance over 3D test volumes and individual 2D test images. The results are averaged over 7 AL cycles and 5 training hyperparameter sets. * indicates the statistical significance of the result with a p-value < 0.05 given a paired permutation test

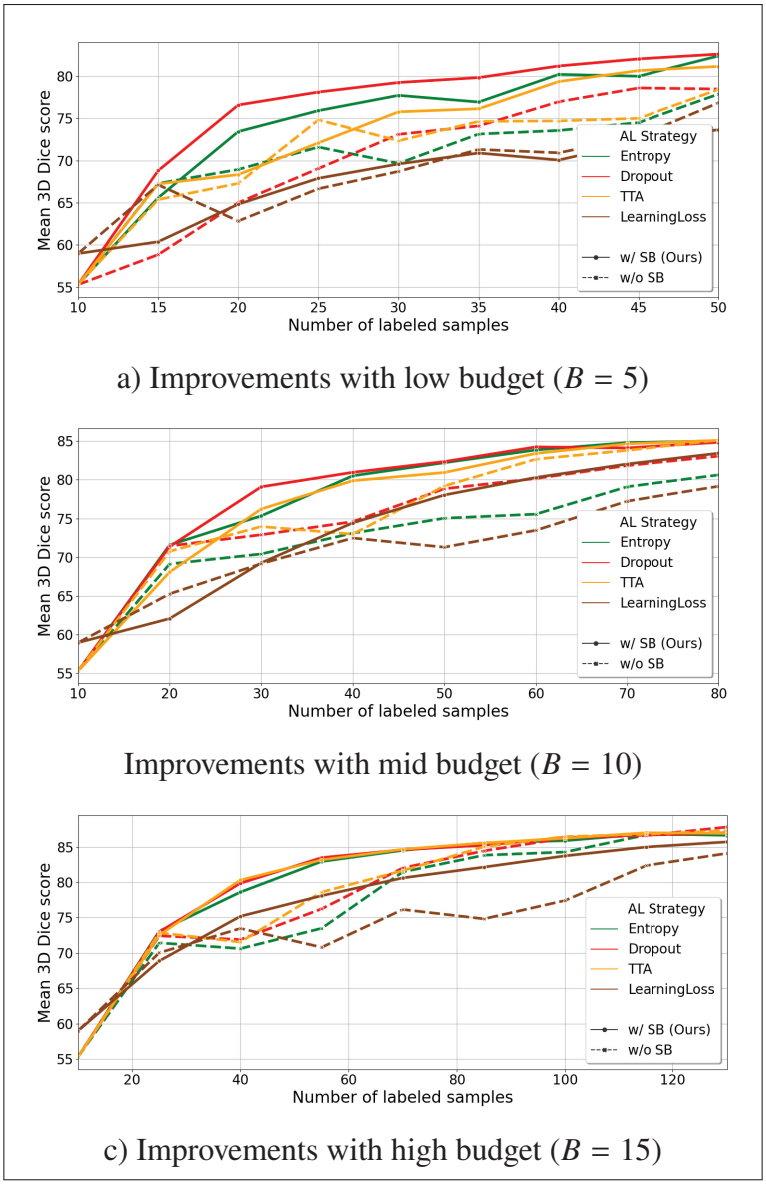
		Entropy		Dropout		TTA		Learning Loss	
	RS	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours
3D DSC (↑ best)	75.57 ±6.48	75.13 ±6.95	**78.44*** ±6.02	76.49 ±7.65	**78.59*** ±6.09	77.33 ±6.92	**78.67*** ±5.53	69.53 ±8.43	**76.25*** ±6.68
2D DSC (↑ best)	68.29 ±6.79	68.90 ±7.34	**71.04*** ±6.51	69.62 ±6.70	**71.08*** ±6.79	70.46 ±7.05	**71.31*** ±5.71	64.27 ±7.23	**69.16*** ±6.80
3D HD95 (↓ best)	7.58 ±3.86	7.87 ±4.28	**6.83*** ±3.31	**6.72** ±2.75	6.74 ±3.29	6.32 ±2.87	**6.13** ±2.82	8.78 ±4.22	**7.85*** ±3.68

The benefit of using our stochastic batches is most evident in the average dice scores over all AL cycles for both test images and volumes, as given in Tab. 2.4. Test-Time Augmentation (TTA) generally performs better with stochastic batches, although the results are not statistically significant for distance-based metrics. This could be due to the fact that we vary the training and regularization hyperparameters while keeping data augmentation parameters fixed for sampling.

2.5.2.3 Impact of sampling budget

We also investigate the robustness of stochastic batches to the sampling budget. Keeping the initial labelled set and training hyperparameters fixed, we run experiments with 5 different sampling budgets, which we keep constant across cycles. In this experiment, since we vary B, images are allowed to be resampled when generating the stochastic batches, and we keep the number of generated batches to a fixed $Q = 100$.

The results shown in Fig. 2.9 reveal that stochastic batches have a more consistent impact on model performance as the budget size increases. With a high budget $B = 15$, the use of stochastic batches constantly improves purely uncertainty-based methods. An improvement is also visible for lower budgets, such as $B = 5$, particularly for the Entropy, Dropout and TTA-based sampling.



a) Improvements with low budget ($B = 5$)

Improvements with mid budget ($B = 10$)

c) Improvements with high budget ($B = 15$)

Figure 2.9 **Improvements with Stochastic Batches given different budget sizes**. Model performance in terms of 3D dice score on test volumes given active learning selection with (solid) and without (dashed) stochastic batches on Prostate data. The results are given for sampling budgets a) $B = 5$, b) $B = 10$ and c) $B = 15$. Depicted are the results for sampling based on Entropy (green) and Dropout (red), TTA (yellow) and Learning Loss (brown). Using stochastic batches during sampling improves the model performance at both low and higher budgets

However, with very low budgets, batch uncertainty is highly influenced by the uncertainty of each individual sample, potentially reducing the benefits of diversity offered by stochastic batches. The selection is dominated by uncertainty, and if the measure for uncertainty is not representative of the true uncertainty of the model, then uninformative samples could be selected and consequently bias the model.

2.5.2.4 Impact of sampling stochastic pool size

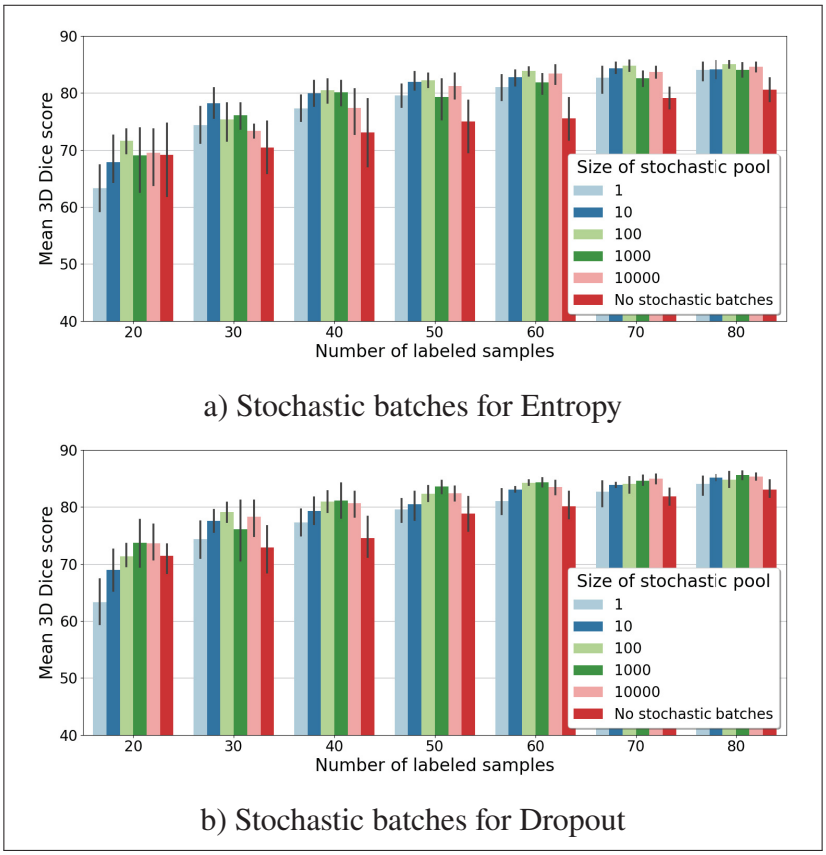


Figure 2.10 **Impact of pool size of Stochastic Batches**. Model performance in terms of 3D dice score on test volumes from Prostate data given stochastic batch pools of different sizes. The error bars (black) corresponds to the 95% confidence interval over 5 experiments with different seed initialization. Depicted are the results 2 popular uncertainty-based AL methods: a) Entropy-based sampling and b) Dropout-based sampling. A medium pool size between 10 to 100 yields some of the most advantageous performances

In our last ablation study, we evaluate the influence of the number of batches in the stochastic pool on the model performance, fixing the initial labelled set, training hyperparameters and sampling budget. Instead of generating $Q = \text{floor}(|\mathcal{D}_u|/B)$ batches, we artificially vary Q. Accordingly, we allow resampling so samples can appear in multiple generated batches. The results for our experiments on Entropy-based and Dropout-based sampling are given in Fig. 2.10. Applying the biggest pool size does not necessarily yield the best performance. On the contrary, the model performs best when the most uncertain batch is selected from a pool containing 10 or 100 different batches. Increasing the pool of choices by 10 or 100 does not lead to significant improvements and can lead to worse performances.

2.6 Discussion

Overall, our results demonstrate that using stochastic batches during uncertainty-based sampling is an efficient strategy to ensure diversity among the selected batch of samples. Furthermore, we experimentally observe that the benefit of using stochastic batches is robust to changes in the initial labelled set, initialization of the model and training hyperparameters, as well as to variations in the sampling budget.

As illustrated in Fig. 2.6, the redundancy of queried samples constitutes one of the main drawbacks of uncertainty-based AL strategies. Their queried samples may indeed convey highly similar information. Hence, the annotation effort on these samples will be suboptimal. If, on the contrary, the most uncertain batches rather than the most uncertain samples are queried, the added diversity within our stochastic batches mitigates the overlap of information and redundancy between samples. Our stochastic scheme adds diversity to the uncertainty-based sampling in AL in a fast, computationally-efficient way, as shown by Tab. 2.2. Our quantitative results demonstrate the advantages of adding such a stochastic scheme in AL in terms of added segmentation accuracy in a low-labelled set regime and reduced number of required training samples.

Previous AL works have observed that the initial labelled pool can significantly impact the training and final performance of AL models (Chen *et al.*, 2022a). Nevertheless, a robust AL method should still perform well regardless of this initial labelled set. The results obtained in our experiment with varying initial labelled sets (Sec. 2.5.1 and Sec. 2.5.2.1) reveal that the performance boost from our stochastic batch sampling strategy is robust to changes in both the initial labelled set and model initialization. On average, selecting the most uncertain batches across AL cycles yields better results than selecting the most uncertain samples. Similarly, Sec. 2.5.2.2 shows that the improvements yielded by stochastic AL batches are also robust to changes in the training and regularization parameters. Hence, our method can maintain efficiency despite changes in the learning environment. These results suggest that using stochastic batches during AL for uncertainty-based sampling can be a reliable and robust AL approach.

Our stochastic batch querying strategy for uncertainty-based AL operates as a balance between a fully random and a purely uncertainty-based selection. While we set $Q = \text{floor}\left(|\mathcal{D}_U|/B\right)$, the stochastic pool size Q can also be directly modified to control the amount of randomness desired in the AL selection. With the smallest pool size ($Q = 1$), our stochastic batch selection is equivalent to random sampling since the single suggested batch will automatically have the highest uncertainty score in the pool. With the biggest pool size ($Q \to \infty$), all possible combinations of samples are available in the pool, and selecting the most uncertain batch of samples is equivalent to selecting the top uncertain samples. In other words, the approach becomes a purely uncertainty-based AL strategy with a larger pool size. As shown in Sec. 2.5.2.4, the benefits of our stochastic batches are apparent in between those extreme Q values, when the sampling strategy combines the informativeness of uncertainty-based sampling with the diversity provided by random sampling. Active learning is an expensive framework to experiment with, given that AL cycles are iterative and that procedures should be repeated to reduce as much as possible the influence of initialization. In this work, we ran multiple experiments with different settings (size and type of initial labelled set, training hyperparameters, stochastic pool size, sampling budget) to test how stable our method was. However, we acknowledge that our experiments do not cover all ranges of possible setups.

2.7 Conclusion

Active learning is particularly relevant in medical image segmentation since manual labelling is highly time-consuming and expensive. This paper addresses three main limitations of AL strategies: the relatively limited literature on AL work for medical image segmentation compared to classification tasks, the tendency of uncertainty-based batch sampling strategies to select very similar samples and the computational burden of diversity-based methods. Instead of employing sample-level uncertainty for candidate selection, we suggest a batch-level approach where uncertainty is computed over randomly generated batches of samples. Using stochastic batches with uncertainty-based sampling is a simple, computational-inexpensive approach to improve the AL candidate selection and, hence, the final model performance. Our method is flexible and easily adaptable to any uncertainty-based AL strategy. In addition, our extensive experiments show that adding stochastic batches improves purely uncertainty-based methods consistently across different experimental setups. Hence, stochastic batching could bring a more reliable advantage over other representative-based works, which have shown significantly varying amounts of robustness in performance (Munjal *et al.*, 2022). Our method could therefore act as a strong baseline to better use the limited annotation time of clinical experts when segmenting medical images.

CHAPTER 3

PROMPT LEARNING WITH BOUNDING BOX CONSTRAINTS FOR MEDICAL IMAGE SEGMENTATION

Mélanie Gaillochet[1,2], Mehrdad Noori[1], Sahar Dastani[1], Christian Desrosiers[1], Hervé Lombaert[2,3]

[1] Department of Software and IT Engineering , École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Mila - Quebec AI Institute,
6666 Rue Saint-Urbain, Montréal, Québec, Canada H2S 3H1
[3] Department of Computer Engineering, Polytechnique Montréal,
2500 Chem. de Polytechnique, Montréal, Québec, Canada H3T 0A3

Article published in *IEEE Transactions on Biomedical Engineering (TBME)*, January 2026

Presentation

The previous chapter addressed the annotation cost by identifying the most informative samples for labelling. It implicitly assumed that dense, pixel-wise annotation was available, which might not be realistic. This chapter proposes to use easy-to-acquire bounding boxes as labels instead of dense segmentation masks. In particular, we propose a prompt learning framework that automates and adapts vision foundation models to medical image segmentation using only weak labels in the form of bounding boxes, replacing the need for test-time prompting and specialization with dense annotations. To compensate for the reduced supervision signal, training combines multiple box-based constraints with consistency regularization, enabling the model to produce accurate segmentation masks competitive with fully supervised approaches.

This article, entitled *"Prompt learning with bounding box constraints for medical image segmentation"*, was published in 2026 in **IEEE Transactions on Biomedical Engineering**. A preliminary work *"Automating MedSAM by learning prompts with weak few-shot supervision"* was presented in 2024 at the **MICCAI Workshop** on Foundation Models for General Medical AI.

Abstract

Pixel-wise annotations are notoriously laborious and costly to obtain in the medical domain. To mitigate this burden, weakly supervised approaches based on bounding box annotations—much easier to acquire—offer a practical alternative. Vision foundation models have recently shown noteworthy segmentation performance when provided with prompts such as points or bounding boxes. Prompt learning exploits these models by adapting them to downstream tasks and automating segmentation, thereby reducing user intervention. However, existing prompt learning approaches depend on fully annotated segmentation masks. This paper proposes a novel framework that combines the representational power of foundation models with the annotation efficiency of weakly supervised segmentation. More specifically, our approach automates prompt generation for foundation models using only bounding box annotations. Our proposed optimization scheme integrates multiple constraints derived from box annotations with pseudo-labels generated by the prompted foundation model. Extensive experiments across multi-modal datasets reveal that our weakly supervised method achieves an average Dice score of 84.90% in a limited data setting, outperforming existing fully-supervised and weakly-supervised approaches. The code is available at https://github.com/Minimel/box-prompt-learning-VFM.git.

3.1 Introduction

The precise delineation of regions of interest is a crucial step in clinical decision-making, affecting the diagnosis, treatment planning and patient outcome. However, manual annotation of medical images remains a labour-intensive and time-consuming process. With an ever-growing volume of medical imaging data, developing efficient and accurate segmentation methods has become a major challenge. Automatic segmentation algorithms (Litjens *et al.*, 2017) have been introduced to alleviate the burden of manual annotation and reduce the impact of expert subjectivity and inadvertent errors. In medical image analysis, segmentation methods based on variants of the UNet (Ronneberger *et al.*, 2015) and lately of the vision transformer architecture (Vaswani *et al.*, 2017; Dosovitskiy *et al.*, 2021), have achieved state-of-the-art performance on multiple tasks and datasets (Isensee *et al.*, 2021; Hatamizadeh *et al.*, 2022; Chen *et al.*,

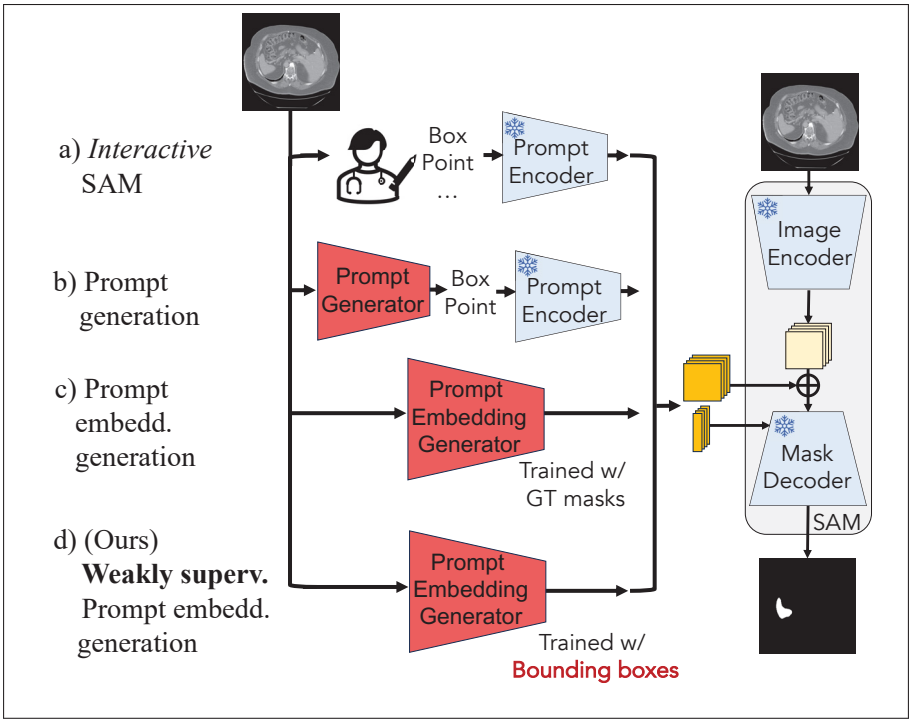


Figure 3.1 **Differences between a) Interactive SAM and b-d) Various adaptive methods for prompt automation**: b) Using SAM's prompt encoder with physical prompts generated by a trained prompt module, c) Directly generating dense and/or sparse prompt embeddings (*dark yellow*) with a prompt module trained with ground truth (GT) masks, and d) Our method, which generates prompt embeddings from the image using only bounding box annotations during training

2024a). However, these models depend on large, fully-annotated datasets for training, and their performance typically deteriorates when training data is scarce. Yet, the prohibitive cost of manually labeling medical images makes the collection of such well-annotated datasets difficult, hampering the development of these data-hungry models.

Recently, large promptable vision models (Kirillov *et al.*, 2023; Zou *et al.*, 2023; Ma *et al.*, 2024) have gained considerable attention for their flexibility and generalization abilities on multiple natural imaging benchmarks. Unlike specialized models trained on domain-specific datasets, these foundation models can generate, for any new image, informative representations which can be leveraged to produce a segmentation mask. Notably, the Segment Anything Model (SAM)

(Kirillov *et al.*, 2023) has showcased impressive zero-shot capabilities by generating accurate segmentation masks for tasks and classes unseen during training.

Current foundation models require user-provided prompts, typically manually crafted, to tailor the segmentation outputs to specific tasks. The zero-shot performance of these models heavily depends on the quality of the user prompt (Mazurowski *et al.*, 2023; Kim, Yun, Yun & Bae, 2025; Yue *et al.*, 2024; Ma *et al.*, 2024). However, prompts can be ambiguous, particularly in medical images where object boundaries are often subtle, leading to poor predictions (see Fig.3.2).

To fully exploit the potential of foundation models and improve scalability, recent efforts have focused on automating prompt generation. In particular, prompt tuning (Shaharabany, 2023; Wu, Zhang & Elbatel, 2023; Chen *et al.*, 2024b; Zhang *et al.*, 2024a; Kim *et al.*, 2025) adapts large models for downstream tasks by optimizing a small set of trainable prompt embeddings, rather than updating millions of model parameters. Prompt tuning searches for the best input prompt to satisfy the target task. Visual prompt tuning (Jia *et al.*, 2022; Wang *et al.*, 2023) has demonstrated remarkable adaptation performance with minimal trainable parameters, and several studies have applied this approach to SAM (Shaharabany, 2023; Wu *et al.*, 2023; Zhang *et al.*, 2024a; Ayzenberg, Giryes & Greenspan, 2025; Yue *et al.*, 2024; Li *et al.*, 2025). Existing methods for medical image segmentation have focused on automatically generating a physical prompt—i.e., point or box (Wu *et al.*, 2023; Ayzenberg *et al.*, 2025) or a prompt embedding (Shaharabany, 2023; Yue *et al.*, 2024; Li *et al.*, 2025) (see Fig.3.1). However, these techniques still heavily rely on fully annotated segmentation masks, which are burdensome to obtain in the medical domain.

In parallel to advanced prompt tuning techniques, weakly supervised learning offers a pragmatic approach to reduce the reliance on exhaustive manual annotations by incorporating more accessible forms of labeling. Weak labels can come in various shapes, such as scribbles (Lin, Dai, Jia, He & Sun, 2016), image tags (Pathak *et al.*, 2015), points (Bearman, Russakovsky, Ferrari & Fei-Fei, 2016) or bounding boxes (Dai, He & Sun, 2015; Khoreva *et al.*, 2017; Hsu, Hsu,

Tsai, Lin & Chuang, 2019; Kervadec *et al.*, 2020). Among these, bounding boxes are particularly appealing due to their simplicity and light storage—in practice, only two corner coordinates are needed to define a bounding box. Bounding boxes have been used as pseudo-labels to generate initial segmentation proposals (Rother *et al.*, 2004; Rajchl *et al.*, 2017), which are refined iteratively until a more precise segmentation is obtained. However, such approaches are subject to error propagation if the initial segmentation is inaccurate, ultimately impairing model performance. To mitigate this issue, alternative methods have introduced attention mechanisms to improve gradient flow (Song, Huang, Ouyang & Wang, 2019; Kulharia, Chandra, Agrawal, Torr & Tyagi, 2020) or imposed constraints on the output probabilities during optimization (Jia, Huang, Chang & Xu, 2017; Kervadec *et al.*, 2020). Building on these advancements, our work integrates constraint-based optimization with bounding box annotations into the framework of visual prompt tuning.

Our work introduces a novel framework that leverages the complementary strengths of foundation models and weakly supervised learning through prompt tuning. We train an auxiliary prompt module using only bounding box annotations—much easier to obtain than ground truth masks—to automatically generate informative prompt embeddings from input images. Relying on weak labels significantly reduces the annotation burden, making the adaptation of foundation models more scalable for medical imaging applications. Building on our preliminary work (Gaillochet, Desrosiers & Lombaert, 2024), we address prior limitations by moving beyond using weak-label constraints that only exploited bounding box tightness. Our updated training strategy now employs a box-based multi-loss optimization framework that integrates predictions from the prompted foundation model with consistency-based regularization, leading to more robust performance. Furthermore, while our earlier work focused exclusively on MedSAM, we now demonstrate that SAM—despite being trained on out-of-domain data—can serve as a general backbone foundation model when coupled with a deeper prompt module and our refined optimization strategy.

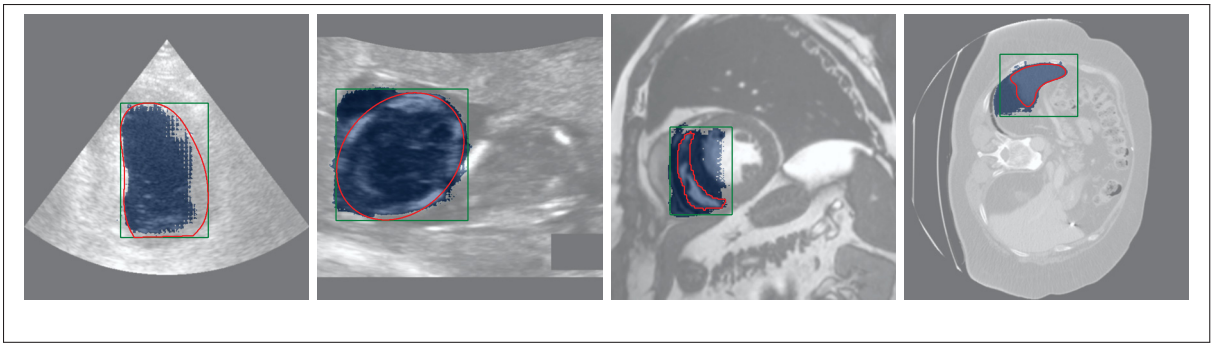


Figure 3.2 **Examples of SAM predictions with varying noise levels in the box prompts.** From left to right: no noise (tight box), 0-1.5%, 1.5-3%, and 3-5% pixel displacement from the original tight box. Ground-truth annotations and associated box prompts are shown in red and green, respectively, with predicted segmentation masks overlaid in blue. Prompt ambiguity and noise, combined with applications to out-of-domain data, can cause vision foundation models to fail in effectively segmenting the target object

3.2 Related work

3.2.1 Vision foundation models for medical image segmentation

Recent years have witnessed the emergence of foundation models with impressive zero-shot capabilities like the Segment Anything Model (Kirillov *et al.*, 2023), trained on a vast dataset of 1B masks and 11M images.

Based on vision transformers (Dosovitskiy *et al.*, 2021), the promptable SAM has demonstrated notable success across various computer vision tasks due to its extensive pre-training.

Because of its success on natural images, SAM's potential for medical image segmentation has been a growing field of study (Zhang, Deng & Lu, 2023; Mazurowski *et al.*, 2023; Huang *et al.*, 2024b). Efforts have been undertaken to adapt SAM for medical imaging and other specific segmentation tasks (Shaharabany, 2023; Gu, Dong, Yang & Mazurowski, 2025). (Cheng *et al.*, 2023; Ma *et al.*, 2024; Gu *et al.*, 2025) explored fine-tuning strategies, whereas (Wu *et al.*, 2023; Shaharabany, 2023; Wu *et al.*, 2025a) investigated externally designed components. However, fine-tuning approaches incur practical challenges in terms of data collection, data annotation

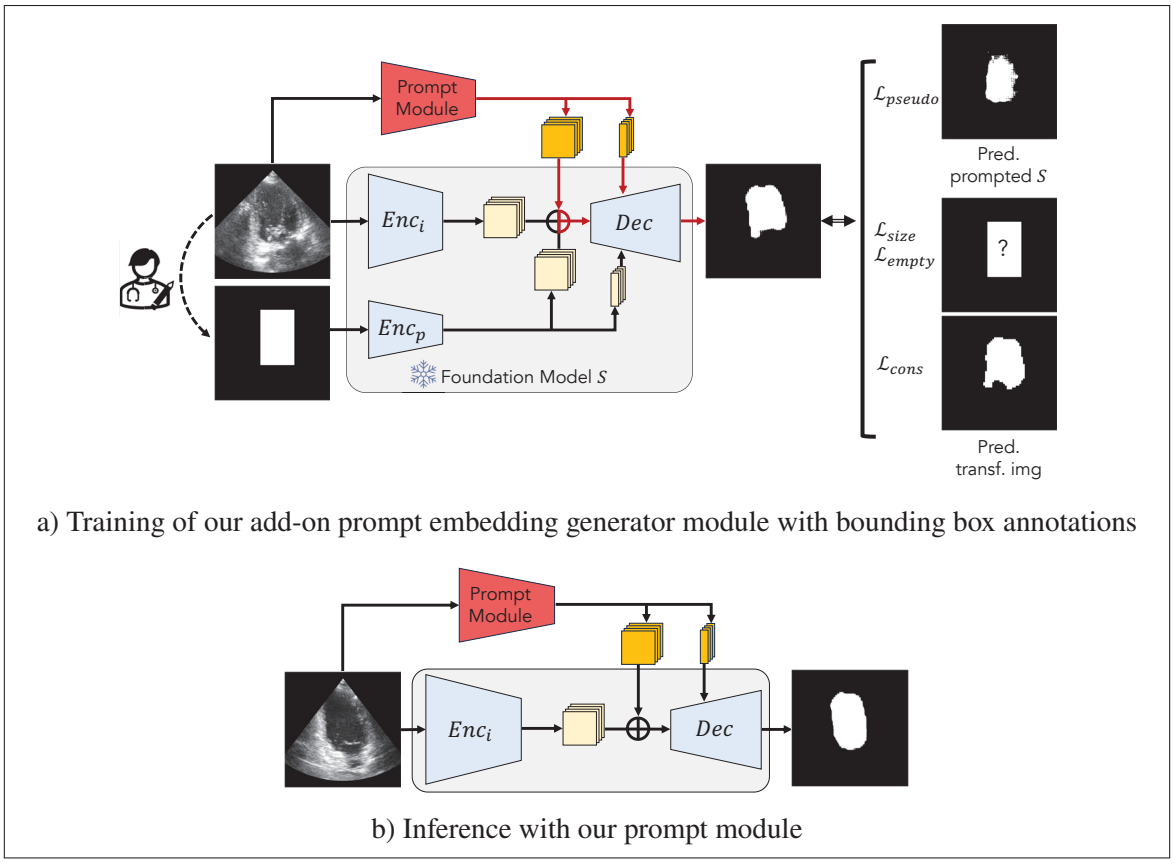


Figure 3.3 **Overview of our framework.** Our prompt module fits on top of promptable foundation models. a) During training, our module (*red*) learns to generate relevant prompt embeddings (*dark yellow*) from the image. Training requires only bounding box annotations and optimizes multiple losses. The components of the backbone foundation model (*blue*) remain frozen during training. Red arrows indicate active gradient propagation. b) At inference, our trained prompt module replaces the original prompt encoder of the foundation model, eliminating user interaction

and computational time. Moreover, they remain interactive models, requiring user input during the inference stage. Hence, alternative approaches have explored the idea of prompt tuning for SAM in the context of medical imaging.

3.2.2 Prompt learning for vision foundation models

Prompt-tuning has recently been applied to large vision models (Jia *et al.*, 2022) to adapt them to medical image segmentation or to automate prompt generation. In the medical domain,

one approach has been to automatically generate point and box prompts from the image by incorporating a detection network into SAM's architecture (Wu *et al.*, 2023; Ayzenberg *et al.*, 2025). These prompts are then converted to embeddings via SAM's prompt encoder (see Fig.3.1.b). An alternative has been to use a self-prompting module to yield prompt embeddings directly, hence replacing the original prompt encoder (Shaharabany, 2023; Yue *et al.*, 2024; Li *et al.*, 2025) (see Fig.3.1.c). PerSAM (Zhang *et al.*, 2024a) and ProtoSAM (Ayzenberg *et al.*, 2025) compare the embeddings of the image, from SAM or an external network, with those of a single image-mask pair and generated appropriate point and box prompts.

Unlike our approach, these prompt learning methods require samples with ground-truth annotation masks, which remains a burden for medical datasets. Recently, (Gaillochet *et al.*, 2024) showed that weak labels could be exploited to automate prompt learning but used a restrictive tight box constraint, limiting the approach to in-domain performance.

3.2.3 Weakly supervised learning with bounding boxes

Weakly supervised segmentation methods have been developed to reduce the cost of manual annotation. Popular among such methods are approaches based on bounding box annotations. These bounding boxes are typically used as initial pseudo-labels for identifying target regions. For instance, the classic GrabCut algorithm (Rother *et al.*, 2004), based on graph cuts (Boykov, Veksler & Zabih, 2001), iteratively separates foregrounds from backgrounds using bounding boxes. DeepCut (Rajchl *et al.*, 2017) extends GrabCut to neural networks. However, a well-known challenge with these iterative methods involves errors appearing in the initial segmentation and being propagated during training. Attention maps were proposed during training to reduce the propagation of incorrect gradients by masking out irrelevant regions (Song *et al.*, 2019) or by handling the label noise assumed to be contained in the bounding boxes (Kulharia *et al.*, 2020). More recently, constraints on tightness and size were applied to guide the segmentation during training, either alone (Kervadec *et al.*, 2020) or in combination with multiple instance learning (MIL) and smooth maximum approximation (Hsu *et al.*, 2019; Wang & Xia, 2021).

3.3 Contributions

This paper introduces a novel framework that leverages the strengths of foundation models and the cost-efficiency of weakly supervised segmentation. More specifically, we *automate and adapt foundation models* by training with only *bounding box annotations* an auxiliary *prompt module* that automatically *generates a relevant prompt embedding* from the input image. Our novel training approach for weakly labeled data exploits multiple pieces of information provided by bounding box annotations. Training our auxiliary prompt module optimizes a loss based on the prediction of the foundation model when prompted by a bounding box, as well as box-based spatial constraints and a consistency-based regularization. The constraints and regularization refine the segmentation of the prompted foundation model, which may contain inaccuracies, and allow for the application of the foundation model to out-of-domain data.

At its core, our method fundamentally replaces the traditional prompt encoder of the foundation model with a new auxiliary prompt embedding generator, which:

1. Requires only **bounding boxes** to train by employing a novel box-based multi-loss optimization strategy,
2. Generalizes effectively to full and **limited training data**,
3. Performs well **across modalities**, as shown by our experiments on MRI, CT and ultrasound images, and
4. Successfully functions on **out-of-domain data**, as shown by our experiments with SAM as a backbone foundation model.

3.4 Method

Let $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ be a 3-channel input image of height H and width W. Suppose we only have access to a bounding box $\mathbf{m}$ enclosing the object of interest, rather than a full ground-truth segmentation mask. Our approach automates and adapts promptable foundation models for new tasks using such weakly labeled data. We train a prompt module to generate a relevant prompt embedding from the input image using an objective function comprising three main

components, detailed in Sections 3.4.2 to 3.4.4. The overall framework is illustrated in Fig.3.3, and the training procedure is outlined in Alg.3.1.

3.4.1 Add-on prompt module to vision foundation model

Although promptable vision foundation models are universal models that can generalize to varied tasks, they lack the ability to automatically segment a specific target object without user interaction. Our end-to-end method eliminates the reliance on user-defined prompts to automatically segment specified objects in medical images.

3.4.1.1 Vision foundation model architecture

The Segment Anything Model (SAM) (Kirillov *et al.*, 2023) is used as our prototypical promptable vision foundation model architecture. SAM consists of three distinct components: an image encoder, a prompt encoder and a mask decoder, which we denote as Enc_i, Enc_p and Dec. As a promptable model, SAM takes as input a set of encoded prompts $\mathbf{P} = \{\mathbf{P}_1 \dots \mathbf{P}_n\}$ where $\mathbf{P}_i \in \mathbb{R}^D$, which can represent any combination of:

1. Point coordinates: $\mathbf{p} \in \mathbb{R}^2$
2. Bounding box boundary coordinates: $\mathbf{b} = [\mathbf{p}_1, \mathbf{p}_2]$ with $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^2$, and
3. A coarse mask: $\tilde{\mathbf{y}} \in \{0, 1\}^{H \times W}$.

Given $\mathbf{x}$ and $\mathbf{P}$, SAM generates an image embedding $\mathbf{Z}_i \in \mathbb{R}^{256 \times 64 \times 64}$, as well as a sparse and dense prompt embedding $\mathbf{Z}_{ps} \in \mathbb{R}^{256}$ and $\mathbf{Z}_{pd} \in \mathbb{R}^{256 \times 64 \times 64}$, following:

$$\begin{aligned} \mathbf{Z}_i &= Enc_i(\mathbf{x}), \\ \mathbf{Z}_p &= \{\mathbf{Z}_{ps}, \mathbf{Z}_{pd}\} = Enc_p(\mathbf{P}). \end{aligned} \tag{3.1}$$

The encoded image and prompts are then fed into the mask decoder to obtain a probability map.

3.4.1.2 Prompt module architecture

We automate the foundation model by training a prompt module g_θ to directly generate the prompt embedding $\mathbf{Z}'_p = \{\mathbf{Z}'_{pd}, \mathbf{Z}'_{ps}\} = g_\theta(\mathbf{x})$ according to the target task. Following (Shaharabany, 2023), our prompt embedding generator module is composed of a Harmonic Dense Net (Chao, Kao, Ruan, Huang & Lin, 2019) pre-trained on ImageNet as well as a decoder to produce an output of shape $256 \times 64 \times 64$. The Harmonic Dense Net takes as input the image $\mathbf{x}$ and comprises six blocks with channel outputs of size 192, 256, 320, 480, 720, and 1280. The decoder consists of two upsampling blocks, each with two convolutional layers. Our experiments focus on $\mathbf{Z}'_{pd}$ and use for $\mathbf{Z}'_{ps}$ SAM's default sparse embedding given an empty input prompt.

3.4.2 Pseudo-label loss from prompted foundation model

Previous works have noted that bounding box prompts provide a less ambiguous spatial context for the object of interest (Mazurowski *et al.*, 2023; Huang *et al.*, 2024b; Ma *et al.*, 2024). Hence, we take advantage of the fact that the provided tight bounding box $\mathbf{m}$ can be converted into a box prompt $\mathbf{b}$ by extracting the lower left and upper right coordinates of $\mathbf{m}$. Given $\mathbf{b}$, the promptable foundation model can generate a segmentation mask, which acts as a coarse pseudo-label to guide the learning process of the prompt module g_θ. With the predicted output probabilities as $S_\theta(\mathbf{x}) = Dec(\mathbf{Z}_i, g_\theta(\mathbf{x}))$ and the pseudo-label obtained from the prompted foundation model as $S(\mathbf{x}, \mathbf{b}) = [\![Dec\big(\mathbf{Z}_i, Enc_p(\mathbf{b})\big) \geq 0.5]\!]$, we define our pseudo-label loss:

$$\begin{aligned} \mathcal{L}_{pseudo}(\alpha, \beta, S(\mathbf{x}, \mathbf{b})) \;=\; & \alpha \cdot \mathrm{CE}\left(S_\theta(\mathbf{x}), S(\mathbf{x}, \mathbf{b})\right) \\ & + \beta \cdot \mathrm{Dice}\left(S_\theta(\mathbf{x}), S(\mathbf{x}, \mathbf{b})\right). \end{aligned} \tag{3.2}$$

3.4.3 Box-based constraints

Since the predictions of the foundation model are only a rough estimate of the true mask and may contain inaccuracies (see Fig.3.2), additional constraints should be used to guide the training

process. Denoting as Ω_I and Ω_O the regions inside and outside the bounding box $\mathbf{m}$ such that $\Omega_I \cup \Omega_O = \Omega \in \mathbb{R}^{H \times W}$, we apply two box-based constraints on the output by exploiting the following facts:

- The size of the bounding box sets a *lower and upper limit* on the *number of foreground pixels*
- Ω_O contains *only background pixels*

3.4.3.1 Size constraint

The bounding box puts constraints on the size of the predicted mask. Thus, the sum of foreground predicted pixels should not be greater than the size of Ω_I. Similarly, we can apply a prior on the size of the predicted foreground, given the size of Ω_I. Setting the ratio $\epsilon_1, \epsilon_2 \in [0, 1]$ of pixels from Ω_I that should be classified as foreground, we obtain

$$\epsilon_1 |\Omega_I| \leq \sum_{(i,j) \in \Omega} S_\theta(\mathbf{x})_{ij} \leq \epsilon_2 |\Omega_I|. \tag{3.3}$$

Once again, $S_\theta(\mathbf{x}) = Dec(\mathbf{Z}_i, g_\theta(\mathbf{x}))$ and $S_\theta(\mathbf{x})_{ij}$ refers to the pixel (i, j) of the output probabilities.

The condition can be translated into a constraint-based loss by applying a penalty function ψ_t, which becomes positive when the condition is not met, like a simple ReLU function. In this work, we employ a pseudo log-barrier function offering more stable optimization (Kervadec *et al.*, 2022). The function $\psi_t(x)$ approximates a hard barrier as $t \to \infty$, where $\psi_t(x) = \infty$ if $x > 0$, and $\psi_t(x) = 0$ otherwise. The size-based loss can then be formulated as

$$\begin{aligned} \mathcal{L}_{size}(\epsilon_1, \epsilon_2, \Omega_I) = \psi_t\left(\epsilon_1 |\Omega_I| - \sum_{(i,j) \in \Omega} S_\theta(\mathbf{x})_{ij}\right) \\ + \psi_t\left(\sum_{(i,j) \in \Omega} S_\theta(\mathbf{x})_{ij} - \epsilon_2 |\Omega_I|\right). \end{aligned} \tag{3.4}$$

3.4.3.2 Emptiness constraint

The bounding box also clearly defines a region that should contain only background pixels (i.e., Ω_O, the regions outside the box). This emptiness constraint inside Ω_O can be expressed as a cross-entropy loss on the background pixels:

$$\mathcal{L}_{empty}(\Omega_O) \;=\; -\sum_{(i,j)\in\Omega_O} \log(1 - S_\theta(\mathbf{x})_{ij}) \tag{3.5}$$

3.4.4 Consistency-based regularization

We propose a regularization strategy based on transformation consistency to alleviate the problem of over-fitting when training with few weakly-annotated samples. As the computational bottleneck of the foundation model is its image encoder, this strategy operates directly on features from the encoder instead of image pixels, as in traditional approaches. Following previous definitions, we denote $\mathbf{Z}_i \in \mathbb{R}^{256\times64\times64}$ the embedding of an image $\mathbf{x}$ obtained with the original image encoder. When training our prompt module g_θ, we randomly sample a geometric transformation T (combination of random rotation/flip) and apply it on the image encoded features to obtain $T(\mathbf{Z}_i)$. We then enforce the predictions of the segmentation decoder to be equivariant using an L2 consistency loss. Noting $S'_\theta\big(T(\mathbf{x})\big) = Dec\big(T(\mathbf{Z}_i), g_\theta\big(T(\mathbf{x})\big)\big)$:

$$\mathcal{L}_{cons}(T) \;=\; \Big\| S'_\theta\big(T(\mathbf{x})\big) \;-\; T\left(S_\theta(\mathbf{x})\right)\Big\|_2^2 . \tag{3.6}$$

3.4.5 Box-based multi-loss optimization framework

In summary, training the prompt module involves optimizing four losses conditioned on a prompt-based pseudo-label, size and emptiness constraints and a consistency regularization. Combining equations (3.2), (3.4), (3.5) and (3.6), the final loss to optimize becomes:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{pseudo} + \lambda_2\mathcal{L}_{size} + \lambda_3\mathcal{L}_{empty} + \lambda_4\mathcal{L}_{cons}, \tag{3.7}$$

where λ_i's are weights applied to each individual loss.

Algorithm 3.1 Prompt Module Training Procedure via our Box-based Multi-loss Optimization Scheme

Input: g_θ: prompt module to train, $\mathcal{D}_{train} = \{\mathbf{x}^{(k)}, \mathbf{m}^{(k)}\}_{k=1}^{N}$: training set, Enc_i, Enc_p, Dec: components of foundation model

Input: $\alpha, \beta, \epsilon_1, \epsilon_2, \psi_t, T, \lambda_1, \lambda_2, \lambda_3, \lambda_4$

for $(\mathbf{x}^{(k)}, \mathbf{m}^{(k)}) \in \mathcal{D}_{train}$ **do**

  $\mathbf{Z}_i \leftarrow Enc_i(\mathbf{x}^{(k)})$;

  Convert $\mathbf{m}^{(k)}$ to the box prompt $\mathbf{b}$;

  $S(\mathbf{x}, \mathbf{b}) \leftarrow [\![Dec(\mathbf{Z}_i, Enc_p(\mathbf{b})) \geq 0.5]\!]$;

  Compute $\mathcal{L}_{pseudo}(\alpha, \beta, S(\mathbf{x}, \mathbf{b}))$ using (3.2) ; `/* Pseudo-label loss */`

  Ω_I: foreground region of $\mathbf{m}^{(k)}$ (inside box);

  Ω_O: background region of $\mathbf{m}^{(k)}$ (outside box);

  Compute $\mathcal{L}_{size}(\epsilon_1, \epsilon_2, \Omega_I)$ using (3.4) ; `/* Constraint-based losses */`

  Compute $\mathcal{L}_{empty}(\Omega_O)$ using (3.5);

  Compute $\mathcal{L}_{cons}(T)$ using (3.6) ; `/* Consistency-based regularization */`

  $\mathcal{L}_{total} \leftarrow \lambda_1 \mathcal{L}_{pseudo} + \lambda_2 \mathcal{L}_{size} + \lambda_3 \mathcal{L}_{empty} + \lambda_4 \mathcal{L}_{cons}$;

  Update weights θ of g

end for

Output: g_θ

3.5 Experiments and results

We evaluate the efficacy of our proposed optimization scheme for prompt automation by comparing it to specialized models (trained exclusively for the target task) and existing SAM-based adaptations which require ground-truth labels, as well as a weakly supervised approach based on bounding boxes (Kervadec *et al.*, 2020; Wang & Xia, 2021). To validate the robustness and generalizability of our approach, we conduct experiments on datasets of different imaging modalities and anatomical structures. We then explore the effectiveness of our method through ablation studies on the impact of trainings set size, individual loss components, backbone foundation model and label noise.

3.5.1 Datasets

To ensure a comprehensive evaluation, we tested our method on ultrasound (US), MRI and CT datasets. Five publicly available medical imaging datasets were used: the ultrasound Head Circumference dataset (HC18) (van den Heuvel *et al.*, 2018), the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) (Leclerc *et al.*, 2019), the MRI Automated Cardiac Diagnosis Challenge (ACDC) (Bernard *et al.*, 2018) and the spleen and liver CT datasets from the Medical Segmentation Decathlon (MSD) (Antonelli *et al.*, 2022). Tasks included segmentation of the head circumference, end-diastole of the left (LV) and right (RV) ventricles, as well as spleen and liver. For uniformity across all datasets, we conducted slice-based segmentation of imaging volumes and filtered out background-only slices, keeping the largest object of size at least 10 pixels.

For HC18, we used 507 images for training, 77 for validation and 148 for testing. For CAMUS, we focused on left ventricle (LV) segmentation and kept 350 training, 50 validation and 100 test images. For ACDC, we used 90, 10 and 50 patients (765, 78 and 470 images) for training, validation and testing. For the MSD-Spleen dataset, we used 4 patients (114 images) for validation, 10 patients (340 images) for testing and the remaining 27 patients (597 images) for training. Finally, for the MSD-Liver dataset containing 131 patients, we utilized 13 patients (1,195 images) for validation and 30 patients (3,466 images) for testing. Following (Ma *et al.*, 2024), our preprocessing involved clipping the intensity values of each 2D image (HC18, CAMUS) or each 3D volume (ACDC, MSD) to the 0.5th and 99.5th percentiles, followed by rescaling to the range [0, 255]. Each 3D volume from the ACDC and MSD datasets was partitioned into 2D images and resampled to a fixed resolution of 1mm x 1mm. We then center-cropped and padded each sample to size 640×640 (HC18), 512×512 (CAMUS and MSD-Spleen) or 256×256 (ACDC and MSD-Liver). Finally, to comply with SAM's requirements, all images were resized to a fixed dimension of 3×1024×1024 before being fed to the foundation model.

3.5.2 Implementation details

Our backbone promptable foundation model used SAM ViT-H, the largest variant of SAM, which was kept frozen. Training of the prompt module operated on 20 samples for 200 epochs using a batch size of 4. The learning rate was set to 0.0001 and reduced by a factor of 0.1 midway through the training process. Additionally, a weight decay of 0.0001 was applied. Our training process optimized the total loss, $\mathcal{L}_{\text{total}}$, composed of four distinct loss terms, each assigned a specific weight. For all experiments, we used the following weight values: $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 0.001$, and $\lambda_4 = 0.001$. These hyperparameters were tuned by optimizing for the best Dice score on the validation set, based on bounding boxes generated from the provided weak annotation and predicted mask. To avoid selecting hyperparameters encouraging hollow segmentation masks, we required the average predicted foreground-to-bounding-box size ratio to be above 50%. Optimization was performed on the ACDC dataset, and the same hyperparameters were then kept fixed throughout the experiments. To impose a strong size constraint, we set $[\epsilon_1, \epsilon_2] = [0.7, 0.9]$ and, following (Kervadec *et al.*, 2020), scaled the parameter t of the log-barrier function ψ_t by a factor 1.1 every 5 epochs. For the consistency loss, transformations included random flips and rotations. Due to the symmetrical nature of images in the MSD-Spleen dataset, we replaced flips with random translations and scaling to better align with the dataset's characteristics. We used the model from the final training epoch at test time.

To minimize computational complexity and speed-up training, we did not perform data augmentation. Doing so allowed us to discard the image encoder during training by using pre-computed image embeddings, reducing the number of model parameters from 141.8M to 45.6M, with 41.6M trainable parameters.

Each experiment was repeated using three randomly selected training subsets and three different initialization seeds to ensure robustness. The results were averaged across all 9 trials. All experiments were conducted using Python 3.8.10 with PyTorch on NVIDIA RTX A6000 GPUs.

3.5.3 Evaluation protocol

3.5.3.1 Evaluation measures

Two evaluation measures assess the performance of our proposed approach: the Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD).

The DSC is defined as follows:

$$DSC = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \tag{3.8}$$

where A and B represent the sets of predicted and ground truth segmentation masks. The DSC quantifies the overlap between the predicted and ground truth masks, with values ranging from 0% (no overlap) to 100% (perfect overlap).

The Average Symmetric Surface Distance measures the average shortest distances between contour C_A to any point on contour C_B, and vice-versa:

$$ASSD(C_A, C_B) = \frac{\sum_{a \in C_A} d(a, C_B) + \sum_{b \in C_B} d(b, C_A)}{|C_A| + |C_B|}, \tag{3.9}$$

with $d(i, C_J) = \min_{j \in C_J} d(i, j)$. The ASSD being undefined for empty ground truths or predictions, we set in such case the distance $d(i, C_j) = \sqrt{H^2 + W^2}$, corresponding to the maximum possible distance in the image.

3.5.3.2 Baselines and comparative methods

We compare our proposed method to two specialized models: a standard UNet (Ronneberger *et al.*, 2015) and TransUNet (Chen *et al.*, 2024a). Additionally, we validate our approach against the original SAM prompted with a tight bounding box based on the ground truth mask, as well as three SAM-based automated approaches: Self-prompting (Wu *et al.*, 2023), AutoSAM (Shaharabany, 2023), and PerSAM (Zhang *et al.*, 2024a). We also do a comparison with two weakly-supervised methods based on a residual UNet trained with bounding box annotations:

the first using box-based constraints (Kervadec *et al.*, 2020) and the other based on generalized MIL and smooth maximum approximation (Wang & Xia, 2021).

The UNet, TransUNet, Self-prompting, PerSAM and AutoSAM models were trained using full segmentation masks. To ensure optimal performance for the baseline models, we increased the batch size to 24 for TransUNet, following (Chen *et al.*, 2024a). The UNet and TransUNet were optimized with a standard Dice cross-entropy loss. The weakly supervised and SAM-based prompt learning baselines followed the best practices outlined in (Kervadec *et al.*, 2020; Wang & Xia, 2021; Shaharabany, 2023; Wu *et al.*, 2023; Zhang *et al.*, 2024a).

3.5.4 Quantitative and qualitative results

Table 3.1 **Model performance on test sets in terms of mean (±std) 2D DSC** (↑), with limited training set size (20 samples except for PerSAM, which uses 1 sample). The first row gives the results of SAM when prompted with a tight bounding box. The best results for weakly supervised approaches are shown in bold while the best fully-supervised results are underlined. * indicates statistical significance with a p-value < 0.05 for all paired permutation tests between our method and each baseline

Train Label	Method	#Train. Params	Ultrasound (US)		MRI		CT	
			HC18	CAMUS	ACDC-RV	ACDC-LV	MSD-Spleen	MSD-Liver
-	Interactive SAM	-	94.18	85.49	90.64	93.71	92.82	93.40
Fully supervised (GT mask)	UNet	6.8 M	70.55 ±2.35	72.97 ±9.47	50.66 ±4.31	67.67 ±3.69	67.12 ±9.49	64.86 ±4.78
	TransUNet	105 M	<u>95.82</u> ±0.29	88.64 ±0.89	66.19 ±3.70	83.48 ±1.97	70.16 ±2.28	75.67 ±2.46
	Self-prompting	257	83.38 ±1.18	74.04 ±1.20	52.27 ±3.65	63.67 ±1.71	70.95 ±0.96	68.15 ±3.55
	AutoSAM	41.6 M	92.14 ±1.80	<u>88.78</u> ±1.84	<u>69.01</u> ±7.02	<u>87.10</u> ±1.82	<u>82.30</u> ±4.01	<u>81.05</u> ±3.42
	PerSAM	-	58.98 ±0.19	36.13 ±0.00	27.64 ±9.48	45.43 ±5.47	12.84 ±6.26	23.40 ±0.67
	PerSAM-f	2	68.67 ±4.32	48.84 ±4.82	28.47 ±12.42	61.13 ±9.16	36.72 ±17.62	55.25 ±6.31
Weakly supervised **(Bounding box)**	Kervadec et al.	18.7 M	76.08 ±5.42	79.56 ±2.46	55.64 ±4.49	72.43 ±4.81	75.64 ±3.82	72.43 ±4.87
	Wang et al.	19.6 M	30.23 ±5.65	72.66 ±7.52	35.30 ±5.13	65.35 ±5.66	26.36 ±26.46	39.46 ±23.67
	Ours	41.6 M	**92.25** ±0.84	**84.21*** ±1.15	**80.77*** ±1.12	**89.82*** ±1.00	**83.89*** ±1.62	**78.47** ±1.59

Our main findings, based on experiments conducted on five distinct medical imaging datasets— ultrasound (HC18 and CAMUS), MRI (ACDC), and CT (MSD-Spleen and MSD-Liver)— are summarized in Tables 3.1–3.2 and visualized in Fig. 3.4.

From Tables 3.1 and 3.2, we observe that our prompt-learning method outperforms by a large margin the other weakly supervised approaches (Kervadec *et al.*, 2020; Wang & Xia, 2021). Compared to fully-supervised methods, our approach outputs results that are on par with the

Table 3.2 **Model performance on test sets in terms of mean (±std) 2D ASSD (↓)**, with limited training set size (20 samples except for PerSAM, which uses 1 sample). The first row gives the results of SAM when prompted with a tight bounding box. The best results for weakly supervised approaches are shown in bold while the best fully-supervised results are underlined. * indicates statistical significance with a p-value < 0.05 for all paired permutation tests between our method and each baseline

Train Label	Method	#Train. Params	Ultrasound (US)		MRI		CT	
			HC18	CAMUS	ACDC-RV	ACDC-LV	MSD-Spleen	MSD-Liver
-	Interactive SAM	-	12.98	13.24	1.60	1.34	1.75	1.82
Fully supervised (GT mask)	UNet	6.8 M	61.90 ±2.74	27.30 ±10.29	36.12 ±5.15	29.31 ±5.55	84.18 ±31.07	26.05 ±5.54
	TransUNet	105 M	<u>8.33</u> ±0.97	15.55 ±4.86	14.55 ±2.07	9.40 ±2.79	<u>14.55</u> ±2.84	11.02 ±3.06
	Self-prompting	257	37.23 ±2.73	25.70 ±1.01	41.97 ±4.94	18.47 ±0.44	28.31 ±5.17	27.18 ±3.82
	AutoSAM	41.6 M	16.50 ±5.51	<u>11.60</u> ±2.73	<u>11.58</u> ±4.1	<u>5.36</u> ±1.39	24.51 ±11.13	<u>8.91</u> ±3.22
	PerSAM	-	106.44 ±0.63	110.88 ±0.01	57.71 ±16.52	43.58 ±3.21	143.20 ±11.95	65.52 ±0.58
	PerSAM-f	2	75.10 ±2.78	72.36 ±10.62	42.80 ±14.26	21.83 ±5.96	55.18 ±21.57	28.15 ±6.82
Weakly supervised	Kervadec et al.	18.7 M	38.98 ±3.53	15.52 ±0.94	31.70 ±7.65	27.84 ±11.11	17.27 ±3.69	30.97 ±14.24
	Wang et al.	19.6 M	32.77 ±5.27	21.63 ±4.19	48.45 ±25.83	35.62 ±13.63	97.23 ±6.53	41.80 ±19.21
(Bounding box)	**Ours**	41.6 M	**18.75** ±3.00	**14.22*** ±0.97	**5.34*** ±0.81	**3.40*** ±0.69	**13.14*** ±4.48	**10.37** ±3.17

best performing specialized method (TransUNet) and SAM-based method (AutoSAM). For half of the tasks, our approach based on bounding boxes is even able to outperform the best fully-supervised approach requiring ground truth masks. These findings are supported visually by Fig.3.4. Interestingly, the figure also shows that our method yields a better segmentation than interactive SAM on the CAMUS test sample, supporting our claim that our proposed approach is able to address the failed predictions of the foundation model.

Table 3.3 **Impact of each loss component on the 2D DSC (↑)**. Results are reported for two settings: a highly complex setting with a small backbone and training set size (SAM ViT-b and 10 samples), and a moderately complex setting with a huge backbone and larger training set size (SAM ViT-H and 20 samples). The importance of each loss component is most apparent in the highly complex setting, where the foundation model is most likely to make errors

Task complexity	Pseudo-label loss	Size constraint	Emptiness constraint	Consistency loss	HC18	CAMUS	ACDC-LV	ACDC-RV	MSD-Spleen
High	✓				38.43 ±3.16	76.56 ±5.96	62.87 ±7.39	80.06 ±1.94	81.49 ±2.18
	✓	✓	✓		77.38 ±3.39	82.65 ±1.18	65.46 ±3.21	79.38 ±3.87	81.19 ±0.86
	✓	✓	✓	✓	**79.68** ±1.83	**83.13** ±0.33	**72.59** ±1.67	**84.33** ±2.02	**81.92** ±0.72
Moderate	✓				90.15 ±1.05	83.54 ±1.14	72.70 ±4.95	87.49 ±0.94	**86.19** ±2.40
	✓	✓	✓		90.46 ±0.66	83.28 ±1.11	70.18 ±7.01	87.44 ±1.55	84.10 ±1.43
	✓	✓	✓	✓	**92.25** ±0.84	**84.21** ±1.15	**80.77** ±1.12	**89.82** ±1.00	83.89 ±1.62

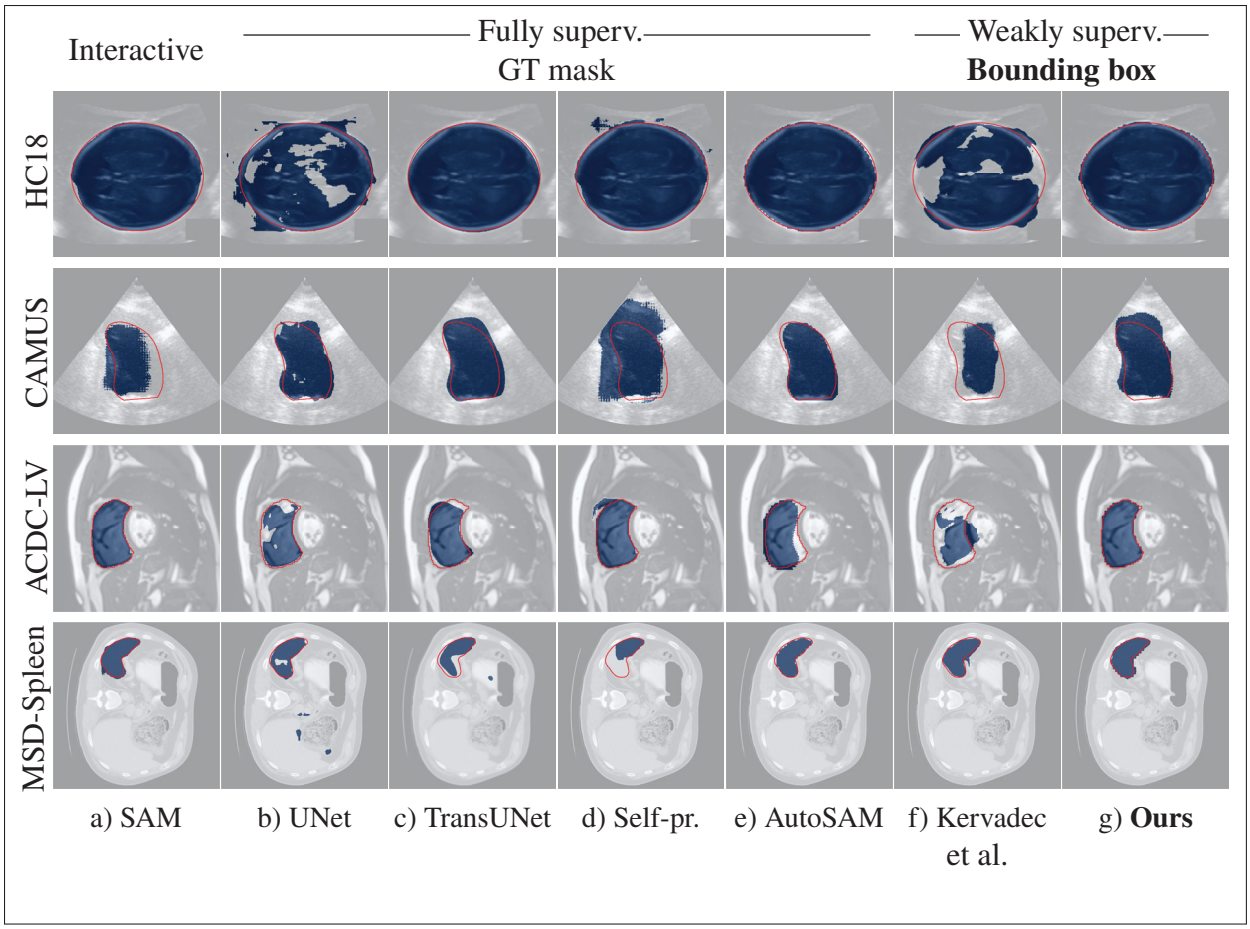


Figure 3.4 **Predicted segmentations on test samples** of the HC18, CAMUS, ACDC-LV and MSD-Spleen datasets. From left to right, a) SAM prompted with a tight box based on the ground truth, b-e) UNet, TransUNet, Self-prompting and AutoSAM, all trained with ground-truth masks, and f) residual UNet trained with tight box constraints and g) our prompt learning method trained with bounding box annotations. All automatic methods are given for the 20-shot setting. Ground-truth annotation is drawn in red, with the predicted segmentation mask overlayed in blue. In most cases, our weakly supervised approach is able to produce better segmentations than methods that require full annotation masks during training

3.5.5 Ablation study

3.5.5.1 Impact of loss components

In the absence of ground-truth masks, our objective function exploits the predictions of the prompted foundation model, box-based constraints, and consistency-based regularization. In this

ablation study, we assessed the contribution of each loss component in both a highly complex setting (with SAM ViT-b and 10 samples) and a moderately complex setting (with SAM ViT-H and 20 samples). Table 3.3, shows that, in highly complex settings where the foundation model is most likely to make errors even when provided with a prompt, our size and emptiness constraints are able to successfully guide the model to produce more accurate predictions. This is especially true for tasks where the pseudo-label loss alone falls short (i.e., ultrasound head segmentation). However, these constraints benefit all tested datasets. Our consistency loss can further boost performance by up to 10.9% by regularizing the prompt module's training. In a moderately complex setting, where the foundation model is more informative and the module is trained on a bigger set, using only our pseudo-label loss already provides reasonable segmentation masks. Additional constraints and consistency loss can nonetheless further improve the performance. Similarly, Table 3.4 validates the strength of our proposed weakly supervised optimization scheme. Using our proposed losses yields up to 32% improvement compared to when using the tightness constraints of (Kervadec *et al.*, 2020) to train our prompt module.

Table 3.4 **Impact of weakly supervised optimization scheme on our prompt learning module training.** The 2D test DSC (↑) is reported after training with 20 samples. Our box-based optimization strategy significantly outperforms existing weakly supervised loss functions

Weakly Supervised Loss	HC18	ACDC-LV	MSD-Spleen
Tightness constraints (Kervadec *et al.*, 2020)	85.68 ±2.54	61.09 ±83.91	73.09 ±1.46
Ours	**92.25** ±0.84	**80.77** ±1.12	**83.89** ±1.62

3.5.5.2 Impact of number of training samples

We evaluated the limits of our approach by experimenting with a smaller and larger training set, respectively 10 samples and the complete training set. We focused on the HC18, ACDC-LV and MSD-Spleen datasets, each representing a different modality: ultrasound, MRI and CT. We kept the same setup as with the main experiments. However, in the full data setting, we reduced the number of epochs to 20, and given the low-intensity contrast typical of ultrasound images, we tightened the size constraint on the HC18 datasets by doubling t every epoch.

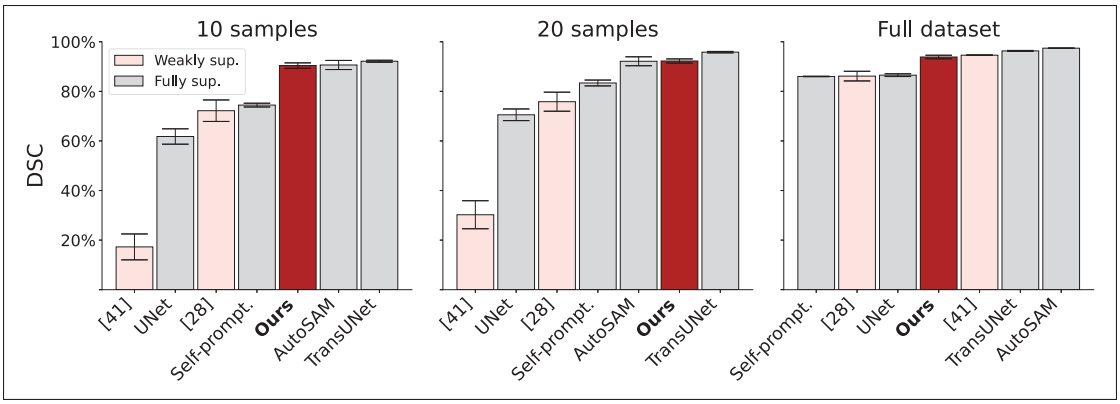


Figure 3.5 **Results on HC18 with 10, 20 and all samples.** Our weakly supervised method (dark red) consistently ranks in the top-3 approaches. In low data regimes, it outperforms other bounding box-based (pink) and fully supervised (grey) approaches by a large margin

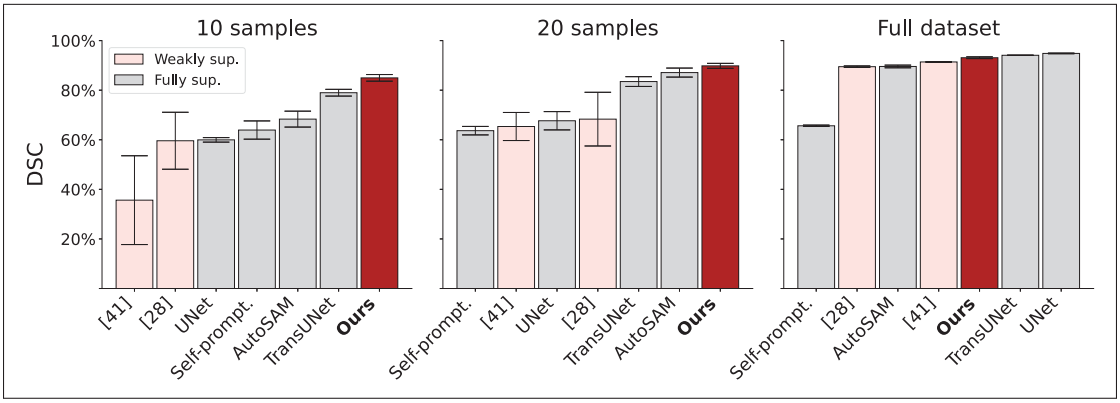


Figure 3.6 **Results on ACDC-LV for increasing training set sizes.** Given only limited data, our approach (dark red) outperforms all other methods—including fully supervised methods with GT masks (grey)

From Fig.3.5, Fig.3.6 and Fig.3.7, we observe that our method always performs better than other weakly supervised approaches (Kervadec *et al.*, 2020; Wang & Xia, 2021), across different training set sizes. In the very low data setting (10 samples), our method remains competitive, outperforming by a large margin UNet and Self-prompting, and even surpassing TransUNet on two out of three tasks. Our prompt learning approach delivers results that are comparable to—and in one case, better than—AutoSAM, without requiring GT masks. These results validate the robustness of our approach, in scenarios with both very limited and full training data.

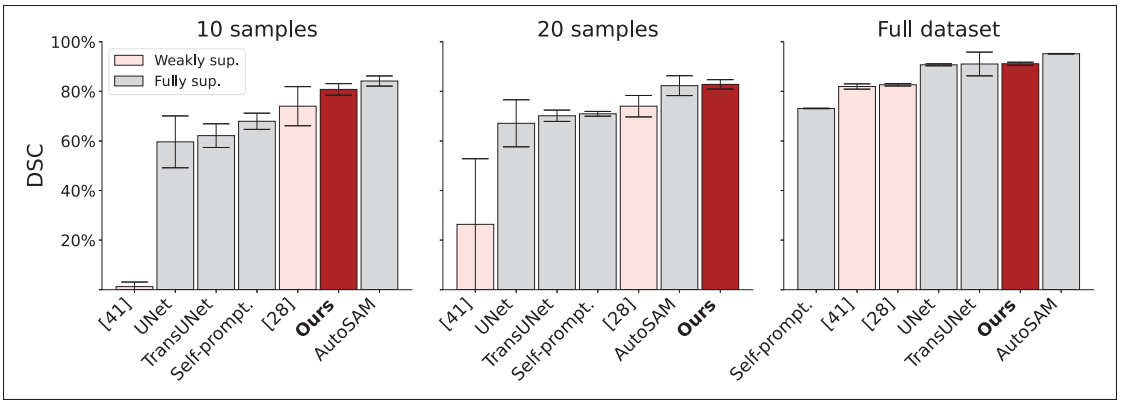


Figure 3.7 **Results on MSD-Spleen with 10, 20 and all samples.** Our method (dark red) constantly ranks top-2, surpassing all box-based (pink) and nearly all mask-based (grey) approaches

3.5.5.3 Impact of foundation model backbone

To assess the generalizability of our framework, we perform evaluations with different vision foundation models, specifically MedSAM (Ma *et al.*, 2024), a specialized version of SAM fine-tuned for medical imaging tasks, as well as another versions of SAM (SAM ViT-b). The results presented in Table 3.5 for the HC18 and ACDC datasets, none of which were used to train MedSAM, demonstrate that our method performs effectively across different foundation model backbones. Notably, using a backbone model tailored explicitly for medical image analysis improves overall performance compared to the model trained on natural images, as evidenced by the higher Dice similarity scores across all tasks. The mean Dice score increases from 79.68% to 91.17% with the HC18 dataset, and similar performance gains are observed for both right and left ventricle segmentation of the ACDC dataset. These results reinforce the advantage of leveraging domain-specific foundation models and confirm the robustness of our approach to different backbone models.

3.5.5.4 Impact of label noise

Finally, we explored a challenging real-case scenario, where bounding boxes are subject to human error. We simulated label noise by randomly displacing in any direction the box boundaries by

Table 3.5 **2D DSC (↑) on the test set with different backbone foundation models**, when trained with 10 samples

Backbone	HC18	ACDC-RV	ACDC-LV
SAM ViT-b	79.68 ±1.83	72.59 ±1.67	84.33 ±2.02
SAM ViT-H	90.40 ±1.09	72.43 ±3.84	84.97 ±1.34
MedSAM	**91.17** ±1.05	**74.52** ±3.12	**86.14** ±1.30

up to 1.5%, 1.5-3% and 3-5% of the total number of image pixels. Visual examples of such noisy prompts are shown in Fig.3.2. The results for three datasets with different tasks, modalities and image sizes are provided in Table 3.6. Despite an expected decrease in performance with increasing variability of the bounding box sizes, our prompt module maintains a competitive performance. For instance, given light human error (less than 1.5% of pixel displacement), our approach only shows a decrease of 0.5-0.7% in the Dice similarity score for HC18 and ACDC-LV.

Table 3.6 **Mean 2D DSC (↑) for different noise levels of the bounding box annotation** (in % of total number of image pixels), when trained with 20 samples

Label pixel displacement	HC18	ACDC-LV	MSD-Spleen
None (tight box)	92.25 ±0.84	89.82 ±1.00	82.82 ±1.90
$< 1.5\%$	91.75 ±0.22	89.20 ±0.07	80.33 ±0.77
$1.5 - 3\%$	89.87 ±0.79	77.45 ±2.14	69.89 ±1.54
$3 - 5\%$	84.99 ±2.42	68.45 ±0.83	58.03 ±1.36

3.6 Conclusion

Visual foundation models have enabled significant progress in medical image segmentation by reducing the burden of manual annotation. Recent prompt learning strategies automate these interactive models by training auxiliary modules to generate prompts directly from images. However, their dependence on pixel-wise annotated datasets remains a major limitation.

In this work, we propose a novel framework that combines the strengths of foundation models with the cost-efficiency of weakly supervised learning. Our approach automates and adapts foundation models through a dedicated prompt module using only bounding box annotations. The module is trained via a multi-loss optimization scheme that integrates the segmentation predictions from the prompted foundation model with box-based spatial constraints and consistency regularization. Our method not only reduces annotation costs but also improves segmentation performance compared to existing weakly supervised approaches. Through extensive experiments across multi-modal datasets—spanning full-data, limited-data and out-of-domain settings—we demonstrate the generalizability and robustness of our method. Although our current implementation focuses on SAM-based models, our proposed framework is readily extendable to other interactive foundation models, and provides a promising direction for future work in multi-class and multi-organ segmentation tasks.

CHAPTER 4

ANATOMICALLY-AWARE CONFORMAL PREDICTION FOR MEDICAL IMAGE SEGMENTATION WITH RANDOM WALKS

Mélanie Gaillochet[1,2], Christian Desrosiers[1], Hervé Lombaert[2,3]

[1] Department of Software and IT Engineering , École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Mila - Quebec AI Institute,
6666 Rue Saint-Urbain, Montréal, Québec, Canada H2S 3H1
[3] Department of Computer Engineering, Polytechnique Montréal,
2500 Chem. de Polytechnique, Montréal, Québec, Canada H3T 0A3

Article submitted to *IEEE Journal of Biomedical and Health Informatics (JBHI)*, January 2026

Presentation

The two previous chapters focused on building more accurate segmentation models under tight annotation budgets. Once a model is trained, however, a question emerges: how can we formally quantify and control the risk of failure at inference time? This chapter takes up that question, introducing a framework for providing rigorous statistical guarantees on model outputs.

We develop an anatomically-aware conformal prediction framework designed for binary medical image segmentation. Our proposed method constructs prediction sets guaranteed to contain the true segmentation with a user-specified probability. We improve the uncertainty quantification process through random walks guided by the rich feature embeddings of vision foundation models. This process enables statistically valid and spatially coherent predictions.

This article, entitled *"Anatomically-aware conformal prediction for medical image segmentation with random walks"*, was submitted to the **IEEE Journal of Biomedical and Health Informatics** in January 2026.

Abstract

The reliable deployment of deep learning in medical imaging requires uncertainty quantification that provides rigorous error guarantees while remaining anatomically meaningful. Conformal prediction (CP) is a powerful distribution-free framework for constructing statistically valid prediction intervals. However, standard applications in segmentation often ignore anatomical context, resulting in fragmented, spatially incoherent, and over-segmented prediction sets that limit clinical utility. To bridge this gap, this paper proposes Random-Walk Conformal Prediction (RW-CP), a model-agnostic framework that can be added on top of any segmentation method. RW-CP enforces spatial coherence to generate anatomically valid sets. Our method constructs a k-nearest neighbor graph from pre-trained vision foundation model features and applies a random walk to diffuse uncertainty. The random walk diffusion regularizes the non-conformity scores, making the prediction sets less sensitive to the conformal calibration parameter $\hat{\lambda}$, ensuring more stable and continuous anatomical boundaries. RW-CP maintains rigorous marginal coverage while significantly improving segmentation quality. Evaluations on multi-modal public datasets show improvements of up to 35.4% compared to standard CP baselines, given an allowable error rate of $\alpha = 0.1$. The code is available at https://github.com/Minimel/RW_ConformalPrediction.git.

4.1 Introduction

While deep neural networks can achieve impressive accuracy, they remain susceptible to failure when encountering distribution shifts, noise, or rare pathological cases. Quantifying and mitigating uncertainty is thus essential for building trust in automated segmentation pipelines. Uncertainty quantification (UQ) methods (Gal & Ghahramani, 2016a; Kendall & Gal, 2017; Guo, Pleiss, Sun & Weinberger, 2017; Lakshminarayanan, Pritzel & Blundell, 2017a; Teye, Azizpour & Smith, 2018; Abdar *et al.*, 2021; Huang, Ruan, Xing & Feng, 2024a) capture aspects of predictive variability, but fail to provide valid prediction sets with theoretical guarantees and are often sensitive to both model design and data distribution.

Conformal Prediction (CP) (Angelopoulos & Bates, 2023) has recently emerged as a powerful alternative that offers statistically valid prediction sets with finite-sample guarantees under minimal assumptions. Unlike heuristic uncertainty measures (Gal & Ghahramani, 2016a; Guo *et al.*, 2017; Lakshminarayanan *et al.*, 2017a), CP allows a user to specify a maximum allowable error rate, or miscoverage level, denoted by $\alpha \in (0, 1)$. The framework then constructs a prediction set $\hat{C}(X)$ that ensures marginal coverage, satisfying the condition $P(Y \in \hat{C}(X)) \geq 1 - \alpha$, where Y represents the ground truth. In the context of medical segmentation, this guarantee ensures that the true anatomical structure is contained within the predicted region with a high probability.

However, directly applying CP to pixel-wise outputs yields two critical limitations that hinder the utility of CP in clinical settings: a lack of spatial context and significant probability miscalibration.

First, standard CP methods often ignore spatial context. Because the guarantee is marginal (averaged over pixels or images), it can result in noisy, fragmented regions and anatomically implausible boundaries. This is particularly problematic in medical imaging, where anatomical structures exhibit strong spatial continuity and clinical utility depends on an accurate, coherent representation of organ boundaries. Moreover, evaluations typically focus on statistical coverage guarantees. Practical evaluation metrics, such as the Dice score or Hausdorff distance, are rarely assessed. In practice, enforcing the coverage guarantee often comes at the cost of severe over-segmentation and excessive volume increase. This trade-off remains largely unquantified.

Second, most CP methods derive their non-conformity score—a heuristic measure of prediction error, almost exclusively from the raw softmax probabilities of the model, which tend to be overly confident and miscalibrated (Guo *et al.*, 2017). Foundation models (Caron *et al.*, 2021; Kirillov *et al.*, 2023; Ma *et al.*, 2024; Siméoni *et al.*, 2025), trained on vast datasets, offer an opportunity to construct more informative and robust non-conformity scores for segmentation by leveraging their rich, high-dimensional feature embeddings.

In this work, we propose a Random-Walk Conformal Prediction (RW-CP) framework for image segmentation to enforce spatial coherence of CP sets via the feature space of foundation models. Our method diffuses the predicted softmax probability map using a random-walk process guided by the feature space of foundation models. Probabilities are propagated through semantically similar neighborhoods, effectively regularizing and denoising uncertainty estimates. Following traditional conformal segmentation, we calibrate a global parameter $\hat{\lambda}$ on these spatially-aware, diffused probabilities. The resulting segmentation region then retains the statistical guarantees of CP while better respecting anatomical boundaries.

Our contribution

This work addresses a critical gap in trustworthy medical image analysis: the lack of spatial coherence in uncertainty quantification. We introduce the Random-Walk Conformal Prediction (RW-CP) framework, the first split-CP approach for binary segmentation that leverages geometric diffusion to generate anatomically-informed valid prediction sets. Our framework is model-agnostic and enhances any base segmentation model by incorporating high-dimensional context from foundation models. We summarize our contributions as follows:

- We introduce a **novel conformal scoring mechanism** that combines foundation model features with **random-walk diffusion** to propagate uncertainty coherently across anatomical structures.
- We provide a **formal theoretical analysis** to prove how regularizing the score-function gradient prevents set-size instability.
- We extend the assessment of conformal sets beyond simple coverage guarantees to include **segmentation accuracy**.
- We demonstrate **state-of-the-art performance** across multiple modalities, achieving superior spatial plausibility and higher efficiency (smaller set sizes) than standard CP baselines.

4.2 Related work

4.2.1 Conformal prediction for medical image analysis

The high-stakes nature of clinical applications requires statistically rigorous uncertainty quantification, making CP a natural fit for medical imaging. Early works in medical imaging primarily applied CP to whole-slide images to assess the confidence of segmented regions, such as in lung tissue analysis (Wieslander *et al.*, 2021). Other CP studies have focused on controlling image-level metrics (Wundram, Fischer, Mühlebach, Koch & Baumgartner, 2024) or improving calibration (Brunekreef, Marcus, Sheombarsing, Sonke & Teuwen, 2024; Chen *et al.*, 2025). In particular, Conformal Performance Range Prediction (Wundram *et al.*, 2024) used CP to predict ranges of expected performance metrics, such as Dice score (DSC) or Intersection over Union (IoU), for image-level quality control. Kandinsky conformal prediction (Brunekreef *et al.*, 2024) clustered pixels based on the similarity of their non-conformity curve, in order to improve model calibration.

4.2.2 Conformal prediction for segmentation

Few works have attempted to apply conformal prediction to segmentation tasks. Initial approaches focused on generalizing the probability thresholding method used in classification via conformal risk control (Angelopoulos, Bates, Fisch, Lei & Schuster, 2024; Angelopoulos, Barber & Bates, 2025; Mossina *et al.*, 2024). Recently, (Mossina & Friedrich, 2025) proposed to generate a margin around the predicted mask, requiring only the predicted segmentation. To provide geometrically informed uncertainty quantification, Spatially-Adaptive Conformal Prediction (Bereska, Karimi & Samavi, 2025) locally weighted the non-conformity scores according to the distance from key interfaces. Alternatively, Feature Conformal Prediction (Teng *et al.*, 2023) constructed confidence sets within the deep feature space rather than the output layer to generate shorter confidence bands. However, all these methods face notable limitations in terms of input requirements or conformal set construction, including white-box access to

model features (Teng *et al.*, 2023), specific image content (Mossina *et al.*, 2024; Bereska *et al.*, 2025), or restriction to uniform margin expansion (Mossina & Friedrich, 2025).

4.2.3 Vision foundation models

Vision Foundation Models (VFMs) based on self-supervised learning (Caron *et al.*, 2021; Oquab *et al.*, 2024; Siméoni *et al.*, 2025), large-scale visual pre-training (Kirillov *et al.*, 2023; Ma *et al.*, 2024), or general vision transformers (He *et al.*, 2022), are trained on massive, diverse datasets. VFMs are primarily developed on natural images (Kirillov *et al.*, 2023; Siméoni *et al.*, 2025). While domain shifts typically necessitate adaptation modules to tailor these models to medical imagery (Dutt, Ericsson, Sanchez, Tsaftaris & Hospedales, 2024; Wu *et al.*, 2025b; Gaillochet, Noori, Dastani, Desrosiers & Lombaert, 2025), recent evidence suggests that the features of VFMs are sufficiently robust for direct application (Liu *et al.*, 2025). Specifically, DINOv3 (Siméoni *et al.*, 2025) has proven to be a highly effective off-the-shelf encoder for medical segmentation tasks.

4.3 Method

4.3.1 Framework overview

Let $X \in \mathbb{R}^{C\times H\times W}$ be an input image and $Y \in \{0,1\}^{\Omega}$ its ground-truth binary segmentation mask for $\Omega = \{1,\dots,H\}\times\{1,\dots,W\}$ denoting the pixel grid. Let f be a segmentation model whose output $S^{(0)} = f(X) \in [0,1]^{\Omega}$ is a pixel-wise foreground probability map. Given a calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$ and a user-specified error level $\alpha \in (0,1)$, our goal is to construct, for a new test point (X_{n+1}, Y_{n+1}), a prediction set $C_{\hat{\lambda}}(X_{n+1}) \subseteq \Omega$ whose expected False Negative Rate is bounded by α while being spatially coherent and as tight as possible.

Our proposed Random-Walk Conformal Prediction (RW-CP) framework produces statistically valid prediction sets in three stages (see Fig. 4.1):

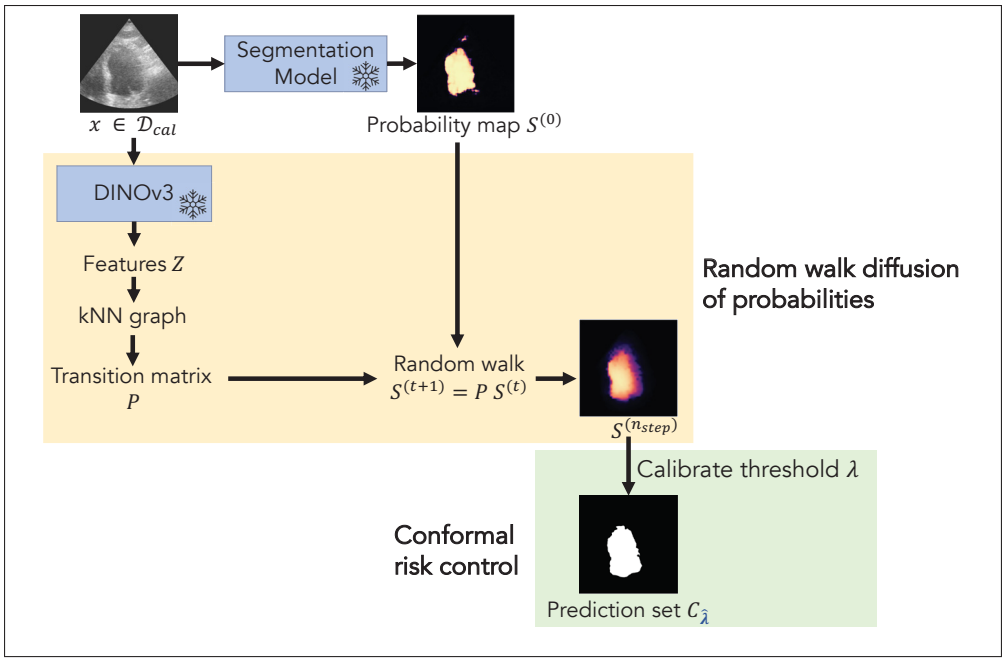


Figure 4.1 **Overview of our proposed RW-CP framework.** Given a calibration image, a trained segmentation model produces a probability map $S^{(0)}$. To diffuse the raw softmax probabilities, a frozen DINOv3 encoder extracts pixel embeddings Z to build a transition matrix P for a k-NN graph. Then, a random walk diffuses $S^{(0)}$ over n_{step} steps to yield the probability map $S^{(n_{\text{step}})}$. Conformal risk control then calibrates the threshold λ to yield the final prediction set $C_{\hat{\lambda}}$ with statistical guarantees

(i) Segmentation prediction. A trained segmentation network f maps X to a pixel-wise probability map $S^{(0)} = f(X)$.

(ii) Feature-guided random-walk diffusion. In parallel, a frozen vision foundation model Φ encodes X into features $Z = \Phi(X)$. These are used to produce a k-NN graph and a transition matrix P. The probabilities $S^{(0)}$ are then diffused over multiple random-walk steps using P.

(iii) Conformal risk control calibration. Using a held-out calibration set, conformal risk control (Angelopoulos *et al.*, 2024) selects the threshold $\hat{\lambda}$ on the probabilities to bound the expected FNR.

RW-CP requires no model retraining or gradient updates, and only needs to be rerun if the target error rate α or the test data distribution changes. It is hence easy to add on top of an existing segmentation pipeline.

4.3.2 Segmentation prediction

The first stage of the pipeline is a segmentation network f that maps an input image X to an initial pixel-wise foreground probability map $S^{(0)} = f(X)$. RW-CP is model-agnostic and treats f as a black box: any architecture that outputs a softmax (or sigmoid) foreground map can be plugged in, and no access to internal features, gradients, or training data is required. Throughout this work, we instantiate f as a standard 4-layer UNet (Ronneberger *et al.*, 2015), trained once per task with a supervised cross-entropy loss and training hyperparameters reported in Section **??**. At calibration and inference time, f is frozen. RW-CP only post-processes its outputs $S^{(0)}$.

4.3.3 Feature-guided random-walk diffusion

Raw network predictions are known to be overconfident: values tend to be compressed near 0 or 1 (Guo *et al.*, 2017). This makes the conformal thresholding of softmax probabilities extremely sensitive to small fluctuations and leads to fragmented prediction sets. We therefore diffuse $S^{(0)}$ through a random-walk process (Grady, 2006) guided by a semantically rich embedding space, in order to spread confident probabilities across anatomically coherent regions. This diffusion process improves the smoothness of the conformal score function which, as shown in Appendix 1, leads to more stable conformal sets.

4.3.3.1 Feature embedding and k-NN graph

A frozen pre-trained representation model Φ (e.g. DINOv3 (Siméoni *et al.*, 2025)) encodes the image into pixel-wise descriptors $Z = \{z_j\}_{j=1}^{M} \in \mathbb{R}^d$, where M is the total number of embedding-space pixels. These embeddings capture local appearance and semantic similarity

beyond raw intensity, enabling propagation between similar pixels rather than strictly adjacent ones.

We then construct a graph where each node represents a pixel. To keep computation tractable, for each pixel j, we restrict the graph to its k nearest neighbours $N(j)$ in the embedding space, based on cosine similarity. We define transition weights between pixels j and k as

$$w_{jk} = \exp\left(-\beta \cdot dist(z_j, z_k)\right), \quad k \in N(j), \tag{4.1}$$

where $dist(\cdot, \cdot)$ denotes the cosine distance.

This exponential kernel assigns higher transition probability to neighbours with similar embeddings while exponentially suppressing the influence of dissimilar pixels.

4.3.3.2 Random-walk diffusion

We then form the row-normalised transition matrix P, whose entry P_{ij} represents the probability of moving from pixel i to a neighbour j:

$$P_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}. \tag{4.2}$$

By construction, each row of P sums to 1, making it a valid probability transition operator. Having retained only the k-nearest neighbours per pixel, P can be stored as a sparse matrix with $O(kHW)$ instead of $O(H^2W^2)$ non-zero entries, enabling tractable computation.

The initial probability map $S^{(0)} = f(X)$ is resampled to match the spatial dimensions of the embedding Z and is then propagated through the feature-guided graph via the linear diffusion rule

$$S^{(t+1)} = P\, S^{(t)}, \tag{4.3}$$

After n_{step} diffusion steps, the resulting probability map $S^{(n_{\text{step}})}$ is resampled back to the original ground-truth dimensions $H \times W$ to construct the final conformal prediction set.

4.3.3.3 Role of hyperparameters

The hyperparameters k, β and n_{step} jointly control the diffusion. The number of neighbors k sets the graph connectivity: larger k accelerates propagation but increases the risk of leakage across anatomical boundaries (Grady, 2006). The scale parameter $\beta > 0$ controls the sharpness of this decay: small values encourage broader diffusion, whereas large values restrict propagation to highly similar regions and, in the limit, reduce P to the identity. The number of steps n_{step} governs a bias-variance trade-off, as repeated application of P drives $S^{(t)}$ toward its stationary distribution. Too few steps leave the map under-smoothed and $\hat{\lambda}$ sensitive to local fluctuations in $S^{(0)}$, whereas too many over-smooth the map and wash out anatomical boundaries. The link between the smoothness of $S^{(n_{\text{step}})}$ and the stability of $C_{\hat{\lambda}}(X)$ is formalized in the Appendix.

4.3.4 Conformal risk control calibration

We now describe how the refined map $S^{(n_{\text{step}})}$ is used to build a statistically valid prediction set.

4.3.4.1 Split conformal prediction

Given a calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$, split conformal prediction (Angelopoulos & Bates, 2023) constructs a post-processed prediction set $C(X)$ that is *marginally valid*. In other words, given a specified error rate $\alpha \in (0, 1)$, the set satisfies

$$\mathbb{P}\left[Y_{n+1} \notin C(X_{n+1})\right] \leq \alpha, \tag{4.4}$$

where (X_{n+1}, Y_{n+1}) is a new test point.

This probability is computed over the randomness of all $n + 1$ samples (i.e. the calibration set and the test point).

4.3.4.2 Conformal risk control

For segmentation, we adopt conformal risk control (CRC) (Angelopoulos *et al.*, 2024), which generalizes (4.4) to any bounded, monotone, non-increasing loss $\mathcal{L}\big(Y_i, C_\lambda(X_i)\big)$ indexed by a scalar threshold $\lambda \in [0, 1]$. The goal is to find the smallest $\hat{\lambda}$ such that the expected risk on future data is bounded by α:

$$\mathbb{E}\big[\mathcal{L}\big(Y_{n+1}, C_{\hat{\lambda}}(X_{n+1})\big)\big] \;\leq\; \alpha. \tag{4.5}$$

This guarantee holds for any finite n, provided the calibration and test data are exchangeable. Letting $\hat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\big(Y_i, C_\lambda(X_i)\big)$ denote the empirical calibration risk and B the upper bound of the loss, (4.5) holds for

$$\hat{\lambda} \;=\; \inf\left\{\lambda : \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}. \tag{4.6}$$

A direct consequence is that attaining a target α requires at least $n \geq B/\alpha - 1$ calibration samples. When $B = 1$, introducing the finite-sample inflated target $\alpha^\star = \frac{n+1}{n}\alpha - \frac{1}{n}$ rewrites (4.6) as

$$\hat{\lambda} \;=\; \inf\left\{\lambda : \hat{R}_n(\lambda) \leq \alpha^\star\right\}. \tag{4.7}$$

When the individual losses $\mathcal{L}_i$ are i.i.d., CRC also admits a lower bound that tightens with n:

$$\mathbb{E}\big[\mathcal{L}\big(Y_{n+1}, C_{\hat{\lambda}}(X_{n+1})\big)\big] \;\geq\; \alpha - \frac{2B}{n+1}. \tag{4.8}$$

4.3.4.3 RW-CP prediction set and non-conformity score

Unlike standard CRC for segmentation which builds the prediction set from raw model probabilities, our RW-CP framework thresholds the spatially-refined map $S^{(n_{\text{step}})}$. For a threshold

Algorithm 4.1 RW-CP Calibration

Input: target α, $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$, output probabilities $\{S_i^{(0)}\}_{i=1}^{n}$, pre-trained encoder Φ, neighbors k, walk steps n_{step}, scale β

for $i = 1 \dots n$ **do**
  $Z_i \leftarrow \Phi(X_i)$;
  Compute w_{jk} for k-nearest neighbours of j using Z_i ; /* (eq. 4.1) */
  Build P_i ; /* (eq. 4.2) */
  Row-normalize P_i;
  $S_i^{(n_{\text{step}})} \leftarrow (P_i)^{n_{\text{step}}} \cdot S_i^{(0)}$; /* Random-walk diffusion (eq. 4.3) */
end for
$\alpha^\star \leftarrow \frac{n+1}{n}\alpha - \frac{1}{n}$; /* Finite-sample inflated target */
$\hat{\lambda} \leftarrow \min_{\lambda \in [0,1]} \frac{1}{n}\sum_i \text{FNR}(Y_i, \{S_i^{(n_{\text{step}})} \geq 1 - \lambda\}) \leq \alpha^\star$;

Output: Threshold $\hat{\lambda}$

λ, the RW-CP prediction set is given by

$$C_\lambda(X) = \left\{p \in \Omega : S_p^{(n_{\text{step}})} \geq 1 - \lambda\right\}. \tag{4.9}$$

As λ grows, $C_\lambda(X)$ becomes larger and more conservative. Following standard practice for segmentation (Angelopoulos *et al.*, 2024), we use the False Negative Rate as the non-conformity score:

$$\mathcal{L}_i(\lambda) = \text{FNR}\big(Y_i, C_\lambda(X_i)\big) = 1 - \frac{|Y_i \cap C_\lambda(X_i)|}{|Y_i|}. \tag{4.10}$$

Since FNR $\in [0, 1]$, the loss is bounded with $B = 1$, so CRC selects $\hat{\lambda}$ via (4.7) such that the set $C_{\hat{\lambda}}(X_{n+1})$ covers, on average, at least $(1 - \alpha) \cdot 100\%$ of the true foreground pixels.

The full RW-CP procedure to calibrate $\hat{\lambda}$ is summarized in Algorithm 4.1, and the procedure to generate the final prediction set $C_{\hat{\lambda}}(X_{n+1})$ for a test image is summarized in Algorithm 4.2.

Algorithm 4.2 RW-CP Inference

Input: Test sample X_{test}, prediction $S_{test}^{(0)}$, pre-trained encoder Φ, k, n_{step}, β, calibrated $\hat{\lambda}$

$Z_{test} \leftarrow \Phi(X_{test})$;
Compute w_{ij} for k-nearest neighbours j of i using Z_{test};
Build P_{test};
Row-normalize P_{test};
$S_{test}^{(n_{\text{step}})} \leftarrow (P_{test})^{n_{\text{step}}} \cdot S_{test}^{(0)}$; `/* Random-walk diffusion */`
$C_{\hat{\lambda}}(X_{test}) \leftarrow \{p : (S_{test}^{(n_{\text{step}})})_p \geq 1 - \hat{\lambda}\}$;

Output: $C_{\hat{\lambda}}(X_{test})$

4.4 Experiment and results

4.4.1 Datasets

We validate our proposed method across three imaging modalities: ultrasound (US), MRI, and CT. We utilized the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset (Leclerc *et al.*, 2019), employing 10 samples for training and 100 for testing. For the Automated Cardiac Diagnosis Challenge (ACDC) MRI dataset (Bernard *et al.*, 2018) and pancreas CT dataset from the Medical Segmentation Decathlon (MSD) (Antonelli *et al.*, 2022), we extract the 2D slice with the largest foreground area from each 3D volume. The ACDC experiments focus on right ventricle (RV) and left ventricle (LV) segmentation using 50 training and 50 test images at end-diastole. For MSD Pancreas, 81 images are used for training and 100 for testing.

In our main experiments, a calibration set of $n = 20$ images is held out from the remaining samples. This represents the minimum sample size required to satisfy the conditions for conformal prediction across all tested significance levels ($\alpha \in \{0.2, 0.1, 0.05\}$).

Preprocessing: Images are resampled to a resolution of 1 mm $\times$ 1 mm. Intensity values are clipped to the 0.5^{th} and 99.5^{th} percentiles to remove outliers and rescaled to the $[0, 255]$

range. For the MSD Pancreas CT images specifically, Hounsfield Units (HU) are clipped to the $[-100, 240]$ range prior to rescaling, following (Man, Huang, Feng, Li & Wu, 2019). Finally, all samples are cropped and resized to a fixed resolution of 512×512 pixels.

4.4.2 Implementation details

4.4.2.1 Segmentation model training

Initial segmentation masks are generated using a standard 4-layer UNet (Ronneberger *et al.*, 2015), serving as a representative proxy for widely adopted medical segmentation backbones. The models are trained independently for each task. Training is performed for 200 epochs with a batch size of 16, using a supervised CE loss with the Adam optimizer (Kingma & Ba, 2015), a learning rate of 0.001, gradual warmup with a cosine annealing scheduler (Loshchilov & Hutter, 2017; Goyal *et al.*, 2018) and a weight decay of 0.0001.

4.4.2.2 Random walk diffusion

We apply our random-walk diffusion on a k-nearest neighbour (kNN) graph constructed from feature maps obtained with the DINOv3 model (Siméoni *et al.*, 2025). To compute the transition matrix P, we use $k = 20$ neighbours and $\beta = 50$. The random walk runs for $n_{\text{step}} = 10$ diffusion steps on the predicted probability map. These hyperparameters were determined by optimizing for the Dice Similarity score on the calibration set.

The output probability map $S^{(0)} = f(X)$ is resampled to match the spatial dimensions of the embedding Z. After n_{step} diffusion steps, the resulting probability map $S^{(n_{\text{step}})}$ is resampled back to the original ground-truth dimensions $H \times W$ to construct the final conformal prediction set.

4.4.2.3 Experimental set-up

We perform each experiment 3 times, with varying initialization seeds. This procedure results in three distinct calibration sets, contributing to the reported standard deviations.

4.4.3 Evaluation metrics

We evaluate both the statistical validity and geometric quality of the generated conformal prediction sets $C_{\hat{\lambda}}(X)$.

4.4.3.1 Statistical validity metrics

These metrics assess whether the conformal set fulfills the coverage guarantee and the associated cost of doing so.

- Empirical coverage measures the average fraction of true foreground pixels in the conformal set:

$$\text{coverage} = \frac{|C_{\hat{\lambda}}(X) \cap Y|}{|Y|} \in [0, 1] \tag{4.11}$$

  Coverage is linked to the conformity score via FNR = 1 − coverage.

- Stretch measures the relative size increase of the conformal set $C_{\hat{\lambda}}(X)$ compared to original predicted mask $\hat{Y}$:

$$\text{stretch} = \frac{|C_{\hat{\lambda}}(X)|}{|\hat{Y}|} \tag{4.12}$$

  The stretch quantifies the cost of achieving coverage. Hence, for similar coverage, lower stretch is better.

4.4.3.2 Geometric quality metrics

These metrics assess the accuracy of the prediction set in terms of overlap and distance to the ground-truth.

- Dice Similarity Coefficient (DSC) quantifies the overlap between $C_{\hat{\lambda}}(X)$ and the ground truth mask:

$$\text{DSC} = \frac{2 \cdot |C_{\hat{\lambda}}(X) \cap Y|}{|C_{\hat{\lambda}}(X)| + |Y|} \tag{4.13}$$

  Values range from 0 (no overlap) to 1 (perfect overlap).

- Average Symmetric Surface Distance (ASSD) measures the average shortest distance between the contours C_C of the conformal set and the contour of the ground truth C_Y, and vice-versa:

$$\text{ASSD} = \frac{1}{|C_C| + |C_Y|}\Big(\sum_{p \in C_C} d(p, C_Y) + \sum_{q \in C_Y} d(q, C_C) \Big) \tag{4.14}$$

  where $d(i, C_J) = \min_{j \in C_J} \|i - j\|_2$ is the shortest Euclidean distance from a point i to the contour C_J. The ASSD provides a balanced measure of contour accuracy.

- Hausdorff Distance (HD) measures the maximum shortest distance from all points on the contour of one shape to the contour of the other:

$$\text{HD} = \max \left\{ \max_{p \in C_C} d(p, C_Y), \max_{q \in C_Y} d(q, C_C) \right\} \tag{4.15}$$

  For stability, we use the 95th percentile Hausdorff Distance (HD95).

4.4.4 Results

We assess the effectiveness of our proposed Random-Walk Conformal Prediction framework by benchmarking it against standard conformal risk control (CRC), applied directly to unmodified output probabilities (Angelopoulos & Bates, 2023; Angelopoulos *et al.*, 2024), as well as against a state-of-the-art split conformal prediction method for image binary segmentation, specifically Consema (Mossina & Friedrich, 2025), which thresholds the number of morphological dilations applied to the predicted mask.

To evaluate robustness and generalizability, we perform extensive experiments across datasets encompassing different modalities and target tissues. Furthermore, we conduct ablation studies to quantify the influence of key design choices, including calibration set size and random-walk hyperparameters, on predictive coverage and segmentation accuracy.

Table 4.1 evaluates the performance of different conformal prediction methods across various risk levels α. Our proposed method, RW-CP, consistently demonstrates strong performance, particularly in terms of segmentation quality metrics. Except for CAMUS at a tight risk level

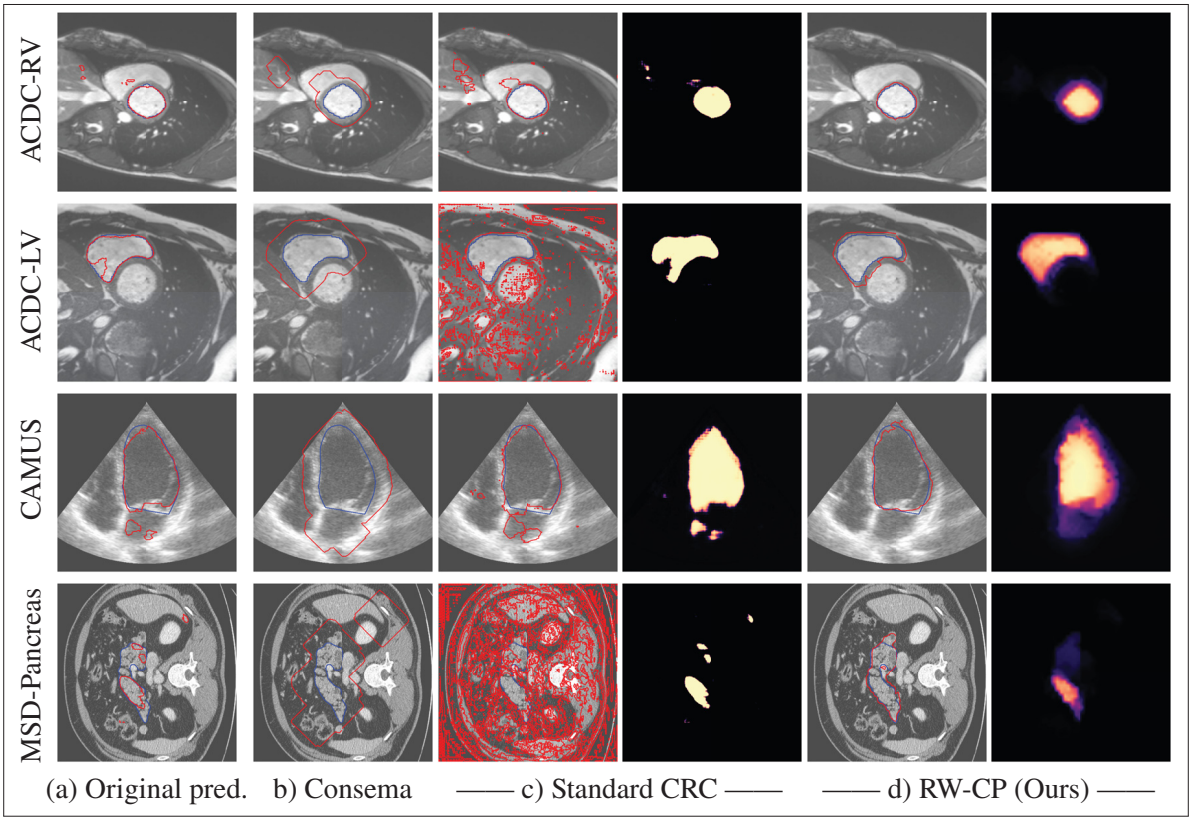


Figure 4.2 **Examples of generated prediction masks (red) and ground-truths (blue), and associated probability maps**. From left to right, (a) the original predicted mask with no conformal guarantees, (b–d) the conformalized prediction masks with $\alpha = 0.1$. Specifically, b) Consema applies a fixed number of dilatations on the original prediction (Mossina & Friedrich, 2025), c) Standard CRC thresholds the model's raw output probabilities (Angelopoulos *et al.*, 2024), and (d), our method RW-CP, thresholds the probabilities after diffusion with a random walk. Columns 4 and 6 show the probability map used, where lighter colour indicates a value close to 1 and darker colour indicates a value close to 0. Our method is able to successfully diffuse the probabilities and uncertainties, making the conformalized prediction masks closer to the ground truth, despite initially incorrect predictions

$\alpha = 0.05$, RW-CP consistently achieves the best DSC, ASSD, and HD95 values while maintaining the required coverage level above $1 - \alpha$. In addition, RW-CP exhibits much lower stretch than standard CRC and Consema, which indicates tighter prediction sets. Furthermore, Fig. 4.3 shows that RW-CP maintains higher performance stability across varying confidence levels, avoiding the sharp accuracy declines of other existing methods.

Table 4.1 **Results summary of split-CP on the test set, for different error rate constraints α.** For each dataset, the top row shows the model performance before applying conformal prediction guarantees. The best CP results are shown in bold. RW-CP achieves the best segmentation performance in terms of overlap- and distance-based metrics, while maintaining the required coverage level above $(1-\alpha)$

Dataset	α	Method	Coverage (↑)	Stretch (↓)	DSC (↑)	ASSD (↓)	HD95 (↓)
ACDC LV	-	-	0.86	-	86.97	21.14	119.31
	0.2 (Moderate)	Standard CRC	0.865 ±0.016	1.00 ±0.04	86.91 ±0.33	**21.76** ±1.74	123.74 ±12.08
		Consema	0.962 ±0.012	1.63 ±0.21	76.71 ±4.70	29.73 ±3.59	145.07 ±2.84
		RW-CP (Ours)	0.856 ±0.030	**0.90** ±0.05	**88.12** ±1.30	30.32 ±8.02	**44.02** ±6.93
	0.1 (Tight)	Standard CRC	0.959 ±0.008	1.88 ±0.55	72.62 ±10.27	91.96 ±38.84	255.36 ±22.00
		Consema	0.972 ±0.003	1.83 ±0.11	72.34 ±2.22	33.09 ±1.72	147.74 ±1.36
		RW-CP (Ours)	0.961 ±0.019	**1.36** ±0.23	**82.86** ±5.75	**26.40** ±4.67	**53.96** ±16.48
	0.05 (Very tight)	Standard CRC	0.998 ±0.001	30.80 ±0.87	8.12 ±0.20	166.38 ±2.47	283.06 ±1.89
		Consema	0.990 ±0.014	4.61 ±2.32	45.63 ±21.46	62.99 ±24.60	172.55 ±20.46
		RW-CP (Ours)	0.990 ±0.005	**2.61** ±0.56	**59.92** ±8.08	**38.49** ±10.22	**112.19** ±22.66
ACDC RV	-	-	0.85	-	83.05	32.36	155.70
	0.2 (Moderate)	Standard CRC	0.867 ±0.019	1.04 ±0.06	82.83 ±0.34	33.89 ±2.39	161.50 ±6.91
		Consema	0.957 ±0.002	1.74 ±0.04	69.61 ±0.75	41.26 ±0.61	181.28 ±0.89
		RW-CP (Ours)	0.878 ±0.020	**0.94** ±0.04	**87.02** ±0.60	**22.74** ±0.71	**59.94** ±8.61
	0.1 (Tight)	Standard CRC	0.957 ±0.008	2.39 ±0.84	59.55 ±11.56	127.51 ±14.14	311.90 ±0.93
		Consema	0.981 ±0.011	2.64 ±0.67	56.17 ±8.43	54.22 ±9.08	195.98 ±8.23
		RW-CP (Ours)	0.962 ±0.013	**1.35** ±0.27	**80.61** ±6.64	**23.55** ±8.05	**113.27** ±29.73
	0.05 (Very tight)	Standard CRC	0.998 ±0.001	19.91 ±0.29	10.83 ±0.14	165.29 ±0.75	330.90 ±3.62
		Consema	0.994 ±0.004	4.14 ±1.34	42.12 ±8.76	73.45 ±16.23	212.61 ±13.46
		RW-CP (Ours)	0.996 ±0.001	**3.17** ±0.38	**50.37** ±3.82	**62.09** ±4.95	**191.08** ±6.33
CAMUS	-	-	0.87	-	80.85	19.66	78.06
	0.2 (Moderate)	Standard CRC	0.845 ±0.013	0.94 ±0.03	81.07 ±0.04	18.53 ±0.53	74.26 ±2.24
		Consema	0.961 ±0.006	1.44 ±0.06	73.23 ±1.23	26.61 ±1.21	93.93 ±1.40
		RW-CP (Ours)	0.848 ±0.010	**0.88** ±0.02	**84.20** ±0.11	**12.02** ±0.03	**39.70** ±0.31
	0.1 (Tight)	Standard CRC	0.952 ±0.005	1.39 ±0.05	74.12 ±1.03	32.05 ±2.09	109.06 ±3.25
		Consema	0.982 ±0.000	1.74 ±0.00	66.83 ±0.00	33.66 ±0.00	101.60 ±0.00
		RW-CP (Ours)	0.954 ±0.006	**1.23** ±0.04	**79.75** ±0.82	**17.19** ±0.88	**55.45** ±2.77
	0.05 (Very tight)	Standard CRC	0.997 ±0.001	8.29 ±0.49	20.85 ±1.07	133.95 ±4.02	242.25 ±4.66
		Consema	0.994 ±0.005	**2.38** ±0.43	**55.75** ±6.75	**50.24** ±11.25	118.25 ±11.04
		RW-CP (Ours)	0.998 ±0.001	2.65 ±0.20	51.73 ±2.64	57.19 ±5.40	**115.76** ±7.07
MSD Pancreas	-	-	0.58	-	63.52	14.23	70.46
	0.2 (Moderate)	Standard CRC	0.818 ±0.031	7.72 ±2.89	45.19 ±8.93	103.03 ±22.09	263.47 ±11.89
		Consema	0.927 ±0.057	8.90 ±3.97	38.05 ±10.15	39.08 ±11.58	102.97 ±11.70
		RW-CP (Ours)	0.780 ±0.030	**1.77** ±0.21	**63.33** ±0.81	**44.81** ±3.60	**80.61** ±2.36
	0.1 (Tight)	Standard CRC	0.936 ±0.017	50.64 ±12.75	11.79 ±2.81	150.53 ±4.10	283.85 ±2.26
		Consema	0.980 ±0.010	17.38 ±4.31	25.23 ±4.18	60.59 ±9.93	124.32 ±9.71
		RW-CP (Ours)	0.923 ±0.037	**4.34** ±1.56	**49.62** ±7.79	**30.96** ±3.76	**81.60** ±7.57
	0.05 (Very tight)	Standard CRC	0.998 ±0.001	140.49 ±4.67	5.17 ±0.12	157.43 ±0.36	300.36 ±0.33
		Consema	0.988 ±0.005	23.15 ±5.91	20.93 ±3.65	72.76 ±11.84	136.32 ±11.71
		RW-CP (Ours)	0.987 ±0.012	**16.00** ±4.52	**25.17** ±5.44	**62.05** ±12.30	**131.10** ±16.64

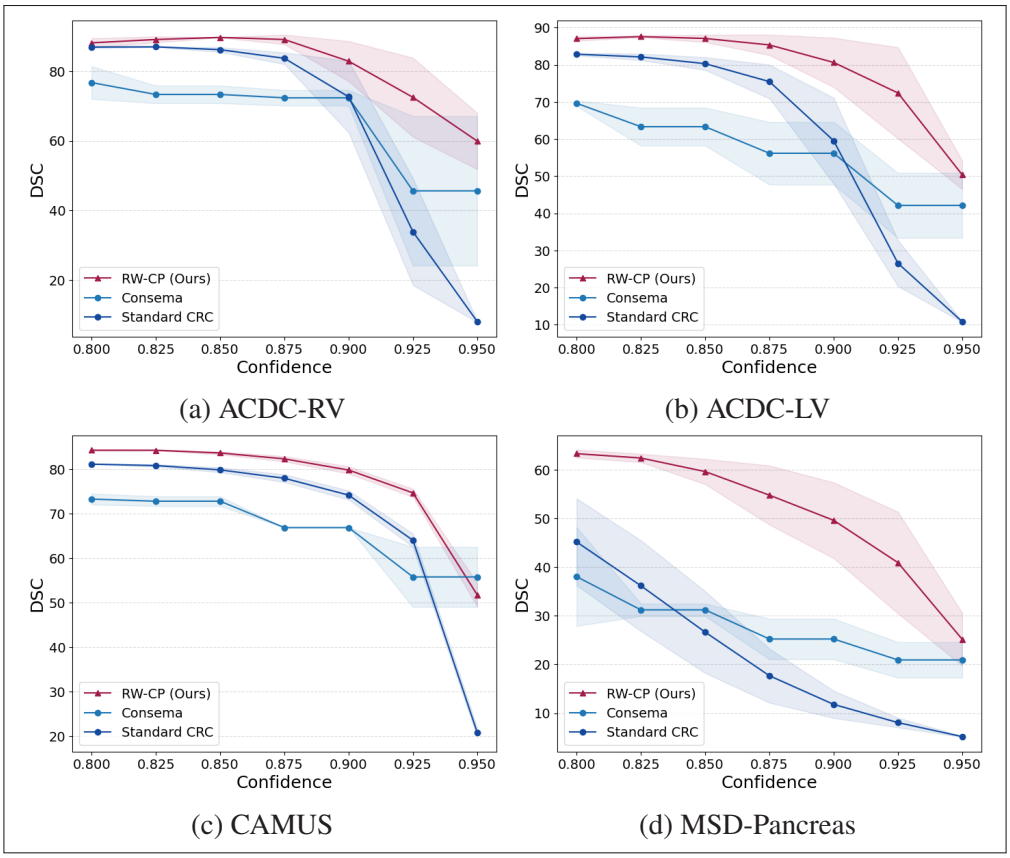


Figure 4.3 **Mean dice score for different confidence levels** $(1-\alpha) \in [0.8, 0.95]$, given a calibration set of 20 samples, for the (a) ACDC-RV, (b) ACDC-LV, (c) CAMUS and (d) MSD-Pancreas datasets. Our method, RW-CP (red), outperforms other CP methods (blue) across different confidence values

Visually, we observe in Fig. 4.2 that RW-CP is able to produce prediction sets much closer to the ground-truth than standard CRC or Consema. Our approach is even able to remove small over-segmented areas, whereas methods such as Consema can only increase the size of the prediction set, even when the segmentation model produced false positives.

We hypothesize that the performance of RW-CP stems from its pre-processing step on probabilities, which mitigates the overconfidence commonly observed in deep learning segmentation models (Guo *et al.*, 2017). Their raw softmax probabilities are often skewed towards either 0 or 1 (see Fig. 4.2.c). This overconfidence compresses the effective range of possible values for

the empirical $\hat{\lambda}$ threshold used in Conformal Risk Control, making the final prediction mask extremely sensitive to small fluctuations in $\hat{\lambda}$. On the contrary, by first applying a diffusion process, the pixel-wise probabilities become more varied and less polarized (as shown in Fig. 4.2.d). This broadened distribution stabilizes the empirical $\hat{\lambda}$ during calibration, leading to a more robust and tighter prediction set, which translates to better overall segmentation metrics.

4.4.5 Ablation study

We assess the impact of calibration set size and random walk hyperparameters by performing an ablation study on the CAMUS test set with an error-rate fixed to $\alpha = 0.1$.

4.4.5.1 Impact of calibration set size

Table 4.2 **Performance of conformal prediction sets with varying calibration set sizes.** The best results are in bold. A larger calibration set results in better overlap- and distance-based segmentation metrics and an empirical coverage closer to the $(1 - \alpha)$ confidence target

# samples	Coverage (↑)	DSC (↑)	ASSD (↓)	HD95 (↓)
10	0.991 ±0.007	64.05 ±7.29	36.84 ±10.12	87.46 ±14.71
20	0.954 ±0.006	79.75 ±0.82	17.19 ±0.88	55.45 ±2.77
100	0.924 ±0.001	**82.69** ±0.06	**14.15** ±0.08	**47.87** ±0.15

We first evaluate the influence of the calibration dataset size on performance. The results, detailed in Table 4.2, compare the performance when using 10, 20, and 100 images for the calibration set. Increasing the calibration set size significantly improves the overall quality of the prediction sets. The coverage is greatest for the smallest set (10 samples) and draws closer to $(1 - \alpha)$ as the set size grows. Larger calibration sets lead to more stable and less conservative coverage guarantees, corroborating (4.8).

4.4.5.2 Impact of random walk hyperparameters

We examine the impact of random walk hyperparameters such as the number of neighbours k, the scaling factor β and the number of diffusion steps n_{step}.

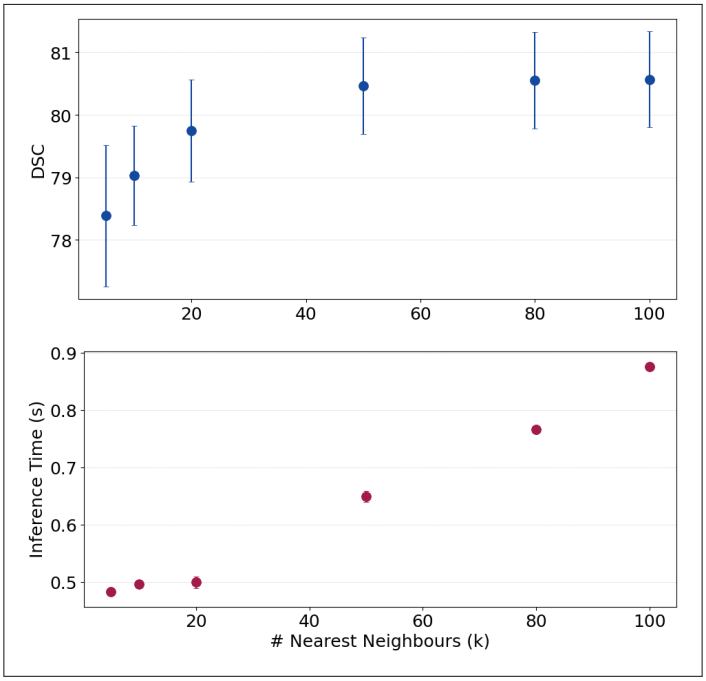


Figure 4.4 **Impact of the number of nearest neighbours** (1, 5, 10, 20, 50, 80 and 100) used to compute the random walk transition matrix on the model performance and inference time per sample. Considering $k = 20$ neighbours for each pixel yields a high dice score while maintaining low inference time

We vary the number of nearest neighbours k used in the random walk and observe the resulting conformal set. Fig. 4.4 shows the change in the Dice Similarity score and inference time with k. Increasing k generally improves the accuracy of the prediction set in terms of Dice Similarity score. However, this performance gain comes at the cost of increased computational complexity and higher inference time, as the size of the sparse transition matrix and the inference time grow linearly with k. We observe that when $k = 20$, we obtain competitive results in terms of segmentation metric and computational efficiency.

Table 4.3 **Impact of β on the conformal prediction set.** Segmentation accuracy of the conformal sets peaks at $\beta = 50$

β	FNR (↓)	DSC (↑)	ASSD (↓)	HD95 (↓)
0.1	0.052 ±0.007	78.22 ±0.86	18.50 ±0.91	56.35 ±2.04
0.5	0.052 ±0.007	78.24 ±0.85	18.48 ±0.91	56.28 ±2.05
1	0.052 ±0.007	78.27 ±0.85	18.46 ±0.92	56.20 ±2.06
5	0.051 ±0.007	78.46 ±0.83	18.24 ±0.90	55.57 ±1.97
10	0.051 ±0.006	78.68 ±0.80	17.98 ±0.86	**54.86** ±1.86
50	0.046 ±0.006	**79.75** ±0.82	**17.19** ±0.88	55.45 ±2.77
100	**0.045** ±0.008	78.30 ±1.20	20.07 ±1.22	72.49 ±2.47

We then investigate the impact of the scaling factor β, which controls the sharpness of the transition kernel (eq. 4.1). As shown in Table 4.3, the segmentation accuracy increases slowly with β, until a peak at $\beta = 50$, which yields the best DSC (79.75%) and ASSD (17.19). Higher β values ($\beta = 100$) restrict the diffusion to an overly local neighborhood, causing performance degradation.

Table 4.4 **Impact of random walk diffusion steps on segmentation performance**

Steps	FNR (↓)	DSC (↑)	ASSD (↓)	HD95 (↓)
1	0.049 ±0.007	77.38 ±1.00	21.69 ±1.07	81.00 ±2.38
5	**0.046** ±0.006	79.33 ±0.88	18.28 ±0.96	62.62 ±1.87
10	**0.046** ±0.006	**79.75** ±0.82	17.19 ±0.88	55.45 ±2.77
20	0.051 ±0.006	79.40 ±0.64	**17.17** ±0.71	**51.30** ±1.28
50	0.058 ±0.004	73.50 ±0.35	23.35 ±0.44	66.11 ±0.79

Finally, we examine the effect of the number of diffusion steps (n_{step}) on the final performance (Table 4.4). The number of steps determines how far the initial uncertainty information ($S^{(0)}$) is propagated through the feature graph. Using a single step ($n_{\text{step}} = 1$) results in poor geometric metrics (DSC 77.38%, HD95 81.00), indicating insufficient spatial smoothing. Performance significantly improves as n_{step} increases, peaking around $n_{\text{step}} = 10$ (DSC 79.75%) and $n_{\text{step}} = 20$ (lowest distance metrics). However, excessive diffusion, such as $n_{\text{step}} = 50$, leads to oversmoothing, where probabilities blend across distinct anatomical boundaries, causing a sharp degradation in DSC (73.50%) and an increase in distance metrics.

4.4.6 Limitations and future work

Building on the current findings, future work could extend the robustness and generalizability of RW-CP. First, to specialize the feature representations, subsequent work could explore domain-specific fine-tuning of foundation models or the integration of multi-scale hierarchies to capture more intricate anatomical nuances. Second, the granularity of the diffusion process could be augmented through spatially-adaptive rates that dynamically adjust diffusion strength based on local intensity gradients. Finally, scaling the framework to 3D volumetric segmentation would enable supporting an even broader range of complex clinical workflows.

4.5 Conclusion

In this work, we tackled the critical challenge of spatial incoherence in trustworthy medical image analysis. While conformal prediction offers statistical validity, traditional applications to segmentation often fail to account for anatomical context, resulting in fragmented, spatially incoherent, and over-segmented prediction sets. To address this issue, we present Random-Walk Conformal Prediction (RW-CP), a novel framework designed to generate statistically valid and anatomically informed prediction sets for medical image segmentation. RW-CP integrates a random-walk diffusion process guided by high-dimensional feature embeddings from pre-trained foundation models (e.g., DINOv3). This process effectively diffuses raw segmentation probabilities across semantically similar regions, creating a more spatially coherent and robust probability map. Our evaluations on MRI, ultrasound and CT datasets across various risk levels α demonstrate that RW-CP consistently achieves better segmentation accuracy (higher DSC, lower ASSD/HD95) compared to standard CRC and state-of-the-art conformal prediction methods. These improvements in anatomical plausibility are achieved while maintaining the required marginal coverage guarantees. By mitigating the issues of fragmentation and over-segmentation, RW-CP provides a path towards uncertainty quantification that is both statistically rigorous and clinically practical.

CONCLUSION

This thesis investigated whether accurate and reliable automatic segmentation of medical images can be achieved with reduced human annotation effort. In this section, we summarize our research contributions across the model development life-cycle. We then examine limitations of our work and propose possible directions for future work.

5.1 Summary of contributions

Objective 1: Prioritize annotation via stochastic batch active learning

Manual data annotation is tedious and expensive to acquire. In Chapter 2, we proposed a new active learning (AL) strategy that identifies the most relevant samples to annotate, in order to maximize segmentation performance with minimal annotation effort. Our stochastic batch AL method randomly generates batches of samples and selects the batch with the highest mean uncertainty score for annotation. Computing uncertainty at the batch level implicitly enforces diversity in the selection without additional computational cost, mitigating the tendency of purely uncertainty-based methods to query redundant samples, which wastes annotation resources and biases the model toward narrow regions of the data distribution. Experiments showed that stochastic batch active learning outperforms both random and purely uncertainty-based sampling, and remains robust to variations in training and sampling hyperparameters. Answering our initial question, we found that randomness serves as an effective criterion for selecting diverse samples and reducing annotation cost without sacrificing performance.

Impact: Identifying the most valuable samples to annotate and use for training enables smaller clinical centers with limited radiologist availability to develop their own high-performing models. In addition, optimizing the annotation of new datasets will make it possible to extend application

of automatic segmentation algorithms to new pathologies and rare clinical scenarios where large-scale annotation is unfeasible.

Objective 2: Improve weakly-supervised segmentation via prompt learning with constraints

Dense pixel-wise annotations are the most expensive form of supervision, yet remain the standard for specializing segmentation models. In Chapter 3, we introduced a prompt learning framework that specializes promptable vision foundation models to clinical tasks using only sparse bounding box labels. We train an auxiliary module to generate prompt embeddings automatically, guided by box-based spatial constraints and consistency regularization to compensate for the reduced label information. This eliminates the need for dense annotations or manual prompting at test time. Our framework generalizes across imaging modalities, dataset sizes and model backbones, establishing weakly-supervised prompt learning as a scalable alternative to fully-supervised specialization of both foundation and non-foundation models. Hence, answering our second question, weak supervision was indeed shown to reach near state-of-the-art performance when combined with promptable foundation models.

Impact: Weakly-supervised prompt learning enables the rapid specialization of large foundation models to clinical tasks without the computational or labeling overhead typically required. This scalability ensures that high-quality automated tools can be developed and deployed across a broader range of imaging modalities and medical centers.

Objective 3: Guarantee coverage of predicted segmentations via anatomically-aware conformal prediction

Reliable clinical deployment requires uncertainty estimates with rigorous guarantees on prediction errors. In Chapter 4, we developed a conformal prediction framework for binary segmentation that constructs post-processed prediction sets containing the true segmentation with a user-

specified probability. Our Random-Walk Conformal Prediction (RW-CP) method leverages the rich representations learned by foundation models to incorporate anatomical context into the uncertainty quantification process. Specifically, it applies a random walk mechanism guided by these representations to diffuse the output probabilities, encouraging spatially coherent and anatomically plausible boundaries in the resulting prediction sets. Our framework is model-agnostic and requires no retraining, making it a practical post-hoc addition to any trained segmentation model. Evaluations on multimodal datasets across various risk levels demonstrate that RW-CP consistently achieves better segmentation accuracy compared to standard and state-of-the-art conformal prediction methods, while maintaining coverage guarantees. Therefore, answering our last question, we found that foundation model features could enrich predicted probability maps with spatial context, improving the resulting conformal sets.

Impact: Enabling reliable uncertainty quantification allows clinicians to identify and prioritize cases that require expert review, ensuring that manual intervention is directed where it is most needed. Furthermore, unlike heuristic uncertainty measures, conformal prediction provides formal coverage guarantees that the medical community and regulators can rely on, helping improve trust and facilitating the broader adoption of computer-aided diagnosis.

5.2 Limitations and future research directions

While the contributions of this thesis address important challenges to reliable and annotation-efficient automatic segmentation, each proposed method carries its own assumptions and constraints. This section examines the key limitations of the presented work and discusses the research directions they motivate.

Fairness: Beyond performance, fairness has become a growing concern in medical image analysis. A segmentation model is considered fair if its performance does not systematically differ across patient groups defined by sensitive attributes such as race or sex. Active learning,

while primarily designed to reduce annotation cost, also shapes the demographic composition of the labeled set. Active learning can therefore implicitly promote model fairness. Recent preliminary work showed encouraging results in reducing group-wise performance disparities in segmentation (Danaee, Gaillochet, Desrosiers, Lombaert & Bouix, 2026). That said, these findings are limited to a specific anatomy and synthetic bias, and it remains unclear how well fairness-aware active learning generalizes to real-world medical tasks. Therefore, a natural next step would be to evaluate these fair AL strategies on datasets with real-world demographic imbalances.

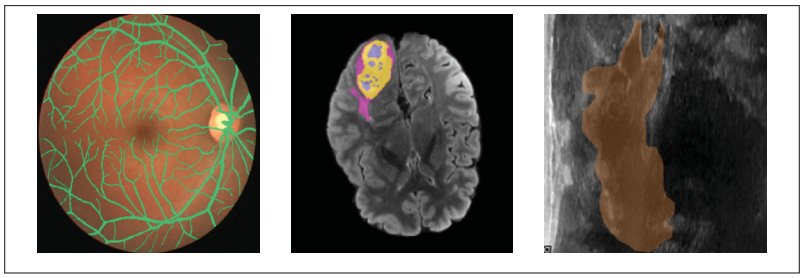


Figure 5.1 **Examples of complex anatomical structures.** The coloured overlays highlight anatomical structures with intricate shapes. For such structures, simple bounding-box annotations can be insufficient for training accurate segmentation models.
Taken from Bougourzi & Hadid (2025)

Complex tissue shape: Using bounding box annotations is substantially faster and less costly than pixel-wise masks, making them an appealing form of weak supervision for medical image segmentation. Bounding boxes are particularly effective when the target structure is compact and roughly convex, such as the spleen or cardiac ventricles, where the box tightly constrains the foreground region and leaves little ambiguity about the structure's extent. However, for anatomical structures with complex, elongated or irregularly shaped structures, a bounding box captures a large background region that provides little information about the true structure boundary (see Fig. 5.1). In such cases, the box-based constraints exploited during training can become loose and uninformative, limiting the guidance they provide. More expressive yet still

lightweight forms of weak supervision such as scribbles or sparse point clouds could be better suited to these geometrically complex structures and represent a natural direction for extending the proposed framework.

Pixel-level guarantees: A known fundamental limitation of conformal prediction is the marginal rather than conditional coverage it guarantees. This means that the coverage holds on average across test images but not necessarily for individual predictions or specific subgroups of patients. Extending the conformal prediction set-up to pixel-level guarantees is a fundamentally harder problem in segmentation, owing to the strong spatial interdependence between pixels. Yet, pixel-level conformal prediction represents a critical direction for future work, as individual guarantees are more actionable in clinical decision-making than population-level ones.

5.3 Overall conclusion

This thesis investigated whether annotation effort could be reduced without sacrificing segmentation quality or prediction reliability. The three contributions presented address this challenge at distinct stages of the model development life-cycle. The first research objective led to an active learning method that exploits stochasticity to identify the most valuable samples *before* training and produce the best segmentation models with minimal annotated data. The subsequent research objective led to the development of a prompt learning framework that specializes and automates vision foundation models using only weak labels *during* training to bypass the need for dense manual labels. The final research objective introduced a framework to produce spatially coherent prediction sets with guaranteed coverage *after* deployment. Beyond their methodological novelty, these contributions have practical implications for clinical workflows, including accelerating radiotherapy planning, supporting cardiac assessment and enabling analysis in rare disease settings where labeled data is inherently limited. Ultimately, this translates into a more reassuring patient experience, with shorter delays between imaging and treatment and greater confidence in diagnostic outcomes.

APPENDIX I

ADDITIONAL MATERIAL FOR CHAPTER 4

Stability of conformal sets and smoothness of the score function

Let $X \in \mathcal{X}$ be an input image defined on a discrete grid $\Omega \subset \mathbb{R}^2$, and $Y \in \{0,1\}^{\Omega}$ be a binary segmentation mask. Denote as $S : \Omega \to [0,1]$ the pixelwise non-conformity score function. For a calibration-derived threshold λ ensuring marginal coverage $1-\alpha$, the conformal prediction set is given by

$$C_\lambda(X) = \big\{p \in \Omega : S(p) \geq 1-\lambda\big\}.$$

We define the *tightness* of the conformal set as the number of pixels it contains:

$$\text{Tightness}(C_\lambda(X)) = |C_\lambda(X)| = \int_\Omega \mathbf{1}_{\{S(p)\geq 1-\lambda\}}\, dp,$$

where the integral corresponds to pixel counting.

Assume that S is Lipschitz continuous. By the coarea formula, for any integrable function $g : \Omega \to \mathbb{R}$,

$$\int_\Omega g(p)\, dp = \int_{-\infty}^{+\infty} \left(\int_{S^{-1}(t)} \frac{g(p)}{\|\nabla S(p)\|}\, d\mathcal{H}^1(p) \right) dt,$$

where $\mathcal{H}^1$ denotes the one-dimensional Hausdorff measure and $S^{-1}(t)$ is the level set for a value $t \geq 1-\lambda$ (i.e., the set of pixels p such that $S(p) = t$). Choosing $g(p) = \mathbf{1}_{\{S(p)\geq 1-\lambda\}}$, we obtain

$$|C_\lambda(X)| = \int_{1-\lambda}^{1} \left(\int_{S^{-1}(t)} \frac{1}{\|\nabla S(p)\|}\, d\mathcal{H}^1(p) \right) dt.$$

Deriving with respect to λ gives

$$\frac{d|C_\lambda(X)|}{d\lambda} = \int_{S^{-1}(\lambda)} \frac{1}{\|\nabla S(p)\|}\, d\mathcal{H}^1(p).$$

This expression indicates that a non-conformity score function with regions of small gradient $\|\nabla S(p)\|$ can produce large variations in the size of the conformal set. In contrast, score functions that change more uniformly over the image tend to generate more stable conformal sets. By spatially diffusing the scores over semantically-related pixels via a random walk, our RW-CP method thus enhances the robustness to the choice of threshold λ using the calibration set.

APPENDIX II

TAAL: TEST-TIME AUGMENTATION FOR ACTIVE LEARNING IN MEDICAL IMAGE SEGMENTATION

Mélanie Gaillochet[1] , Christian Desrosiers[1] , Hervé Lombaert[1]

[1] Department of Software and IT Engineering , École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Published in the *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections (MICCAI-DALI)*, September 2022

Presentation

This appendix presents the article *"TAAL: Test-time Augmentation for Active Learning in Medical Image Segmentation"* presented in 2022 at the **MICCAI Workshop** on Data Augmentation, Labelling, and Imperfections. The article precedes and motivates the active learning contribution of Chapter 2. It introduces a semi-supervised active learning approach that uses prediction disagreement across augmented views as a sample selection criterion.

Abstract

Deep learning methods typically depend on the availability of labeled data, which is expensive and time-consuming to obtain. Active learning addresses such effort by prioritizing which samples are best to annotate in order to maximize the performance of the task model. While frameworks for active learning have been widely explored in the context of classification of natural images, they have been only sparsely used in medical image segmentation. The challenge resides in obtaining an uncertainty measure that reveals the best candidate data for annotation. This paper proposes Test-time Augmentation for Active Learning (TAAL), a novel semi-supervised active learning approach for segmentation that exploits the uncertainty information offered by data transformations. Our method applies cross-augmentation consistency during training and inference to both improve model learning in a semi-supervised fashion and

identify the most relevant unlabeled samples to annotate next. In addition, our consistency loss uses a modified version of the JSD to further improve model performance. By relying on data transformations rather than on external modules or simple heuristics typically used in uncertainty-based strategies, TAAL emerges as a simple, yet powerful task-agnostic semi-supervised active learning approach applicable to the medical domain. Our results on a publicly-available dataset of cardiac images show that TAAL outperforms existing baseline methods in both fully-supervised and semi-supervised settings. Our implementation is publicly available on https://github.com/melinphd/TAAL.

3. Introduction

The performance of deep learning-based models improves as the number of labeled training samples increases. Yet, the burden of annotation limits the amount of data that can be labeled. One solution to that problem is offered by active learning (AL) (Settles, 2009). Based on the hypothesis that all data samples have a different impact on training, active learning aims to find the best set of candidate samples to annotate in order to maximize the performance of the task model. In such context, medical image segmentation emerges as a remarkably relevant task for active learning. Indeed, medical images typically require prior expert knowledge for their analysis and annotation, an expensive and time-consuming task. Initial attempts have explored active learning in medical imaging (Budd *et al.*, 2021), but their methodology either relied on simple uncertainty heuristics (Top *et al.*, 2011; Konyushkova *et al.*, 2019) or required heavy computations during sampling (Sourati *et al.*, 2019; Nath, Yang, Landman, Xu & Roth, 2020) or training (Yang *et al.*, 2017).

Deep active learning Active learning has been extensively explored for the classification (Ash *et al.*, 2020; Beluch *et al.*, 2018; Gal *et al.*, 2017; Sener & Savarese, 2018; Wang *et al.*, 2017; Yoo & Kweon, 2019) or segmentation (Vezhnevets, Buhmann & Ferrari, 2012; Siddiqui, Valentin & Nießner, 2020; Casanova, Pinheiro, Rostamzadeh & Pal, 2019) of natural images. Recent deep active learning approaches based on entropy (Wang *et al.*, 2017) or ensembles (Beluch *et al.*, 2018) adapted traditional uncertainty-based AL strategies to deep learning

models. Similarly, DBAL (Gal *et al.*, 2017) combined measures such as entropy or mutual information with Monte-Carlo dropout to suggest which samples to annotate next. Core-set selection (Sener & Savarese, 2018) aimed to find the best batch sampling strategy for CNNs in classification, but did not scale well to high-dimensional data.

The use of auxiliary modules (Yoo & Kweon, 2019; Sinha *et al.*, 2019; Kim *et al.*, 2021) has been similarly explored to improve AL sampling strategies. The loss prediction module of Yoo & Kweon (2019) measured model uncertainty with intermediate representations. Likewise, a VAE was used in VAAL (Sinha *et al.*, 2019) to learn the latent representation of the unlabeled dataset and distinguish between labeled and unlabeled samples. While these state-of-the-art methods have improved previous approaches, their dependence on auxiliary modules reduces their flexibility and increase the burden of hyperparameter tuning.

Semi-supervised AL Semi-supervised learning (SSL) exploits the representations of unlabeled data to improve the performance of the task model. Since semi-supervised learning and active learning are closely connected, recent works in AL have attempted to combine both domains (Wang *et al.*, 2017; Sinha *et al.*, 2019; Kim *et al.*, 2021; Huang, Wang, Xiong, Huan & Dou, 2021b). For instance, CEAL (Wang *et al.*, 2017) used pseudo-labeling of unlabeled samples to enhance the labeled set during training. VAAL (Sinha *et al.*, 2019) and TA-VAAL (Kim *et al.*, 2021) employed a VAE to learn a latent representation of labeled and unlabeled data. The Mean Teacher framework of Huang *et al.* (2021b) combined a supervised loss on labeled data with an unsupervised loss on unlabeled data based on Temporal Output Discrepancy (TOD), evaluating the distance between the model's output at different gradient steps. The model used a variant of TOD at sampling time to identify the most uncertain samples to annotate. However, these semi-supervised AL methods solely focused on classification tasks or the segmentation of natural images in very large quantities, which is a different context than medical imaging. Another recent work comparable to ours combined AL and SSL via consistency regularization (Gao *et al.*, 2020). The consistency loss adopted during training employed MixMatch (Berthelot *et al.*, 2019) and sample selection measured inconsistency across input perturbations. However, as opposed to our work, Gao *et al.* (2020) kept the consistency loss used during training and the

AL inconsistency metric used for sample selection independent of each other, and the latter was quantified through variance. Furthermore, the method was only validated on classification tasks.

Test-time augmentation Data augmentation is a well-known regularization technique to improve generalization in low-data regimes. These augmentation techniques are particularly essential in medical imaging where datasets tend to be smaller than those of natural images. Yet most recent attempts in active learning do not exploit data augmentation during training (Ash *et al.*, 2020; Nath *et al.*, 2020), or only use random horizontal flipping (Sinha *et al.*, 2019; Kim *et al.*, 2021). Recent learning methods (Ayhan & Berens, 2018; Wang *et al.*, 2019) have also investigated the use of augmentation at test-time in order evaluate prediction uncertainty. Randomly augmented test images yield different model outputs. Combining these outputs can improve the overall predictions as well as generate uncertainty maps for these predictions. Uncertainty estimated through test-time augmentation was shown to be more reliable than model uncertainty measures such as test-time dropout or entropy of the output (Wang *et al.*, 2019).

Motivated by the limitations of current active learning methods for medical image segmentation and the unused potential of active augmentation, this paper proposes a novel semi-supervised active learning strategy called Test-time Augmentation for Active Learning (TAAL).

Our contribution: Our method leverages the uncertainty information provided by data augmentation during both training and test-time sample selection phases. More specifically, TAAL employs a cross-augmentation consistency loss both to train the model in a semi-supervised fashion *as well as* to identify the most uncertain samples to annotate at the next cycle. TAAL comprises three key features:

1. a semi-supervised framework based on cross-augmentation consistency that exploits unlabeled samples during training and sampling;
2. a flexible task-agnostic sample selection strategy based on test-time augmentation;
3. a novel uncertainty measure based on a modified Jensen-Shannon divergence (JSD), which accounts for both cross-augmentation consistency and prediction entropy, and leads to improved performance.

4. Method

Cross-augmentation consistency training We consider a semi-supervised setting where we train a multi-class segmentation model $f_\theta(\cdot)$ parameterized by θ with N labeled samples and M unlabeled samples. We denote the labeled set as $\mathcal{D}_L = \{(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\}_{j=1}^{N}$ and the unlabeled set as $\mathcal{D}_U = \{\mathbf{x}_u^{(j)}\}_{j=1}^{M}$, with data $\mathbf{x}, \mathbf{x}_u \in \mathbb{R}^{H\times W}$ and segmentation mask $\mathbf{y} \in \mathbb{R}^{C\times H\times W}$ (C is the number of classes).

The overall loss that we optimize, $\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_c$, is a combination of a supervised segmentation loss $\mathcal{L}_s$ and an unsupervised consistency loss $\mathcal{L}_c$ weighted by a factor λ. More explicitly, the objective is defined as

$$\mathcal{L} = \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_s\big(f_\theta(\mathbf{x}^{(j)}), \mathbf{y}^{(j)}\big) + \frac{\lambda}{M}\sum_{j=1}^{M}\mathcal{L}_c\big(f_\theta(\mathbf{x}_u^{(j)}), \Gamma\big), \tag{A II-1}$$

where Γ are the transformations applied to $\mathbf{x}_u^{(j)}$. At each iteration, we apply a series of random transformations $\{\Gamma_1, ..., \Gamma_K\}$ to $\mathbf{x}_u$. $\mathcal{L}_c$ measures the variability of segmentation predictions for different augmentations of $\mathbf{x}_u$ measured by a function $\mathcal{D}iv$:

$$\mathcal{L}_c\big(f_\theta(\mathbf{x}_u^{(j)}), \Gamma\big) = \mathcal{D}iv\big\{\Gamma_1^{-1}[f_\theta(\Gamma_1(\mathbf{x}_u^{(j)}))], \ \dots, \Gamma_K^{-1}[f_\theta(\Gamma_K(\mathbf{x}_u^{(j)}))]\big\}. \tag{A II-2}$$

While different measures can be used for $\mathcal{D}iv$ (Camarasa *et al.*, 2020), our consistency loss builds on the Jensen Shannon divergence (JSD),

$$JSD(P_1, ..., P_K) = H\big(\frac{1}{K}\sum_{i=1}^{K}P_i\big) - \frac{1}{K}\sum_{i=1}^{K}H(P_i), \tag{A II-3}$$

where $H(P_i)$ is the Shannon entropy (Shannon, 1948) for the probability distributions P_i. Minimizing the JSD reduces the entropy of the average prediction (making the predictions more similar to each other) while increasing the average of individual prediction entropies (ensuring confident predictions). In AL we typically want to select samples which have a high output

entropy (Wang *et al.*, 2017). Selecting samples with highest JSD would thus have the opposite effect. To avoid this issue, and to control the relative importance of average prediction entropy versus entropy of individual predictions, we propose a weighted version of JSD with parameter α.

$$JSD_\alpha(P_1, ..., P_K) = \alpha H\Big(\frac{1}{K}\sum_{i=i}^{K} P_i\Big) - \frac{(1-\alpha)}{K}\sum_{i=i}^{K} H(P_i). \tag{A II-4}$$

Note that using $\alpha = 0.5$ is equivalent to using the standard JSD.

Test-time augmentation sampling In active learning, the goal is to select the best unlabeled samples to annotate after each training cycle to augment the next labeled training set. Hence, after each cycle, we apply our active learning strategy based on test-time augmentation to select the next samples to annotate.

For each sample $\mathbf{x}_u \in \mathcal{D}_U$, we apply a series of transformations $\{\Gamma'_1, \ldots, \Gamma'_{K_s}\}$, and we compute an uncertainty score $U_{\Gamma'}$ based on the same divergence function as the consistency loss:

$$U_{\Gamma'} = JSD_\alpha\big(\Gamma'^{-1}_1[f_\theta(\Gamma'_1(\mathbf{x}_u))], \ldots, \Gamma'^{-1}_{K_s}[f_\theta(\Gamma'_{K_s}(\mathbf{x}_u))]\big). \tag{A II-5}$$

The samples with highest uncertainty are annotated and added to the labeled training set. After sample selection, the model goes through a new training cycle.

5. Experiments and results

5.1 Implementation details

5.1.1 Dataset

The publicly available ACDC dataset (Bernard *et al.*, 2018) comprises cardiac 3D cine-MRI scans from 100 patients. These are evenly distributed into 5 groups (4 pathological and 1 healthy subjects groups). Segmentation masks identify 4 regions of interest: right-ventricule cavity,

left-ventricule cavity, myocardium and background. For comparative purposes, our experiments focus on the MRI scans at the end of diastole. Preprocessing of the volumes includes resampling to a fixed $1.0\,\text{mm} \times 1.0\,\text{mm}$ resolution in the x- and y-directions as well as a 99^{th} percentile normalization. The 3-dimensional dataset of volumes are converted to a 2-dimensional dataset of images by extracting all the z-axis slices for each volume. Each image is downsampled to 128×128 pixels. Testing is performed on 181 images taken from 20 different patients, ensuring subjects are not split up across training and testing sets. The validation uses 100 randomly selected images. The same validation set is used for all experiments. In total, the available training set, both labeled and unlabeled, thus comprises 660 images.

5.1.2 Implementation and training

We employ a standard 4-layer UNet (Ronneberger *et al.*, 2015) for our backbone segmentation model with dropout ($p = 0.5$), batch normalization and a leaky ReLU activation function. For a fairer comparison in our experiments, we keep the number of training steps fixed during all cycles. We train our models for 75 epochs, each iterating over 250 batches, with $BS = 4$. We use the Adam optimizer (Kingma & Ba, 2015), with $LR = 10^{-6}$ and weight decay $w = 10^{-4}$. To improve convergence, we apply a gradual warmup with a cosine annealing scheduler (Loshchilov & Hutter, 2017; Goyal *et al.*, 2018), increasing the learning rate by a factor 200 during the first 10 epochs. During training, we apply data augmentation, using transformations similar to those utilized for the consistency loss.

In this work, we model the transformations Γ as a combination of f, r and ϵ, where f is the random variable for flipping the image along the horizontal axis, r is the number of 90° rotations in 2D, and ϵ models Gaussian noise. We set $f \sim \mathcal{U}(0, 1)$, $r \sim \mathcal{U}(0, 3)$ and $\epsilon \sim \mathcal{N}(0, 0.01)$, and use $K = 3$ transformations to compute the consistency loss during training.

We use the standard Dice loss as our supervised loss. In the semi-supervised case, following (Cui *et al.*, 2019), we ramp-up the unsupervised component weight using a Gaussian ramp-up

curve such that $\lambda = \exp(-5(1 - t/t_R)^2)$, where t is the current epoch. We use a ramp-up length t_R of 10 epochs, corresponding to the learning rate gradual warmup length.

We repeat each experiment 5 times, each with a different seed determining different initialization of our model weights. For all experiments, the same initial labeled set is used for the first cycle. Experiments were run on NVIDIA PV100 GPU with CUDA 10.2 and Python 3.8.10. We implemented the methods using the PyTorch framework.

5.1.3 Evaluation metrics

To evaluate the performance of the trained models, we employ the standard Dice similarity score, averaged over all non-background channels. We compute both the mean 3D Dice on test volumes and mean 2D Dice on the individual images from these volumes. We give the results as the mean Dice obtained over the repeated experiments.

5.2 Active learning setup

We begin each experiment with 10 labeled samples chosen uniformly at random in the training set and use a sampling budget of 1, meaning that we select one new sample to be labeled after each cycle. Following previous active learning validation settings (Sener & Savarese, 2018), we retrain the model from scratch after each annotation cycle. We use the same types of augmentations during training and sample selection. For test-time augmentation (TTA) sampling, $\{\Gamma'_1, \ldots, \Gamma'_{K_s}\}$ comprises all 8 combinations of flip and rotation augmentations, in order to apply similar transformations to all images, and adopts the same augmentation Gaussian noise parameters as for training. For comparative purposes, with dropout-based sampling, we also run 8 inferences with dropout to obtain different predictions. Both TTA and dropout-based sampling then evaluate uncertainty with $U_{\Gamma'}$ computed on the different generated predictions. We set $\alpha = 0.75$ in TAAL's weighted JSD.

5.3 Comparison of active learning strategies

Our aim is to evaluate the effectiveness of our proposed semi-supervised active learning approach on a medical image segmentation task. In our active learning experiments, we compare TAAL and its unweighted version (with standard JSD) with random sampling, entropy sampling, sampling based on dropout and core-set selection. Entropy-based sampling selects the most uncertain samples based on the entropy of the output probabilities. Dropout-based sampling (Gal *et al.*, 2017) identifies the samples with the highest JSD given multiple inferences with dropout. Finally, core-set selection (Sener & Savarese, 2018) aims to obtain the most diverse labeled set by solving the maximum cover-set problem.

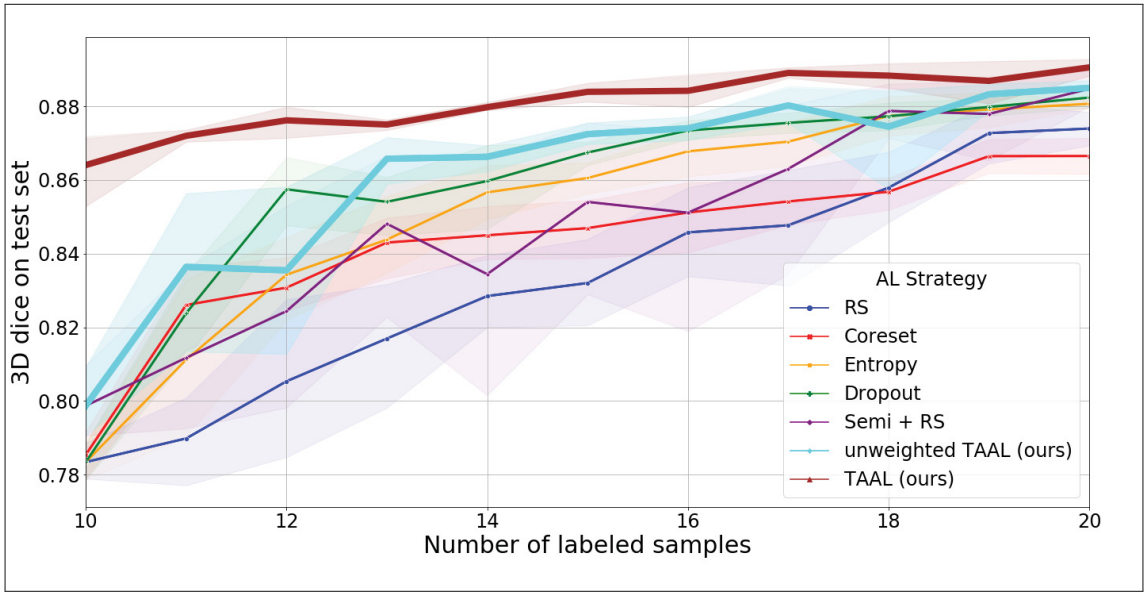


Figure-A II-1 **Active learning results on the ACDC dataset**, given as the mean 3D Dice scores on the test set and corresponding 95% confidence interval. In a fully-supervised setting: random sampling (RS), core-set selection (Coreset), uncertainty-based sampling based on entropy of output probabilities (Entropy), and uncertainty-based sampling based on JSD given multiple inferences with dropout (Dropout). In a semi-supervised setting: random sampling (Semi + RS), TAAL with standard JSD (unweighted TAAL), and TAAL with weighted JSD (TAAL). Our approach TAAL demonstrates significant improvements for low-data regimes in both fully and semi-supervised segmentation

Figure II-1 shows the segmentation performance of our proposed method with its 2 variants along with other existing active learning methods. TAAL consistently outperforms the other baselines by a large margin. We observe that our semi-supervised approach based on cross-augmentation consistency (Semi + RS) noticeably improves the fully-supervised vanilla model (RS). We

notice that our unweighted version of TAAL (with standard JSD, $\alpha = 0.5$) already improves the performance of the semi-supervised model (Semi + RS) by selecting the most uncertain samples based on their cross-augmentation consistency loss. With higher $\alpha = 0.75$, our proposed TAAL with weighted JSD yields the highest performance gain compared to the fully-supervised vanilla model with random sampling (RS).

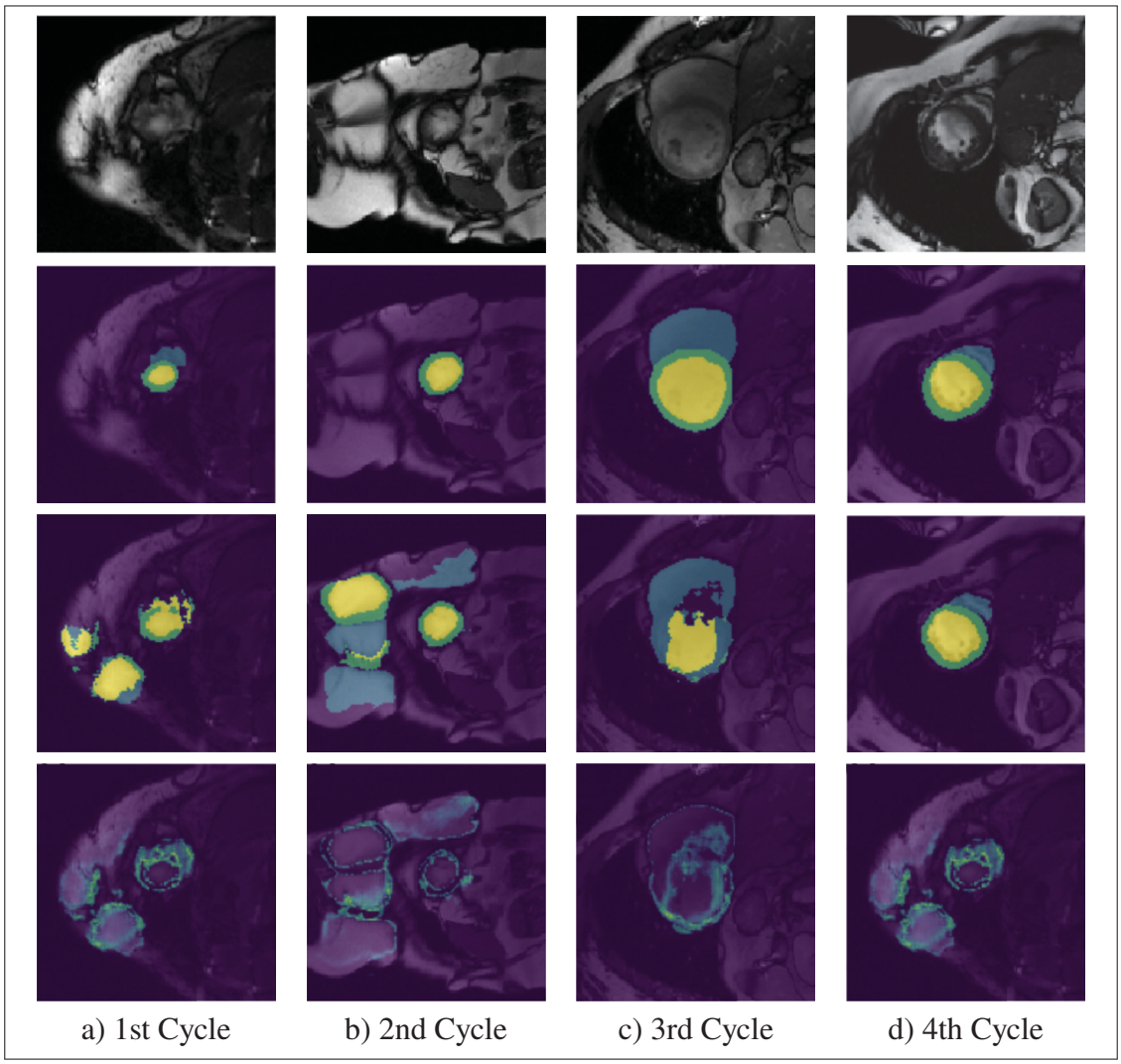


Figure-A II-2 **Examples of images sampled by TAAL at different AL cycles.** Are depicted the image sampled (row 1), the ground-truth segmentation (row 2), the segmentation prediction (row 3), and the JSD map given the different predictions from the augmented image (row 4). We observe that TAAL initially selected images with a large amount of hallucinated inaccurate predictions

Figure II-2 shows examples of images sampled by TAAL during the first 4 annotation cycles. TAAL initially selects image slices which show the apex of the heart. These samples are more difficult to learn in early stages since the areas to segment are much smaller than in the central slices of the heart and the image qualities are typically of lesser quality due to partial volume effects. Thus, we see that the choice of TAAL is first directed at samples yielding highly inaccurate predictions. The previous model has in fact even hallucinated multiple false segmentations for these samples as seen on the third row of subfigures II-2a and II-2b. In the next cycles, TAAL selects more central cardiac slices, which have improved predictions when compared to the ground-truth annotations. Hence, TAAL seems to first focus on correcting inaccurate predictions, before sharpening its predictions on a fine-grained level for slices with more prominent areas to segment.

Table-A II-1 **Active learning performances after doubling the number of initial labeled samples.** We show the mean 2D and mean 3D Dice scores. 'Fully': Fully-supervised vanilla UNet. 'Semi': Proposed semi-supervised training with standard ($\alpha = 0.5$) or weighted ($\alpha = 0.75$) JSD. 'RS': Random sampling. 'TTA': Sampling with Test-time augmentation. 'unweighted TAAL': Our proposed method with standard JSD. 'TAAL': Our proposed method with weighted JSD, which finds the best candidate image to annotate

Metric	Fully					Semi ($\alpha = 0.5$)		Semi ($\alpha = 0.75$)
	RS	Coreset	Entropy	Dropout	TTA	RS	unweighted TAAL	TAAL
2D Dice	80.69	79.95	80.99	81.32	81.67	81.51	81.90	**82.51**
3D Dice	87.40	86.65	88.07	88.24	88.48	88.48	88.50	**89.06**

Table II-1 gathers the model's segmentation performance after 10 cycles in terms of mean 2D Dice and mean 3D Dice scores over whole test volumes. In the fully-supervised setting, test-time augmentation-based sampling (TTA) outperforms random sampling, core-set selection, entropy sampling and sampling based on dropout. Similarly, unweighted TAAL and TAAL outperform random sampling in both semi-supervised and fully-supervised settings. After labeling 10 extra samples, the mean 3D Dice score attains 89.06% with TAAL while only reaching respectively 87.40% and 88.48% with random sampling in fully- and semi-supervised settings. Similar results were observed with 2D Dice on test images.

6. Conclusion

In this paper, we presented a simple, yet effective semi-supervised deep active learning approach for medical image segmentation. Our method, Test-time Augmentation for Active Learning (TAAL), employs a cross-augmentation consistency framework that produces both an improved training due to its unsupervised consistency loss, and a better sampling method through the uncertainty measure it provides. TAAL also uses a modified JSD that significantly improves the model's performance. Our results on the ACDC cardiac segmentation dataset show that, with TAAL, the trained model can reach up to 89.06% 3D Dice with 20 labeled samples when it only reaches 87.40% with random sampling. Because our approach exploits standard augmentation techniques already used in medical image segmentation tasks, TAAL emerges as a simple, yet efficient semi-supervised active learning strategy. While our method highly depends on the presence of disagreeing predictions for augmented inputs to identify the most informative samples, our observed improvements on a cardiac MRI dataset highlight promising avenues for future work, notably the investigation of more complex datasets and types of augmentations.

BIBLIOGRAPHY

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

Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. (2020). Dataset of Breast Ultrasound Images. *Data in Brief*, 28, 104863.

Angelopoulos, A., Bates, S., Fisch, A., Lei, L. & Schuster, T. (2024). Conformal Risk Control. *International Conference on Representation Learning (ICLR)*, pp. 55198–55218.

Angelopoulos, A. N. & Bates, S. (2023). Conformal Prediction: A Gentle Introduction. *Foundations and Trends® in Machine Learning*, 16(4), 494–591.

Angelopoulos, A. N., Barber, R. F. & Bates, S. (2025). Theoretical Foundations of Conformal Prediction. *arXiv preprint arXiv:2411.11824*.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M. J., Heckers, S. H., Huisman, H., Jarnagin, W. R., McHugo, M. K., Napel, S., Pernicka, J. S. G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J. A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A. L., Maier-Hein, L. & Cardoso, M. J. (2022). The Medical Segmentation Decathlon. *Nat Commun*, 13(1), 4128.

Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Van Beeke, E. J. R., Yankelevitz, D., Biancardi, A. M., Bland, P. H., Brown, M. S., Engelmann, R. M., Laderach, G. E., Max, D., Pais, R. C., Qing, D. P. Y., Roberts, R. Y., Smith, A. R., Starkey, A., Batrah, P., Caligiuri, P., Farooqi, A., Gladish, G. W., Jude, C. M., Munden, R. F., Petkovska, I., Quint, L. E., Schwartz, L. H., Sundaram, B., Dodd, L. E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A. V., Gupte, S., Sallamm, M., Heath, M. D., Kuhn, M. H., Dharaiya, E., Burns, R., Fryd, D. S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S. & Croft, B. Y. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical Physics*, 38(2), 915–931.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J. & Agarwal, A. (2020). Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *Eighth International Conference on Learning Representations (ICLR)*.

Ayhan, M. S. & Berens, P. (2018). Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. *Medical Imaging with Deep Learning (MIDL)*.

Ayzenberg, L., Giryes, R. & Greenspan, H. (2025). ProtoSAM for Automated One Shot Medical Image Segmentation Using Foundational Models. *Scientific Reports*, 15(1), 41482.

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

Beluch, W. H., Genewein, T., Nurnberger, A. & Kohler, J. M. (2018). The Power of Ensembles for Active Learning in Image Classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bengar, J. Z., van de Weijer, J., Twardowski, B. & Raducanu, B. (2021). Reducing Label Effort: Self-Supervised meets Active Learning. *International Conference on Computer Vision (ICCV) Workshops*.

Bereska, J. I., Karimi, H. & Samavi, R. (2025). SACP: Spatially-Adaptive Conformal Prediction in Uncertainty Quantification of Medical Image Segmentation. *Medical Imaging with Deep Learning (MIDL)*.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V. A., Krishnamurthi, G., Rohé, M.-M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K. H., Full, P. M., Wolf, I., Engelhardt, S., Baumgartner, C. F., Koch, L. M., Wolterink, J. M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O. & Jodoin, P.-M. (2018). Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*, 37(11), 2514–2525.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. & Raffel, C. A. (2019). MixMatch: A Holistic Approach to Semi-Supervised Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Bougourzi, F. & Hadid, A. (2025). Recent Advances in Medical Imaging Segmentation: A Survey. *arXiv preprint arXiv:2505.09274*.

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

Brunekreef, J., Marcus, E., Sheombarsing, R., Sonke, J.-J. & Teuwen, J. (2024). Kandinsky Conformal Prediction: Efficient Calibration of Image Segmentation Algorithms. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4135–4143.

Budd, S., Robinson, E. C. & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71, 102062.

Burmeister, J.-M., Rosas, M. F., Hagemann, J., Kordt, J., Blum, J., Shabo, S., Bergner, B. & Lippert, C. (2022). Less Is More: A Comparison of Active Learning Strategies for 3D Medical Image Segmentation. *ICML Workshop on Adaptive Experimental Design and Active Learning in the Real World (ICML ReALML)*.

Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., Lugt, A. v. d. & Bruijne, M. d. (2020). Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis (MICCAI-UNSURE)* (pp. 32–41).

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *International Conference on Computer Vision (ICCV)*, pp. 9630–9640.

Casanova, A., Pinheiro, P. O., Rostamzadeh, N. & Pal, C. J. (2019). Reinforced active learning for image segmentation. *International Conference on Learning Representations (ICLR)*.

Chaitanya, K., Erdil, E., Karani, N. & Konukoglu, E. (2020). Contrastive Learning of Global and Local Features for Medical Image Segmentation with Limited Annotations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 12546–12558.

Chao, P., Kao, C.-Y., Ruan, Y., Huang, C.-H. & Lin, Y.-L. (2019). HarDNet: A Low Memory Traffic Network. *International Conference on Computer Vision (ICCV)*, pp. 3551–3560.

Chen, D., Liu, Z., Yang, C., Wang, D., Yan, Y. & Xu, Y. (2025). ConformalSAM: Unlocking the Potential of Foundational Segmentation Models in Semi-Supervised Semantic Segmentation with Conformal Prediction. *International Conference on Computer Vision (ICCV)*.

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A. & Zhou, Y. (2024a). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97, 103280.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z. & Shi, Z. (2024b). RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–17.

Chen, L., Bai, Y., Huang, S., Lu, Y., Wen, B., Yuille, A. L. & Zhou, Z. (2022a). Making Your First Choice: To Address Cold Start Problem in Vision Active Learning. *NeurIPS Workshop on Human in the Loop Learning*.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *International Conference on Machine Learning (ICML)*, pp. 1597–1607.

Chen, Z., Wang, T., Wu, X., Hua, X.-S., Zhang, H. & Sun, Q. (2022b). Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 959–968.

Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M. & Qiao, Y. (2023). SAM-Med2D. *arXiv preprint arXiv:2308.16184*.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M. & Pérez, P. (2019). Addressing Failure Prediction by Learning Model Confidence. *Advances in Neural Information Processing Systems (NeurIPS)*.

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 (IPMI)*, pp. 554–565.

Dai, J., He, K. & Sun, J. (2015). BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. *International Conference on Computer Vision (ICCV)*, pp. 1635–1643.

Danaee, G., Gaillochet, M., Desrosiers, C., Lombaert, H. & Bouix, S. (2026). Exploring Entropy-based Active Learning for Fair Brain Segmentation. *Medical Imaging with Deep Learning (MIDL)*.

DeVries, T. & Taylor, G. W. (2018). Learning Confidence for Out-of-Distribution Detection in Neural Networks. *arXiv preprint arXiv:1802.04865*.

Doersch, C., Gupta, A. & Efros, A. A. (2015). Unsupervised Visual Representation Learning by Context Prediction. *International Conference on Computer Vision (ICCV)*, pp. 1422–1430.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*.

Dumoulin, V. & Visin, F. (2018). A Guide to Convolution Arithmetic for Deep Learning. *arXiv preprint arXiv:1603.07285*.

Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S. A. & Hospedales, T. (2024). Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity. *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, pp. 406–425.

Eber, J., Bockel, S., Antoni, D., Khamphan, C., Noël, G. & Le Fèvre, C. (2025). Delineation of organs at risk in radiotherapy and perspectives. *Cancer/Radiothérapie*, 29.

Fernández-Moreno, M., Lei, B., Holm, E. A., Mesejo, P. & Moreno, R. (2023). Exploring the Trade-off between Performance and Annotation Complexity in Semantic Segmentation. *Engineering Applications of Artificial Intelligence*, 123, 106299.

Gaillochet, M., Tezcan, K. C. & Konukoglu, E. (2020). Joint Reconstruction and Bias Field Correction for Undersampled MR Imaging. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 44–52.

Gaillochet, M., Desrosiers, C. & Lombaert, H. (2023). Active Learning for Medical Image Segmentation with Stochastic Batches. *Medical Image Analysis*, 90, 102958.

Gaillochet, M., Desrosiers, C. & Lombaert, H. (2022). TAAL: Test-time Augmentation for Active Learning in Medical Image Segmentation. *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections (MICCAI-DALI)*.

Gaillochet, M., Desrosiers, C. & Lombaert, H. (2024). Automating MedSAM by Learning Prompts with Weak Few-Shot Supervision. *International Workshop on Foundation Models for General Medical AI (MICCAI-MedAGI)*, pp. 61–70.

Gaillochet, M., Noori, M., Dastani, S., Desrosiers, C. & Lombaert, H. (2025). Prompt learning with bounding box constraints for medical image segmentation. *IEEE Transactions on Biomedical Engineering*, 1–10.

Gal, Y. & Ghahramani, Z. (2016a). Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. *International Conference on Learning Representations (ICLR) workshop*.

Gal, Y. & Ghahramani, Z. (2016b). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *International Conference on International Conference on Machine Learning (ICML)*.

Gal, Y., Islam, R. & Ghahramani, Z. (2017). Deep Bayesian Active Learning with Image Data. *Conference on Machine Learning (ICML)*.

Gao, M., Zhang, Z., Yu, G., Arik, S. O., Davis, L. S. & Pfister, T. (2020). Consistency-Based Semi-supervised Active Learning: Towards Minimizing Labeling Cost. *European Conference on Computer Vision (ECCV)*.

Gidaris, S., Singh, P. & Komodakis, N. (2018). Unsupervised Representation Learning by Predicting Image Rotations. *International Conference on Learning Representations (ICLR)*.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y. & He, K. (2018). Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv preprint arXiv:1706.02677*.

Grady, L. (2006). Random Walks for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11), 1768–1783.

Gu, H., Dong, H., Yang, J. & Mazurowski, M. A. (2025). How to Build the Best Medical Image Segmentation Algorithm Using Foundation Models: A Comprehensive Empirical Study with Segment Anything Model. *Machine Learning for Biomedical Imaging (MELBA)*, 3, 88–120.

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *International Conference on Machine Learning (ICML)*, 70, 1321–1330.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R. & Xu, D. (2022). UNETR: Transformers for 3D Medical Image Segmentation. *Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758.

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988.

Held, K., Kops, E., Krause, B., Wells, W., Kikinis, R. & Muller-Gartner, H.-W. (1997). Markov Random Field Segmentation of Brain MR Images. *IEEE Transactions on Medical Imaging*, 16(6), 878–886.

Hsu, C.-C., Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y. & Chuang, Y.-Y. (2019). Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Hsu, W.-N. & Lin, H.-T. (2015). Active Learning by Learning. *AAAI Conference on Artificial Intelligence*, 29(1).

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E. & Weinberger, K. Q. (2017). Snapshot Ensembles: Train 1, Get M for Free. *International Conference for Learning Representations (ICLR)*.

Huang, L., Ruan, S., Xing, Y. & Feng, M. (2024a). A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97, 103223.

Huang, S., Wang, T., Xiong, H., Huan, J. & Dou, D. (2021a). Semi-Supervised Active Learning with Temporal Output Discrepancy. *International Conference on Computer Vision (ICCV)*.

Huang, S., Wang, T., Xiong, H., Huan, J. & Dou, D. (2021b). Semi-Supervised Active Learning With Temporal Output Discrepancy. *International Conference on Computer Vision (ICCV)*, pp. 3447–3456.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.-P., Dong, F. & Ni, D. (2024b). Segment anything model for medical images? *Medical Image Analysis*, 92, 103061.

Ilse, M., Tomczak, J. & Welling, M. (2018). Attention-Based Deep Multiple Instance Learning. *International Conference on Machine Learning (ICML)*, pp. 2127–2136.

Isensee, F., Jaeger, P. F., Kohl, S. A. 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.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B. & Lim, S.-N. (2022). Visual Prompt Tuning. *European Conference on Computer Vision (ECCV)*, 13693, 709–727.

Jia, Z., Huang, X., Chang, E. I.-C. & Xu, Y. (2017). Constrained Deep Weak Supervision for Histopathology Image Segmentation. *IEEE Transactions on Medical Imaging*, 36(11), 2376–2388.

Jonsson, B. A., Bjornsdottir, G., Thorgeirsson, T. E., Ellingsen, L. M., Walters, G. B., Gudbjartsson, D. F., Stefansson, H., Stefansson, K. & Ulfarsson, M. O. (2019). Brain Age Prediction Using Deep Learning Uncovers Associated Sequence Variants. *Nature Communications*, 10(1), 5409.

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 (MICCAI)*, pp. 492–502.

Jungo, A. & Reyes, M. (2019). Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 48–56.

Jungo, A., Balsiger, F. & Reyes, M. (2020). Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. *Frontiers in Neuroscience*, 14, 282.

Kendall, A. & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

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 (MIDL)*, pp. 365–381.

Kervadec, H., Dolz, J., Yuan, J., Desrosiers, C., Granger, E. & Ayed, I. B. (2022). Constrained deep networks: Lagrangian optimization via log-barrier extensions. *EUSIPCO*, pp. 962–966.

Khoreva, A., Benenson, R., Hosang, J., Hein, M. & Schiele, B. (2017). Simple Does It: Weakly Supervised Instance and Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1665–1674.

Kim, H.-I., Yun, K., Yun, J.-S. & Bae, Y. (2025). Task-Specific Adaptation of Segmentation Foundation Model via Prompt Learning. *European Conference on Computer Vision Workshops (ECCVW)*, 15640, 236–252.

Kim, K., Park, D., Kim, K. I. & Chun, S. Y. (2021). Task-Aware Variational Adversarial Active Learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *International Conference for Learning Representations (ICLR)*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. & Girshick, R. (2023). Segment Anything. *International Conference on Computer Vision (ICCV)*, pp. 3992–4003.

Kirsch, A., van Amersfoort, J. & Gal, Y. (2019). BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Konyushkova, K., Sznitman, R. & Fua, P. (2015). Introducing Geometry in Active Learning for Image Segmentation. *International Conference on Computer Vision (ICCV)*, pp. 2974–2982.

Konyushkova, K., Sznitman, R. & Fua, P. (2019). Geometry in active learning for binary and multi-class image segmentation. *Computer Vision and Image Understanding*, 182, 1–16.

Kulharia, V., Chandra, S., Agrawal, A., Torr, P. & Tyagi, A. (2020). Box2Seg: Attention Weighted Loss and Discriminative Feature Learning for Weakly Supervised Segmentation. *European Conference on Computer Vision (ECCV)*, pp. 290–308.

Kumar, R. R., Shankar, S. V., Jaiswal, R., Ray, M., Budhlakoti, N. & Singh, K. N. (2025). Advances in Deep Learning for Medical Image Analysis: A Comprehensive Investigation. *Journal of Statistical Theory and Practice*, 19(1), 9.

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

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017a). Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017b). Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E. A. R., Jodoin, P.-M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L. & Bernard, O. (2019). Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Transactions on Medical Imaging*, 38(9), 2198–2210.

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

LeCun, Y., Kavukcuoglu, K. & Farabet, C. (2010). Convolutional Networks and Applications in Vision. *International Symposium on Circuits and Systems*, pp. 253–256.

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

Lee, D.-H. (2013). Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *International Conference on Machine Learning (ICML) Workshop*, pp. 6.

Li, C., Sultan, R. I., Khanduri, P., Qiang, Y., Indrin, C. & Zhu, D. (2025). AutoProSAM: Automated Prompting SAM for 3D Multi-Organ Segmentation. *Winter Conference on Applications of Computer Vision (WACV)*, pp. 3570–3580.

Li, H., Wei, D., Cao, S., Ma, K., Wang, L. & Zheng, Y. (2020). Superpixel-Guided Label Softening for Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 227–237.

Li, H. & Yin, Z. (2020). Attention, Suggestion and Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Li, X., Xia, M., Jiao, J., Zhou, S., Chang, C., Wang, Y. & Guo, Y. (2023). HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image segmentation. *Medical Image Analysis*, 88, 102862.

Lin, D., Dai, J., Jia, J., He, K. & Sun, J. (2016). ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3159–3167.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikos, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P. E., Maan, B., van der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H. & Madabhushi, A. (2014). Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2), 359–373.

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

Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., Bercea, C. I., Ouyang, C., Chen, C., Xiong, Z., Wiestler, B., Wachinger, C., Rueckert, D., Bai, W. & Arcucci, R. (2025). Does DINOv3 Set a New Medical Vision Standard? *arXiv preprint arXiv:2509.06467*.

Lloyd, M. C., Monaco, J. P. & Bui, M. M. (2016). Image Analysis in Surgical Pathology. *Surgical Pathology Clinics*, 9(2), 329–337.

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

Ma, J., He, Y., Li, F., Han, L., You, C. & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1), 654.

Mahapatra, D., Bozorgtabar, B., Thiran, J.-P. & Reyes, M. (2018). Efficient Active Learning for Image Classification and Segmentation Using a Sample Selection and Conditional Generative Adversarial Network. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Man, Y., Huang, Y., Feng, J., Li, X. & Wu, F. (2019). Deep Q Learning Driven CT Pancreas Segmentation With Geometry-Aware U-Net. *IEEE Transactions on Medical Imaging*, 38(8), 1971-1980.

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N. & Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89, 102918.

Mittal, S., Tatarchenko, M., Çiçek, O. & Brox, T. (2019). Parting with Illusions about Deep Active Learning. *arXiv preprint arXiv:1912.05361*.

Mossina, L. & Friedrich, C. (2025). Conformal Prediction for Image Segmentation Using Morphological Prediction Sets. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Mossina, L., Dalmau, J. & Andéol, L. (2024). Conformal Semantic Image Segmentation: Post-hoc Quantification of Predictive Uncertainty. *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3574–3584.

Munjal, P., Hayat, N., Hayat, M., Sourati, J. & Khan, S. (2022). Towards Robust and Reproducible Active Learning Using Neural Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nath, V., Yang, D., Landman, B. A., Xu, D. & Roth, H. R. (2020). Diminishing Uncertainty within the Training Pool: Active Learning for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(10), 2534–2547.

Nath, V., Yang, D., Landman, B. A., Xu, D. & Roth, H. R. (2021). Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(10), 2534–2547.

Nath, V., Yang, D., Roth, H. R. & Xu, D. (2022). Warm Start Active Learning with Proxy Labels and Selection via Semi-supervised Fine-Tuning. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B. & Rueckert, D. (2018). Attention U-Net: Learning Where to Look for the Pancreas. *arXiv preprint arXiv:1804.03999*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A. & Bojanowski, P. (2024). DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P. & Goksel, O. (2018). Active Learning for Segmentation by Optimizing Content Information for Maximal Entropy. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.

Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C. & Goksel, O. (2021). Active learning for segmentation based on Bayesian sample queries. *Knowledge-Based Systems*, 214, 106531.

Pathak, D., Krahenbuhl, P. & Darrell, T. (2015). Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. *International Conference on Computer Vision (ICCV)*, pp. 1796–1804.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T. & Efros, A. A. (2016). Context Encoders: Feature Learning by Inpainting. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544.

Petralia, G., Zugni, F., Summers, P. E., Bellomi, M., Alessi, S., Bonello, L., Conte, G., Radice, D., Raimondi, S., Vanzulli, A. & Padhani, A. R. (2021). Whole-body magnetic resonance imaging (WB-MRI) for cancer screening: recommendations for use. *La Radiologia Medica*, 126(11), 1434–1450.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning (ICML)*, pp. 8748–8763.

Rajchl, M., Lee, M. C. H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M. A., Hajnal, J. V., Kainz, B. & Rueckert, D. (2017). DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 36(2), 674–683.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X. & Wang, X. (2021). A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9), 180:1–180:40.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Rother, C., Kolmogorov, V. & Blake, A. (2004). "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3), 309–314.

Sarvamangala, D. R. & Kulkarni, R. V. (2022). Convolutional Neural Networks in Medical Image Understanding: A Survey. *Evolutionary Intelligence*, 15(1), 1–22.

Scatliff, J. H. & Morris, P. J. (2014). From Röntgen to Magnetic Resonance Imaging: The History of Medical Imaging. *North Carolina Medical Journal*, 75(2).

Sener, O. & Savarese, S. (2018). Active Learning for Convolutional Neural Networks: A Core-Set Approach. *International Conference on Learning Representations (ICLR)*.

Settles, B. (2009). *Active Learning Literature Survey*.

Shaharabany, T. (2023). AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder. *British Machine Vision Conference (BMVC)*.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379–423.

Shi, F., Hu, W., Wu, J., Han, M., Wang, J., Zhang, W., Zhou, Q., Zhou, J., Wei, Y., Shao, Y., Chen, Y., Yu, Y., Cao, X., Zhan, Y., Zhou, X. S., Gao, Y. & Shen, D. (2022). Deep Learning Empowered Volume Delineation of Whole-Body Organs-at-Risk for Accelerated Radiotherapy. *Nature Communications*, 13(1), 6566.

Siddiqui, Y., Valentin, J. & Nießner, M. (2020). Viewal: Active learning with viewpoint entropy for semantic segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9433–9443.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P. & Bojanowski, P. (2025). DINOv3. *arXiv preprint arXiv:2508.10104*.

Sinha, S., Ebrahimi, S. & Darrell, T. (2019). Variational Adversarial Active Learning. *International Conference on Computer Vision (ICCV)*.

Sistaninejhad, B., Rasi, H. & Nayeri, P. (2023). A Review Paper about Deep Learning for Medical Image Analysis. *Computational and Mathematical Methods in Medicine*, 2023, 7091301.

Smailagic, A., Costa, P., Young Noh, H., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P. & Campilho, A. (2018). MedAL: Accurate and Robust Deep Active Learning for Medical Image Analysis. *International Conference on Machine Learning and Applications (ICMLA)*, pp. 481–488.

Song, C., Huang, Y., Ouyang, W. & Wang, L. (2019). Box-Driven Class-Wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3131–3140.

Sourati, J., Gholipour, A., Dy, J. G., Kurugol, S. & Warfield, S. K. (2018). Active Deep Learning with Fisher Information for Patch-Wise Semantic Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA)* (vol. 11045, pp. 83–91).

Sourati, J., Gholipour, A., Dy, J. G., Tomas-Fernandez, X., Kurugol, S. & Warfield, S. K. (2019). Intelligent Labeling Based on Fisher Information for Medical Image Segmentation Using Deep Learning. *IEEE Transactions on Medical Imaging*, 38(11), 2642–2653.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958.

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 (NeurIPS)*, 30, 1195–1204.

Teng, J., Wen, C., Zhang, D., Bengio, Y., Gao, Y. & Yuan, Y. (2023). Predictive Inference with Feature Conformal Prediction. *International Conference on Learning Representations (ICLR)*.

Teye, M., Azizpour, H. & Smith, K. (2018). Bayesian Uncertainty Estimation for Batch Normalized Deep Networks. *International Conference on Machine Learning (ICML)*, pp. 4907–4916.

Top, A., Hamarneh, G. & Abugharbieh, R. (2011). Active Learning for Interactive 3D Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

van den Heuvel, T. L. A., de Bruijn, D., de Korte, C. L. & van Ginneken, B. (2018). Automated measurement of fetal head circumference using 2D ultrasound images. *Plos One*, 13(8), e0200412.

van Engelen, J. E. & Hoos, H. H. (2020). A Survey on Semi-Supervised Learning. *Machine Learning*, 109(2), 373–440.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 6000–6010.

Vezhnevets, A., Buhmann, J. M. & Ferrari, V. (2012). Active learning for semantic segmentation with expected change. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3162–3169.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S. & Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338, 34–45.

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., Liu, T. & Zhang, S. (2023). Review of large vision models and visual prompt engineering. *Meta-Radiology*, 1(3), 100047.

Wang, J. & Xia, B. (2021). Bounding Box Tightness Prior for Weakly Supervised Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 526–536.

Wang, K., Zhang, D., Li, Y., Zhang, R. & Lin, L. (2017). Cost-Effective Active Learning for Deep Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27, 2591–2600.

Wang, Z., Wu, Z., Agarwal, D. & Sun, J. (2022). MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3876–3887.

Wenzel, F., Snoek, J., Tran, D. & Jenatton, R. (2020). Hyperparameter Ensembles for Robustness and Uncertainty Quantification. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6514–6527.

Wieslander, H., Harrison, P. J., Skogberg, G., Jackson, S., Fridén, M., Karlsson, J., Spjuth, O. & Wählby, C. (2021). Deep Learning With Conformal Prediction for Hierarchical Analysis of Large-Scale Whole-Slide Tissue Images. *IEEE Journal of Biomedical and Health Informatics*, 25(2), 371–380.

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.

Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M. & Jin, Y. (2025a). Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *Medical Image Analysis*, 102, 103547.

Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M. & Jin, Y. (2025b). Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 102, 103547.

Wu, Q., Zhang, Y. & Elbatel, M. (2023). Self-prompting Large Vision Models for Few-Shot Medical Image Segmentation. *MICCAI Workshop on Domain Adaptation and Representation Transfer (MICCAI-DART)*, pp. 156–167.

Wundram, A. M., Fischer, P., Mühlebach, M., Koch, L. M. & Baumgartner, C. F. (2024). Conformal Performance Range Prediction for Segmentation Output Quality Control. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (MICCAI-UNSURE)*, pp. 81–91.

Xia, Q., Zheng, H., Zou, H., Luo, D., Tang, H., Li, L. & Jiang, B. (2025). A Comprehensive Review of Deep Learning for Medical Image Segmentation. *Neurocomputing*, 613, 128740.

Xie, B., Yuan, L., Li, S., Liu, C. H. & Cheng, X. (2022). Towards Fewer Annotations: Active Learning via Region Impurity and Prediction Uncertainty for Domain Adaptive Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, C., Guo, X., Wang, T., Yang, Y., Ji, N., Li, D., Lv, H. & Ma, T. (2019). Automatic Brain Tumor Segmentation Method Based on Modified Convolutional Neural Network. *IEEE Engineering in Medicine and Biology Conference (EMBC)*, pp. 998–1001.

Yang, L., Zhang, Y., Chen, J., Zhang, S. & Chen, D. Z. (2017). Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 10435.

Yang, Q., Fu, Y., Giganti, F., Ghavami, N., Chen, Q., Noble, J. A., Vercauteren, T., Barratt, D. & Hu, Y. (2020). Longitudinal Image Registration with Temporal-Order and Subject-Specificity Discrimination. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 243–252.

Yoo, D. & Kweon, I. S. (2019). Learning Loss for Active Learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J. & Wang, Z. (2024). SurgicalSAM: Efficient Class Promptable Surgical Instrument Segmentation. *AAAI*, 38, 6890–6898.

Zhang, L., Deng, X. & Lu, Y. (2023). Segment Anything Model (SAM) for Medical Image Segmentation: A Preliminary Review. *International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4187–4194.

Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P. & Li, H. (2024a). Personalize Segment Anything Model with One Shot. *International Conference on Learning Representations (ICLR)*.

Zhang, R., Isola, P. & Efros, A. A. (2016). Colorful Image Colorization. *European Conference on Computer Vision (ECCV)*, pp. 649–666.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M. P., Naumann, T., Wang, S. & Poon, H. (2024b). A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2(1).

Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q. & Ooi, B. C. (2022). BoostMIS: Boosting Medical Image Semi-Supervised Learning With Adaptive Pseudo Labeling and Informative Active Annotation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, S., Song, J. & Ermon, S. (2019). InfoVAE: Balancing Learning and Inference in Variational Autoencoders. *AAAI Conference on Artificial Intelligence*, 33(01), 5885–5892.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J. & Lee, Y. J. (2023). Segment Everything Everywhere All at Once. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 19769–19782.