



University of
Nottingham
UK | CHINA | MALAYSIA

Efficient and Generalised Deep Learning Models for Medical Image Segmentation

Thesis submitted to the University of Nottingham for the degree of
Doctor of Philosophy, January 2026.

Zhendi Gong

20298949

Supervised by

Dr Xin Chen
Prof. Andrew P. French

Signature Zhendi Gong

Date 13 / Jan / 2026

Abstract

Medical image segmentation is a critical task in computer-assisted diagnosis and therapy planning. Accurate and efficient fully automated segmentation methods are essential for accelerating decision-making processes. Compared with natural image segmentation methods, current deep learning-based segmentation models for medical imaging face challenges related to data scarcity, inconsistent annotations, and high computational cost. Consequently, this thesis aims to develop efficient and generalised medical image segmentation models, which are fundamental to the clinical adoption and industrial deployment of this technology in healthcare settings.

This thesis systematically addresses these challenges through a series of progressively advanced deep learning models. The foundational work, CTranS, forms a baseline medical image segmentation model that combines a convolutional neural network (CNN) based encoder with a Transformer-based decoder. This initial contribution leverages the strengths of CNNs for local feature extraction and Transformers for global contextual understanding, achieving state-of-the-art (SOTA) performance.

Subsequently, to enhance the generalisability, MO-CTranS is introduced, which is capable of learning from a mixture of multiple partially labelled datasets. MO-CTranS employs a CNN encoder and Transformer decoder with task-specific tokens to handle label inconsistencies and class imbal-

ances, tackling the challenges of training a single unified model from fragmented and heterogeneously annotated medical data. This innovation moves beyond single-task, fully-annotated model training towards a more scalable and clinically realistic approach.

Despite the superior performance of CTranS and MO-CTranS, the computational efficiency of the Transformer blocks hinders their clinical application. Thus, CRFTrans is proposed to enhance the model efficiency. As a novel recursive Vision Transformer layer designed to replace the commonly used cascaded Transformer layers, this approach represents a fundamental shift from simply stacking layers to a recursive refinement process, demonstrating that comparable performance can be achieved with significantly reduced computational complexity, thus facilitating deployment in resource-constrained clinical environments.

However, these methods still depend on extensive manual annotation, which remains challenging and time-consuming in clinical practice. Thus, to further enhance both efficiency and generality, a novel semi-supervised medical image segmentation framework, referred to as SSL-MedSAM2, has been developed. SSL-MedSAM2 integrates a training-free few-shot segmentation branch for pseudo labelling and a fully supervised learning branch for label refinement. This design allows the model to leverage a small fraction of labelled images together with a large unlabelled set, dramatically decreasing the annotation effort.

Together, these contributions result in deep learning models that are significantly more data-efficient and broadly applicable than conventional approaches. The overall impact of this work is to facilitate the reliable and scalable deployment of segmentation models in real-world clinical settings by bridging the gap among accuracy, efficiency, and generalisability.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr Xin Chen and Prof. Andrew P. French. Their valuable guidance and inspiring encouragement have supported me from various aspects throughout my PhD journey. Xin’s expertise in deep learning and medical imaging provided invaluable direction for my research, while Andy’s insights in computer vision and his careful feedback on my work helped me refine my ideas and strengthen my technical skills. Their patience, constructive criticism, and enthusiasm have continually motivated me to push the boundaries of my own abilities.

I am also grateful to my colleagues and friends in the Computer Vision Lab (CVL) group and the Scientific Imaging Group for Medical Applications (SIGMA) group. Their help during informal discussions and regular lab meetings made the office hours both productive and enjoyable. I especially thank those who offered help and cheered me on during moments of doubt when I encountered difficult problems. The collaborative atmosphere in CVL and SIGMA has not only enriched my research but also developed many valued friendships.

I am eternally grateful to my parents for their unwavering love and encouragement. From the very beginning, they believed in my goals and provided me with the emotional and financial support necessary to pursue this degree. Finally, I want to thank my girlfriend for her patience, understanding, and constant encouragement. Her firm belief in me—especially during challenging phases—has been a source of strength and comfort. Thank you for reminding me to balance work with joy and for celebrating every milestone, no matter how small.

To everyone mentioned here and to those who have silently supported me, this thesis is dedicated to you. Your guidance, friendship, and love have made this achievement possible.

List of publications

Gong, Z., Nicholas, R., Francis, S.T. and Chen, X., 2022, July. Thigh and Calf Muscles Segmentation Using Ensemble of Patch-Based Deep Convolutional Neural Network on Whole-Body Water-Fat MRI. In Annual Conference on Medical Image Understanding and Analysis (pp. 262-270). Cham: Springer International Publishing.

Gong, Z., French, A.P., Qiu, G. and Chen, X., 2024, May. CTranS: A Multi-Resolution Convolution-Transformer Network for Medical Image Segmentation. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE. [**Chapter 3**]

Gong, Z., Francis, S., Cox, E., Sotiropoulos, S.N., Auer, D.P., Qiu, G., French, A.P. and Chen, X., 2025, April. MO-CTranS: A unified multi-organ segmentation model learning from multiple heterogeneously labelled datasets. In 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE. [**Chapter 4**]

Gong, Z. and Chen, X., 2025. SSL-MedSAM2: A Semi-supervised Medical Image Segmentation Framework Powered by Few-shot Learning of SAM2. arXiv preprint arXiv:2512.11548. (Accepted by CARE 2025 MICCAI Challenge, in publication process) [**Chapter 6**]

Gong, Z., Qiu, G., and Chen, X., 2025, November. CRFTrans: A Recur-

sive Vision Transformer Layer Inspired by Conditional Random Field for Medical Image Segmentation. (Submitted to Neural Networks, Elsevier, under review) [**Chapter 5**]

Contents

Abstract	i
Acknowledgements	iii
List of publications	iii
List of Tables	xii
List of Figures	xiii
Chapter 1 Introduction	1
1.1 Background, Challenges, and Contributions	1
1.1.1 Challenge: Difficulty of Segmentation and Limited Dataset	2
1.1.2 Challenge: Heterogeneously Labelled Datasets	4
1.1.3 Challenge: Significant Annotation Efforts	6
1.2 Research Questions and Objectives	8
1.3 Contributions	10
1.4 Thesis Structure	12
Chapter 2 Literature Review	15
2.1 Deep Learning based Medical Image Segmentation Models: from CNNs to Transformers	17
2.1.1 CNN-based Medical Image Segmentation Models . .	17
2.1.2 CNN-Transformer Hybrid Medical Image Segmenta- tion Models	19
2.1.3 Simplifying Transformer Blocks for Clinical Deploy- ment	24

2.2	Multi-Organ Segmentation on Multiple Partially Labelled Datasets	26
2.2.1	Unified Model Design with Task-Specific Modularity	28
2.2.2	Pseudo Mask Generation and Multi-Stage Training .	29
2.3	Large Foundation Models for Medical Image Segmentation .	31
2.3.1	Segment Anything Model and Medical Adaptations .	32
2.3.2	Extending to Videos: Segment Anything Model 2 . .	35
2.4	Learning from Limited Annotations in Medical Image Seg- mentation	38
2.4.1	One-Shot and Few-Shot Learning for Medical Image Segmentation	40
2.4.2	Semi-Supervised Learning for Medical Image Segmen- tation	43
2.5	Conclusions	49
Chapter 3	CTranS: A Multi-resolution Convolution- Transformer Network for Medical Image Seg- mentation	53
3.1	Introduction	53
3.2	Methodology	56
3.2.1	CNN-based Encoder	56
3.2.2	Pyramid Feature Fusion	57
3.2.3	Transformer-based Decoder	59
3.3	Experiments and Results	60
3.3.1	Datasets and Evaluation Metrics	60
3.3.2	Parameter Settings	62
3.3.3	Ablation Study	63
3.3.4	Comparison to SOTA Methods	65
3.4	Discussion and Conclusions	70

Chapter 4	MO-CTranS: A Unified Multi-Organ Segmentation Model Learning from Multiple Heterogeneously Labelled Datasets	73
4.1	Introduction	73
4.2	Methodology	76
4.2.1	Problem Definition and Model Overview	76
4.2.2	Task-Specific Token	76
4.3	Experiments and Results	78
4.3.1	Datasets and Evaluation Metrics	78
4.3.2	Experimental Design and Parameter Settings	79
4.3.3	Comparison to SOTA Methods	81
4.4	Discussion and Conclusions	82
Chapter 5	CRFTrans: A Recursive Vision Transformer Layer Inspired by Conditional Random Field for Medical Image Segmentation	86
5.1	Introduction	86
5.2	Methodology	88
5.2.1	Conditional Random Field	89
5.2.2	Transformer	91
5.2.3	Relationship of CRF and Transformer	93
5.2.4	CRFTrans	97
5.2.5	CRFTrans Plug and Play	99
5.3	Experiments and Results	100
5.3.1	Datasets and Evaluation Metrics	101
5.3.2	Parameter Settings	103
5.3.3	Ablation Study	105
5.3.4	Comparison to SOTA Methods	107
5.4	Discussion and Conclusions	112

Chapter 6	SSL-MedSAM2: A Semi-supervised Medical Image Segmentation Framework Powered by Few-shot Learning of SAM2	115
6.1	Introduction	115
6.2	Methodology	118
6.2.1	Preliminaries	118
6.2.2	Overview of SSL-MedSAM2	119
6.2.3	TFFS-MedSAM2	120
6.2.4	FSL-nnUNet	122
6.3	Experiments and Results	124
6.3.1	Semi-Supervised Learning on the Care-LiSeg Dataset	125
6.3.2	Few-Shot Learning on the Local Dataset	130
6.4	Discussion and Conclusions	135
Chapter 7	Concluding Remarks and Future Work	137
7.1	Contributions	139
7.2	Limitations	144
7.3	Future Work	147
Bibliography		153

List of Tables

3.1	Evaluation of model performance on OASIS-1 dataset ($C_{base} = 32$, $m = 4$). SGM indicates Subcortical-Gray-Matte, WM indicates White-Matter, and CSF indicates cerebrospinal fluid. Mean and Standard deviation of Dice Coefficient (DC) and Average Symmetric Surface Distance (ASSD) are reported for different classes and the Average of all classes	64
3.2	Evaluation of model complexity ($C_{base} = 32$, $m = 4$) on OASIS-1 dataset. The mean value of the number of parameters (NP), memory consumption (GM) and training time per step are reported	64
3.3	Comparison of model performance on five public datasets. Mean and Standard deviation of Dice coefficient (DC) and Average Symmetric Surface Distance (ASSD) are reported. * indicates that the method used 256×256 as the input size, otherwise 224×224 was used. Numbers in bold are the best methods in each input size. † indicates that the result is significantly worse than the best results ($p < 0.01$ measured by WSRT).	66
3.4	Comparison on model complexity on binary segmentation tasks. The mean value of the number of parameters (NP), memory consumption (GM) and training time per step (TS) are reported	68

4.1	Results of method comparison. Bold indicates the best performer, and † indicates statistical significance compared to the bold results ($p < 0.05$ measured by WSRT). The value in brackets for ASSD is the number of empty predictions. . . .	81
5.1	Results of using different iteration times, expressed by $T + 1$. +/- indicates better/worse average performer than $T = 2$ (Bold ones), and † indicates statistical significance ($p < 0.05$ measured by WSRT).	104
5.2	Results of ablation study. +/- indicates better/worse average performer than original setting (Bold ones), and † indicates statistical significance ($p < 0.05$ measured by WSRT). 106	
5.3	Dice of segmentation of different models. ★ indicates no statistical significance compared to CRFTrans ($p > 0.05$ measured by WSRT). +/- indicates better or worse performance for the results that have a statistically significant difference. 109	
5.4	ASSD of segmentation of different models. ★ indicates no statistical significance compared to CRFTrans ($p > 0.05$ measured by WSRT). +/- indicates better or worse performance for the results that have a statistically significant difference.	109
5.5	Results of model complexity comparison. NP stands for number of parameters (in millions), GM stands for GPU memory usage (in GB), and TS stands for training time per step (in seconds).	110
6.1	The data distribution of the CARE-LiSeg dataset. The value in brackets indicates the amount of annotated data in that subset.	126

6.2	Evaluation results of the validation set in terms of GED4 and T1 reported on the leaderboard. Top 5 methods together with TFFS-MedSAM2 are shown (other methods are represented by their team names).	127
6.3	Evaluation results of the test set in terms of GED4 and T1 reported on the leaderboard. “In-Distribution” contains data from vendor A, B1, B2. “Out-Of-Distribution” contains data from vendor C. Top 5 methods are shown (other methods are represented by their team names).	128
6.4	Mean Dice and HD of the few-shot learning methods on the Cardiac dataset. Both 1-shot and 5-shot settings are evaluated. Bold ones have the best mean performance. † indicates statistical significance compared to the bold results (p<0.01 measured by WSRT). TFFS and TFFS* indicate TFFS-MedSAM2 and TFFS-MedSAM2*, respectively.	133
6.5	Average model inference time (MIT, in seconds) and GPU memory usage (GM, in GB) of generating the pseudo mask for one request volume. TFFS and TFFS* indicate TFFS-MedSAM2 and TFFS-MedSAM2*, respectively.	134

List of Figures

- 3.1 Figure A is the overall model structure of CTranS. The example uses an input of a 3-channel colour image and contains 4 levels. Figure B shows the details of the Pyramid Feature Fusion module. The example shows the process in level 0 and level 1 and uses a $256 \times 256 \times 3$ image as input and assumes $C = 64$ and $m = 4$. H and W indicate the height and width of the input image, * indicates value multiplication, and $H \times W$ means the feature dimension of H “by” W 57
- 3.2 The architecture of a Transformer layer. FFN indicates feed-forward networks, Norm indicates layer normalisation, MHA indicates multi-head attention layer, and Concat indicates feature concatenation. Q , K , and V indicates the query, key, and value, and H is the number of heads. 59
- 3.3 The distributions of DC values of different methods on the CVC dataset 69
- 3.4 Some predictions on CVC-ClinicDB that show a large difference between ResUNet and CTranS. The images from left to right are the original data, ground truth, predictions of CTranS, ResUNet, FAT-net, UT-net, SegTran, TransUnet, and SwinUnet. ‘Decoder’, ‘PureC’, ‘Encoder’, ‘Fusion’, ‘Bridge’, ‘PureT’ indicate the category *Transformer in decoder*, *Pure CNN*, *Transformer in encoder*, *Fused CNN and Transformer*, *Transformer as a bridge*, *Pure Transformer* respectively . . . 69

4.1	Overview of MO-CTranS. * indicates value multiplication, and $x \times y$ means the feature dimension of x “by” y	77
4.2	Qualitative examples of different methods. The arrows denote the targets with no training annotations.	83
5.1	The network architecture of a Transformer layer, copied from Section 3.2.3 for convenience.	92
5.2	Key transform steps in mean-field inference. (a) is the 2D feature map view, and (b) is the 1D feature vectors view. Arrows denote ‘contribute to’ In stage 1, W is represented by the pairwise term; In stage 2, W' is represented by the compatibility transform function.	94
5.3	The network architecture of a CRFTrans layer. U is the input unary potential. The recursive process is expressed by H_1 and H_2 , indicating the hidden states in an RNN layer, and G_1 and G_2 , indicating the gating functions.	97
5.4	The network architecture of CTranS. The red blocks are the cascaded Transformer layers that can be replaced directly by CRFTrans.	100
5.5	The network architecture of CRFResidual (a), TransResidual (b), and gRecur-Trans (c).	105
6.1	The workflow of the composition of labelled data and unlabelled data, and the prediction on the unlabelled data from the composition. The inserting location is denoted by the red line in the unlabelled data. MedSAM2 is frozen in the whole process.	120

Chapter 1

Introduction

1.1 Background, Challenges, and Contributions

Medical image segmentation stands as a cornerstone for numerous clinical tasks, providing the quantitative analysis of anatomical structures in imaging modalities like Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasound, which is essential for applications ranging from precise radiation dose calculation in oncology to advanced intra-operative navigation systems during surgery (Yao et al., 2024). Accurate segmentation enables precise measurement of organ volumes, localisation of abnormalities, and improved guidance for diagnosis or treatment planning. However, compared to natural images, medical images mostly have high intra-class variability (due to varying patient anatomy or pathology) and low inter-region contrast (due to the similarity of the intensity between different tissues and organs). Additionally, medical images are typically grayscale or volumetric, making boundaries harder to discern (Gao et al.,

2025). Moreover, acquiring large annotated medical datasets is difficult due to patient privacy constraints and the need for expert manual annotation (Yao et al., 2024). Thus, fully automated segmentation methods that are both accurate and efficient are essential for the advancement of healthcare applications, especially for accelerating decision-making processes.

1.1.1 Challenge: Difficulty of Segmentation and Limited Dataset

One major challenge is how to achieve accurate segmentation with significantly more difficult and less available data than natural images. Conventional automatic segmentation methods like active contours (Kass et al., 1988), atlas-based methods (Gee et al., 1993), and graph cuts (Boykov and Jolly, 2000) require detailed domain knowledge from humans and task-specific designs, which hinders their practical application. Due to the requirement of explicit domain knowledge and task-specific parametrisation on different imaging conditions (e.g., intensity variations, presence of noise, etc.), these methods were outperformed by deep-learning-based methods, particularly convolutional neural networks (CNNs) (Fukushima and Miyake, 1982; Krizhevsky et al., 2012; LeCun et al., 1989; Zhou et al., 2021).

Over the past decade, CNNs have been widely adopted for image segmentation tasks, proving highly effective in various computer vision tasks. For medical image segmentation, the U-Net architecture (Ronneberger et al., 2015), characterised by its symmetric encoder-decoder structure and multi-scale skip connections, has become a dominant force in the domain. Built upon Fully Convolutional Networks (FCNs) (Long et al., 2015), skip connections between the encoder and decoder significantly reduce information

loss during the feature upsampling process. U-Net’s (Ronneberger et al., 2015) ability to learn effectively from relatively smaller datasets further guaranteed its position as a state-of-the-art (SOTA) medical image segmentation method. Subsequent efforts had led to numerous enhancements of the vanilla U-Net. While CNN models leveraged localised convolution kernels to capture fine details, excelling at modelling high-frequency and edge information in medical images, such methods struggled to capture long-range dependencies (Chen et al., 2021b). In segmentation of complex anatomical structures, this limitation can lead to fragmented or spatially inconsistent predictions when global context is needed, hindering recognition of large-scale or context-dependent structures in medical scans.

Transformers have recently emerged as a powerful alternative due to their ability to model long-range relationships through self-attention. Originally developed for Natural Language Processing (NLP) (Vaswani et al., 2017) tasks, Transformer revolutionised sequence modelling and subsequently found application in computer vision. For example, Vision Transformer (Dosovitskiy et al., 2020) inherently captures global context by attending to all parts of the input image by treating the input image as a sequence of image tokens. This property is attractive for medical segmentation, where understanding the global anatomical layout can improve performance, for example, by relating distant regions or ensuring consistent labelling of an organ across slices.

However, despite their advances in capturing global context information, Transformers often lack the inductive bias required for efficient low-level local feature extraction (Gong et al., 2024b). Thus, recognising the complementary strengths and weaknesses of CNNs and Transformers, recent works focus on the combination of CNNs and Transformers. For instance, TransUNet (Chen et al., 2021b) uses Transformers as a bridge between

the CNN encoder and decoder, FAT-Net (Wu et al., 2022) employs a dual CNN-Transformer encoder, and SegTran (Li et al., 2021a) integrates Transformers within its decoder. A key motivation of this thesis is how to combine CNN and Transformer in a more effective manner to fully utilise both advantages, aiming to leverage CNNs for robust local feature extraction while employing Transformers for capturing global context.

While Transformers demonstrate superior ability in modelling long-range dependencies and global context, their clinical deployment faces significant hurdles even with the hybrid architectures. These include quadratic computational complexity, where the $O(n^2)$ complexity (with n being the number of tokens) brings infeasibility for high-resolution medical images, which is a common scenario for MRI and CT data (Liu et al., 2021). Furthermore, prevailing Transformer architectures often rely on deeply cascaded layers, resulting in substantial redundancy and excessive memory footprints, which pose practical challenges given the limited computational resources available in most clinical settings. Thus, another key motivation of this thesis is how to simplify the Transformer used in medical image segmentation to enhance the efficiency without harming the segmentation performance, enabling wider clinical usage in practice.

1.1.2 Challenge: Heterogeneously Labelled Datasets

Another major challenge in medical image segmentation is learning from multiple heterogeneously labelled datasets, which indicates a collection of multiple datasets with disjoint or overlapping label sets. In practice, large fully-annotated datasets with all structures of interest are rare. Instead, as a common scenario, multiple small datasets from different sources are often provided, each annotated for a different subset of organs. For ex-

ample, one abdominal MRI dataset might label the liver and spleen, while another dataset only labels the kidneys. It is more efficient to train a unified model on the mixture of all datasets simultaneously than the intuitive approach to train different models on different datasets separately. However, model generalisability is the key challenge in building a unified model, which needs to be robust to variations in image sources, acquisition views, and heterogeneous labelling practices across different datasets and clinical sites. Especially, the problem of label conflict, where an organ labelled in one dataset might be unannotated (treated as background) in another, may confuse the model. Additionally, the imbalance in dataset sizes or organ prevalence can bias the training and lead to suboptimal results for under-represented classes.

Existing solutions for this heterogeneous labelling problem fall into two categories. First are architectural approaches that condition or partition the model for different tasks. Although these methods (e.g. Med3D’s multi-head architecture (Chen et al., 2019) and conditional nnU-Net variants (Zhang et al., 2021a)) avoid the parameter growth and memory overhead associated with maintaining separate decoders, they still face the challenge of high computational cost and poor scalability as the number of datasets grows. The second category uses label completion via pseudo-labelling: one can train single-organ models on each dataset and then generate pseudo masks for the other organs, finally combining all pseudo-labelled data to train a multi-organ model (Lian et al., 2023; Huang et al., 2020; Liu et al., 2024). While effective, these require training multiple models or iterations, which is computationally expensive. In summary, current solutions either complicate the model (multiple heads or conditional modules) or the training process (multiple stages of pseudo-label generation). Thus, how to overcome these issues to build a simpler generalised segmentation model

based on multiple heterogeneously labelled datasets that works across multiple organs and datasets without incurring huge complexity is another key motivation of this thesis.

1.1.3 Challenge: Significant Annotation Efforts

A third key challenge is the substantial annotation effort required for training medical segmentation deep learning models, further critically affecting the efficiency and generalisability of the segmentation methods. Training deep segmentation networks typically demands a wealth of pixel-level annotations, which in the medical scenario must be provided by skilled experts (radiologists, pathologists) in a time-consuming process. For example, a single abdominal CT scan can contain hundreds of image slices, and manually annotating all organs or lesions of interest is enormously labour-intensive (Yao et al., 2024).

Semi-supervised learning (SSL) approaches tackle this by leveraging unlabelled data alongside a small labelled set for model training. SSL methods have shown great potential to reduce reliance on large annotated datasets (Jiao et al., 2024). They effectively tap into abundant unannotated scans to learn organ boundaries, thus cutting down the number of ground-truth labels needed from experts. In medical segmentation, recent SSL techniques achieve performance close to fully-supervised models that are trained on fully-annotated datasets, while using a fraction of labelled data (Jiao et al., 2024). However, existing widely used SSL methods still face several challenges. For example, the overall performance of pseudo-labelling approaches highly relies on the robustness of the initial model, and inaccuracies in pseudo labels can be quickly amplified during the training process. Nevertheless, the effectiveness of consistency regularisation approaches de-

depends on the strength and type of perturbations; weak ones yield little benefit, while overly strong ones can confuse the model and degrade performance.

Another emerging paradigm to tackle the challenge of annotation costly is the use of foundation models for interactive segmentation, such as the Segment Anything Model (SAM) (Kirillov et al., 2023). SAM is a promptable model trained on over 1 billion masks, designed to generalise to any new image or object with minimal prompting. It can produce accurate masks given simple prompts (points, boxes, or text) and has demonstrated impressive zero-shot segmentation performance even on unseen datasets, significantly reducing the annotation efforts. Based on SAM, SAM2 (Ravi et al., 2024) is proposed for video segmentation. SAM2’s streaming memory mechanism for video segmentation is particularly robust for 3-Dimensional (3D) medical volumes, as the spatial continuity between adjacent slices parallels the temporal coherence of successive video frames. Deploying such foundation models in medical imaging holds immense potential for rapid learning from very limited labelled data due to their reliable predictions from few-shot learning.

In summary, semi-supervised learning and promptable foundation models are two complementary methods to alleviate the large annotation requirements, integrating these approaches into a cohesive solution and ensuring their reliability on medical data are active areas of research. While SAM and SAM2 have been integrated into SAM-based SSL methods to select reliable pseudo-labels (e.g., SAMDSK (Zhang et al., 2023b), CPC-SAM (Miao et al., 2024)) or assist consistency regularisation (e.g., SemiSAM (Zhang et al., 2024b)), a critical limitation is that most existing SAM-based SSL methods still require some form of training or complex prompting to transfer valuable knowledge from labelled to unlabelled data. Thus,

how to effectively fully leverage the few-shot capabilities of SAM or SAM2 for initial pseudo-label generation to further enhance both efficiency and generalisability is the final motivation of this thesis.

These challenges are not isolated, addressing one often influences the others. For instance, highly generalisable models might come with increased computational demands, or efficient models might sacrifice some degree of generalisability. This complex interplay requires a comprehensive model development strategy, where each contribution systematically integrates with and complements previous efforts to achieve a collaborative outcome.

1.2 Research Questions and Objectives

Based on the challenges explained above, the general research question guiding this thesis is: How can deep learning models be designed to simultaneously achieve high efficiency, strong generalisability, and minimal dependence on large annotated datasets for robust medical image segmentation in diverse and evolving clinical environments? Specifically, this thesis aims to address the following research questions (RQs):

RQ1. How can a hybrid CNN-Transformer architecture be designed to fully utilise both advantages for effective medical image segmentation?

RQ2. How can a unified medical image segmentation model be developed to enhance generalisability in the scenario where multiple heterogeneous and heterogeneously labelled datasets are provided?

RQ3. How can the efficiency of the Transformer-based medical image segmentation models be improved in the scenario where the computa-

tional resource is limited in clinical applications?

RQ4. How can semi-supervised learning be effectively integrated with the few-shot capabilities of large foundation models, enabling robust segmentation with minimally labelled data and truly training-free (inference-time adaptation without gradient updates) initial knowledge transfer?

To address these comprehensive questions, the thesis pursues a series of specific research objectives:

- **Objective 1:** To develop a robust baseline medical image segmentation model that effectively combines the strengths of CNNs and Transformers to capture both local and global contextual information, particularly in the context of limited medical datasets. This objective provides the structural basis for future developments.
- **Objective 2:** To improve the generalisability of medical image segmentation models by enabling unified learning from multiple heterogeneously and heterogeneously labelled datasets, effectively addressing challenges such as label conflict and data imbalance. This objective recognises and confronts the issue of diverse and separated data sources in clinical practice.
- **Objective 3:** To significantly enhance the computational efficiency of Transformer-based medical image segmentation models without compromising segmentation accuracy, thereby facilitating deployment in resource-constrained clinical settings. This objective aims to resolve the computational bottleneck in Transformer models.
- **Objective 4:** To further advance the efficiency and generalisability of medical image segmentation frameworks by effectively combining

semi-supervised learning methods and prompt-based large foundation models. This objective aims to minimise the annotation burden and reduce reliance on manual labelling in practical clinical environments.

The progression of these objectives, from establishing a baseline to enhancing generalisability for heterogeneous data, then to improving computational efficiency, and finally to further efficiency and generalisability enhancement through SSL combined with a large foundation model, forms a coherent and logical research structure for the thesis.

1.3 Contributions

This thesis proposes a series of innovative deep-learning-based medical image segmentation methods, each designed to address the specific challenges and solve the RQs stated above. Brief introductions of each method are summarised below:

- *CTranS: A Multi-resolution Convolution-Transformer Network for Medical Image Segmentation* (Baseline): CTranS serves as the foundational baseline model for this thesis. It is designed to address the inherent limitations of pure CNNs and pure Transformers. CTranS directly addresses RQ 1 by optimally combining CNN and Transformer strengths. The key innovation of CTranS lies in its employment of a CNN-based encoder to efficiently extract multi-resolution local image features and a Transformer-based decoder to capture long-range dependencies across these features. A novel Pyramid Feature Fusion module connects the encoder and decoder in a multi-resolution manner, ensuring efficient learning and minimal information loss.

- *MO-CTrans: A Unified Multi-Organ Segmentation Model Learning from Multiple Heterogeneously Labelled Datasets* (Generalisability Enhancement): While CTrans provides a strong baseline for single-task segmentation, real-world clinical data often presents multiple, heterogeneously and heterogeneously labelled datasets. MO-CTrans extends the generalisability of deep learning models to this complex scenario. It directly addresses RQ 2, focusing on how a single model can robustly learn from diverse, heterogeneously labelled datasets. Based on the CTrans architecture, MO-CTrans introduces task-specific tokens within the Transformer-based decoder. These tokens enable the model to effectively differentiate between various organ segmentation tasks and handle label discrepancies when learning from a mixture of datasets. This approach avoids the need for computationally expensive multi-head networks or complex pseudo-mask generation strategies, offering a simpler and lighter solution.
- *CRFTrans: A Recursive Vision Transformer Layer Inspired by Conditional Random Field for Medical Image Segmentation* (Efficiency Enhancement): CRFTrans is proposed to tackle the significant challenge of Transformer efficiency, focusing on significantly improving the computational efficiency of Transformer blocks, directly addressing RQ 3. The key innovation of CRFTrans is its reformulation of the self-attention mechanism as a pairwise potential optimisation process, drawing inspiration from Conditional Random Field theory. It introduces a single, recursive Transformer layer that can effectively replace multiple cascaded layers. This results in models with substantially fewer parameters, reduced memory usage, and faster training, while maintaining or even improving segmentation accuracy.
- *SSL-MedSAM2: A Semi-Supervised Medical Image Segmentation Frame-*

work Powered by Few-Shot Learning of SAM2 (Further Improvement of Efficiency and Generalisability): All the aforementioned methods (CTranS, MO-CTranS, and CRFTrans) are primarily fully-supervised, requiring fully annotated datasets. SSL-MedSAM2 addresses the critical bottleneck of large annotation efforts and further enhances efficiency and generalisability by tackling semi-supervised tasks, directly addressing RQ 4. The key innovation of SSL-MedSAM2 is its novel framework comprising two main branches: a training-free few-shot learning branch based on the pre-trained large foundation model for initial pseudo-label generation, and an iterative fully-supervised learning branch for pseudo-label refinement, which can flexibly incorporate different fully-supervised models like CTranS or MO-CTranS.

In summary, CTranS establishes the foundational hybrid architecture, MO-CTranS enhances model generalisability, CRFTrans improves computational efficiency, and SSL-MedSAM2 further advances both efficiency and generalisability. This structured progression collectively contributes to a comprehensive solution to the overarching challenge of achieving efficient and generalised medical image segmentation.

1.4 Thesis Structure

This thesis consists of 7 chapters, including “Introduction”, “Literature Review”, 4 main technical chapters, and “Concluding Remarks and Future Work”. This chapter provides the background and motivation for the research, outlines the key challenges in medical image segmentation, states the research questions and objectives, briefly introduces the proposed solutions, and details the overall thesis structure.

The remainder of this thesis is structured as follows:

- *Chapter 2:* This chapter provides a comprehensive review of existing research related to deep learning-based medical image segmentation, tracing its evolution from traditional convolutional architectures to large foundation models. The chapter systematically reviews relevant work on the combination of CNNs and Transformers, efficiency optimisation of Transformer architectures, multi-organ segmentation from heterogeneously labelled datasets, and label-efficient paradigms such as few-shot and semi-supervised learning, thereby identifying critical research gaps that motivate the contributions of this thesis.
- *Chapter 3:* This chapter is dedicated to the initial foundational contribution of the thesis. It provides the context for the CTranS model, introduces its core concepts, details the proposed architectural approach, presents and discusses the experimental results, and concludes with a summary of its findings and limitations.
- *Chapter 4:* This chapter addresses the generalisability aspect of the research work. It introduces the context and motivation for MO-CTranS, details its proposed unified approach for learning from diverse and heterogeneously labelled datasets, presents the experimental results and their analysis, and concludes on its effectiveness in enhancing model generalisability.
- *Chapter 5:* This chapter focuses on the efficiency enhancement aspect of the research. It outlines the context and introduction of CRFTrans, describes the novel proposed approach for a recursive Transformer layer based on conditional random fields, presents the experimental results and their discussion, and concludes on its contributions to model efficiency.

- *Chapter 6*: This chapter details the SSL-MedSAM2 framework, explaining its training-free few-shot learning branch leveraging SAM2 for pseudo-label generation and its iterative refinement process, showcasing its efficacy in reducing annotation efforts and improving generalisability in semi-supervised settings.
- *Chapter 7*: This final chapter provides a comprehensive summary of the entire thesis, highlighting the key contributions made by the research. It also discusses the limitations of the presented work and outlines promising directions for future research in the field of efficient and generalised deep learning models for medical image segmentation.

Chapter 2

Literature Review

Medical image segmentation is fundamentally the task of partitioning an image into meaningful regions – typically isolating anatomical structures or pathologies from background tissue (Gao et al., 2025). In a clinical context, the outputs of segmentation algorithms can be used to compute volumetric measurements (e.g. tumour volume), aid in visualisation (highlighting target structures), or serve as inputs to diagnostic models and treatment planning systems. Early segmentation approaches relied on hand-crafted techniques. For example, thresholding to separate foreground/background, edge detection to find boundaries, region growing, morphological operations, and deformable models (Yao et al., 2024; Gao et al., 2025). While these traditional methods were intuitive and sometimes effective, they often struggled with the diversity and complexity of medical images. Hand-crafted algorithms lack adaptability, often requiring extensive tuning for each new task and showing sensitivity to imaging noise or anatomical variations. As a result, the performance and robustness of classical methods were limited, especially when confronted with the full range of variability encountered in clinical practice (Gong et al., 2024b).

To overcome the difficulties in segmenting the complex objects in medical imaging, deep learning has revolutionised medical image segmentation over the past decade, yielding significant performance improvements over prior manual or classical methods (Gao et al., 2025). This chapter reviews the evolution of medical image segmentation approaches with an emphasis on deep learning models and the emerging directions to address current limitations.

This chapter begins by discussing the progression from Convolutional Neural Networks (CNNs) based models (Section 2.1.1) to CNN-Transformer hybrid models (Section 2.1.2), highlighting how each new category of models aimed to overcome shortcomings of the previous. Additionally, this chapter reviews efforts to improve the efficiency of Transformer-based methods in Section 2.1.3, focusing on approaches that simplify or optimise Transformers for faster and lighter segmentation networks. Moreover, Section 2.2 examines techniques to improve generalisability, specifically multi-organ segmentation across multiple partially labelled datasets. Furthermore, Section 2.3 covers the recent trend of Transformer-based large foundation models in vision and their application to medical segmentation, representing a paradigm shift toward general-purpose models. Last, Section 2.4 surveys few-shot learning and semi-supervised learning approaches that address the data scarcity issue by incorporating unlabelled data into model training.

Throughout this literature review, the insights are prioritised from recent survey papers and high-impact studies in order to capture the state-of-the-art (SOTA) methods and key trends. Each section not only recounts major developments but also identifies open research gaps, thereby underlining the need for the solutions proposed in Chapters 3–6.

2.1 Deep Learning based Medical Image Segmentation Models: from CNNs to Transformers

The landscape of medical image segmentation has been profoundly reshaped by the advent of deep learning, moving beyond traditional image processing techniques to sophisticated neural network architectures. This section traces the evolution of deep learning models, from the foundational CNNs to the more recent Transformer-based architectures, highlighting how each paradigm emerged to address the limitations of its predecessors and the persistent challenges in the field.

2.1.1 CNN-based Medical Image Segmentation Models

Convolutional Neural Networks have underpinned or dominated most medical image segmentation methods since the mid-2010s. This section highlights the main developments and trends in CNN-based models, as well as their limitations, which led to the exploration of transformer architectures.

The earliest deep segmentation models were adaptations of CNN classifiers into fully-convolutional networks (FCN) that output segmentation maps. Long et al. (2015) introduced the FCN architecture, which replaces fully-connected layers with convolutional layers to enable dense pixelwise prediction. FCNs demonstrated that CNN features could be repurposed for segmentation by upsampling the feature maps to the original image size. However, the initial FCN had coarse outputs, subsequent refinements like FCN-8s added skip connections to use shallow layer features for sharper

results.

Building on these ideas, U-Net (Ronneberger et al., 2015) provided a more elegant encoder-decoder structure with symmetric skip connections at multiple scales, effectively mitigating information loss during upsampling and enabling precise localisation. This design allowed U-Net to learn effectively from relatively small datasets, quickly making it the SOTA method for many medical image segmentation tasks. U-Net’s effectiveness was evident across modalities; it achieved top results on tasks in microscopy, and was soon successfully applied to segment organs in CT, MRI, and ultrasound scans (Yao et al., 2024). The architecture is flexible and data-efficient, often yielding good performance even with only dozens of training images (augmented heavily), which is a common scenario in medical applications.

The success of U-Net spurred numerous efforts focused on improving its vanilla design. For instance, ResUNet (Zhang et al., 2018) applied residual connections (He et al., 2016) to address vanishing gradients and improve feature propagation. DRU-Net (Jafari et al., 2020) combined principles from ResNet (He et al., 2016) and DenseNet (Huang et al., 2018) for enhanced feature learning. UNet++ (Zhou et al., 2018) introduced nested skip pathways and dense connectivity between decoder layers to better propagate features and ease gradient flow. Attention U-Net (Oktay et al., 2018) added attention gates on skip connections to suppress irrelevant regions and highlight salient features, which was beneficial for segmenting small structures in cluttered images. The V-Net (Milletari et al., 2016) (for volumetric MRI segmentation) extended the U-Net to fully 3D convolutions, capturing 3D context in tasks like prostate or brain segmentation.

A significant advancement in the practical deployment of U-Net-like networks was nnU-Net (Isensee et al., 2021), a self-configuring framework that

automatically adapts preprocessing, network architecture, and training hyperparameters for new datasets. This automation simplified the deployment of CNNs in diverse medical imaging scenarios, achieving SOTA performance across multiple organ segmentation tasks without manual tuning.

CNN models excel at learning local patterns and invariances due to the convolution operation, which is suitable for capturing small-scale textures and edges that define organ or lesion boundaries. As a result, CNN segmenters achieve high precision in delineating structures that have clear local contrast. However, their weakness lies in modelling global relationships, resulting in the loss of long-range connections. Increasing depth or using bigger kernels can partially address this, but with diminishing returns and higher computational costs (Yao et al., 2024). This inability of CNNs to effectively capture global context created a significant research gap, motivating the exploration of architectures capable of modelling long-range dependencies, ultimately leading to the adoption of Transformer-based models, as will be discussed next.

2.1.2 CNN-Transformer Hybrid Medical Image Segmentation Models

The Transformer architecture was first introduced by Vaswani et al. (2017) for machine translation tasks within the domain of natural language processing (NLP). As the core of this architecture, the self-attention mechanism is designed to capture long-range dependencies among input tokens, such as words in a sentence. Self-attention computes a weighted sum of input values, where the weights are determined by the similarity between query and key vectors. In the context of machine translation, these vectors correspond to the embedded representations of the input words.

Unlike convolutional layers, which have a limited receptive field and do not directly relate all positions in the input and output, self-attention establishes global connections between all input positions in a single layer (Vaswani et al., 2017). Furthermore, the Transformer employs a multi-head attention (MHA) mechanism, which projects query, key, and value vectors through multiple learnable linear transformations and processes them in parallel. This allows the model to attend to information from different representation subspaces simultaneously, thereby enhancing its capacity to capture diverse contextual relationships. Due to its ability to selectively focus on relevant parts of the input sequence, the attention mechanism has become a foundational component in a wide range of neural transduction models (Chorowski et al., 2015; Devlin et al., 2018; Wu et al., 2016).

Building upon the Transformer encoder architecture, Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT) for image classification tasks. Unlike earlier approaches that incorporated attention mechanisms alongside image-specific inductive biases, ViT adopted a pure Transformer architecture as originally proposed by Vaswani et al. (2017), without incorporating any domain-specific modifications. In their method, the input image is first divided into equally sized, non-overlapping patches. These 2D patches are then flattened into 1D vectors and linearly projected into a lower-dimensional embedding space to serve as input tokens. Consequently, the query, key, and value vectors in the attention mechanism are derived from these embedded patch representations. Thus, originating in NLP, the Transformer architecture can directly capture long-range dependencies by allowing each pixel (or patch) to attend to others globally. This capability to model global context has sparked substantial interest in applying Transformers to medical image segmentation, where context and spatial relationships are often crucial (Yao et al., 2024).

However, when applied to images, pure Transformers are less inductive-bias-efficient at capturing local features compared to CNNs, resulting in the requirement of extremely large training data to generalise well (Chen et al., 2021b). In medical imaging, datasets are often limited, and retaining precise local information (e.g. lesion boundaries) is crucial. To leverage the strengths of both CNNs and Transformers, researchers began combining them in hybrid models, aiming to integrate CNN’s ability in fine detail localisation with Transformers’ global context understanding (Gong et al., 2024b).

Hybrid models integrate the two architectures in various configurations. Some models apply Transformer layers in the encoder for initial image feature extraction. For instance, Hatamizadeh et al. (2022) introduced UNETR, a Transformer-based framework for 3D medical image segmentation. In this architecture, the input 3D volume is partitioned into patches and fed into a Transformer encoder. The multi-scale representations produced by the Transformer are subsequently refined through a convolutional encoder-decoder structure with skip connections at various feature levels. Wu et al. (2022) proposed FAT-Net, a feature-adaptive Transformer network that employs a dual-encoder design combining both convolutional and Transformer-based modules for feature extraction. A feature-adaptive module is incorporated within the skip connections to enable effective feature fusion, and a memory-efficient decoder is utilised for segmentation prediction. Similarly, Qi et al. (2019) developed X-Net, a hybrid architecture comprising a CNN and a Transformer branch. The CNN branch processes the original input image for reconstruction, while the Transformer branch encodes patch-wise representations and leverages convolutional layers shared with the CNN branch for decoding. Additionally, Shen et al. (2022b) proposed COTR-Net for automated kidney tumour segmentation,

where Transformer layers are integrated after each CNN block within the encoder of a U-Net-like structure.

Different from using Transformer in the encoder to replace or assist the early convolutional blocks, other approaches use Transformer layers as a bridge, which operates on bottleneck features only, to connect a CNN-based encoder and decoder. For instance, TransUNet (Chen et al., 2021b) incorporates an attention module consisting of 12 Transformer layers to capture long-range dependencies within the hidden features extracted by a CNN-based encoder. The decoder employs convolutional upsampling layers, and skip connections are used to directly link the encoder and decoder, facilitating the preservation of spatial information. Expanding on this design, TransAttUNet (Chen et al., 2021a) integrates both Transformer self-attention and global spatial attention modules between the encoder and decoder, along with multi-scale skip connections to enhance the performance of the U-shaped architecture. In a similar vein, Ji et al. (2021) proposed the Multi-Compound Transformer, which introduces two distinct Transformer modules between the encoder and decoder: a self-attention module for capturing multi-scale contextual information and a cross-attention module with a proxy-embedding mechanism to model category dependencies and feature interactions. Zhang et al. (2021c) also employed Transformer blocks as a bridging component between the CNN-based encoder and decoder for cell segmentation. To further improve segmentation accuracy, particularly along object boundaries, they introduced a dual-loss strategy that combines losses on the cell body and edges, promoting edge precision and local consistency. In contrast to these approaches that place Transformer modules between the encoder and decoder, AttUNet (Oktay et al., 2018) utilises attention gates (AGs) directly within the skip connections to selectively filter the encoder features. This mechanism aims to suppress

irrelevant activations and reduce false positives, especially for small object segmentation, by enhancing feature selectivity.

In contrast to the previously discussed approaches, some methods employ Transformer modules within the decoder, following a CNN-based encoder, to capture global contextual information from the encoded feature maps. Li et al. (2021a) introduced SegTran, a medical image segmentation framework that utilises a pre-trained ResNet as the encoder to extract hierarchical features. The decoder incorporates a squeeze-and-expansion Transformer architecture, where the squeeze block is designed to compute attention matrices and the expansion block learns diverse feature representations to enhance segmentation accuracy. Similarly, Li et al. (2021b) proposed a modified U-Net architecture by replacing conventional convolutional up-sampling layers in the decoder with a window attention upsampling module. In this design, different resolution levels within the decoder are connected through a combination of window-based attention and bilinear upsampling, the latter serving as a residual connection to preserve fine-grained spatial details.

Although CNN–Transformer hybrid models have shown promise in medical image segmentation, a critical limitation across many existing approaches lies in their treatment of feature fusion. Most models perform feature interaction either at a single resolution or through relatively shallow connections, where CNN features are injected into a Transformer encoder at one scale, or Transformer outputs are fused back into CNN decoders without considering multi-scale consistency. As a result, these methods may fail to fully exploit the complementary strengths of CNNs and Transformers, leading to suboptimal representation learning and insufficient contextual reasoning. A more effective integration strategy would require multi-scale alignment, enabling richer semantic exchange. By incorporating cross-scale

fusion mechanisms, the hybrid models could achieve more comprehensive feature representations, ultimately improving segmentation accuracy and robustness.

Furthermore, despite the advancements offered by the CNN-Transformer hybrid models, Transformers suffer from high computational complexity caused by the large number of input tokens and the common practice of stacking Transformer layers (e.g., 12 layers in ViT-Base (Dosovitskiy et al., 2020)), hindering practical deployment in resource-constrained clinical environments.

2.1.3 Simplifying Transformer Blocks for Clinical Deployment

Transformer-based segmentation models, especially in 3D or high-resolution imaging, face efficiency challenges. Representative examples include UNETR (Hatamizadeh et al., 2022), which uses a Transformer encoder over 3D patches with U-shaped decoding, and nnFormer (Zhou et al., 2023), which interleaves convolution and self-attention while introducing volume-based attention variants to improve 3D context modelling. However, the vanilla self-attention mechanism has quadratic complexity in the number of image tokens, leading to extreme computational and memory costs when modelling full-resolution medical images (Xie et al., 2021).

Recent studies focused on the efficiency improvement of the Transformer-based models for clinical deployment. For example, the Swin Transformer (Liu et al., 2021) uses fixed-size windowed attention, by computing self-attention only within small windows and then shifting window positions, it achieves linear scaling and still allows cross-window interaction. Incorpor-

rating Swin attention into segmentation networks (as in Swin-Unet (Cao et al., 2021)) greatly reduces computational costs while maintaining accuracy, enabling processing of large images with fine detail. Similarly, CoTr (Xie et al., 2021) introduced a deformable Transformer module (DeTrans) to bridge a CNN and a Transformer for 3D segmentation. Instead of attending to all patch embeddings, DeTrans selects a small set of key sampling points for each query and focuses attention there. Additionally, efficient self-attention mechanisms, as seen in Efficient Transformer (Xu et al., 2021b), incorporated optimised multi-head self-attention and novel MLP-based heads to lower computational complexity without sacrificing accuracy.

Beyond modifying the attention mechanism itself, some methods focus on decoder-centric simplification, such as Segmenter (Strudel et al., 2021), which applies a point-wise linear decoder to patch embeddings, thereby leveraging global context without incurring excessive computational cost in the final segmentation head.

While reducing spatial complexity, these methods often overlook depth-wise over-parameterisation, retaining a long stack of Transformer layers. This means that even if individual attention computations are made more efficient, the sheer number of stacked Transformer layers still contributes to high parameter counts and extensive computational graphs, limiting the overall efficiency gains. Moreover, such spatially localised simplifications tend to sacrifice global anatomical consistency for efficiency. In contrast, several studies have explored recurrence in depth by reusing the same Transformer block multiple times to enhance efficiency. The Universal Transformer (UT) (Dehghani et al., 2018) was the first to explicitly reformulate the Transformer as a recurrent sequence model along the depth dimension. Instead of stacking multiple independent layers, UT repeatedly

applies a shared block comprising self-attention and feed-forward sublayers to token representations, enabling each position to evolve iteratively across recurrent steps.

In computer vision, the Sliced Recursive Transformer (SReT) (Shen et al., 2022a) introduces parameter sharing across layers to improve parameter efficiency while maintaining accuracy. Similarly, Recurrent Vision Transformers (RViTs) (Gehrig and Scaramuzza, 2023) adapt the idea of recurrence for visual reasoning and event-stream processing, replacing static feed-forward passes with iterative refinement over the same image or sequence features. In medical imaging, recent works have incorporated iterative Transformer refinements to capture contextual dependencies more effectively. For instance, DSIT-UNet (Al Hasan et al., 2025) integrates iterative Transformer modules within a U-Net encoder–decoder framework for brain-tumour segmentation, demonstrating the effectiveness of recursive attention mechanisms in enhancing medical image analysis.

2.2 Multi-Organ Segmentation on Multiple Partially Labelled Datasets

The fragmented and inconsistently labelled nature of real-world clinical datasets, which directly impacts model generalisability, is one of the main challenges in medical image segmentation. In clinical practice, obtaining a single, large-scale medical dataset where all organs of interest are consistently and fully annotated is exceptionally difficult due to expertise and time constraints. Instead, researchers and clinicians often have access to multiple smaller datasets from different sources, which are “heterogeneously labelled” (i.e., different organs are labelled in different datasets)

and “partially labelled” (i.e., not all organs of interest are annotated in every image). For example, one dataset might contain annotations only for the liver, while another, from a different hospital, might only have annotations for the spleen, even if both are abdominal CT scans.

Training individual models for each of these datasets is inefficient and computationally costly, as it requires maintaining and deploying multiple models. Furthermore, this approach fails to fully utilise common information shared across datasets from the same imaging modality, leading to redundant learning and suboptimal performance. A key challenge in training a single, unified model on such mixed datasets is “label conflict,” where an unlabelled organ in one dataset might be considered background in another, confusing the model during the learning process. For instance, if a model is trained on a dataset where the kidney is always background, but then encounters a dataset where the kidney is a foreground object, it can struggle to adapt. Additionally, data imbalance across different organ labels (e.g., many liver annotations but few kidney annotations) can cause performance inconsistency among organs, leading to models that perform well on common organs but poorly on rarer ones.

Existing approaches to enhance generalisability in this multi-organ, multi-dataset learning problem generally fall into two main categories, which are unified model design with task-specific modularity, and pseudo-labelling with multi-stage training.

2.2.1 Unified Model Design with Task-Specific Modularity

The first branch of approaches modifies the network architecture to handle multiple datasets/tasks within one model. A simple example is multi-head networks, where a shared encoder learns common features and multiple decoder heads output segmentations for different label sets. Med3D (Chen et al., 2019) adopted a main component sharing model weights across different datasets, while individual task-specific layers or “heads” are separately trained for each distinct segmentation task. While this allows for a single backbone, such methods often suffer from high computational costs and poor scalability when more datasets or organ types are included, as each new task requires additional dedicated layers and increased memory footprint.

Another line of work uses conditional networks, which introduce task-specific conditions or parameters inside the network to inform it which organ to segment. For instance, Dmitriev and Kaufman (2019) used conditions from label classes integrated with intermediate activation signals within convolutional layers to facilitate specific organ prediction. Similarly, conditional nn-Unet (Zhang et al., 2021a) leveraged conditional information of organ classes in each resolution level of the decoder for specific organ segmentation. These methods aim to make the network “aware” of the target organ, guiding its feature learning. These conditioned models avoid having completely separate heads for each task, but still require careful design to handle conflicting labels.

Building on the concept of conditioned networks, an influential work addressing partial labels is DoDNet by Zhang et al. (2021b). DoDNet (Zhang et al., 2021b) introduced a dynamic convolutional head. Instead of mul-

multiple static decoders, it has a single segmentation head whose weights are dynamically generated based on a task embedding (which organ to segment). In practice, a small controller network takes the task code and image features to output the appropriate convolution kernels on-the-fly. This allows one network to flexibly perform different segmentations without switching heads, greatly improving efficiency. DoDNet (Zhang et al., 2021b) showed strong results and even demonstrated that a model pre-trained on multiple partially labelled sets could be fine-tuned to a fully labelled multi-organ dataset to achieve state-of-the-art performance. Extensions of this idea include TransDoDNet (Xie et al., 2023), which swapped the convolution-based kernel generator with a Transformer-based generator to better capture inter-organ relationships. Overall, unified model designs focus on architectural means to learn from all data simultaneously, using shared representations and minor per-task adaptations to resolve label conflicts.

2.2.2 Pseudo Mask Generation and Multi-Stage Training

The second category of methods handles label disparities by generating pseudo labels for missing annotations, effectively filling in the gaps so that a standard multi-class training is possible. A representative approach is to first train individual single-organ models on each dataset and then use their predictions as pseudo-labels on the other datasets. Lian et al. (2023) adopted the strategy by training a model for each organ with its own dataset, applied it to other datasets to get pseudo masks where that organ was not labelled, and then combined all data with real and pseudo labels to train a multi-organ model. Huang et al. (2020) similarly trained separate

networks and then employed a co-training scheme, where two models were trained on the union of datasets using both ground truth and each other’s outputs as pseudo-annotations, with weight averaging to stabilise learning.

These approaches manage to assemble a “fully” labelled training set (in terms of all organs), but at the cost of training many networks and trusting the pseudo-label quality. They also tend to be computationally heavy, since multiple rounds of training are involved. A more recent advance in this direction is COSST (Comprehensive Supervision and Self-Training) by Liu et al. (2024). COSST (Liu et al., 2024) is a two-stage framework that first trains an initial unified model on the available ground-truth labels (treating unlabelled organs as background for that step), and then iteratively improves the model by self-training with pseudo labels. In the self-training phase, COSST (Liu et al., 2024) generates predictions for unlabelled organs and adds those as training targets, but with a twist: it evaluates the reliability of each pseudo-label (e.g. by detecting outliers in the feature space) and discards the most unreliable predictions in each iteration. By progressively including only high-confidence pseudo labels, COSST (Liu et al., 2024) avoids the trap of reinforcing errors and achieves significantly better multi-organ accuracy than earlier methods.

Overall, the existing works introduced in the above sections have substantially improved the ability to learn a generalisable multi-organ segmentation model from fragmented annotations. In practice, a unified model can indeed be trained on several partially labelled datasets by either network architectural design or iterative pseudo-label generating. This unlocks the synergistic value of diverse datasets, which effectively increases the training data without the cost of full annotations for every image. However, they frequently introduce significant computational overhead, require training multiple models, or involve complex multi-stage processes. This creates a

critical gap in developing a simple, lightweight, and unified model that can effectively differentiate tasks, handle label conflicts, and leverage shared information across diverse datasets without sacrificing efficiency or scalability. Such models should inherently handle label conflicts and data imbalance without requiring task-specific layers, separate decoder heads, or computationally expensive multi-stage pseudo-mask generation, thereby improving generalisability with less overhead.

2.3 Large Foundation Models for Medical Image Segmentation

The field of computer vision has recently witnessed the emergence of large foundation models, trained on unprecedented scales of data, demonstrating remarkable generalisation capabilities. Among these methods, Segment Anything Model (SAM) (Kirillov et al., 2023) is the most widely applied foundation model for image segmentation due to its superior zero-shot and few-shot learning capacity. This represents a significant paradigm shift, moving towards models that can perform a wide range of tasks with minimal task-specific fine-tuning. These models, often pre-trained on vast and diverse datasets, learn a rich, generalised representation of visual information that can then be adapted to various downstream tasks (Zhang et al., 2024a). This paradigm is very attractive for medical image segmentation, where traditionally a new model must be trained for each organ, modality, or disease. A general segmentation model that can adapt to any medical image with minimal user input could greatly expedite clinical workflows. Researchers are thus exploring how these large vision foundation models can be applied or adapted to segment medical images (Sengupta et al.,

2025).

The promise of a general, prompt-driven segmentation tool, which could delineate any anatomical structure or pathology with a simple click or box, has spurred many efforts to harness SAM-style models for medical use. However, applying the foundation models pretrained on natural images to the medical domain is not straightforward. Medical images (e.g. radiology scans, microscopy, ultrasound) differ significantly from natural images in terms of contrast, texture, and object appearance (Zhang et al., 2024a). The features learned from natural images may not directly transfer to medical structures (e.g. a CT scan of soft tissue). Early studies found that SAM (Kirillov et al., 2023) often struggles with medical images, necessitating domain-specific adaptation (Sengupta et al., 2025).

2.3.1 Segment Anything Model and Medical Adaptations

SAM is a promptable image segmentation model proposed by Kirillov et al. (2023), trained on an unprecedented scale of 11 million images and 1.1 billion segmentation masks, encompassing billions of 2D natural image segmentation tasks. SAM employs a transformer-based architecture with three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder is a ViT-based feature extractor that splits the image into patches and produces high-level feature embeddings. The prompt encoder accepts various user inputs, like points indicating foreground/background, bounding boxes, or even free-form masks, and embeds these prompts into the same feature space. This flexible prompt interface allows users to specify the target object in different ways (e.g. clicking a point inside a tumour, or drawing a box around an organ).

Finally, the mask decoder combines the image features with the prompt information to predict a segmentation mask. Notably, the mask decoder uses a two-way transformer design to refine the segmentation by iteratively exchanging information between image and prompt embeddings. Additionally, it includes an IoU prediction head that estimates how good the mask is, which can help in selecting the best mask when multiple masks are predicted. This architecture enables SAM to produce accurate object masks in a single forward pass given any prompt, making it a general-purpose segmentation tool.

SAM (Kirillov et al., 2023) is particularly suitable for medical image segmentation due to its strong generalisation capabilities and promptable interface (Zhang et al., 2023a). The ability to perform zero-shot or few-shot segmentation means it can potentially delineate anatomical structures or lesions with very few examples or minimal user interaction, significantly reducing the annotation burden that plagues medical imaging (Zhang et al., 2023a). This interactive and adaptable nature makes it a powerful tool for accelerating medical image analysis workflows and potentially democratising access to advanced segmentation capabilities (Sengupta et al., 2025).

However, a significant challenge in directly applying SAM (Kirillov et al., 2023), which is primarily trained on natural images, to medical imaging is the inherent domain gap between natural and medical images. Medical images often exhibit distinct visual characteristics, including different textures, intensity distributions, and anatomical structures, compared to the diverse but typically non-medical imagery seen during SAM’s pre-training (Sengupta et al., 2025). This domain shift makes direct application difficult and often leads to suboptimal performance. To bridge this domain gap and adapt SAM for high-performance medical image segmentation, numerous efforts have focused on specialised strategies.

For example, MedSAM (Ma et al., 2024) and SAM-Med (Ye et al., 2023) fine-tuned the original SAM on extensive medical segmentation tasks. This process involves continuing SAM’s (Kirillov et al., 2023) training on large datasets of medical images with corresponding annotations, allowing the model to specialise in medical image characteristics, learn domain-specific features, and significantly improve accuracy on medical datasets. These fine-tuned versions aim to retain SAM’s (Kirillov et al., 2023) powerful generalisation abilities while adapting to the unique visual properties and anatomical contexts of medical data. Other studies kept SAM’s (Kirillov et al., 2023) weights mostly frozen but added adapter modules or modified the prompt mechanisms. For example, Wu et al. (2025a) proposed a Medical SAM Adapter that attaches small learned layers to SAM, successfully improving organ and tumour segmentation without full re-training.

Since SAM (Kirillov et al., 2023) is originally designed for 2D image segmentation, researchers have developed methods to extend its capabilities to 3D segmentation, which is crucial for accommodating the volumetric nature of many medical images (e.g., CT, MRI scans). For example, Sammed3d (Wang et al., 2024) used a purely 3D network—comprising a 3D image encoder, 3D prompt encoder, and 3D mask decoder—that preserves spatial coherence throughout the volume, and trained on a massive multi-source 3D dataset. Ma-sam (Chen et al., 2024a) embedded lightweight 3D adapter layers into the transformer blocks of SAM’s image encoder, enabling it to extract volumetric or temporal context while keeping most of SAM’s pre-trained weights frozen. Additionally, Gong et al. (2024a) developed a 3D SAM-Adapter to holistically adapt SAM from 2D to 3D segmentation, bridging the gap between slice-wise predictions.

In summary, SAM (Kirillov et al., 2023) provides a strong foundation and a new way of thinking about medical image segmentation (as a promptable,

interactive task), but it typically needs adaptation or careful use to handle the uniqueness of medical imaging. Researchers continue to refine SAM-based approaches to address these limitations.

2.3.2 Extending to Videos: Segment Anything Model

2

Building on the success of SAM (Kirillov et al., 2023), the Segment Anything Model 2 (SAM2) (Ravi et al., 2024) further extended these capabilities, particularly for video segmentation. SAM2’s architecture modifies SAM’s core components, including an Image Encoder for embeddings, a Prompt Encoder for user prompts, and a Mask Decoder that produces masklets (spatio-temporal masks) by utilising features across multiple frames. The technical innovation of SAM2 lies in its lightweight memory bank, which stores feature embeddings from previous frames in a streaming manner, typically organised as a first-in-first-out (FIFO) queue to maintain a bounded and efficient memory footprint. This memory bank enables efficient retrieval and fusion of historical embeddings with current-frame features, allowing the model to reference past context and object states during segmentation. This streaming memory allows SAM2 to retain object identity and propagate segmentations automatically as an object moves or deforms across a sequence, ensuring temporal coherence and consistency in its predictions. This is a significant advancement over SAM, which primarily operates on individual 2D images without explicit temporal or sequential awareness. The introduction of the memory mechanism also enhances robustness. SAM2 has been shown to handle challenges like object occlusion or motion between frames by leveraging the memory of prior frames to disambiguate the current segmentation (Zhang and Shen, 2024).

It also offers faster inference on images, due to the efficient reuse of temporal information, making it a strong general-purpose segmentation model for both images and videos (Zhang and Shen, 2024).

This streaming memory mechanism makes SAM2 particularly robust and beneficial for adaptation to 3D medical volumes. The inherent spatial continuity between adjacent slices in a 3D medical image (e.g., a CT or MRI scan) closely parallels the temporal coherence of successive video frames (Shen et al., 2024). By treating a 3D volume data as a video sequence, where each slice is a “frame,” SAM2 can naturally propagate segmentations across slices, reducing the need for per-slice prompts and potentially producing high-quality, consistent masks for entire volumes from minimal input (e.g., a few prompts on a single slice or sparse prompts across a few slices). This inherent suitability for volumetric data, combined with its impressive promptable interface and few-shot capabilities, makes SAM2 a powerful tool for medical image segmentation, promising to significantly reduce the manual effort typically required for comprehensive 3D annotation, especially for large datasets and complex anatomical structures (Shen et al., 2024).

For example, Shen et al. (2024) applied SAM2 in a zero-shot interactive setting for 3D medical image segmentation, where a user provides prompts on a few slices and SAM2 propagates the segmentation through the volume. They reported that SAM2 could indeed leverage slice-to-slice continuity to reduce the annotation burden in segmenting complex structures, though its accuracy still fell short of fully supervised 3D models in their experiments. To explore the effect of different usage of SAM2, Dong et al. (2024) conducted a comprehensive evaluation of SAM2 on a collection of 18 medical imaging datasets spanning various modalities. Their findings highlighted that SAM2’s performance in the medical domain is highly variable and sen-

sitive to the details of the usage pipeline. Factors such as the order in which slices are segmented, the direction of propagation, and how intermediate predictions are chosen or corrected can significantly affect the result.

Similar to SAM, SAM2 also faces the challenge of the “domain gap” when applied to medical images, as its primary training is on natural image and video data, and without adaptation it may mis-segment or require extensive prompting in the medical setting. To bridge this gap and leverage SAM2’s capabilities in the medical domain, fine-tuning on medical datasets has been a key strategy. The major development is the introduction of MedSAM2 (Ma et al., 2025). Ma et al. (2025) fine-tuned SAM2 on an extensive medical dataset comprising hundreds of thousands of 3D images and a large number of medical video frames. This fine-tuning, effectively training a foundation model for 3D and video medical segmentation, produced a model that substantially outperformed previous approaches across a wide range of organs, lesion types, and imaging modalities. MedSAM2 was shown to significantly narrow the gap between foundation models and specialist models. It achieved near state-of-the-art accuracy on various benchmarks while maintaining the generality of a promptable model. MedSAM2 enables more efficient annotation of scans, the authors reported that using MedSAM2 in a human-in-the-loop annotation pipeline reduced the required manual effort by a large margin compared to traditional slice-by-slice annotation.

In summary, foundation models like SAM and SAM2 have introduced a new, generalist approach to medical image segmentation, showing remarkable ability to adapt across tasks. Through architectural innovations (prompt mechanisms, memory modules) and large-scale training, they achieve a level of generalisation that was previously unattainable with single-purpose models. Adapted variants such as MedSAM (Ma et al., 2024) and MedSAM2 (Ma et al., 2025) demonstrate that with the right

fine-tuning, these models can also attain high accuracy tailored to medical needs, while easing the burden of manual annotation.

2.4 Learning from Limited Annotations in Medical Image Segmentation

Over the past decade, the development of deep learning-based medical image segmentation has progressed through several key phases. Section 2.1 outlined the evolution from CNNs, which excel in local feature extraction but struggle with capturing long-range dependencies, to vision transformers and hybrid CNN-Transformer models that better integrate global context with local detail. Further refinements have targeted architectural efficiency (Section 2.1.3), reducing the computational and memory footprint of transformer-based frameworks without sacrificing segmentation accuracy. Beyond architectural improvements, Section 2.2 addressed the critical need for generalisability in real-world settings, where models must operate across heterogeneous imaging modalities and partially labelled multi-organ datasets. Strategies such as cross-dataset training and domain adaptation have been explored to enhance robustness under these conditions. Section 2.3 then examined the paradigm shift brought by large foundation models such as SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2024), which demonstrate unprecedented generalisation and promptable segmentation capabilities. Their introduction marks a transition from task-specific, fully supervised models to more flexible, pre-trained architectures that can be adapted to a range of segmentation tasks with reduced fine-tuning.

Despite these advances, a common assumption underlying the majority of methods in Sections 2.1–2.3 is the availability of large, high-quality, fully

annotated datasets to train the models. In practice, this assumption rarely holds in clinical environments. The process of generating pixel-level medical image annotations is resource-intensive, requiring specialised expertise and substantial time investment. Even with the emergence of foundation models, domain adaptation to medical imagery often still requires curated labelled data to achieve clinically acceptable performance. Consequently, a major bottleneck in translating segmentation models from research to deployment is the high cost and effort of manual annotation, which directly limits the scalability and applicability of these models.

To address this bottleneck, research has increasingly shifted toward label-efficient learning paradigms. Among these, one-/few-shot learning aims to enable a model to generalise from one or a small number of annotated examples, offering a direct solution to extreme annotation scarcity. Such methods can be used in isolation or as part of larger adaptation pipelines, including foundation models. Complementary to few-shot approaches, semi-supervised learning (SSL) leverages the abundant unlabelled medical images available in clinical archives, pairing them with a limited labelled set to guide the model toward improved performance without proportional increases in annotation cost. These paradigms are not mutually exclusive; in fact, recent trends increasingly combine their strengths, for example, by integrating promptable foundation models into semi-supervised frameworks.

This section examines these two approaches in detail. Section 2.4.1 surveys the development of one-shot and few-shot segmentation methods, beginning with early metric-learning and meta-learning frameworks and progressing to recent techniques that incorporate large foundation models. Section 2.4.2 reviews semi-supervised segmentation methods, with a focus on two branches of methods: pseudo-labelling and consistency regularisation. Thus, this section completes the broader literature review by transitioning

from the fully supervised methods to the label-efficient strategies.

2.4.1 One-Shot and Few-Shot Learning for Medical Image Segmentation

One-shot learning and few-shot learning represent powerful paradigms aimed at enabling models to learn new concepts or tasks from an extremely limited number of labelled examples, typically one (one-shot) or a handful (few-shot). The motivation for few/one-shot learning in medical imaging is compelling, directly addressing the challenges of data scarcity and annotation burden. This capability is vital for accelerating medical research and clinical deployment in data-limited environments.

One early approach to one-shot medical segmentation was to synthesise training data from a single example. For instance, Zhao et al. (2019) proposed an automated data augmentation method that learns realistic transformations from unlabelled images and applies them to a single labelled scan. In their framework, a model is trained to generate combined spatial and intensity transformations that mimic anatomical variability and imaging differences. Applying these learned deformations to one labelled MRI brain scan produced many plausible training examples, enabling a conventional segmentation network to achieve significantly improved one-shot accuracy.

Most few-shot segmentation methods for medical images follow a meta-learning paradigm with a support-query training architecture. During training, the model is presented with many episodes, each consisting of a small support set (e.g. one annotated image) and a query image from base classes, to simulate the one/few-shot scenario (Bai et al., 2024). A

common strategy is metric-based learning: the support image and mask are encoded into class prototypes (feature vectors summarising the target class and background), and the query image’s pixels are classified by comparing their features to these prototypes (Bai et al., 2024). For example, Roy et al. (2020) developed a 3D few-shot segmentation network with two parallel encoder–decoder streams for support and query volumes. They introduced “squeeze-and-excite” attention modules between the streams to recalibrate the query feature maps using information from the support embedding. This guided attention mechanism helped the model focus on features relevant to the target organ in the query scan. Additionally, Sun et al. (2022) proposed a global correlation network that computes dense similarity maps between support and query feature embeddings, yielding a more discriminative representation for segmenting the query. Overall, these metric-learning approaches leverage class-specific feature comparisons and have proven effective in segmenting structures with very few annotations.

More recently, UniverSeg (Butoi* et al., 2023) introduced a universal few-shot segmentation framework designed to handle highly diverse medical imaging tasks using a single model. Unlike traditional meta-learning approaches trained on a narrow range of base classes, UniverSeg learns a universal conditioning mechanism that allows it to generalise to new anatomical structures, imaging modalities, and segmentation protocols without task-specific retraining. By conditioning the segmentation process on both image and mask pairs in a support-query format, it provides strong generalisation across unseen datasets and organ types.

In recent years, transformer architectures have been applied to few-shot medical segmentation to model long-range dependencies and complex interactions between support and query features. Zhu et al. (2023) proposed a region-enhanced prototypical Transformer that embeds both support and

query feature maps into a transformer module. By enhancing prototype representations with attention across the whole support-query pair, this method achieved improved accuracy on few-shot organ segmentation tasks. Similarly, Lin et al. (2023) developed a Cross Attention Transformer Network (CAT-Net) that explicitly mines the correlations between support and query features through masked cross-attention. Experiments on multiple 2D datasets showed that these transformer-based models outperform earlier convolution-based few-shot models, validating the benefit of attention mechanisms in the few-shot setting.

Finally, this section discusses the emerging trend of adapting foundation models like SAM and SAM2 for few-shot medical segmentation. As explained in Section 2.3, directly applying these foundation models to medical segmentation often yields suboptimal results, necessitating adaptation. However, even after fine-tuning, the inference typically remains prompt-based, meaning the model expects a user-provided point or bounding box on each image to initiate segmentation (Bai et al., 2024). This is not ideal in fully automated scenarios. To overcome these limitations, researchers have started exploring prompt-free few-shot adaptation of SAM/SAM2. Instead of retraining SAM on a large medical dataset, the idea is to use SAM/SAM2’s pre-trained capabilities in a one-shot or few-shot manner so that only a handful of annotated medical images are needed for a new task. Bai et al. (2024) demonstrated this concept with FS-MedSAM2, a framework that uses SAM2’s memory attention module to perform few-shot segmentation on medical scans without any model fine-tuning. In their approach, a single support image with its mask is provided as a mask prompt to SAM2’s prompt encoder. The model can take the support mask embedding and apply it to segment the query images by retrieving and refining the relevant memory representations. Similarly, Zu et al. (2025) introduced

a novel, training-free framework for few-shot medical image segmentation with SAM2. Starting from a single support image (and its annotation), Zu et al. (2025) generated a diverse set of augmented versions via geometric and colour transformations. For each query slice (frame) in the volume, the system automatically selects the most perceptually similar augmented support slice. This matched pair—support image and its mask is used as a mask prompt for SAM2, avoiding any retraining of the model.

2.4.2 Semi-Supervised Learning for Medical Image Segmentation

While one-shot and few-shot learning paradigms focus on adapting models to new, unseen classes with minimal examples, semi-supervised learning (SSL) addresses the broader challenge of leveraging large amounts of unlabelled data to improve the performance of models for existing classes, thereby reducing the overall annotation burden. Both few/one-shot learning and SSL contribute significantly to label-efficient learning, but from slightly different angles: few/one-shot learning excels at rapid adaptation to novel categories, whereas SSL aims to boost performance on known categories by exploiting the vast pool of readily available unlabelled data. This synergy makes SSL a crucial component in developing robust and practical medical image segmentation solutions.

The core objective of SSL is to efficiently learn valuable knowledge from the unlabelled data to achieve performance comparable to fully-supervised settings, thereby significantly reducing dependence on large, expert-labelled datasets (Solatidehkordi and Zualkernan, 2022). Two main branches of semi-supervised segmentation approaches have been dominant: (1) pseudo-labelling methods, which generate and use artificial labels for unlabelled

data, and (2) consistency regularisation methods, which encourage the model to produce stable predictions under various perturbations. These are often combined in practice, but provide a useful conceptual division.

Pseudo Labelling

Pseudo-labelling is a technique where the model’s own predictions on unlabelled data are treated as surrogate labels for further training (Ran et al., 2024). In the simplest form, one would train an initial segmentation network on the small labelled set, use this network to predict masks for unlabelled images, then add those predictions as “pseudo-labels” to the training set and continue training a new model. This process can be iterated multiple times, gradually improving the quality of the pseudo-labels as the model becomes more robust (Ran et al., 2024). A critical aspect of pseudo-labelling methods is ensuring the pseudo-labels are reliable, since using them naively can reinforce the model’s errors. To address this, researchers focus on filtering out predictions with low confidence or uncertain regions before using pseudo-labels for training (Yu et al., 2019).

Pseudo labels with higher confidence levels have been shown to be more effective in guiding model training (Wang et al., 2022). As a result, numerous confidence- or uncertainty-aware approaches have been proposed to produce more stable and reliable pseudo labels. For instance, Ssa-net (Wang et al., 2022) introduced a trust module designed to re-evaluate the pseudo labels generated by the model. This module applies a confidence-based thresholding mechanism to retain only high-confidence predictions, thereby enhancing the reliability of the selected pseudo labels. Additionally, Yao et al. (2022) involves quantifying prediction confidence by measuring the stability of a model’s output under perturbation. In their method,

an input image is perturbed (e.g., via Fourier transformation), and the pixel-wise Kullback-Leibler (KL) divergence between the predictions for the original and perturbed images is calculated to estimate variance, which is then used to compute a confidence score, guiding the selection of the most trustworthy pseudo-labels for inclusion in the loss calculation.

The efficacy of the pseudo-labelling strategy is fundamentally constrained by the quality of the pseudo labels themselves. The central challenge remains the management of confirmation bias, where a model that has overfit to the initial small labelled set is unable to correct its own mistakes. This can initiate a damaging cycle where incorrect predictions are not only amplified but are also assigned high confidence by the model, leading to the generation of confident but erroneous pseudo labels that progressively corrupt the training process. While the pseudo labelling methods effectively encourage low-density separation between classes by reinforcing high-confidence predictions, this becomes a liability when that confidence is misplaced. Therefore, the critical considerations for employing this technique are the judicious selection of which pseudo labels to trust and the careful determination of the optimal number of training iterations to maximise the benefits of data augmentation without succumbing to the propagation of systemic error.

Consistency Regularisation

Another dominant paradigm in SSL for medical image segmentation involves incorporating unlabelled data directly into the model’s training loop via unsupervised regularisation. In contrast to multi-stage pseudo labelling, this approach augments a standard supervised loss function with an additional unsupervised loss term, which is optimised simultaneously. This

allows the model to learn from both scarce labelled data and abundant unlabelled data in an end-to-end manner. The efficacy of this approach is based on the core assumptions of SSL, which posit that a well-generalised decision boundary should lie in low-density regions of the feature space (Tarvainen and Valpola, 2017). The overarching principle used to enforce this is consistency regularisation, which asserts that a model’s prediction should remain robust and stable when subjected to small, semantically meaningless perturbations of its input or internal state. This section provides a comprehensive review of the primary strategies that embody this principle.

Consistency learning is the most direct application of this regularisation principle. It operates by enforcing the invariance of model predictions for the same input under different stochastic perturbations, thereby encouraging the decision boundary to avoid high-density data regions. The unsupervised consistency loss is calculated as a distance measure (e.g., Mean Squared Error or KL-Divergence) between the outputs of two differently perturbed versions of an input (Jiao et al., 2024).

A SOTA approach is the Mean Teacher model (Tarvainen and Valpola, 2017). In mean teacher, a student model is forced to match the teacher model’s predictions on the same image under different perturbations, where the teacher is often the exponential moving average of the student’s weights. The consistency loss is calculated between the prediction of the student and the more stable prediction of the teacher for the same perturbed input. This on-the-fly target generation makes the process highly efficient.

The choice of perturbation is critical to the success of consistency learning. Perturbations must be strong enough to challenge the model but not so strong as to alter the image’s semantic content. The input-level pertur-

bations include transformations applied directly to the input image, such as Gaussian noise, blurring, random rotation, scaling, and contrast variations (Bortsova et al., 2019). Additionally, feature-level perturbations can be introduced into the network’s intermediate feature maps. Examples include injecting feature noise, applying dropout to feature channels, or even designing parallel decoders with different architectural properties (Ouali et al., 2020; Xu et al., 2021a). Furthermore, task-level regularisation involves adding auxiliary tasks to the network, such as predicting a signed distance map (SDM) or reconstructing the input image (Luo et al., 2021; Zhang and Zhang, 2021; Shi et al., 2023). The model is then regularised by enforcing consistency between the output of the main segmentation task and the outputs of these auxiliary tasks. Despite its effectiveness, finding the right balance for perturbations is non-trivial. Weak perturbations may not provide a sufficient learning signal, while overly strong perturbations can confuse the model and degrade performance.

SAM/SAM2-Enhanced SSL

Recently, some methods use SAM/SAM2 to select reliable pseudo-labels. For instance, SAMDSK (Zhang et al., 2023b) utilised SAM with an empty prompt on unlabelled data and then relied on a CNN, trained on the labelled dataset, to select reliable predictions. This approach essentially uses SAM as an external pseudo-labelling tool. Similarly, Li et al. (2023) proposed a consistency filtering strategy: they first obtain a prediction from a trained CNN on an unlabelled image, then use that prediction to prompt SAM (e.g. by giving SAM the bounding box of the CNN-predicted mask) and obtain SAM’s segmentation. If the SAM output agrees well with the CNN output, it is deemed a reliable pseudo-label and used for training. Zhang et al. (2024b) took this further in SemiSAM, a framework where a

Mean Teacher CNN is trained on the labelled data and periodically used to generate prompt points or boxes for SAM on unlabelled images. SAM then produces a segmentation which acts as an additional supervision signal for the student network during training. By plugging SAM’s output into the semi-supervised training loop, they reported notable improvements, especially in extremely low-label settings. Based on SemiSAM (Zhang et al., 2024b), SemiSAM+ (Zhang et al., 2025) leveraged one or multiple promptable foundation models as “generalist” models and a trainable task-specific segmentation model as a “specialist” model. It aims to efficiently learn from limited labelled data by exploiting the pre-trained knowledge of the foundation model. These approaches demonstrated that SAM could boost the learning of a medical segmentation model by providing extra pseudo-labels or guidance, effectively enlarging the annotated set with SAM-provided annotations.

However, a limitation observed in those initial attempts is their domain gap, since SAM is originally pretrained on natural images. Chen et al. (2024b) introduced a method where SAM itself is fine-tuned on the fly within a semi-supervised loop. In their approach, SAM’s encoder is partially fine-tuned on the small labelled set (to learn medical image features) and then used to segment unlabelled images with prompts generated from the student model. To modify the prompt generation process, CPC-SAM (Miao et al., 2024) employs cross-prompting consistency, it sets up two parallel segmentation branches that both utilise SAM as the core, and these branches prompt each other in a criss-cross manner with unlabelled data. This design not only adapts SAM through fine-tuning in the loop but also tackles the prompt variability issue by making the model robust to how prompts are given.

In summary, both one-/few-shot learning and SSL represent complementary strategies to address the persistent challenge of limited annotations in

medical image segmentation. At their core, both paradigms share the common goal of maximising knowledge extraction from scarce labelled data while exploiting unlabelled or weakly labelled information. The recent emergence of foundation models such as SAM and SAM2 provides new opportunities for integrating these paradigms. SAM’s prompt-driven capabilities naturally align with few-shot settings, as accurate segmentations can be obtained from minimal annotations or even interactive clicks. Meanwhile, SAM2 extends this utility by enabling temporal or volumetric propagation, making it particularly suited for SSL frameworks where unlabelled data are abundant. By serving as reliable pseudo-label generators or teachers, these models can substantially enhance both few-shot generalisation and semi-supervised consistency, thereby reducing annotation dependence even further.

Nevertheless, important limitations remain. First, both few-shot and SSL frameworks heavily rely on the quality of supervision signals: in few-shot learning, overfitting to the scarce labelled set remains a critical risk, while in SSL, error propagation from noisy pseudo labels can degrade performance. Additionally, current SAM-based SSL and few-shot pipelines often depend on heuristic prompting strategies that may not generalise across modalities or pathologies. Furthermore, most methods are still constrained by slice-level segmentation strategies, with limited exploration of fully 3D-consistent frameworks.

2.5 Conclusions

This chapter has reviewed the evolution of deep learning-based medical image segmentation from classical convolutional models to the lat-

est Transformer- and foundation-model-driven frameworks. The review revealed that while remarkable progress has been achieved, current approaches still face several intertwined challenges concerning data efficiency, model generalisation, and computational scalability—all of which limit their direct applicability in real clinical settings.

Early research built upon CNNs such as FCN (Long et al., 2015) and U-Net (Ronneberger et al., 2015) demonstrated strong capability in extracting local texture and boundary information from limited data. However, their inherent locality restricted the capture of long-range contextual dependencies, often resulting in fragmented or spatially inconsistent segmentations in complex anatomical regions. The emergence of Transformer architectures and their adaptation to vision tasks introduced the ability to model global context through self-attention, leading to a new class of CNN–Transformer hybrid models. These hybrids successfully integrated the complementary strengths of both paradigms but brought substantial increases in parameter size and computational cost, constraining their feasibility for high-resolution or real-time clinical use.

Recent studies have therefore shifted focus toward improving the efficiency and generalisability of the methods. Efforts to simplify Transformer blocks have reduced spatial complexity but often sacrifice contextual knowledge learning and lack theoretical support. Simultaneously, research on multi-organ and cross-dataset segmentation has sought to address the fragmented nature of clinical data, where annotations are often incomplete or heterogeneous across sources. Existing strategies, such as multi-head networks, conditional decoders, or pseudo-labelling pipelines, improved performance but at the expense of training complexity and scalability. Consequently, there remains a need for unified, lightweight frameworks that can learn effectively from partially labelled, domain-diverse datasets without extensive

manual curation.

Parallel to these developments, the field has witnessed the rise of large vision foundation models, such as SAM and SAM2, which demonstrate powerful zero- and few-shot segmentation capability. These models signal a paradigm shift from task-specific to general-purpose learning, offering a potential route toward annotation-efficient and rapidly adaptable medical segmentation. Nevertheless, due to the domain gap between natural and medical images, their direct application still requires careful adaptation and integration with traditional training schemes. Few-shot and semi-supervised learning methods complement this trend by leveraging unlabelled data and minimal annotations, but their success depends heavily on reliable pseudo-labels and stable consistency regularisation.

In summary, the literature reveals four persistent research gaps that motivate the contributions of this thesis:

- The need for a hybrid architecture that balances CNNs’ local precision with Transformers’ global reasoning while remaining computationally tractable.
- The need for a lightweight unified segmentation framework capable of learning from multiple partially labelled or domain-heterogeneous datasets.
- The need for an efficient and theory-supported Transformer mechanism that reduces computational overhead without compromising contextual understanding.
- The need for annotation-efficient frameworks that effectively combine the foundation models and semi-supervised learning to alleviate manual labelling demands.

To address these identified gaps, the subsequent chapters introduce a series of progressively developed models—CTranS, MO-CTranS, CRFTrans, and SSL-MedSAM2—that collectively advance medical image segmentation toward improved efficiency, enhanced generalisation, and greater practicality for real-world clinical deployment.

Chapter 3

CTranS: A Multi-resolution Convolution- Transformer Network for Medical Image Segmentation

3.1 Introduction

Accurate image segmentation is an essential and foundational task for many clinical applications, as it provides critical information for tasks like diagnosis, treatment planning, and monitoring disease progression. The development of successful deep learning-based segmentation models hinges on their ability to learn and integrate both global contextual information and fine-grained local details from medical images (Gong et al., 2024b). This presents a significant challenge, as traditional architectures have inherent limitations in capturing both types of information simultaneously.

Convolutional Neural Networks (CNNs), particularly the U-Net (Ronneberger et al., 2015) architecture, have long been the state-of-the-art (SOTA) for medical image segmentation due to their effectiveness in learning hierarchical, multiresolution local features from small datasets. However, their primary weakness lies in the small, local receptive fields of their convolutional kernels, which fundamentally limit their ability to capture long-range spatial dependencies and model global contextual relationships across an entire image (Gong et al., 2024b). This limitation can lead to fragmented or inconsistent segmentations, especially for large or irregularly shaped anatomical structures. In contrast, Transformer-based methods, inspired by their success in natural language processing, utilise self-attention mechanisms to explicitly capture long-range relationships between image patches, effectively addressing the global context problem of CNNs (Gong et al., 2024b). Despite their strengths, pure Transformer models are notoriously data-hungry and computationally intensive, scaling quadratically with input size and requiring a large number of learnable parameters. These characteristics make them less suitable for the data-limited and resource-constrained environment of medical imaging, where obtaining large-scale annotated datasets is a significant challenge.

This chapter introduces CTranS, a novel model designed to address these core challenges by optimally combining the strengths of both paradigms. CTranS directly responds to the research question posed in Chapter 1: How can a hybrid CNN-Transformer architecture be designed to fully utilise both advantages for effective medical image segmentation? The method aims to fulfil Objective 1 of this thesis: To establish a robust baseline medical image segmentation model that effectively combines the strengths of CNNs and Transformers to capture both local and global contextual information, particularly in the context of limited medical datasets. By

jointly modelling global context and local features in a multi-resolution manner, CTranS aims to establish a strong baseline for the subsequent efficiency- and generalisability-focused developments in this thesis.

Inspired by CNNs and ViT (Dosovitskiy et al., 2020), CTranS utilises the advantages of both methods. Due to the effective feature learning property provided by CNN, CTranS applies CNN to extract multi-resolution image features, which is considered as an encoder. These learned features represent a rich characteristic of local image patches, which are then input to a set of attention blocks in a Transformer, which forms the decoder of the proposed segmentation model. Both the encoder and decoder are multi-resolution, which are interconnected seamlessly via computationally efficient re-sampling and re-shaping operations. This novel architecture design enables both effective local feature learning and long-range contextual information interaction for image segmentation. For method comparison, CTranS and several other state-of-the-art approaches were evaluated on multiple publicly available datasets for binary-class segmentation, including skin lesion dataset (ISIC2018 (Codella et al., 2019)), cell dataset (Bowl: <https://www.kaggle.com/c/data-science-bowl-2018>), polyp dataset (CVC-ClinicDB (Bernal et al., 2015)), and multi-class segmentation, including cell dataset (Pannuke (Gamper et al., 2019)) and brain tissue dataset (OASIS-1 (Marcus et al., 2007)).

The contributions of this work are summarised as follows. (1) Novel model architecture: this work applies CNN as an encoder to extract multiresolution image features and applies Transformer as a decoder to learn the long-range dependency in multiresolution. To reduce the information loss in the feature upsampling process, simplify the model and increase the training speed when applying Transformer, this work proposes computationally efficient pyramid patch-based operations to connect the encoder and decoder

in multiresolution. (2) Fewer model parameters without pretraining: compared to other Transformer-based methods, the proposed method has the fewest model parameters without the requirement of model pretraining to achieve better performance.

3.2 Methodology

As illustrated in Figure. 3.1(A), the proposed framework comprises three key components: a CNN-based encoder (CE), a Transformer-based decoder (TD), and a Pyramid Feature Fusion (PFF) module. Drawing inspiration from the effectiveness of U-Net (Ronneberger et al., 2015) in handling small-scale datasets, the encoder and decoder are connected across multiple resolutions to preserve spatial information. At each resolution level, the CE extracts local feature maps from N image patches, which are then processed by the PFF module into N one-dimensional image tokens. These tokens serve as inputs to the TD, where they pass through three sequential Transformer blocks (denoted as Trans block $\times 3$). To incorporate spatial context, positional encoding (POS) is added to the input tokens at the deepest level of the decoder, enabling the model to capture positional relationships among the image patches.

3.2.1 CNN-based Encoder

Similar to ResUNet (Zhang et al., 2018), each ResConv block consists of two convolutional layers, each followed by a batch normalisation layer and a ReLU activation layer, and a skip connection consisting of a convolutional layer and a batch normalisation layer is added between the output and the input. The first convolutional layer and the convolutional layer in the

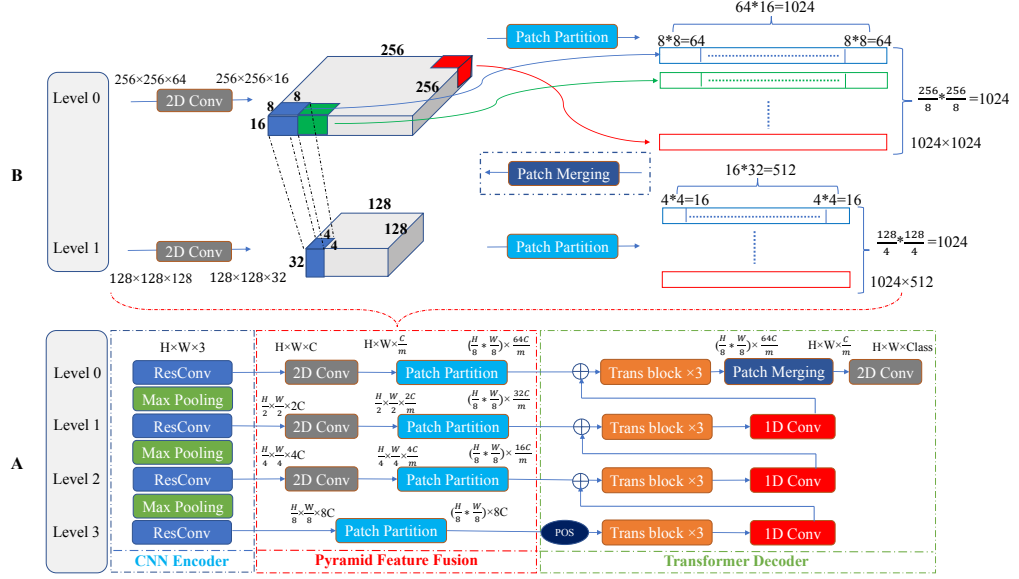


Figure 3.1: Figure A is the overall model structure of CTranS. The example uses an input of a 3-channel colour image and contains 4 levels. Figure B shows the details of the Pyramid Feature Fusion module. The example shows the process in level 0 and level 1 and uses a $256 \times 256 \times 3$ image as input and assumes $C = 64$ and $m = 4$. H and W indicate the height and width of the input image, $*$ indicates value multiplication, and $H \times W$ means the feature dimension of H “by” W .

skip connection change the number of feature channels to C in the first level and double the number in other levels. The second convolutional layer maintains the channel number. Except for the deepest level, a max pooling layer with the kernel size of 2×2 is applied after ResConv to downsample the feature map. Each feature in the lower-resolution feature map represents a larger region of the input image. In CTranS, the size of the image patch is determined by the size of the field of view that a single feature can represent at the deepest level (8×8 in the example shown in Figure. 3.1 (B)).

3.2.2 Pyramid Feature Fusion

The Pyramid Feature Fusion (PFF) module is designed to facilitate efficient learning of multi-resolution CNN features within the Transformer decoder

(TD). Rather than converting every spatial feature in the CNN feature maps into individual input tokens for the TD—a process that would significantly increase computational cost due to the large token count, PFF fixes the number of image patches in each resolution level as the spatial size of the CNN-extracted feature map in the deepest level ($H/8 \times W/8$ in the example case shown in Figure. 3.1). As depicted in Figure. 3.1(B), the blue regions across different resolution levels correspond to the same field of view represented by a single feature at the deepest level (i.e., an image patch). Directly flattening the CNN features of an image patch and feeding them into the Transformer would result in high-dimensional tokens, thereby increasing the computational burden of the Transformer blocks. To address this, PFF reduces the number of feature channels by a factor of m using a 2D convolutional layer. It is important to note that the patch merging operation illustrated in the dashed box is not part of the PFF module itself; rather, it is the inverse of the patch partition process and is applied at the final stage of the TD.

As an example, the feature map from CE has the dimension of $128 \times 128 \times 128$ in level 1 (as shown in Figure. 3.1(B)). The down-sampling linear layer in level 1 then first reduces the dimension of the feature channels from 128 to 32 for dimensionality reduction purpose. After channel size reduction, the $128 \times 128 \times 32$ feature map is divided into 1024 non-overlapping image patches, each with a size of $4 \times 4 \times 32$, by “Patch Partition”. Meanwhile, during “Patch Partition”, each 3D image patch is flattened into a one-dimensional feature embedding with size 512, resulting in 1024 image embeddings with 1×512 size.

3.2.3 Transformer-based Decoder

The Transformer was initially introduced by Vaswani et al. (2017) for machine translation tasks in natural language processing. Building upon the encoder architecture of the Transformer, Dosovitskiy et al. (2020) proposed the Vision Transformer (ViT) for image classification tasks. For clarity, **the Transformer layer referenced throughout this thesis refers to the fundamental building block used in ViT, as illustrated in Figure. 3.2.** Each resolution level within the Transformer-based decoder (TD) consists of three consecutive Transformer blocks (Trans blocks), which are responsible for modelling long-range dependencies based on the image tokens generated by the Pyramid Feature Fusion (PFF) module. To enable multi-resolution feature propagation, a 1D 1×1 convolutional layer (1D Conv) is used to project tokens from level i to level $i - 1$, and the resulting output is added to the input tokens at level $i - 1$. As shown in Figure. 3.2, each Trans block comprises a multi-head self-attention mechanism (MHA)

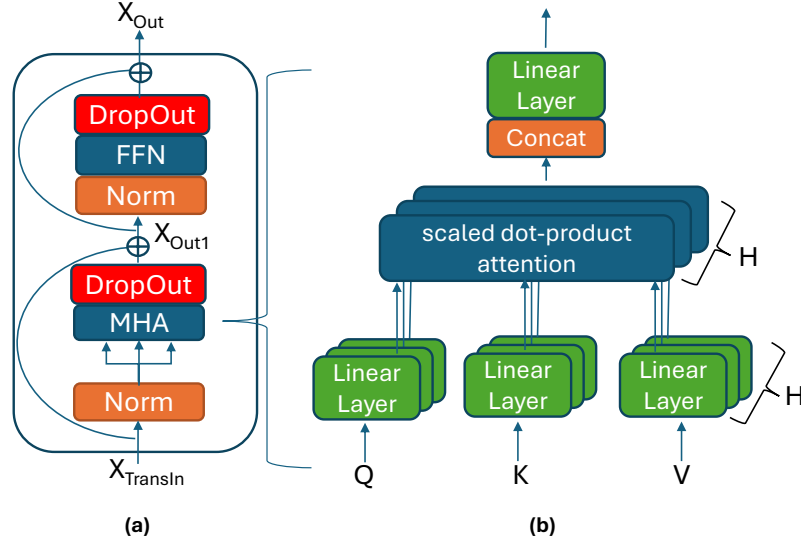


Figure 3.2: The architecture of a Transformer layer. FFN indicates feed-forward networks, Norm indicates layer normalisation, MHA indicates multi-head attention layer, and Concat indicates feature concatenation. Q , K , and V indicates the query, key, and value, and H is the number of heads.

and a position-wise fully connected feedforward network (FFN). Layer normalisation is applied before the image embeddings are input to MHA and FFN, while dropout is applied after each component. Skip connections are also incorporated between the input and the output of the dropout layer in both MHA and FFN, ensuring stable gradient flow and improved training dynamics.

Rather than relying on a single attention operation, MHA splits the input tokens into multiple subspaces and computes scaled dot-product attention (SDPA) in parallel across these subspaces. Learnable linear transformations are applied before and after SDPA to project the input and aggregate the attention outputs. At the finest level (level 0) of the TD, the output tokens from the final Trans block are reshaped into a 2D feature map of dimensions $W \times H \times C/m$ via a patch merging operation. This reshaped feature map is then passed through a 2D 1×1 convolutional layer to generate the final segmentation map. The model is trained using a composite loss function, defined as the average of the soft Dice loss and cross-entropy loss, which jointly optimise boundary alignment and pixel-wise classification accuracy.

3.3 Experiments and Results

3.3.1 Datasets and Evaluation Metrics

To evaluate CTranS and compare it to other state-of-the-art methods, five publicly available medical imaging datasets that cover four medical applications are used:

- (1) Skin lesion segmentation (**ISIC**): ISIC2018 (Codella et al., 2019) (binary-

class), which was published by the International Skin Imaging Collaboration challenge, containing 2594 RGB dermoscopy images with different image sizes.

(2) Cell segmentation: 2018 Data Science Bowl (**Bowl**): <https://www.kaggle.com/c/data-science-bowl-2018> (binary-class) and **Panuke** (Gamber et al., 2019) (5 classes). There are 670 320×256 RGB images in the 2018 Data Science Bowl and 7901 256×256 RGB images in Pannuke. The 5 classes for the segmentation mask in Pannuke are Neoplastic (Neo), Inflammatory (Inf), Connective (Con), Dead, and Epithelial (Epi), respectively.

(3) Polyp segmentation (**CVC**): CVC-ClinicDB (Bernal et al., 2015) (binary-class) contains 612 384×288 RGB colonoscopy images from 31 colonoscopy sequences.

(4) Brain tissue segmentation (**OASIS**): OASIS-1 (Marcus et al., 2007) (4 classes) contains 414 192×192 grey scale images, the 4 classes for the segmentation mask are Cortex, Subcortical-Gray-Matter (SGM), White-Matter (WM), and Cerebrospinal fluid (CSF) respectively.

Although CTranS can accept any size of input image (as long as it can be divided by 2^l), in the experiments, all images were reshaped to 224×224 since most of the methods for comparison used such a resolution as the input size (due to they used pre-trained models). Especially, when CTranS was compared with UTnet (Gao et al., 2021), the images were reshaped to 256×256 , since UTnet used the input size of 256×256 , and the input size of 224×224 cannot fit in the model by their default parameter settings. The percentage of training, validation and testing sets is 70%, 10% and 20%, respectively, for all tasks. All models were trained on a 12GB GeForce GTX 1080 Ti GPU with Adam optimiser. Additionally, the learning rate was set to 0.0001, and the models were all trained for 50 epochs. During

each training phase, the model with the smallest validation loss was saved for testing.

In model evaluation, Dice Coefficient (DC) and Average Symmetric Surface Distance (ASSD) are the main metrics, and the Wilcoxon Signed Rank Test (WSRT) was applied to evaluate the statistical significance of the difference between two methods. This work also compared the number of learnable parameters (NP), the GPU memory usage (GM), and the average training time per step (TS) to evaluate the model complexity and training efficiency.

It is worth noting that, for the Pannuke dataset, some classes have a large number of empty ground truth. The initial results show that the model tends to consistently generate empty predictions on most images in such classes. Thus, this work only calculated the loss on the prediction of the class whose ground truth was not empty in the training process for all methods to avoid model over-fitting caused by unbalanced labels. Similarly, by including these empty classes in the evaluation, it results in very high DC scores. It makes the measure less sensitive to the errors in images with a non-empty foreground. Hence, this work also only calculated DC and ASSD on non-empty classes (in the ground truth mask) in the evaluation.

3.3.2 Parameter Settings

Before comparing CTranS with other state-of-the-art methods, the sensitivities on the choice of hyperparameters in CTranS were tested by performing parameter tuning. Specifically, $l = 3$ was set for the depth of CE and TD, so that each feature vector processed in TD is a representation of an 8×8 image patch. Additionally, this work tested and compared the method performance of using 1, 2, 3, 4, 5 for n (number of Transformer blocks in

TD), 32, 64 for C_{base} (number of output feature channels of the first convolutional block) and 2, 4, 8 for m (down-sampling factor). ISIC2018 was selected as a binary-class dataset and OASIS-1 as a multi-class dataset to evaluate the above parameter settings. Considering the trade-off between model performance and model complexity in both datasets, the optimal settings were $n = 3$, $C_{base} = 64$, and $m = 8$. These settings were then consistently used for evaluating all other datasets and compared to other methods.

3.3.3 Ablation Study

The two key novelties of CTranS design are the integrated multi-level feature fusion and the down-sampling linear layers in skip connections to reduce the model’s complexity. Firstly, to illustrate the benefit of utilising the output feature maps from CE in multiresolution rather than only using the output of level l , CTranS was evaluated with and without the skip connections (SC) between CE and TD. Secondly, CTranS was evaluated with and without the down-sampling linear layers (DSL). The OASIS-1 dataset is chosen to use in these experiments, since it is a more generic multiclass dataset which includes both large and small objects of interest. Smaller C_{base} and m values were used instead of the parameter settings that are concluded in section 3.3.2. Because the model (with SC without DSL) requires significantly larger memory that can not be accommodated by a 12 GB GPU. Thus, C_{base} was changed to 32 instead of 64, which decreases the number of parameters for all models so that the models can be trained on the same GPU. By setting m to 4 instead of 8, it ensures that the model with DSL will not lose much information by the down-sampling process.

The mean and standard deviation of DC and ASSD are reported in Table

Table 3.1: Evaluation of model performance on OASIS-1 dataset ($C_{base} = 32$, $m = 4$). SGM indicates Subcortical-Gray-Matte, WM indicates White-Matter, and CSF indicates cerebrospinal fluid. Mean and Standard deviation of Dice Coefficient (DC) and Average Symmetric Surface Distance (ASSD) are reported for different classes and the Average of all classes

SC	DSL	Cortex	SGM	WM	CSF	Mean
DC						
Yes	Yes	0.904 ± 0.015	0.941 ± 0.014	0.957 ± 0.006	0.929 ± 0.032	0.933 ± 0.009
Yes	No	0.906 ± 0.015	0.939 ± 0.016	0.957 ± 0.006	0.919 ± 0.036	0.930 ± 0.010
No	\	0.878 ± 0.016	0.924 ± 0.016	0.947 ± 0.006	0.901 ± 0.035	0.912 ± 0.010
ASSD						
Yes	Yes	0.514 ± 0.075	0.516 ± 0.170	0.453 ± 0.100	0.396 ± 0.131	0.470 ± 0.083
Yes	No	0.491 ± 0.062	0.520 ± 0.188	0.432 ± 0.080	0.442 ± 0.131	0.472 ± 0.084
No	\	0.675 ± 0.079	0.838 ± 0.195	0.538 ± 0.081	0.687 ± 0.207	0.685 ± 0.100

Table 3.2: Evaluation of model complexity ($C_{base} = 32$, $m = 4$) on OASIS-1 dataset. The mean value of the number of parameters (NP), memory consumption (GM) and training time per step are reported

SC	DSL	NP (m)	GM (GB)	TS (s)
Yes	Yes	11.84	2.41	0.126
Yes	No	138.34	9.35	0.423
No	\	11.83	2.39	0.129

4.1, and the evaluations of model complexity are reported in Table 3.2. Firstly, it is seen from Table 4.1 that compared to the model without SC, SC (with and without DSL) can increase the performance measured by both DC and ASSD significantly in all classes ($p < 0.01$ measured by WSRT). Additionally, the use of DSL further increased the performance on the mean class results measured by DC significantly ($p < 0.01$ measured by WSRT) and by ASSD non-significantly ($p = 0.38$ measured by WSRT), compared to the model without DSL.

By achieving the best segmentation accuracy, the model with SC and DSL has very similar nP, GM and TS, compared to the model without SC and DSL, as shown in Table 3.2. The model with SC but without DSL requires

significantly larger NP, GM and longer TS, although it has a similar performance to the model with both SC and DSL. Both Table 4.1 and 3.2 show the necessity of SC and DSL in achieving good segmentation performance with lower model complexity, less memory consumption and quicker training speed.

3.3.4 Comparison to SOTA Methods

Next, this work compared the performance of CTranS with the SOTA methods. The experiments chose one representative method with highly reported performance from each of the categories described in section 2.1.2. These methods are FAT-Net (Wu et al., 2022) from the category *Transformer in encoder*, SegTran (Li et al., 2021a) from the category *Transformer in decoder*, TransUnet (Chen et al., 2021b) from the category *Transformer as a bridge*, and UTnet (Gao et al., 2021) from the category *Fused CNN and Transformer*, SwinUnet (Cao et al., 2021) from the category *Pure Transformer*. Additionally, CTranS was also compared with the pure CNN method ResUNet (Zhang et al., 2018). For a fair comparison, data augmentation was not applied in the training process in any of the methods. Additionally, besides using the input size of 224×224 , CTranS and ResUNet were also trained with an input size of 256×256 for comparison with UTnet (Gao et al., 2021). The mean and standard deviation of DC and ASSD are reported in section 3.3.4 to compare the segmentation accuracy of all methods. Additionally, NP, GM and TS are reported in section 3.3.4 to compare the models' complexity and training efficiency. Finally, some qualitative results are shown in section 3.3.4 to demonstrate the advantages and drawbacks of each method visually.

Segmentation Accuracy

Table 3.3: Comparison of model performance on five public datasets. Mean and Standard deviation of Dice coefficient (DC) and Average Symmetric Surface Distance (ASSD) are reported. * indicates that the method used 256×256 as the input size, otherwise 224×224 was used. Numbers in bold are the best methods in each input size. † indicates that the result is significantly worse than the best results ($p < 0.01$ measured by WSRT).

Model	ISIC	Bowl	Pannuke	CVC	OASIS
DC					
ResUNet	0.831 ± 0.195 †	0.914 ± 0.073	0.512 ± 0.190 †	0.774 ± 0.225 †	0.937 ± 0.011
SwinUnet	0.877 ± 0.145 †	0.905 ± 0.059 †	0.516 ± 0.186 †	0.792 ± 0.222 †	0.933 ± 0.010 †
FAT-Net	0.884 ± 0.139 †	0.901 ± 0.069 †	0.495 ± 0.197 †	0.894 ± 0.163	0.775 ± 0.008 †
SegTran	0.874 ± 0.161 †	0.895 ± 0.077 †	0.526 ± 0.188 †	0.835 ± 0.253 †	0.906 ± 0.010 †
TransUnet	0.894 ± 0.120	0.900 ± 0.063 †	0.500 ± 0.182 †	0.872 ± 0.173 †	0.917 ± 0.011 †
CTS (ours)	0.900 ± 0.110	0.917 ± 0.065	0.551 ± 0.180	0.900 ± 0.137	0.933 ± 0.010 †
ResUNet*	0.813 ± 0.197 †	0.919 ± 0.063 †	0.525 ± 0.187 †	0.781 ± 0.227 †	0.943 ± 0.011
UTnet*	0.879 ± 0.128 †	0.837 ± 0.243 †	0.553 ± 0.189 †	0.862 ± 0.184 †	0.936 ± 0.012 †
CTS (ours)*	0.885 ± 0.152	0.925 ± 0.066	0.564 ± 0.179	0.889 ± 0.158	0.943 ± 0.010
ASSD					
ResUNet	9.432 ± 20.595 †	1.485 ± 3.648 †	25.422 ± 25.227 †	16.605 ± 39.002 †	0.438 ± 0.108
SwinUnet	5.279 ± 6.862 †	1.482 ± 2.827 †	27.921 ± 28.299 †	7.066 ± 9.806 †	0.454 ± 0.095 †
FAT-Net	5.153 ± 6.742 †	1.444 ± 2.303 †	23.567 ± 25.987 †	7.330 ± 29.772 †	0.933 ± 0.095 †
SegTran	6.736 ± 18.198 †	1.436 ± 2.076 †	24.392 ± 24.193 †	14.854 ± 48.232 †	0.725 ± 0.132 †
TransUnet	4.654 ± 6.320 †	1.406 ± 1.750 †	59.111 ± 26.916 †	5.886 ± 12.744 †	0.601 ± 0.103 †
CTS (ours)	4.211 ± 4.751	1.125 ± 1.642	21.535 ± 19.320	4.003 ± 6.878	0.469 ± 0.095 †
ResUNet*	10.909 ± 17.974 †	1.475 ± 2.620 †	26.681 ± 29.764 †	16.643 ± 39.871 †	0.465 ± 0.121 †
UTnet*	6.405 ± 12.958 †	5.083 ± 26.011 †	27.328 ± 29.459 †	8.567 ± 26.379	0.466 ± 0.102 †
CTS (ours)*	5.501 ± 7.742	1.167 ± 1.695	22.590 ± 20.328	5.202 ± 9.038	0.430 ± 0.087

From the results reported in Table 3.3, it is seen that CTranS achieved the best mean DC and mean ASSD on four out of five datasets (except for OASIS) using both input image sizes. CTranS ranked second after ResUNet and third after ResUNet and SwinUnet on the OASIS dataset using DC and ASSD measurements respectively, when the input image size is 224×224 . When a larger input image size (256×256) is used, CTranS

achieved the same performance as ResUnet and performed better than UT-net using DC measure, and it performed the best using ASSD measure. In conclusion, CTranS (CTS) is overall the best method when evaluated on different datasets with varying input image size based on DC and ASSD measures.

Statistical tests (WSRT) were also performed to compare each method to the best method for each dataset indicated by † in Table 3.3. It can be seen that CTranS is the best method in most datasets. Although it is not always significantly better ($p < 0.01$ by WSRT) than the second-best method, the second-best method is not always the same in these datasets.

Model Complexity

Table 3.4 presents the measurements of NP, GM and TS for comparing the number of parameters, memory consumption and training speed respectively. It is seen that CTranS requires the least number of learning parameters compared to other Transformer based methods. Although the pure Transformer SwinUnet consumed the least GPU memory and had the quickest training speed, it performed worse than CTranS in 4 out of 5 datasets and had similar performance on the remaining dataset (OASIS) in Table 3.3. More importantly, although CTranS is not the best method in terms of model complexity and training speed, pre-training is not needed in CTranS. The only Transformer based method that does not require pre-training is UTnet, but it has many more learnable parameters and performs worse in all datasets than CTranS.

Table 3.4: Comparison on model complexity on binary segmentation tasks. The mean value of the number of parameters (NP), memory consumption (GM) and training time per step (TS) are reported

Model	nP/m	GM/GB	TS (s)
ResUNet	13.04	1.39	0.120
SwinUnet	27.17	0.94	0.110
FAT-Net	28.76	1.01	0.132
SegTran	50.47	0.96	0.131
TransUnet	105.28	2.46	0.174
CTS (ours)	21.48	2.21	0.139
ResUNet*	13.04	1.39	0.161
UTnet*	57.46	2.67	0.250
CTS (ours)*	21.60	3.03	0.161

Qualitative Analysis

To identify the dataset exhibiting the largest performance discrepancy among different methods, the sum of pairwise absolute differences of the Dice coefficient (DC) values across methods was calculated for each dataset, as reported in Table 3.3. The CVC-ClinicDB dataset was identified using such a measure, hence some examples were extracted from this dataset for qualitative comparison.

To visualise the performance differences between different methods, the distributions of DC values of all methods on the CVC dataset are plotted in Figure. 3.3. The plot illustrates that CTS (CTranS) and FAT-net have higher-quality segmentation results ($DC > 0.9$) than other methods. ResUNet and SegTran have more low-quality segmentation results ($DC < 0.1$) than other methods. SegTran also uses the Transformer in the decoder, but different from CTranS, it reveals the strength of the proposed new model structure.

From both Table 3.3 and Figure 3.3, it can be seen that ResUNet performed much worse than all other Transformer-based methods on the CVC dataset. The CVC dataset is for polyp detection, which contains many challenging

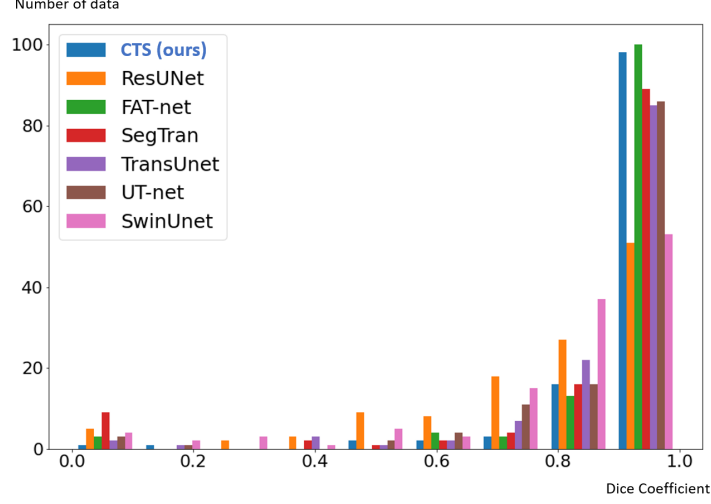


Figure 3.3: The distributions of DC values of different methods on the CVC dataset

cases where the foreground and background share very similar image intensities and feature patterns. In this case, long-range image dependencies and interactions play an important role in distinguishing foreground from background when they share similar features.

	Input	Ground Truth	Decoder	PureC	Encoder	Fusion	Decoder	Bridge	PureT
			CTS (ours)	ResUNet	FAT-net	UT-net	SegTran	TransUnet	SwinUnet
(a)	Data	GT	Dice: 0.669	Dice: 0.0	Dice: 0.57	Dice: 0.648	Dice: 0.0	Dice: 0.974	Dice: 0.952
(b)	Data	GT	Dice: 0.888	Dice: 0.333	Dice: 0.564	Dice: 0.741	Dice: 0.353	Dice: 0.34	Dice: 0.256
(c)	Data	GT	Dice: 0.148	Dice: 0.514	Dice: 0.898	Dice: 0.477	Dice: 0.89	Dice: 0.436	Dice: 0.325
(d)	Data	GT	Dice: 0.465	Dice: 0.549	Dice: 0.0	Dice: 0.0	Dice: 0.0	Dice: 0.0	Dice: 0.0

Figure 3.4: Some predictions on CVC-ClinicDB that show a large difference between ResUNet and CTranS. The images from left to right are the original data, ground truth, predictions of CTranS, ResUNet, FAT-net, UT-net, SegTran, TransUnet, and SwinUnet. ‘Decoder’, ‘PureC’, ‘Encoder’, ‘Fusion’, ‘Bridge’, ‘PureT’ indicate the category *Transformer in decoder*, *Pure CNN*, *Transformer in encoder*, *Fused CNN and Transformer*, *Transformer as a bridge*, *Pure Transformer* respectively

Furthermore, several representative cases with large standard deviations in

the Dice coefficient (DC) values among different methods were selected, as shown in Figure. 3.4. From Figure. 3.4 (a), it can be seen that when the background is dark, ResUNet produced no predictions. CTranS was influenced by the noise (the reflection part), which resulted in over-segmentation in the bottom part. The results from both TransUnet and SwinUnet were more similar to the ground truth mask. However arguably, it is a challenging case even for human annotators to produce a reliable ground truth. In Figure. 3.4 (b), the target object shares very similar image intensity with a large part of the background. CTranS produced the best results compared to other methods. Additionally, Figure 3.4 (c) represents a case of a large region of interest, where the foreground is not continuous in the raw input image. In this case, CTranS under-segmented the target with the lowest DC value, since multi-scale feature learning would introduce some bias when the different regions of the foreground look different. Larger patch size might help with such case, but the accuracy in boundary detection could be affected. The example in Figure. 3.4 (d) can be considered as the situation where the target is difficult to distinguish from the background, where only ResUNet and CTranS can make some correct predictions in the target region. Although ResUNet has a higher dice coefficient, CTranS captures the location of the target more correctly. Overall, SegTran (also using Transformer as decoder) performs the worst in most cases, which reveals the advantage of the proposed method of using Transformer as decoder.

3.4 Discussion and Conclusions

This study proposed CTranS, a hybrid segmentation framework that integrates convolutional neural networks (CNNs) with Transformer components. By leveraging the strengths of both architectures—namely, the local

feature extraction capabilities of CNNs and the global context modelling provided by attention mechanisms, CTranS effectively captures long-range dependencies across multi-resolution feature representations. Comprehensive evaluations of CTranS were conducted across five publicly available medical image segmentation datasets, covering diverse tasks. The performance was compared against pure CNN models, pure Transformer-based models, and other CNN-Transformer hybrid approaches in terms of segmentation accuracy, model complexity, and training efficiency.

The experimental results demonstrate that CTranS achieves the best overall performance in terms of both Dice Coefficient (DC) and Average Symmetric Surface Distance (ASSD), as detailed in Section 3.3.4. Notably, CTranS attains this accuracy while maintaining the lowest number of learnable parameters and exhibiting fast convergence during training. Qualitative analysis on the CVC-ClinicDB dataset further illustrates the superiority of CTranS, particularly in challenging scenarios where the target structures share similar intensities and patterns with the background. Nonetheless, the model exhibits limitations in handling noisy regions, where it tends to produce false positives, as shown in Figure. 3.4(a). Addressing this limitation remains an important direction for future work.

In conclusion, unlike existing Transformer-based approaches, CTranS employs a CNN as the encoder to extract multi-resolution feature representations, which are subsequently utilised by the Transformer-based decoder. This design enables the decoder to effectively learn from local features at varying spatial scales (i.e., both large and small image patches), providing rich contextual information for accurate segmentation. The use of multi-scale features enhances the model’s capacity to generalise, even when trained on limited datasets. As a result, CTranS does not rely on large-scale pretraining to achieve competitive or superior performance. Furthermore,

the architecture supports flexible input sizes, making it more adaptable to different medical imaging scenarios.

For the 3D application, two adjacent 2D slices, together with the current slice, can currently be used as three input channels of the 2D input to generate the prediction for the current slice. In future work, the attention mechanism employed in CTranS will be enhanced, and the model will be extended to 3D applications using volumetric inputs while maintaining computational efficiency.

Chapter 4

MO-CTranS: A Unified Multi-Organ Segmentation Model Learning from Multiple Heterogeneously Labelled Datasets

4.1 Introduction

While Chapter 3 successfully established a foundation model for medical image segmentation, generalisability, which is a significant problem in medical image segmentation, has not been explored yet. Real-world clinical data rarely conforms to the ideal of large, fully-annotated datasets. Instead, the data landscape is often fragmented, consisting of multiple smaller datasets collected from different sources, with each dataset focusing on the annotation of a distinct subset of organs or anatomical structures.

As discussed in the Literature Review (Section 2.2), this reality presents two major challenges for developing a single, unified segmentation model that can learn from such diverse data: “label conflict” and “data imbalance.” Label conflict challenge during cross-dataset training refers to the problem that arises when multiple datasets use incompatible, ambiguous, or inconsistent label definitions. For example, an unlabelled organ in one dataset might be implicitly considered background, but is explicitly a foreground object in another. This problem is particularly severe with imbalanced datasets, where models tend to overfit to the most common labels and perform poorly on under-represented ones, leading to performance inconsistency among different organs. The intuitive solution of training individual models for each dataset is not only computationally costly and time-consuming but also fails to leverage common anatomical and imaging information shared across these datasets.

Existing solutions to this problem, as previously reviewed in Section 2.2, have attempted to address these issues through various methods, but each introduces its own set of limitations. Approaches focused on network design, such as multi-head networks like Med3D (Chen et al., 2019), suffer from high computational costs and poor scalability, as every new dataset or task requires additional dedicated layers. Methods relying on pseudo mask generation are often computationally expensive and prone to error propagation, where inaccuracies in the initial pseudo-labels can be amplified during subsequent training stages. There is a clear need for a simpler, lighter, and more unified model architecture that can effectively differentiate tasks, handle label conflicts, and leverage shared information without such trade-offs.

To overcome these limitations, this chapter introduces MO-CTranS (Multi-Organ Convolution-Transformer Network), a unified model capable of learn-

ing from multiple heterogeneously labelled datasets simultaneously. Building on the hybrid design principles established in CTranS, MO-CTranS incorporates specialised strategies for cross-dataset learning under incomplete annotations. This directly responds to the second research question (RQ2) of this thesis: How can a unified medical image segmentation model be developed to enhance generalisability in the scenario where multiple heterogeneous and partially labelled datasets are provided?

To this end, this study proposes MO-CTranS, a unified multi-organ segmentation model that builds upon the CNN-Transformer architecture proposed in Chapter 3, by introducing a simple yet powerful innovation: task-specific tokens. The key novelty of the proposed method lies in the insertion of a task-specific token into the Transformer-based decoder, which enables effective differentiation between various organ segmentation tasks and facilitates the handling of label discrepancies. This approach allows a single model to learn from multiple datasets simultaneously without requiring separate decoder heads, task-specific layers, or complex multi-stage pseudo mask generation. MO-CTranS thus provides a lightweight and scalable solution to a major generalisability challenge in medical imaging, offering a significant improvement over existing state-of-the-art (SOTA) solutions.

The proposed method was evaluated against SOTA approaches on four partially labelled local abdominal MRI datasets. The experimental results demonstrate that MO-CTranS achieves superior performance in most cases, with statistical significance.

4.2 Methodology

4.2.1 Problem Definition and Model Overview

It is assumed that multiple datasets are collected from different sources, resulting in heterogeneous label definitions, and that their acquisition views (e.g., axial, coronal, or sagittal) may also differ. The k^{th} dataset is denoted as $D_k = (x_i^k, y_i^k)$, where x_i^k is the i^{th} image sample and y_i^k is its corresponding label, with $y_i^k \in \mathcal{Y}_k \subset \mathcal{Y}$. Here, $\mathcal{Y}$ denotes the total label set, where the same anatomical structure imaged in different views is considered as distinct labels. The objective is to train a unified model $f(x_i^k, \theta)$ using all datasets D_k . At inference time, given an input image x , the model should be capable of predicting all labels in $\mathcal{Y}^{(v)}$, where v denotes the acquisition view of the current input.

As illustrated in Figure. 4.1, MO-CTranS consists of three main components: a CNN-based encoder (CE), a Pyramid Feature Fusion module (PFF), and a Transformer-based decoder (TD). The CE extracts multi-scale features from the input images, the PFF downsamples and reshapes these features into image tokens, and the TD captures long-range dependencies among the tokens. For training, a composite loss function defined as the average of soft Dice loss (Milletari et al., 2016) and cross-entropy loss (Bengio et al., 2017) is adopted.

4.2.2 Task-Specific Token

The CNN-based encoder and the pyramid feature fusion module follow the same design as CTranS, which has been explained in Chapter 3. However, different from CTranS, this study introduces a task-specific token in the

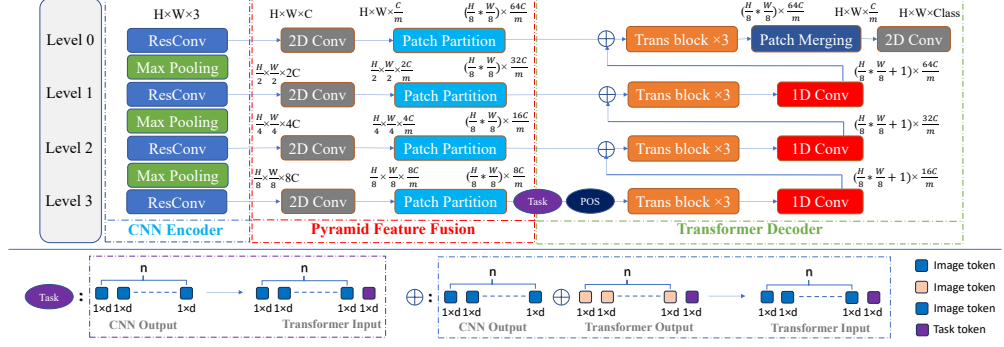


Figure 4.1: Overview of MO-CTrans. * indicates value multiplication, and $x \times y$ means the feature dimension of x “by” y .

Transformer-based decoder to distinguish different segmentation tasks.

At the deepest level, the output image tokens of size $n \times d$ from the PFF are augmented with a 1D task-specific token of size $1 \times d$, as illustrated in Figure. 4.1, where n denotes the number of image tokens and d represents the token dimension. The resulting $n + 1$ tokens are then combined with positional embeddings of size $(n + 1) \times d$ before being passed into the Transformer blocks. The task token is predefined according to the task index (e.g., filled with 1 for task 1) and remains fixed during model training. Notably, the task token is only inserted once in the deepest level then integrated with the image tokens by the attention operations in the Transformer blocks. Currently, for 2D slices acquired from different views, the same organ segmentation is considered as different tasks for different views.

By introducing the task token, the image tokens are able to interact with it through the self-attention layers of the Transformer blocks, thereby encoding task-specific contextual information. As shown in Figure. 4.1, the outputs of the PFF are only fused with the updated image tokens from the Transformer blocks, while the task token itself remains unchanged. Consequently, the output image tokens can further interact with multi-resolution CNN features, reducing information loss across scales. Finally,

only the image tokens are reshaped back into spatial feature maps using Patch Merging (the inverse of Patch Partition) and processed with a 2D convolution layer to generate the segmentation output. In this way, the model can accommodate multiple tasks by simply providing different task tokens as prompts.

4.3 Experiments and Results

4.3.1 Datasets and Evaluation Metrics

For method evaluation, this study constructed a new dataset collected at the Sir Peter Mansfield Imaging Centre, University of Nottingham, UK.

The datasets, denoted as D1–D4, consist of multi-slice balanced turbo field echo (bTFE) localizer MRIs, each containing approximately 35 slices with a resolution of $1.75 \times 1.75 \times 7mm^3$. Specifically, D1 (18 subjects) was acquired in the axial view for liver segmentation; D2 (17 subjects) was acquired in the axial view for both liver and spleen segmentation; D3 (20 subjects) was acquired in the coronal view for spleen segmentation; and D4 (100 subjects) was acquired in the coronal view for kidney segmentation. Label conflicts occur in datasets acquired in the same view. Furthermore, D4 includes a substantially larger number of subjects (100) compared to D3 (20), enabling investigation into whether dataset imbalance exacerbates the label conflict problem during training.

It is also worth noting that ground-truth annotations for the spleen are available in D1; however, these labels were used only for evaluation and not for training. This setup allows us to assess the model’s ability to segment unlabelled organs.

In evaluation, all predictions were resized back to the original image size, and the mean Dice coefficient (DC) and average symmetric surface distance (ASSD) were reported. Especially, ASSD is only calculated on the non-empty predictions, the number of empty predictions is shown in Table 4.1. The Wilcoxon Signed Rank Test (WSRT) was used to assess the statistical significance between the results of two methods. The number of learnable parameters (NP (million)), GPU memory consumption (GM (GB)), and training time per step (TS (seconds)) are reported to demonstrate the model’s complexity and training efficiency.

4.3.2 Experimental Design and Parameter Settings

MO-CTranS without the task token is used as the baseline model (**Base**). For fair comparison, instead of directly benchmarking against SOTA models, the core ideas of representative SOTA solutions on top of **Base** were re-implemented. Four segmentation tasks (denoted as $\mathcal{T}$) were considered in the experiments: axial liver, axial spleen, coronal spleen, and coronal kidney segmentation. The compared methods were implemented as follows:

1. *Base_{single}*: Three binary-class **Base** models (for D1, D3, and D4) and one multi-class **Base** model (for D2) were trained separately on the four datasets.
2. *Base_{multi}*: A single **Base** model was trained jointly on the mixture of all four datasets. To mitigate label conflicts, the label space was standardised to five classes: four foreground classes corresponding to $\mathcal{T}$ and one background class.
3. *Dec_{multi}*: Following Med3D (Chen et al., 2019), a single **Base** model

with separate decoders was trained for each task on the mixture of all datasets.

4. *Dec_{dynam}*: Inspired by TransDoDNet (Xie et al., 2023), a dynamic head was integrated into **Base**. Specifically, the 2D Conv layer in the decoder of **Base** was replaced with dynamic convolutional layers. In addition, a learnable task-specific embedding ($4 \times d$) was introduced and processed by a task-image cross-attention module, where it interacted with the output image tokens ($n \times d$) from the Transformer block at the deepest level of **Base**. The task embedding was further refined by an FFN, and each task token was then used to generate the weights and biases for a dynamic segmentation head corresponding to each task. A single *Dec_{dynam}* model was trained on the mixture of all four datasets.

It should be noted that during the training of the proposed method, *Dec_{multi}*, and *Dec_{dynam}*, each sample was ensured to correspond to a single task in $\mathcal{T}$ by dividing the multi-class dataset (D2) into multiple binary-class datasets. Furthermore, when training on the mixture of datasets, subjects from the larger dataset (D4) were assigned smaller weights in the loss calculation to alleviate the effects of data imbalance.

All models were implemented in a 2D setting, where three consecutive slices were used as input and the prediction was generated for the middle slice in a sliding-window fashion. All images were resampled to 256×256 , and only slices with non-empty labels were included in training and evaluation. Model performance was assessed using 5-fold cross-validation.

The following parameters of all models were the same: $C_{base} = 64$, $m = 4$, and the number of levels was 4. All models were trained on a 12GB GeForce GTX 1080 Ti GPU with Adam optimiser with a learning rate of 0.0001 for

Table 4.1: Results of method comparison. **Bold** indicates the best performer, and † indicates statistical significance compared to the bold results ($p < 0.05$ measured by WSRT). The value in brackets for ASSD is the number of empty predictions.

	D1 (Axial)		D2 (Axial)		D3 (Coronal)	D4 (Coronal)
	liver	spleen	liver	spleen	spleen	kidney
	DC					
<i>Base_{single}</i>	†0.832	/	†0.885	†0.831	†0.757	†0.877
<i>Base_{multi}</i>	†0.855	†0.374	†0.909	†0.359	†0.257	†0.768
<i>Dec_{multi}</i>	†0.868	†0.882	†0.921	†0.887	†0.799	†0.883
<i>Dec_{dynam}</i>	0.875	†0.890	0.923	0.891	†0.805	†0.876
Ours	0.875	0.903	0.923	0.894	0.817	0.883
	ASSD					
	†3.244(26)	/	†3.598(6)	†3.468(6)	†7.884(0)	†2.979(13)
	†2.782(20)	†4.195(131)	†2.378(3)	†3.844(149)	†10.168(152)	†5.671(69)
<i>Base_{single}</i>	2.597(15)	†3.201(0)	1.970(3)	†1.902(0)	†4.604(6)	3.128(6)
<i>Base_{multi}</i>	†2.815(12)	†3.817(0)	†2.513(0)	†2.172(1)	†5.415(0)	†3.270(2)
<i>Dec_{multi}</i>	2.255(19)	2.845(0)	1.809(3)	1.862(0)	3.205(10)	2.966(7)
<i>Dec_{dynam}</i>	Complexity & training efficiency					
Ours	NP		GM		TS	
<i>Base_{single}</i>	40.64		4.27		0.30	
<i>Base_{multi}</i>	40.64		4.29		0.32	
<i>Dec_{multi}</i>	143.17		5.63		0.90	
<i>Dec_{dynam}</i>	40.86		4.28		0.25	
Ours	40.64		4.40		0.30	

100 epochs. The model with the best performance on the validation set was saved for testing.

4.3.3 Comparison to SOTA Methods

The results of the 5-fold cross-validation for different methods are reported in Table 4.1. First, MO-CTranS consistently outperforms the baseline solutions *Base_{single}* and *Base_{multi}* across all datasets, with statistical significance. Notably, *Base_{multi}* fails to effectively address the data imbalance problem (e.g., Dice = 0.257 on D3), as it predominantly focuses on predicting the dominant class (i.e., kidney in D4). In contrast, MO-CTranS successfully mitigates the label conflict issue (e.g., spleen in D1 and D2), which substantially affects the performance of *Base_{multi}*. Second, when compared to the SOTA-inspired solutions *Dec_{multi}* and *Dec_{dynam}*, MO-CTranS achieves significantly better performance in most cases. The per-

formance gap is particularly pronounced for spleen segmentation in D1 and D3, highlighting the superior ability of MO-CTranS to leverage complementary information from multiple datasets—both for predicting unlabelled classes (spleen in D1) and for improving performance on relatively small datasets (spleen in D3).

Finally, the complexity and training efficiency results, also shown in Table 4.1, indicate that MO-CTranS and Dec_{dynam} maintain substantially lighter architectures compared to Dec_{multi} , with nearly identical complexity to a single **Base**.

Figure. 4.2 presents qualitative examples that highlight the advantages of MO-CTranS compared to the baseline models and SOTA-inspired solutions. As shown, MO-CTranS is capable of predicting all organs learned from different datasets within the same view—an ability that none of the baseline models can achieve. For instance, in D3, the $Base_{multi}$ model suffers from label conflict, particularly under dataset imbalance, incorrectly predicting kidney instead of spleen. Although both Dec_{multi} and Dec_{dynam} are able to segment unlabelled organs, their overall performance remains inferior to that of the proposed method.

4.4 Discussion and Conclusions

This chapter has introduced MO-CTranS, a novel model designed to tackle the significant generalisability challenges that arise from using multiple heterogeneously and partially labelled datasets in medical image segmentation. By building upon the foundational CNN-Transformer architecture of CTranS, MO-CTranS successfully demonstrates that a single, unified model can effectively learn from diverse data sources by introducing a sim-

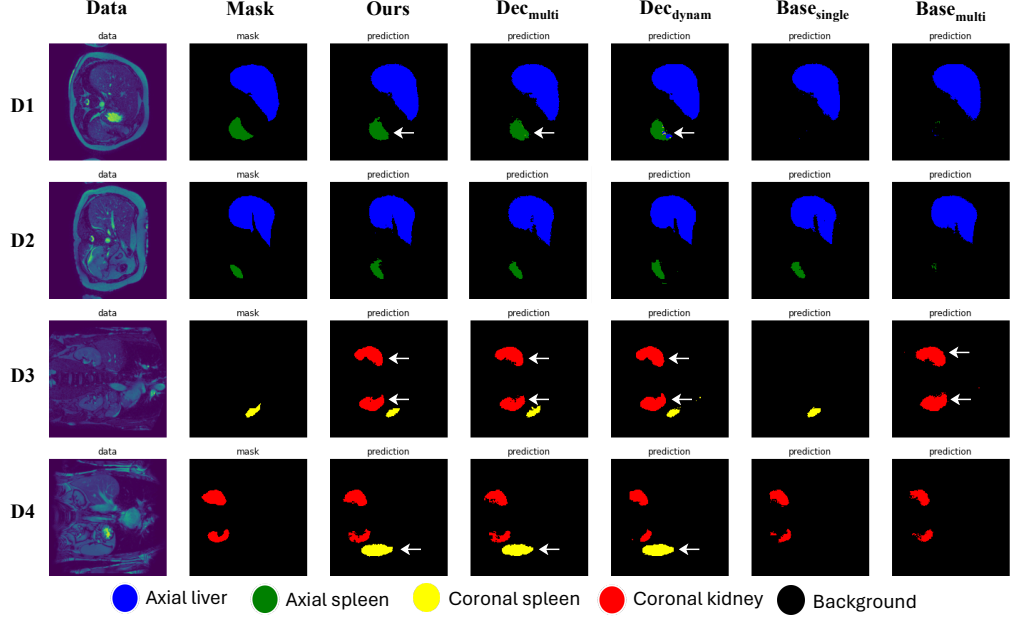


Figure 4.2: Qualitative examples of different methods. The arrows denote the targets with no training annotations.

ple yet powerful innovation: task-specific tokens.

The methodology of MO-CTranS enables a single network to differentiate between various segmentation tasks and gracefully handle label discrepancies, such as when an organ is considered background in one dataset but a foreground object in another. The experimental evaluation against several state-of-the-art methods on a complex, multi-centre, and imbalanced MRI dataset confirmed the efficacy of this approach. The results showed that MO-CTranS achieved superior segmentation accuracy and, importantly, was able to robustly handle label conflicts and data imbalance, an area where other unified models often struggle. Crucially, MO-CTranS accomplishes this with a lightweight network, avoiding the computational overhead and poor scalability of multi-head architectures or the risk of error propagation inherent in multi-stage pseudo mask generation pipelines.

While the experimental evaluation of MO-CTranS focuses on four datasets, the proposed architecture is explicitly designed to scale to a substantially

larger number of tasks. The key design choice enabling this scalability is the use of task-specific tokens rather than task-specific decoder heads or dataset-dependent architectural branches. Adding a new task in MO-CTranS only requires introducing an additional token embedding, while the encoder and decoder parameters remain unchanged. Consequently, the parameter growth is linear in the number of tasks only at the embedding level, which is negligible compared to the overall model size. Additionally, as the number of tasks increases, shared anatomical knowledge is expected to become richer rather than diluted, particularly for under-represented organs, which can benefit from cross-task feature sharing. A potential challenge when the number of datasets increase is the task categorisation problem (e.g, same organ in the same view but from different scanners), the generalisation capability of MO-CTranS to cross-domain settings will be our future research focus.

This work represents a significant step in the progression of this thesis. Chapter 3 established a strong, supervised baseline with CTranS, demonstrating a synergistic blend of CNN and Transformer architectures for accurate segmentation. Now, with MO-CTranS, this architectural paradigm has been extended to solve a major generalisability problem in real-world clinical data. Moving from the baseline model that can only be trained separately on each dataset, the cross-dataset knowledge sharing of MO-CTranS enables unified training on multiple partially labelled datasets, enhancing the feasibility of clinical deployment. However, generalisability alone does not fully address the broader challenges of medical image segmentation. In real-world scenarios, the efficiency of model deployment is another key challenge. To tackle this challenge, the next chapter introduces CRFTrans, a novel recursive vision Transformer layer designed to reduce the computational cost of the SOTA Transformer-based models without

losing segmentation accuracy.

Chapter 5

CRFTrans: A Recursive Vision Transformer Layer Inspired by Conditional Random Field for Medical Image Segmentation

5.1 Introduction

While Chapter 4 successfully extend the foundation model introduced in Chapter 3 for cross-dataset training on multiple partially labelled datasets, which explores the generalisability issues of the medical segmentation models, this chapter continues to push the boundaries of deep learning for medical image segmentation by addressing a critical challenge: the computational inefficiency of Transformer-based models.

As discussed in the Literature Review (Section 2.1.3), most Transformer-based models still rely on the conventional approach of stacking multiple

identical layers (e.g., 12 layers in a ViT-Base) to achieve robust long-range dependency modelling. This design leads to “depth-wise over-parameterisation”, where the redundancy of cascaded layers contributes to a significant number of learnable parameters and a high computational burden. The computational complexity of self-attention scales quadratically with the input size ($O(n^2)$), which is particularly problematic for high-resolution 2D and 3D medical images, where the number of tokens is often very large. Consequently, these models can be resource-intensive and impractical for deployment in clinical settings with limited computational power, such as on-premise servers or edge devices. This persistent issue presents a major bottleneck for the widespread clinical adoption of this technology.

The work here is a direct continuation of the previous contributions, building upon the effectiveness of the CNN-Transformer hybrid model while specifically targeting the inefficiency of its Transformer component. This chapter responds to the third research question (RQ3) of this thesis: How can the efficiency of Transformer-based medical image segmentation models be improved in the scenario where the computational resource is limited in clinical applications?

As explained in Section 2.1.3, existing efforts to simplify Transformers mainly focus on spatial complexity reduction and recurrence for layer simplification. This study extends this line of research by introducing CRF-Trans, a CRF-inspired recursive Transformer layer, but it differs from the common recurrence in two key aspects. First, unlike the generic weight-sharing strategies employed in UT and SReT, which repeatedly apply the same Transformer block across iterations, CRFTrans fixes the query (Q) and key (K) representations while updating only the value (V). This mechanism closely resembles the mean-field message passing process in conditional random fields (CRFs). Supported by the theory of CRF, it results in

more precise and stable boundary refinement. Second, compared to task-specific recurrent Vision Transformers (e.g., RViTs), CRFTrans is plug-and-play for static medical images and can be seamlessly integrated into standard encoder-decoder segmentation architectures without task-specific customisation.

Based on four Transformer-based methods, by comparing the results on five different segmentation tasks, including polyp segmentation, skin lesion segmentation, abdominal organs segmentation, cardiac segmentation, and vessel segmentation respectively, this study demonstrates the redundancy and the over-parameterisation of deep layer-stacking in the classical Transformer model for the image segmentation task. The contributions are summarised below.

1. This study is the first to explore the relationship between the mean-field approximation in Conditional Random Field (CRF) theory and the design of Transformer operations.
2. A novel model, CRFTrans, is proposed by reformulating the mean-field approximation as a single recursive Transformer layer, which replaces cascaded Transformer layers and significantly simplifies the Transformer blocks without sacrificing segmentation accuracy.

5.2 Methodology

Both fully-connected CRF and Transformer are able to learn the long-range pairwise relationships for a given input. In the fully-connected CRF, all cliques are connected with each other, and most studies applied pre-defined functions to model the pairwise relationships. Similarly, in the Transformer

layer, the pairwise relationships of all tokens are modelled by the attention layers. The following sections explain how CRF theory connects with the Transformer by introducing both architectures, followed by introducing our proposed CRFTrans layer.

5.2.1 Conditional Random Field

For an image $\mathbf{I} = \{I_1, \dots, I_N\}$ and labels $\mathbf{X} = \{X_1, \dots, X_N\}$ (N is the number of pixels in $\mathbf{I}$) over classes $\mathcal{L}$, a CRF models the Gibbs energy (Lafferty et al., 2001):

$$E(x) = \sum_i \Phi_u(x_i) + \sum_{i \neq j} \Phi_p(x_i, x_j), \quad (5.1)$$

where Φ_u is the unary energy calculated on individual pixel independently, measuring the cost of labelling pixel i as x_i . Φ_p is the pairwise energy calculated between pairs of different pixels, measuring the cost of labelling i and j as x_i and x_j simultaneously. Intuitively, the unary term measures how likely a pixel belongs to a particular class based on local evidence, while the pairwise term encourages spatial consistency by penalising unlikely label assignments between pairs of pixels. The balance between these two terms allows CRFs to combine local appearance cues with global contextual coherence, which is particularly important for delineating object boundaries in medical images. A common fully connected form, initially proposed by Krähenbühl and Koltun (2011), uses a mixture of Gaussian kernels:

$$\Phi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^M \omega^{(m)} k_G^{(m)}(f_i, f_j), \quad (5.2)$$

with features f_i (e.g., intensity and position), Gaussian kernels k_G , weights $\omega^{(m)}$, and label compatibility function $\mu(\cdot, \cdot)$.

Direct optimisation of the fully connected CRF energy is computationally intractable due to the dense pairwise connections. Therefore, mean-field approximation is commonly used to approximate the true posterior distribution by an iterative update of a simpler, factorised distribution (Krähenbühl and Koltun, 2011). Each iteration refines the label probability of a pixel by aggregating information from all other pixels, weighted by their similarity. By minimizing the KL-divergence $D(Q||P)$, the approximate distribution Q can be used to represent the target distribution,

$$P(X) = \frac{1}{Z} \tilde{P}(X) = \frac{1}{Z} \exp\left(\sum_i \Phi_u(x_i) + \sum_{i \neq j} \Phi_p(x_i, x_j)\right) \quad (5.3)$$

where $Z = \sum_x \tilde{P}(X)$ is the partition function. Derived by Krähenbühl and Koltun (2011), the following iterative update equation is proposed:

$$Q(x_i = l) = \frac{1}{Z_i} \exp\left\{-\Phi_u(x_i) - \sum_{j \neq i} E_{\theta_j \sim Q_j}[\Phi_p(x_i, \theta_j)]\right\} \quad (5.4)$$

$$= \frac{1}{Z_i} \exp\left\{-\Phi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^M \omega^{(m)} \sum_{j \neq i} k_G^{(m)}(f_i, f_j) Q_j(l')\right\} \quad (5.5)$$

where l and l' indicate the labels defined in the label space $\mathcal{L}$.

Algorithm 1 shows the implementation of Eqn. (5.5). The algorithm contains an initialisation step to obtain the initial output probability map from the unary term, and iterative refinement steps to improve the initial probability map. As a result, each update step can be interpreted as a message-passing process in which pixels exchange information with all others based on feature similarity, followed by a compatibility transformation in label space.

Algorithm 1 Mean-field inference process in fully-connected CRFs (Zheng et al., 2015)

```

 $Q_i(l) \leftarrow \frac{1}{Z_i} \exp(U_i(l))$  for all  $i$  ▷ Initialisation
while not converged do
     $\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(f_i, f_j) Q_j(l)$  for all  $m$  ▷ Message Passing
     $\check{Q}_i(l) \leftarrow \sum_m \omega^{(m)} \tilde{Q}_i^{(m)}(l)$  ▷ Weighting Filter Outputs
     $\hat{Q}_i(l) \leftarrow \sum_{l' \in \mathcal{L}} \mu(l, l') \check{Q}_i(l')$  ▷ Compatibility Transform
     $\check{Q}_i(l) \leftarrow U_i(l) - \hat{Q}_i(l)$  ▷ Adding Unary Potentials
     $Q_i(l) \leftarrow \frac{1}{Z_i} \exp(\check{Q}_i(l))$  ▷ Normalizing
end while
    
```

5.2.2 Transformer

The Transformer structure was briefly introduced in Chapter 3, and this section explains it with more details, illustrating the connection with CRFs. As shown in Figure. 5.1 (a), the Transformer block consists of a multi-head self-attention mechanism (MHA) and a position-wise fully connected feed-forward network (FFN). Additionally, a normalisation layer (Norm) is applied before both the MHA and FFN, while a dropout layer (DropOut) follows each of them. A skip connection is included between the input and the output of the DropOut layer to support stable gradient flow and improve training dynamics. The MHA mechanism enables hierarchical context integration, which is critical for medical image segmentation, by employing H parallel attention heads.

As illustrated in Figure. 5.1 (b), $Norm(X_{TransIn})$ is projected into Q (Query), K (Key), and V (Value) $\in \mathbb{R}^{N \times d_{model}}$, where Q denotes the query tokens (distinct from $Q_i(l)$ in Section 5.2.1), N is the number of image tokens (derived from image patches), and d_{model} is the embedding dimension (i.e., the feature length of each patch). Each head h computes the scaled

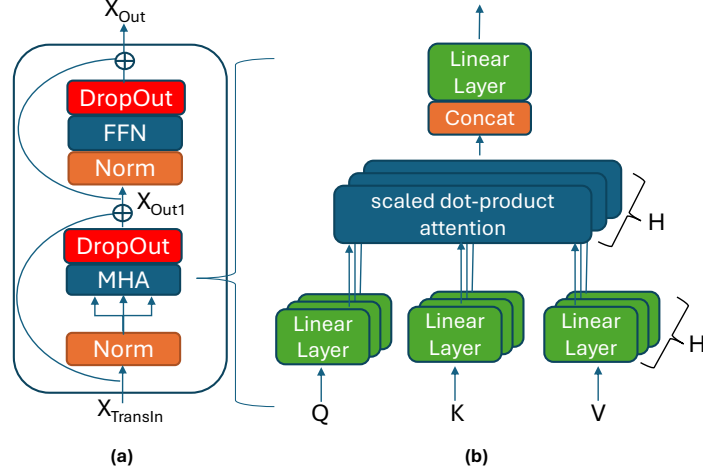


Figure 5.1: The network architecture of a Transformer layer, copied from Section 3.2.3 for convenience.

dot-product attention (SDPA) in parallel:

$$P_h = \text{Softmax} \left(\frac{QW_h^Q (KW_h^K)^T}{\sqrt{d_k}} \right) \quad (5.6)$$

$$head_h = P_h(VW_h^V) \quad (5.7)$$

Here, P_h represents the pairwise attention weights in head h ; $W_h^Q, W_h^K \in \mathbb{R}^{d_{model} \times d_k}$, and $W_h^V \in \mathbb{R}^{d_{model} \times d_v}$ are learnable projection matrices, where $d_v = d_k = d_{model}/H$ in self-attention. By updating V using the attention scores computed from Q and K , SDPA allows each token to selectively attend to others within the sequence, effectively capturing spatial dependencies regardless of position.

The outputs of all H attention heads are then concatenated and passed through a linear projection to obtain the final MHA output:

$$MHA(Q, K, V) = \text{Concat}(head_1, head_2, \dots, head_H)W^O \quad (5.8)$$

where $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$ is a learnable projection matrix that maps the concatenated features into a unified anatomical embedding space. The

Concat operation denotes feature-wise concatenation along the channel dimension.

In addition to capturing global contextual relationships through MHA, the feed-forward network (FFN) provides non-linear feature transformation, which is essential for distinguishing subtle patterns and enhancing the representation of each token:

$$FFN(X) = ReLU(XW_1 + b_1)W_2 + b_2 \quad (5.9)$$

Here, $ReLU(X) = \max(0, X)$ is an activation function that introduces non-linearity into the network while maintaining computational efficiency. The weight matrices $W_1 \in \mathbb{R}^{d_{model} \times Ld_{model}}$ and $W_2 \in \mathbb{R}^{Ld_{model} \times d_{model}}$ form an expansion-contraction structure, where the expansion ratio $L > 1$ increases the feature capacity before projecting back to the original embedding dimension.

5.2.3 Relationship of CRF and Transformer

Both MF inference in CRF and the Transformer block perform conceptually similar two-stage transformations, as illustrated in Figure. 5.2. In MF, these correspond to spatial and channel-wise refinements, which can be closely mapped to the MHA and FFN operations in a Transformer layer.

Stage 1: Spatial-wise transformation. According to the MF inference steps in Algorithm 1, the first stage performs a spatial-wise linear transformation that captures long-range contextual dependencies. It combines the *Message Passing (MP)* and *Weighting Filter Outputs (WFO)* operations, where the updated probability $Q_i(l)$ is computed as the weighted sum of

all other probabilities $Q_{j \neq i}(l)$ within each class channel c_l :

$$Q_i(l) \leftarrow \sum_{j \neq i} W_{i,j} Q_j(l),$$

where the weight $W_{i,j}$ in Figure. 5.2 (a) corresponds to the pairwise term $k_G(f_i, f_j)$ based on intensity and spatial distance. This weight remains constant across channels. In this stage, each pixel is refined according to all other pixels within the same class channel. In vector form shown in Figure. 5.2 (b), each feature vector is updated as a weighted sum of all other vectors, integrating global spatial context.

Stage 2: Channel-wise transformation. The second stage performs channel-wise linear transformation to refine contextual information in the label space. It corresponds to the *Compatibility Transform* step in MF,

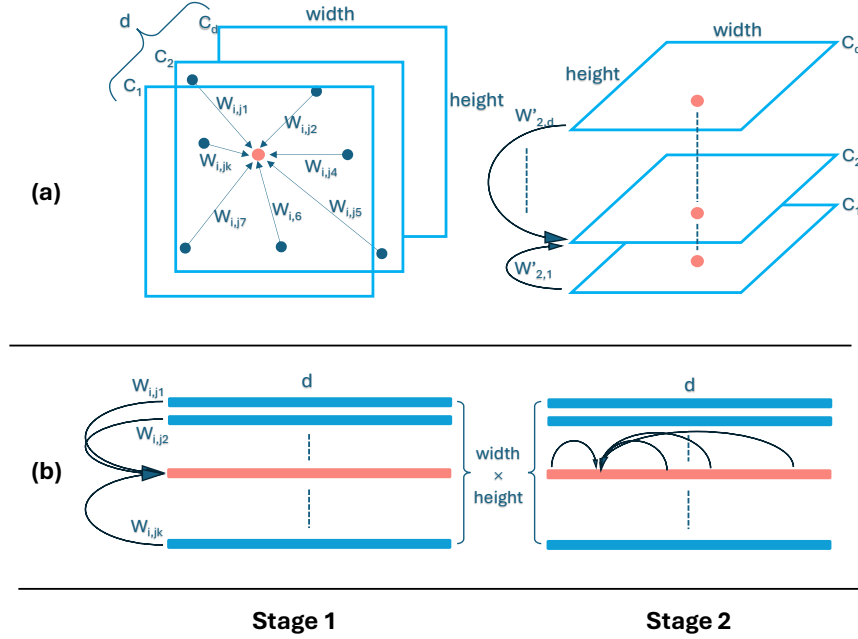


Figure 5.2: Key transform steps in mean-field inference. (a) is the 2D feature map view, and (b) is the 1D feature vectors view. Arrows denote ‘contribute to’. In stage 1, W is represented by the pairwise term; In stage 2, W' is represented by the compatibility transform function.

which linearly mixes feature responses across channels:

$$Q_i(l) \leftarrow \sum_{l' \in \mathcal{L}} W'_{l,l'} Q_i(l'),$$

where $W'_{l,l'}$ in Figure. 5.2 (a) denotes the label compatibility function $\mu(l, l')$. This transformation captures inter-class relationships, refining each pixel's representation using the responses of all channels at the same spatial location.

After the two transformations, a residual connection is added between the unary term and the updated probability map to obtain the refined unary potential $\check{Q}_i(l)$. The refined map then serves as input for the next iteration until convergence, forming an iterative contextual refinement process.

Transformer analogy. In a Transformer layer, SDPA in MHA performs a similar spatial refinement to Stage 1. Each token $v'_i \in \mathbb{R}^{1 \times d_{model}}$ is updated as a weighted sum of all tokens v'_j :

$$v'_i \leftarrow \sum_j P_{i,j} v'_j,$$

where $P_{i,j}$ is the learned attention weight computed from Q' and K' (shown in Eqn. (5.6)). Interpreting each feature vector in Figure. 5.2(b) as an image token, the SDPA operation corresponds directly to the spatial aggregation step in MF inference. Likewise, the FFN in a Transformer performs channel-wise transformation analogous to Stage 2, where each token is refined through linear projections in the embedding dimension d_{model} , analogous to label-space mixing in CRF.

Thus, the structural correspondence of CRF and Transformer can be concluded below. In the CRF view, the feature map in Figure. 5.2(a) repre-

sents a probability map with d channels (number of classes), whereas in the Transformer view, d corresponds to the embedding dimension d_{model} and the number of tokens is $N = width \times height$. A fully connected CRF models pairwise relationships among all pixels, while a Transformer limits connectivity to patch tokens. Hence, the sparsity of dependencies in a Transformer is determined by the chosen patch or window size.

Furthermore, in the CRF perspective, $W_{i,j}$ in the Stage 1 of Figure. 5.2 is represented by $k_G(f_i, f_j)$, while in the view of the Transformer, $W_{i,j}$ is represented by the pairwise weight P in Eqn. (5.6). Additionally, in the view of Transformer, $W'_{i,j}$ in the stage 2 is represented by the layer parameters in FFN rather than the compatibility function $\mu(l_i, l_j)$.

Compared with MF inference, MHA and FFN offer several advantages. In Stage 1, the pairwise term in CRF relies on predefined Gaussian kernels based on intensity and position, which restricts its ability to model complex semantic relationships. MHA instead learns pairwise affinities in high-dimensional feature space, providing richer and more adaptive similarity measures. The multi-head mechanism further enables the model to capture diverse relation patterns across feature subspaces. In Stage 2, while CRF employs a single linear compatibility transformation, the FFN adopts a two-layer expansion–contraction structure with nonlinear activation, substantially enhancing representation power. Overall, the Transformer layer generalises the MF process, offering superior flexibility and expressiveness for modelling long-range and cross-channel dependencies.

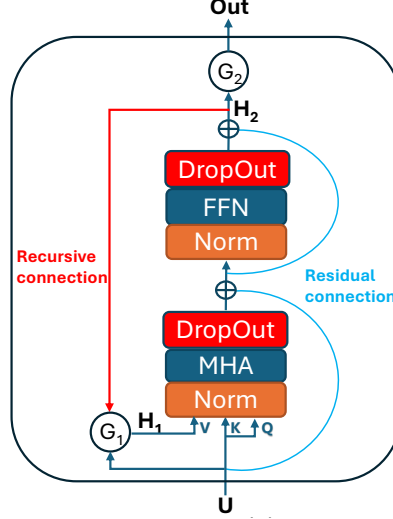


Figure 5.3: The network architecture of a CRFTrans layer. U is the input unary potential. The recursive process is expressed by H_1 and H_2 , indicating the hidden states in an RNN layer, and G_1 and G_2 , indicating the gating functions.

5.2.4 CRFTrans

Building on the theoretical equivalence established above, this study introduces CRFTrans—a recursive Transformer layer that reformulates the mean-field (MF) inference process as iterative refinement of hidden feature representations, as illustrated in Figure. 5.3. Unlike traditional CRF-MF approaches, which iteratively update segmentation probability maps as a post-processing step, CRFTrans focuses on refining latent feature representations within the model architecture itself. This design enables end-to-end feature learning during training, shifting the role of MF from output refinement to integrated representation learning.

As shown in Figure. 5.3, the key difference from a normal Transformer block shown in Figure. 5.1 is the recursive connection between the output and input, resulting in an RNN manner expressed by the following

Algorithm 2 CRFTrans-Recursive Transformer Layer

```

1:  $V_i(l) \leftarrow U_i(l)$  for all  $i$  ▷ Initialisation
2: while  $t < T$  do
3:    $\tilde{V}_i(l) \leftarrow \text{Dropout} \left( MHA \begin{pmatrix} \text{Norm}(U_i(l)), \\ \text{Norm}(U_i(l)), \\ \text{Norm}(V_i(l)) \end{pmatrix} \right)$  ▷ MP+WFO
4:    $\check{V}_i(l) \leftarrow U_i(l) + \tilde{V}_i(l)$  ▷ Residual connect
5:    $\hat{V}_i(l) \leftarrow \text{Dropout}(FFN(\text{Norm}(\check{V}_i(l))))$  ▷ CompTrans
6:    $V_i(l) \leftarrow \check{V}_i(l) + \hat{V}_i(l)$  ▷ Residual connect
7: end while
    
```

equations:

$$H_1(t) = \begin{cases} U, & \text{if } t = 0, \\ H_2(t-1), & \text{if } 0 < t \leq T, \end{cases} \quad (5.10)$$

$$H_2(t) = f_\theta(U, H_1(t)), \quad (5.11)$$

$$Out(t) = \begin{cases} 0, & \text{if } 0 \leq t < T, \\ H_2(t), & \text{if } t = T. \end{cases} \quad (5.12)$$

where $t = 0, 1, \dots, T$ is the iteration index, and f_θ represents the operation of one iteration in CRFTrans (θ denotes the parameters in CRFTrans). G_1 and G_2 shown in Figure. 5.3 are expressed as function (5.10) and (5.12). Especially, CRFTrans becomes a single Transformer layer when $T = 0$ (one forward pass without recursive process).

The detailed procedure of CRFTrans is presented in Algorithm 2. Inspired by the mean-field (MF) formulation, CRFTrans employs a fixed pairwise term by keeping Q and K (derived from the unary term) constant during the recursive updates. This design ensures that image token pairs with higher attention scores are progressively pulled closer in the feature space throughout the recursive refinement process.

In contrast to conventional CRF-MF, which only connects the output of each iteration back to its input, CRFTrans retains the internal residual connections characteristic of Transformer blocks, as illustrated in Figure. 5.3. These residual paths help stabilise training by mitigating the risk of vanishing gradients during iterative updates. Furthermore, the recursive linkage to the initial input U distinguishes CRFTrans from traditional cascaded Transformer layers. Instead of generating entirely new representations at each layer, CRFTrans reuses the initial embedding U across iterations, enabling progressive refinement of features while preserving original tissue-specific signatures and integrating global contextual information.

Thus, in each forward pass of an iteration, CRFTrans processes and aggregates three distinct hidden states: $H^{(0)}$, $H^{(MHA)}$, and $H^{(FFN)}$, which are combined as follows:

$$X_{out} = H^{(0)} + H^{(MHA)} + H^{(FFN)} \quad (5.13)$$

where $H^{(0)} = U$, $H^{(MHA)} = \tilde{V}$, and $H^{(FFN)} = \hat{V}$, as defined in Algorithm 2.

5.2.5 CRFTrans Plug and Play

CRFTrans is designed as a single-layer Transformer module, making it straightforward to replace conventional cascaded Transformer layers with CRFTrans in existing architectures. To demonstrate the simplicity and flexibility of this integration, the previously proposed method, CTranS (explained in Chapter 3, was used as an example. The original cascaded Transformer layers in CTranS, which are represented by red blocks in Figure. 5.4, can be directly replaced by CRFTrans without altering the input

or output shapes. This is possible because both CRFTrans and standard Transformer blocks share similar structural components, including multi-head self-attention (MHA), feed-forward networks (FFN), and normalisation layers (Norm), and can operate on the same input token dimensions.

Owing to its architectural compatibility, CRFTrans can be seamlessly integrated into a variety of model backbones by substituting existing Transformer blocks. This chapter evaluates and compares the performance of these configurations in the following section.

5.3 Experiments and Results

In the experiments, the performance of conventional cascaded Transformer layers (referred to as Trans-stack) was compared with CRFTrans across four state-of-the-art (SOTA) Transformer-based medical image segmentation models, evaluated on five diverse medical segmentation tasks. To assess the generalisability and effectiveness of integrating CRFTrans into different architectural configurations, representative SOTA models that utilise Transformer layers in distinct positions within their pipelines were selected. Specifically, three hybrid models were included, including FAT-Net (Wu et al., 2022), which incorporates Trans-stack in the encoder; TransUNet

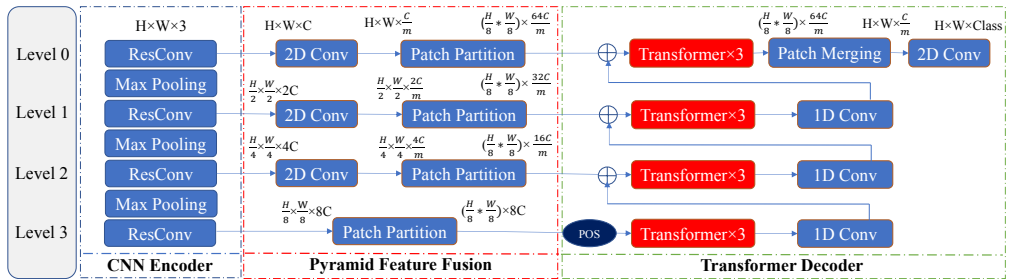


Figure 5.4: The network architecture of CTranS. The red blocks are the cascaded Transformer layers that can be replaced directly by CRFTrans.

(Chen et al., 2021b), which employs Trans-stack as a bridging module between encoder and decoder; and CTranS (Gong et al., 2024b), which places Transformer layers in the decoder.

In addition to these hybrid CNN-Transformer models, two further baselines for comparison were included, which are ResUNet, a pure CNN-based architecture, and SwinUNet (Cao et al., 2021), a model composed entirely of Transformer layers without CNN components. This comprehensive evaluation setup allows us to systematically examine the performance benefits and adaptability of CRFTrans across a range of model structures and medical imaging tasks.

Moreover, the number of Transformer layers within each model varies depending on the specific architectural design. Let l_{Tr} denote the number of Trans-stack modules, and n_{Tr} represent the number of Transformer layers within a single Trans-stack. The total number of Transformer layers in a model is then given by $l_{Tr} \times n_{Tr}$. According to this formulation, FAT-Net and TransUNet each contain 1×12 Transformer layers, CTranS includes 4×3 , and SwinUNet comprises 7×2 layers. To ensure consistency across all evaluations, the replacement procedure described in Section 5.2.5 was consistently applied to substitute each model’s original Trans-stack with the proposed CRFTrans module.

5.3.1 Datasets and Evaluation Metrics

The following five public datasets were used for evaluation.

(1) Skin lesion segmentation (**ISIC**): ISIC2018 (Codella et al., 2019) (binary-class), which was published by the International Skin Imaging Collaboration challenge, containing 2594 RGB dermoscopy images with different

image sizes.

(2) Polyp segmentation (**CVC**): CVC-ClinicDB (Bernal et al., 2015) (binary-class) contains $612\,384 \times 288$ RGB colonoscopy images from 31 colonoscopy sequences.

(3) Abdominal organs segmentation (**CHAOS**): CHAOS challenge dataset (Kavur et al., 2019) (multi-class), which aims to segment abdominal organs from CT and MRI data. The T2-SPIR sequence of the MRI data was used, including annotations for the liver, spleen, and right and left kidneys, comprising 20 subjects with 3D volumes of varying sizes.

(4) Cardiac segmentation (**ACDC**): Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018), which was created from real clinical exams acquired at the University Hospital of Dijon, containing 150 patients (3D data in different volume sizes) with the labels of left ventricle (LV), the myocardium (MYO), and the right ventricle (RV).

(5) Vessel Segmentation (**FIVES**): FIVES (Jin et al., 2022) dataset consists of 800 high-resolution RGB fundus images with pixel-wise manual annotation on the vessels.

All models evaluated in this study are 2D-based. For 3D volumetric applications, A sliding-window strategy was adopted, in which three consecutive slices were input to the model, and the prediction was generated for the middle slice. In the experiments, all input images were resized to 224×224 , in accordance with the default resolution used in most compared methods. An exception was made for the fundus dataset, where images were resized to 512×512 , as aggressive downsampling in vessel segmentation leads to substantial information loss compared to other segmentation tasks.

Model performance was assessed using 5-fold cross-validation. All models

were trained using the Adam optimiser on a single 11GB GeForce GTX 2080 Ti GPU. For high-resolution inputs (512×512), training was conducted on a 48GB NVIDIA RTX A6000 GPU. The learning rate was set to 0.0001, and each model was trained for 150 epochs. During training, the model achieving the lowest validation loss was selected for testing.

For evaluation, all predicted segmentation masks were resized back to their original image dimensions. The mean Dice Coefficient (DC) and Average Symmetric Surface Distance (ASSD) were used to measure segmentation accuracy. To assess statistical significance between competing methods, the Wilcoxon Signed Rank Test (WSRT) was employed. Additionally, model complexity and training efficiency were quantified using the number of learnable parameters (NP, in millions), GPU memory consumption (GM, in GB), and training time per step (TS, in seconds).

5.3.2 Parameter Settings

CTranS with CRFTrans (as described in Section 5.2.5) trained on the CVC-ClinicDB dataset is used as a case study for parameter tuning and ablation analysis. Specifically, the effect of varying the number of iterations ($T + 1$) in CRFTrans is examined, where $T = 0$ corresponds to a standard single-layer Transformer without recursive refinement. The segmentation accuracy results for different values of T are presented in Table 5.1.

As shown in Table 5.1, the configuration with $T = 2$ consistently yields better performance across all datasets in terms of the Dice coefficient, with several results demonstrating statistical significance. For the ASSD metric, $T = 2$ also performs favorably in most cases. In the few instances where $T = 2$ does not outperform other configurations, the differences are not

statistically significant. Based on these observations, $T = 2$ is adopted as the default setting for CRFTrans in all subsequent model comparison experiments.

Importantly, comparing $T = 0$ and $T = 2$ reveals that $T = 0$ leads to significantly lower performance in most scenarios, underscoring the value of recursive refinement. Although using a single Transformer layer may limit the model’s capacity compared to cascaded architectures, the iterative updates inspired by CRF inference substantially enhance the segmentation capability of the model.

Table 5.1: Results of using different iteration times, expressed by $T + 1$. +/- indicates better/worse average performer than $T = 2$ (**Bold** ones), and $\dagger$ indicates statistical significance ($p < 0.05$ measured by WSRT).

	CVC	ISIC	CHAOS	ACDC	FIVES
	Dice				
T=0	89.06 ⁻ $\dagger$	88.33 ⁻ $\dagger$	86.08 ⁻ $\dagger$	89.04 ⁻	84.48 ⁻
T=2	89.79	88.68	88.31	89.68	84.84
T=4	89.27 ⁻	88.43 ⁻	87.63 ⁻	87.44 ⁻ $\dagger$	83.95 ⁻ $\dagger$
T=6	89.64 ⁻	88.52 ⁻	86.45 ⁻	89.16 ⁻	84.51 ⁻ $\dagger$
	ASSD				
T=0	6.22 ⁻ $\dagger$	5.06 ⁻	3.98 ⁻	0.51 ⁻	16.74 ⁻ $\dagger$
T=2	5.93	5.04	3.68	0.50	14.43
T=4	6.03 ⁻	5.12 ⁻	4.29 ⁻	1.90 ⁻ $\dagger$	20.24 ⁻ $\dagger$
T=6	5.87 ⁺	5.05 ⁻	3.59 ⁺	0.49 ⁺	17.10 ⁻

Furthermore, when comparing $T = 0$ with higher iteration settings such as $T = 4$ and $T = 6$, the results indicate that performance may degrade as the number of iterations increases in certain cases. This suggests that additional recursive steps do not necessarily lead to improved segmentation performance and may even have a detrimental effect in some scenarios. Increasing the number of iterations results in further refinement of the Transformer output ($H_2(t)$) based on similarity scores computed from Q and K . While insufficient refinement may limit the model’s capacity to capture complex contextual information, excessive refinement can lead to over-reliance on the representations of Q and K —which, in this case,

originate from the CNN encoder—thus constraining the model’s ability to adapt to diverse features.

5.3.3 Ablation Study

The ablation study first evaluates the impact of the internal skip connections between the MHA and FFN modules by designing a variant called **CRFResidual**, which adopts a CRF-style residual structure. As shown in Figure 5.5(a), this variant removes the internal residual connections and instead uses a single residual connection from the final output to the initial input. Specifically, the two residual connections (line 4 and line 6) in Algorithm 2 are replaced with a single update rule: $V_i(l) \leftarrow U_i(l) - \hat{V}_i(l)$, applied after the Compatibility Transform step, as detailed in Algorithm 3.

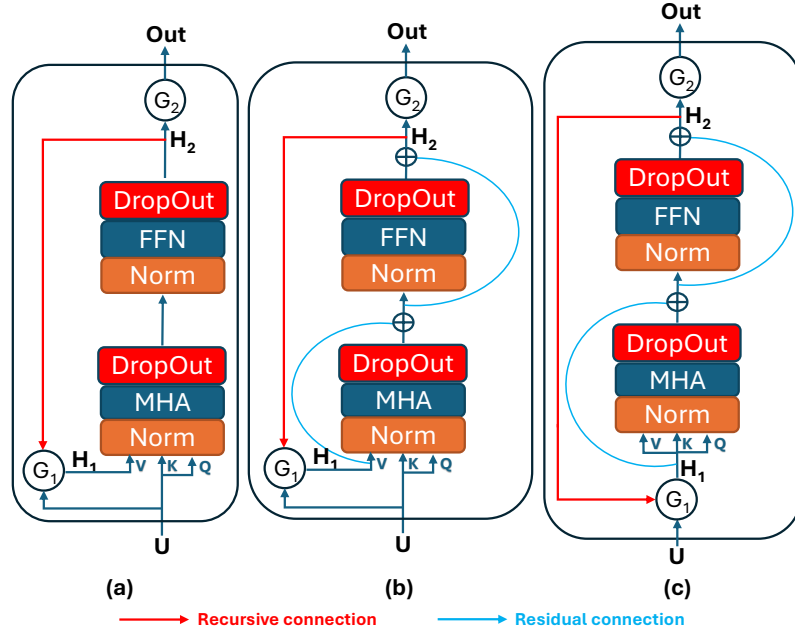


Figure 5.5: The network architecture of CRFResidual (a), TransResidual (b), and gRecur-Trans (c).

To assess the effect of maintaining access to the original input features

throughout the recursive process, another variant **TransResidual** is designed, illustrated in Figure. 5.5(b). This version mimics the residual structure of a standard Transformer by removing the connection to the initial input $U_i(l)$ during recursion. Concretely, the update rule in Algorithm 2 is modified from $\tilde{V}_i(l) \leftarrow U_i(l) + \tilde{V}_i(l)$ to $\tilde{V}_i(l) \leftarrow V_i(l) + \tilde{V}_i(l)$. The segmentation performance of these ablation variants is presented in Table 5.2.

Last, this section also compares CRFTrans with the general Recursive Transformer (denoted as **gRecur-Trans** shown in Figure. 5.5 (c)), similar to the block used in UT (Dehghani et al., 2018) and SReT (Shen et al., 2022a) to explore whether fixing Q and K during the iterative refinement effectively helps. The results are shown in the Table. 5.2.

Algorithm 3 CRFResidual

- 1: $V_i(l) \leftarrow U_i(l)$ for all i ▷ Initialisation
 - 2: **while** $t < T$ **do**
 - 3: $\tilde{V}_i(l) \leftarrow \text{Dropout} \left(MHA \left(\begin{matrix} Norm(U_i(l)), \\ Norm(U_i(l)), \\ Norm(V_i(l)) \end{matrix} \right) \right)$ ▷ MP+WFO
 - 4: $\hat{V}_i(l) \leftarrow \text{Dropout}(FFN(Norm(\tilde{V}_i(l))))$ ▷ CompTrans
 - 5: $V_i(l) \leftarrow U_i(l) - \hat{V}_i(l)$ ▷ Residual connect
 - 6: **end while**
-

As shown in Table 5.2, in most cases, CRFTrans has better mean perfor-

Table 5.2: Results of ablation study. +/- indicates better/worse average performer than original setting (**Bold** ones), and † indicates statistical significance ($p < 0.05$ measured by WSRT).

	CVC	ISIC	CHAOS	ACDC	FIVES
	Dice				
CRFTrans	89.79	88.68	88.31	89.68	84.84
CRFResidual	88.85 ^{-†}	88.44 ^{-†}	86.71 ^{-†}	88.87 ⁻	84.21 ^{-†}
TransResidual	89.27 ⁻	88.66 ⁻	87.77 ⁻	89.58 ⁻	85.03 ^{+†}
gRecur-Trans	89.42 ⁻	88.86 ⁺	88.00 ⁻	89.35 ⁻	84.02 ^{-†}
	ASSD				
CRFTrans	5.93	5.04	3.68	0.50	14.43
CRFResidual	7.17 ^{-†}	4.99 ^{+†}	4.45 ⁻	0.61 ⁻	17.01 ⁻
TransResidual	6.40 ⁻	4.99 ⁺	3.67 ⁺	0.51 ⁻	16.34 ^{-†}
gRecur-Trans	8.38 ^{-†}	5.16 ^{-†}	4.37 ⁻	0.45 ⁺	24.88 ^{-†}

mance than the Recursive Transformer, with more statistical significance in the ASSD measure, demonstrating the advantage of CRFTrans in sharper boundary detection compared with generic weight sharing. Additionally, CRFResidual performs significantly worse than CRFTrans in most cases, indicating that the removal of internal residual connections adversely affects model performance. In contrast, TransResidual shows no statistically significant difference compared to CRFTrans, suggesting that omitting the connection to the initial input in each recursive step has a limited impact on segmentation accuracy.

These findings demonstrate that the internal skip connections applied to both the $MP + WFO$ and CompTrans stages provide meaningful improvements over the CRF-style residual connection alone, as used in CRFResidual. On the other hand, the presence or absence of access to the original input features in each recursive iteration does not significantly influence performance in most scenarios.

Nevertheless, the full CRFTrans configuration—combining both CRF-style and Transformer-style residual connections—is adopted in all subsequent experiments. This choice is motivated by its overall superior average performance and the observed stability it provides during training, particularly in settings requiring a greater number of recursive iterations.

5.3.4 Comparison to SOTA Methods

After determining the optimal model settings, this section compares the performance of CRFTrans with that of the conventional Trans-stack. For a fair comparison, CRFTrans is configured to use the same parameter settings as a single layer within the Trans-stack (e.g, the hidden dimension

and number of heads were kept identical). Additionally, although the original implementations of FAT-Net, TransUNet, and SwinUNet employ pre-trained Transformer blocks by default, all layers were trained from scratch in this study to ensure consistency and fairness in comparison with CRFTrans.

The segmentation performance, measured by Dice coefficient and Average Symmetric Surface Distance (ASSD), is presented in Table 5.3 and Table 5.4, respectively. The results show that, across the three CNN–Transformer hybrid models (FAT-Net, TransUNet, and CTranS), there is no statistically significant difference between Trans-stack and CRFTrans in terms of Dice scores and in most cases of ASSD. This indicates that CRFTrans can effectively simplify hybrid architectures without compromising segmentation accuracy, regardless of the position where the Transformer is applied within the model pipeline.

Notably, CRFTrans achieves statistically significant improvements in vessel segmentation performance on the FIVES dataset across all hybrid models, with respect to the Dice score (and in some cases, ASSD). This suggests that for more challenging segmentation tasks involving small or thin structures, the recursive refinement mechanism in CRFTrans—guided by token similarity—can offer a meaningful advantage over cascaded Transformer layers.

However, as shown in Table 5.3 and Table 5.4, CRFTrans does not perform effectively when integrated into SwinUNet, exhibiting decreased performance in many cases. This result highlights the inapplicability of CRFTrans to purely Transformer-based models. The effectiveness of recursive refinement in CRFTrans is highly dependent on the quality of the Q and K representations, as the similarity matrix computed from these terms

Table 5.3: Dice of segmentation of different models. $\star$ indicates no statistical significance compared to CRFTrans ($p > 0.05$ measured by WSRT). $+/-$ indicates better or worse performance for the results that have a statistically significant difference.

	CVC	ISIC	CHAOS	ACDC	FIVES
	ResUNet				
	76.07	76.37	84.95	88.35	86.59
	FAT-Net				
Trans-stack	88.07 $\star$	87.17 $\star$	87.03 $\star$	89.84 $\star$	87.52
CRFTrans	89.20	87.29	86.35	89.10	88.53 ⁺
	TransUNet				
Trans-stack	86.19 $\star$	87.99 $\star$	85.89 $\star$	88.69 $\star$	87.91
CRFTrans	85.65	88.08	86.17	88.00	89.64 ⁺
	SwinUNet				
Trans-stack	75.99	87.15	76.53 $\star$	82.96	86.14
CRFTrans	73.73 ⁻	85.69 ⁻	74.37	80.97 ⁻	85.16 ⁻
	CTranS				
Trans-stack	90.01 $\star$	89.07 $\star$	88.51 $\star$	89.68 $\star$	84.98 $\star$
CRFTrans	89.79	88.68	88.31	89.68	84.84

guides the refinement process. Therefore, a reliable unary term is essential for CRFTrans to function properly.

In hybrid models such as TransUNet and CTranS, this unary term is generated by a CNN-based encoder, which provides robust feature representations to support the refinement process in CRFTrans. Although FAT-Net

Table 5.4: ASSD of segmentation of different models. $\star$ indicates no statistical significance compared to CRFTrans ($p > 0.05$ measured by WSRT). $+/-$ indicates better or worse performance for the results that have a statistically significant difference.

	CVC	ISIC	CHAOS	ACDC	FIVES
	ResUNet				
	18.99	12.13	1.65	0.53	16.05
	FAT-Net				
Trans-stack	7.24 $\star$	5.80 $\star$	3.30 $\star$	0.31 $\star$	16.30
CRFTrans	7.46	5.87	3.29	0.38	17.94 ⁻
	TransUNet				
Trans-stack	11.43 $\star$	5.53 $\star$	3.23 $\star$	0.39 $\star$	14.12
CRFTrans	10.61	5.74	1.25	0.40	15.83 ⁻
	SwinUNet				
Trans-stack	17.24 $\star$	5.80	2.11 $\star$	1.63	9.94
CRFTrans	14.87	6.46 ⁻	4.13	0.85 ⁺	7.82 ⁺
	CTranS				
Trans-stack	5.82 $\star$	4.72 $\star$	3.72 $\star$	0.47 $\star$	14.88
CRFTrans	5.93	5.04	3.68	0.50	14.43 ⁺

applies CRFTrans within the encoder, its dual-encoder architecture mitigates over-reliance on a single representation source, thereby preserving performance. In contrast, when CRFTrans is used in a pure Transformer architecture such as SwinUNet, the output of each CRFTrans layer becomes the sole source for subsequent Q and K representations. This compounding dependency on the shallower CRFTrans layers limits the model’s capacity to learn rich and discriminative features, thereby diminishing overall performance compared to Trans-stack.

Moreover, comparing the baseline models—pure CNN (ResUNet) and pure Transformer (SwinUNet)—with the hybrid models reveals that the latter generally achieve superior performance. This underscores the advantage of combining CNNs and Transformers: CNNs excel at capturing local, high-level spatial features, while Transformers effectively model long-range dependencies. Together, these complementary strengths result in more accurate and robust segmentation across diverse tasks.

Table 5.5: Results of model complexity comparison. NP stands for number of parameters (in millions), GM stands for GPU memory usage (in GB), and TS stands for training time per step (in seconds).

	NP	GM	TS
	ResUNet		
	13.04	1.42	0.03
	FAT-Net		
Trans-stack	26.98	1.03	0.14
CRFTrans	23.72	0.90	0.05
	TransUNet		
Trans-stack	105.28	2.46	0.13
CRFTrans	27.31	1.11	0.07
	SwinUNet		
Trans-stack	27.17	0.90	0.09
CRFTrans	15.41	0.91	0.08
	CTranS		
Trans-stack	21.48	2.14	0.09
CRFTrans	11.75	2.07	0.06

Last but not least, the model complexity, GPU memory consumption, and training efficiency—measured by the number of parameters (NP), GPU

memory usage (GM), and training time per step (TS)—are summarised in Table 5.5. The results show that CRFTrans significantly simplifies the models across all settings, achieving substantial reductions in both parameter count and memory usage. Specifically, CRFTrans reduces the total number of Transformer layers from $l_{Tr} \times n_{Tr}$ to l_{Tr} by replacing each Trans-stack with a single CRFTrans layer.

As the computational complexity is significantly influenced by the number of stacks in Trans-stack, the memory consumption and training speed are primarily influenced by the number of recursive operations in CRFTrans. Theoretically, when the number of iterations is equal to the number of stacks, although CRFTrans still includes significantly less parameters, CRFTrans and Trans-stack would have similar computational complexity. In the experiments, the number of operations per CRFTrans is set to 3 ($T = 2$), which is considerably fewer than that used in FAT-Net and TransUNet (12 Transformer layers). For CTranS and SwinUNet, the number of operations per Trans-stack is 3 and 2, respectively—comparable to the number used in CRFTrans. Consequently, the efficiency improvements in terms of memory and speed are especially pronounced in FAT-Net and TransUNet.

For practical applications involving the replacement of Trans-stack with CRFTrans, it is recommended to start with $T = 0$ and incrementally increase it only as needed—until CRFTrans achieves segmentation performance comparable to or better than Trans-stack. This strategy ensures maximal model simplification, as increasing T leads to decreased efficiency and, as shown in the experiments, does not necessarily yield performance gains.

5.4 Discussion and Conclusions

This chapter has introduced CRFTrans, a novel recursive Vision Transformer layer, specifically designed to address the critical problem of computational inefficiency and depth-wise over-parameterisation in conventional Transformer-based models for medical image segmentation. Building on the theoretical relationship between the self-attention mechanism and the iterative refinement process of the mean-field approximation approach in CRFs, this chapter proposed a single, recursive Transformer layer to replace the deep stack of cascaded layers typically found in existing architectures.

The methodology hinges on reformulating the self-attention mechanism as a pairwise potential optimisation process, which allows a single CRFTrans layer to iteratively refine feature representations in an effective and computationally lightweight manner. Experimental evaluations, conducted by plugging CRFTrans into four SOTA Transformer-based models across five different medical segmentation tasks, demonstrated the efficacy of CRFTrans. The results show that CRFTrans-based hybrid models can achieve comparable or even superior segmentation accuracy to their cascaded-layer counterparts, particularly in complex tasks like vessel segmentation, where recursive refinements can outperform a long stack of layers.

Crucially, CRFTrans significantly reduces model complexity, leading to substantially fewer learnable parameters, lower memory consumption, and faster training speeds, thereby making powerful Transformer-based models more feasible for deployment in resource-constrained clinical environments. However, a core limitation of the method was also identified: its performance is highly dependent on the quality of the initial features (“unary terms”) provided by a CNN-based encoder. This dependence renders CRFTrans inapplicable for use as an encoder and causes its performance to be

influenced by the effectiveness of the underlying feature extractor.

In summary, this work provides a new, fundamental component for building lightweight yet powerful segmentation models. By simplifying the Transformer architecture without sacrificing performance, CRFTrans represents a significant step toward developing more efficient and clinically deployable deep learning solutions. However, current experiments focus on the plug-and-play in the 2D-based networks, feasibility of replacing 3D blocks (e.g., in UNETR (Hatamizadeh et al., 2022)) with CRFTrans will be evaluated in future work. Moreover, the broader applicability of CRFTrans beyond segmentation will be explored, including its extension to tasks such as image classification, to further investigate the generalisability of this recursive paradigm as a backbone for various vision tasks.

From CTranS, MO-CTranS, to CRFTrans, despite their collective contributions, these three methods share a main limitation - the dependence on a fully-supervised setting, requiring substantial labelled data for effective training. Even MO-CTranS, though capable of leveraging partially labelled datasets, still depends heavily on the availability of labelled organs within each dataset. In real-world clinical applications, however, annotation remains a major bottleneck, none of these models fully capitalise on the vast quantities of unlabelled data that are readily available in medical imaging archives, which represent an untapped source of information for improving efficiency and robustness.

To further realise efficient, scalable, and broadly generalisable medical image segmentation systems, it is essential to move toward annotation-efficient learning frameworks that integrate semi-supervised and few-shot learning strategies. The next chapter introduces SSL-MedSAM2, a framework that unites semi-supervised learning principles with the prompting

capabilities of SAM2, representing the next step in this thesis toward highly efficient and generalisable medical image segmentation.

Chapter 6

SSL-MedSAM2: A Semi-supervised Medical Image Segmentation Framework Powered by Few-shot Learning of SAM2

6.1 Introduction

As discussed in Section 2.4 of the Literature Review, semi-supervised learning (SSL) has emerged as a powerful paradigm to leverage large collections of unlabelled scans alongside a small labelled subset. Techniques such as pseudo-labelling and consistency regularisation have shown strong potential to extend supervision from annotated to unannotated data, thereby reducing annotation costs. Similarly, few-shot learning approaches aim to generalise segmentation models to new tasks or classes using only a handful

of labelled samples. Recent progress in large foundation models, notably the Segment Anything Model (SAM) and its successor SAM2, has significantly advanced the feasibility of few-shot and prompt-based segmentation in medical imaging by providing powerful prompt-driven segmentation capabilities that align naturally with both SSL and few-shot paradigms. SAM2’s ability to propagate segmentations across sequential frames aligns particularly well with 3D volumetric medical data, where spatial continuity between slices parallels temporal continuity in video.

However, existing SSL approaches face notable limitations. Pseudo-labelling strategies often propagate initial labelling errors, leading to degraded performance, while consistency regularisation methods are highly sensitive to the choice of perturbations. Moreover, most SSL frameworks train models from scratch on small labelled datasets, which restricts their capacity to capture generalisable features early in training. Although SAM- and SAM2-based SSL methods have recently been proposed, they often rely on heuristic prompt generation or domain-specific adaptations, and they cannot fully exploit the training-free few-shot capabilities of foundation models.

To overcome these challenges, this chapter presents SSL-MedSAM2, a novel framework that integrates semi-supervised learning with the few-shot generalisation power of SAM2, adapted to medical imaging through MedSAM2. SSL-MedSAM2 contains two synergistic branches: a training-free few-shot learning branch (TFFS-MedSAM2), which leverages MedSAM2 to generate ensemble pseudo-labels for unlabelled volumetric data without requiring prompts or additional training, and an iterative fully supervised branch (FSL-nnUNet), which refines the pseudo-labels via iterative retraining on both labelled and pseudo-labelled data, ensuring robustness and improved segmentation quality.

The key novelty of SSL-MedSAM2 lies in its integration of training-free pseudo-label generation with iterative refinement, allowing the framework to combine the strengths of large foundation models with the adaptability of task-specific segmentation networks. Unlike prior SSL pipelines, SSL-MedSAM2 does not depend solely on the reliability of an initial lightweight model; instead, it exploits SAM2’s strong zero-/few-shot capabilities to bootstrap pseudo-labels and iteratively refines them through nnU-Net. This design addresses both the annotation bottleneck and the domain generalisation problem, achieving state-of-the-art (SOTA) results on challenging benchmarks such as the MICCAI 2025 CARE-LiSeg challenge (Liu et al., 2025; Gao et al., 2023; Wu and Zhuang, 2023).

In summary, SSL-MedSAM2 directly addresses the fourth research question (RQ4) of this thesis: How can semi-supervised learning be effectively integrated with the few-shot capabilities of large foundation models, enabling robust segmentation with minimally labelled data and truly training-free initial knowledge transfer? Building upon the fully supervised architectures introduced in earlier chapters, SSL-MedSAM2 represents the final step in this thesis toward an annotation-efficient, robust, and clinically practical segmentation framework.

The contributions of this work are summarised as follows:

1. A novel semi-supervised medical image segmentation framework, SSL-MedSAM2, is proposed. To the best of current knowledge, this framework is among the first to integrate a large foundation segmentation model adapted for medical images (MedSAM2) in a training-free manner with a task-specific segmentation network (nnU-Net) within a semi-supervised learning (SSL) pipeline.
2. A novel training-free few-shot learning branch, TFFS-MedSAM2, is

introduced in an ensemble prediction manner. This branch leverages the pre-trained MedSAM2 model for efficient and high-quality pseudo-label generation on unlabelled volumetric data without requiring any user prompts.

In the remainder of this chapter, the technical design and evaluation of the proposed SSL-MedSAM2 framework are presented. First, Section 6.2 details the architectural design and workflow of SSL-MedSAM2, emphasising how the integration of semi-supervised learning principles with the few-shot capabilities of MedSAM2 enhances both efficiency and generalisability. Section 6.3.1 reports experimental results on the MICCAI 2025 CARE-LiSeg challenge, demonstrating the effectiveness of SSL-MedSAM2. Beyond the overall framework, particular attention is devoted to TFFS-MedSAM2, the training-free few-shot segmentation module, whose performance critically influences the final segmentation outcomes. Section 6.3.2 therefore compares the performance of TFFS-MedSAM2 and its variant with the SOTA few-shot medical segmentation method UniverSeg (Butoi* et al., 2023).

6.2 Methodology

6.2.1 Preliminaries

SSL aims to train high-performing models by leveraging a small amount of labelled data together with a large amount of unlabelled data. The SSL task is formulated as follows. Given a dataset $\mathcal{D} = \{I_i\}_{i=1}^N$, each image $I_i \in \mathbb{R}^{H \times W \times C}$ is defined over a continuous spatial domain $\Omega \subset \mathbb{R}^2$. $\mathcal{D}$ is composed of two subsets, $\mathcal{D}^l$ and $\mathcal{D}^u$, such that $\mathcal{D} = \{\mathcal{D}^l, \mathcal{D}^u\}$. The labelled subset $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^M$ consists of M fully annotated cases, while

the unlabelled subset $\mathcal{D}^u = \{x_i^u\}_{i=1}^N$ contains N unannotated cases ($M \ll N$). Here, x_i^l and x_i^u denote the labelled and unlabelled input images, respectively, and y_i^l represents the ground-truth mask corresponding to x_i^l , $x \in \mathbb{R}^{H \times W \times D}$ is defined over a continuous spatial domain $\Omega \subset \mathbb{R}^3$, and $y_i \in \{0, 1, \dots, K\}^{H \times W \times D}$, K denotes the number of foreground classes.

The SSL-MedSAM2 framework aims to generate reliable pseudo masks $\hat{y}_i^u$ for the unlabelled set $\mathcal{D}^u$, thereby constructing a pseudo-labelled dataset $\mathcal{D}_p^u = (x_i^u, \hat{y}_i^u)_{i=1}^N$ using TFFS-MedSAM2. The combined dataset $\mathcal{D}^l, \mathcal{D}_p^u$ is then used for fully supervised training with FSL-nnUNet, which iteratively refines the pseudo labels $\hat{y}_i^u$ to improve segmentation performance.

6.2.2 Overview of SSL-MedSAM2

SSL-MedSAM2 consists of two complementary branches: a few-shot learning branch, TFFS-MedSAM2, built upon the pretrained foundation model MedSAM2 (Ma et al., 2025), and an iterative fully supervised learning branch, FSL-nnUNet, based on the nnUNet framework (Isensee et al., 2021). The overall workflow of SSL-MedSAM2 is outlined in Algorithm 4, where f_θ denotes MedSAM2 and g_θ denotes nnUNet. The first stage—Pseudo Label Initialisation—is carried out by TFFS-MedSAM2, while the subsequent iterative refinement stages are performed by FSL-nnUNet. Specifically, MedSAM2 generates initial pseudo masks for unlabelled data using prompts derived from the labelled subset, and nnUNet iteratively refines these pseudo masks to improve reliability and accuracy. During inference, only the refined nnUNet (g_θ) is employed for segmentation. The details of TFFS-MedSAM2 and FSL-nnUNet are provided in the following sections.

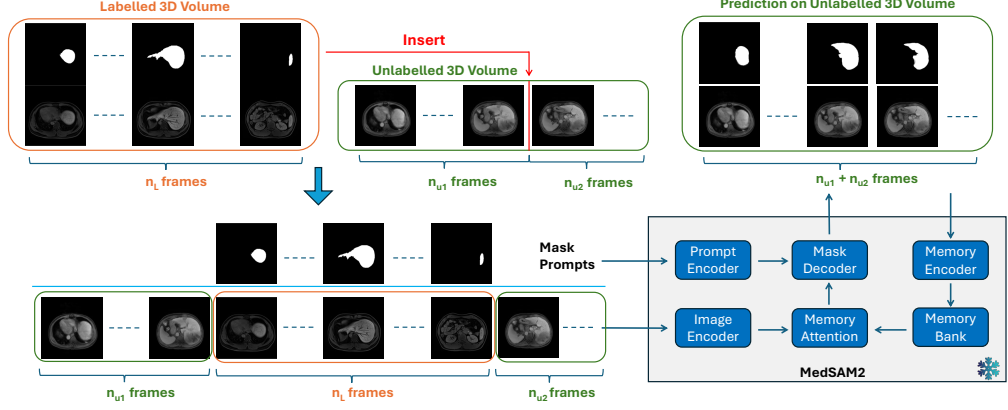


Figure 6.1: The workflow of the composition of labelled data and unlabelled data, and the prediction on the unlabelled data from the composition. The inserting location is denoted by the red line in the unlabelled data. MedSAM2 is frozen in the whole process.

6.2.3 TFFS-MedSAM2

SAM2 employs a Transformer-based backbone augmented with a streaming memory module, enabling efficient processing of frame sequences. Whereas the original SAM supports point, box, and mask prompts for single images, SAM2 extends these capabilities to frame sequences, automatically propagating and refining segmentation masks as objects move or deform (Ravi et al., 2024). The key innovation of SAM2 is a lightweight memory bank that stores embeddings from previous frames, allowing the model to reference historical context when segmenting the current frame. This streaming memory design makes SAM2 particularly suitable for medical imaging, as it naturally propagates segmentations across slices of a 3D volume, thereby

Algorithm 4 Training procedure of SSL-MedSAM2.

$\hat{Y}^u \leftarrow f_{\theta}(X^u \mid \mathcal{D}^l = (X^l, Y^l))$	▷ Pseudo Label Initialisation
while not converge do	
$\mathcal{D} \leftarrow \{\mathcal{D}^l, (X^u, \hat{Y}^u)\}$	▷ Training Set Building
$g_{\theta} \leftarrow \text{train on } \mathcal{D}$	▷ Model Refinement
$\hat{Y}^u \leftarrow g_{\theta}(X^u)$	▷ Pseudo Label Updating
end while	
return Refined model g_{θ}	

reducing the need for slice-by-slice prompting. By exploiting temporal coherence, SAM2 can generate consistent and accurate masks for entire volumes from minimal manual input.

In the proposed framework, the labelled and unlabelled volumes are combined together, and the image slices are treated as continuous frames of a video sequence. Subsequently, the ground truth masks of the labelled slices are used as the mask prompts to guide SAM2 in predicting segmentations for the unlabelled slices. Specifically, MedSAM2 (Ma et al., 2025), a variant fine-tuned on large-scale medical image segmentation tasks, was adopted, in place of the original SAM2, to fully leverage domain-specific prior knowledge encoded in the pretrained foundation model.

As illustrated in Figure. 6.1, the input 3D volume data is treated as a sequence of video frames. To facilitate pseudo-label generation, a fully labelled 3D volume is randomly inserted into an unlabelled volume at arbitrary slice positions. The combined frame sequence is then fed into the pretrained and frozen MedSAM2, where labelled slices provide mask prompts, while unlabelled slices receive no prompts. Leveraging its memory attention mechanism, MedSAM2 stores contextual information from both the labelled prompts and predictions of adjacent slices. This enables the model to recognise similar image features in the slices of the unlabelled volume and propagate consistent predictions through its mask decoder.

The ensemble pseudo-labelling process for each unlabelled volume using TFFS-MedSAM2 is summarised in Algorithm 5. Here, R denotes the number of random insertion locations for each unlabelled volume, x_i^c represents the labelled–unlabelled combined data (purple box in Figure. 6.1), and p_i^u denotes the probability map of the unlabelled input x_i^u , generated by f_θ (MedSAM2). In TFFS-MedSAM2, each of the M labelled volumes is

Algorithm 5 Pseudo-Labeling of unlabelled data x_j^u by TFFS-MedSAM2.

Input: Unlabelled data x_j^u and labelled dataset $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^M$
Output: Pseudo-label $\hat{y}_j^u$ for x_j^u

```

1:  $\hat{p}_j^u \leftarrow 0$  ▷ Zero Initialisation
2: for  $i \leftarrow 1$  to  $M$  do
3:   Randomly generate  $R$  insertion locations on the unlabelled frames
4:   for  $r \leftarrow 1$  to  $R$  do
5:      $\{x_i^c, y_i^l\} \leftarrow \text{Insert } \{x_i^l, y_i^l\} \text{ to } x_j^u \text{ between the frame index } F^{(r)}$ 
       and  $F^{(r)} + 1$ 
6:      $\hat{p}_j^u \leftarrow \hat{p}_j^u + \text{Sigmoid}(f_\theta(x_j^u | \{x_i^c, y_i^l\}))$ 
7:   end for
8: end for
9:  $\hat{p}_j^u \leftarrow \hat{p}_j^u / (R \times M)$  ▷ Get Average
10:  $\hat{y}_j^u \leftarrow \text{threshold}(\hat{p}_j^u)$  ▷ Get Binary Mask

```

inserted into every unlabelled volume at R randomly chosen slice locations. Consequently, $R \times M$ probability maps are generated for each unlabelled sample. The final pseudo mask is obtained by averaging these probability maps, thereby improving robustness through prediction ensembling.

Although the MedSAM2 model is pre-trained, the entire adaption procedure is training-free, and the pseudo-labelling of each unlabelled sample is performed independently, requiring no additional user prompts and no reliance on other unlabelled data. This design yields a robust and efficient few-shot learning capability, effectively exploiting limited labelled data for large-scale pseudo mask generation.

6.2.4 FSL-nnUNet

After generating pseudo masks for the unlabelled dataset using TFFS-MedSAM2, the second branch of SSL-MedSAM2 performs an iterative fully supervised learning process, FSL-nnUNet, based on nnU-Net. This stage refines the pseudo masks $\hat{Y}^u$ by training on the combined dataset $\mathcal{D} = \mathcal{D}^l, (X^u, \hat{Y}^u)$.

To avoid overestimation when updating pseudo labels during inference, a 5-fold cross-validation strategy on the unlabelled dataset is adopted. Specifically, the unlabelled data are partitioned into five folds. In each round of training, one fold of unlabelled data is held out as the validation set, while the remaining four folds, together with the labelled dataset, constitute the training set. After completing all five rounds, predictions from each fold are aggregated, producing updated pseudo masks for the entire unlabelled dataset.

Initially, nnU-Net is trained on a small set of labelled images together with the pseudo labels generated by MedSAM2 for the unlabelled data. Although these pseudo labels may be imperfect, the strong prior knowledge encoded in the foundation model provides a useful initialisation that guides the segmentation network. The newly trained nnU-Net is then used to re-predict segmentation masks for the unlabelled data, producing refined masks that surpass the original MedSAM2 outputs due to the network’s adaptation to the specific task domain. These refined predictions are treated as updated pseudo labels, and nnU-Net is retrained using the improved annotations.

This iterative self-training cycle can be repeated multiple times, progressively improving both pseudo-label quality and segmentation performance. For inference on unseen test data, the final refined models obtained from 5-fold cross-validation are used, and predictions are averaged across the five folds to yield the final segmentation. Importantly, FSL-nnUNet is a flexible framework in which nnU-Net can be readily replaced with other fully supervised medical image segmentation architectures.

6.3 Experiments and Results

This section presents a comprehensive evaluation of the proposed SSL-MedSAM2 framework and its few-shot learning branch TFFS-MedSAM2. The experiments are organised into two parts to systematically assess the framework under different supervision regimes and to highlight the contribution of each component. In the first part, the experiments evaluate SSL-MedSAM2 under a semi-supervised learning (SSL) setting using the MICCAI 2025 CARE-LiSeg dataset (Liu et al., 2025; Gao et al., 2023; Wu and Zhuang, 2023), which contains binary-class segmentation tasks. This setting enables quantitative analysis of segmentation performance when only a small subset of data is annotated. This experiment verifies the effectiveness of combining the foundation model MedSAM2 with the iterative nnU-Net refinement in leveraging both labelled and unlabelled data for robust segmentation.

TFFS-MedSAM2 plays a crucial role in determining the overall performance of SSL-MedSAM2, since the quality of pseudo labels generated by TFFS-MedSAM2 directly affects the accuracy and stability of subsequent iterative refinements in the nnU-Net branch. Poor initial pseudo masks can lead to error propagation and suboptimal convergence during semi-supervised training. Therefore, it is essential to thoroughly evaluate and quantify the independent performance of TFFS-MedSAM2 to understand its contribution to the entire framework.

To this end, in the second part, the experiments further investigate the few-shot learning capability of the TFFS-MedSAM2 branch, focusing on more challenging multi-class segmentation tasks. Specifically, additional experiments are conducted to compare TFFS-MedSAM2 with the SOTA medical image few-shot segmentation method UniverSeg (Butoi* et al., 2023).

UniverSeg is chosen as a baseline because it represents one of the most general and scalable few-shot segmentation frameworks, capable of segmenting previously unseen anatomical structures by learning from diverse support-query pairs through a universal conditioning mechanism. Its strong generalisation ability makes it an appropriate and competitive benchmark for assessing the effectiveness and robustness of TFFS-MedSAM2 under both 1-shot and 5-shot scenarios on the local multi-class dataset.

All experiments were deployed on a GPU instance with about 10 GB of memory from an Nvidia A100 GPU.

6.3.1 Semi-Supervised Learning on the Care-LiSeg Dataset

Dataset and Experimental Settings

The CARE-LiSeg dataset (Liu et al., 2025; Gao et al., 2023; Wu and Zhuang, 2023) consists of T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), and Gadolinium ethoxybenzyl diethylenetriamine pentaacetic acid (Gd-EOB-DTPA)-enhanced dynamic MRI scans. The Gd-EOB-DTPA-enhanced dynamic MRIs include five phases: non-contrast (T1), arterial, venous, delayed, and hepatobiliary (GED4). Within the dataset, only a limited amount of ground-truth annotation is available for GED4 MRI. Accordingly, the Contrast-Enhanced subtask focuses on the segmentation of GED4, while the Non-Contrast subtask focuses on T1, T2WI, and DWI modalities.

In total, the dataset contains 610 patients diagnosed with liver fibrosis, each undergoing multi-phase MRI scans performed on three different scanners:

Table 6.1: The data distribution of the CARE-LiSeg dataset. The value in brackets indicates the amount of annotated data in that subset.

	Vendor A	Vendor B1	Vendor B2	Vendor C
Training	130 (10)	170 (10)	60 (10)	0
Validation	20	20	20	0
Testing	40	40	40	70

Philips Ingenia 3.0T, Siemens Skyra 3.0T, and Siemens Aera 1.5T (denoted Vendor A, B1, and B2, respectively). This design results in a diverse multi-phase, multi-centre dataset. Importantly, unlike the training and validation sets, the testing set also includes cases from an unseen vendor (Vendor C), enabling evaluation of both in-distribution (ID) and out-of-distribution (OOD) performance. A detailed summary of the dataset composition is provided in Table 6.1, which shows that 360 cases are used as training dataset, about 8% of them are manually annotated.

In TFFS-MedSAM2, all input frames were resized to 512×512 to match the default configuration of the pretrained MedSAM2. After inference, the predicted masks were resized back to the original image resolution for pseudo-label generation. In FSL-nnUNet, the training configurations were automatically determined by nnU-Net. Each nnU-Net model was trained for 1000 epochs, and FSL-nnUNet was executed three times for the Contrast-Enhanced subtask and twice for the Non-Contrast subtask. When generating pseudo masks for non-contrast sequences using TFFS-MedSAM2, the labelled GED4 data were still combined with them, despite the difference in MRI sequences. For evaluation, the mean Dice coefficient (DC) and Hausdorff distance (HD) across the test sets are reported.

Results and Discussion

For the Non-Contrast subtask of the CARE 2025 challenge, results were submitted only for the T1 sequence, as the initial pseudo masks gener-

ated by TFSS-MedSAM2 for T2WI and DWI sequences were of insufficient quality for direct use in FSL-nnUNet training. This limitation likely arises from the larger appearance discrepancy between GED4 and T2WI/DWI compared to GED4 and T1, given that T1 and GED4 represent two distinct phases of the same Gd-EOB-DTPA-enhanced acquisition, thus sharing more consistent intensity and structural patterns.

The average Dice coefficient and HD on the validation datasets, as reported on the official leaderboard, are summarised in Table 6.2. For comparison, the performance of the initial pseudo masks generated directly by TFSS-MedSAM2 is also reported, prior to any refinement by FSL-nnUNet. The results on the test datasets are presented in Table 6.3. Compared to the validation results, the testing phase also reported the In-Distribution (ID) and Out-Of-Distribution (OOD) performance separately to evaluate the robustness of the methods on the unseen vendor data. However, due to restrictions on the number of submissions permitted during the testing phase, the results of TFSS-MedSAM2 were not evaluated on the test server.

As shown in Table 6.2, SSL-MedSAM2 ranks third in both Dice and HD for the GED4 validation data, while achieving the best overall performance for the T1 validation data. Notably, the TFSS-MedSAM2 results indicate

Table 6.2: Evaluation results of the validation set in terms of GED4 and T1 reported on the leaderboard. Top 5 methods together with TFSS-MedSAM2 are shown (other methods are represented by their team names).

Method	Dice	HD	Method	Dice	HD
TanglangSled	97.30	31.01	SSL-MedSAM2	95.74	41.14
MHIL	97.05	19.55	TFSS-MedSAM2	95.46	40.65
SSL-MedSAM2	97.04	21.28	LidaYang	94.88	49.10
Monster	97.04	19.28	MHIL	94.65	62.84
InteliLiver	97.00	21.34	Sigma	94.19	59.32
TFSS-MedSAM2	95.92	24.71	CitySJTU	93.81	44.98

(a) GED4.
(b) T1.

that even without the iterative refinement stage, TFFS-MedSAM2 alone can generate high-quality pseudo masks through its training-free few-shot segmentation mechanism — even outperforming other participants on the T1 sequence. This demonstrates the strong initialisation capability of the few-shot branch and its effectiveness in transferring structural priors from limited annotated volumes. Furthermore, the inclusion of FSL-nnUNet refinement substantially improves segmentation accuracy for the GED4 sequence, whereas the gain is less pronounced for T1.

Table 6.3: Evaluation results of the test set in terms of GED4 and T1 reported on the leaderboard. “In-Distribution” contains data from vendor A, B1, B2. “Out-Of-Distribution” contains data from vendor C. Top 5 methods are shown (other methods are represented by their team names).

Method	Dice	HD	Method	Dice	HD
Monster	97.03	21.40	SuperIdols	97.70	14.65
SSL-MedSAM2	96.84	22.97	Monster	97.65	15.17
xcj	96.47	25.03	SSL-MedSAM2	97.54	15.11
BI	96.44	22.63	BI	97.54	14.37
TeamSpace	96.41	24.32	TeamSpace	97.53	15.72
(a) In-Distribution GED4.			(b) Out-Of-Distribution GED4.		
Method	Dice	HD	Method	Dice	HD
SSL-MedSAM2	96.08	19.87	SSL-MedSAM2	97.16	25.58
BIGS2	94.34	38.06	BIGS2	95.48	22.11
Sigma	93.18	58.99	Sigma	95.03	29.04
CitySJTU	92.62	27.80	CitySJTU	94.44	25.54
BioDreamer	91.96	31.83	BioDreamer	94.18	28.24
(c) In-Distribution T1.			(d) Out-Of-Distribution T1.		

On the testing datasets, SSL-MeSAM2 consistently ranks among the top three approaches across all leaderboard categories. As shown in Table 6.3 (a) and (b), SSL-MedSAM2 achieves top-three performance on the GED4 data in both ID and OOD settings with respect to both Dice and HD metrics. More notably, Table 6.3 (c) and (d) show that SSL-MeSAM2 obtains the highest Dice score and lowest HD among all participants for most cases on the T1 testing data, except for the OOD HD metric, where

it ranks the third.

Overall, these results confirm that SSL-MedSAM2 delivers SOTA performance across both validation and testing datasets. In particular, its superior accuracy on the T1 sequence demonstrates excellent knowledge transfer between MRI sequences captured in different acquisition phases of the same contrast agent (Gd-EOB-DTPA-enhanced). Moreover, the model exhibits strong OOD generalisation, often performing better on unseen vendor data than on the training-domain data. This robustness to domain shifts highlights the effectiveness of combining a foundation segmentation model (MedSAM2) with iterative semi-supervised refinement, making SSL-MedSAM2 a promising solution for practical clinical applications where data heterogeneity is inevitable.

In summary, the above results validate the effectiveness and robustness of SSL-MedSAM2 in semi-supervised medical image segmentation. The framework achieves competitive or superior performance across both in-distribution and out-of-distribution evaluations, particularly excelling on the T1 MRI data, where its pseudo-labelling and refinement mechanisms demonstrate excellent domain transfer capability. The strong results of TFFS-MedSAM2 further highlight its capacity to generate high-quality pseudo masks in a training-free few-shot manner, serving as a powerful initialisation for semi-supervised refinement. These findings not only confirm the overall design effectiveness of SSL-MedSAM2 but also motivate a deeper investigation into the few-shot learning capability of the TFFS-MedSAM2 branch under more challenging and data-scarce conditions, as explored in the following section.

6.3.2 Few-Shot Learning on the Local Dataset

In the first part of the experiments on the MICCAI 2025 CARE-LiSeg dataset, the TFFS-MedSAM2 branch was provided with ten labelled volumes for each MRI sequence, enabling the generation of high-quality pseudo masks that effectively guided the semi-supervised learning process. However, real-world clinical applications often face far more constrained annotation conditions, where only a handful of annotated examples are available for new imaging modalities or rare anatomical structures. To evaluate the robustness and adaptability of the proposed framework under such extreme data scarcity, the second part of the experiments focuses on assessing the few-shot learning capability of TFFS-MedSAM2. Specifically, its performance in 1-shot and 5-shot settings is tested, where only one or five labelled volumes are available as support data, respectively. This experiment aims to investigate how effectively TFFS-MedSAM2 can generalise from extremely limited supervision and to explore its potential extension to broader and more challenging clinical scenarios where large-scale annotation is impractical.

Dataset and Experimental Settings

A **Cardiac** dataset comprises 400 late gadolinium enhancement (LGE) 3D MRI scans collected from a single clinical site at Nottingham University Hospital between 2017 and 2021. Three anatomical structures—left ventricular blood pool (LVBP), left ventricular myocardium (LVM), and right ventricular blood pool (RVBP)—were manually annotated by expert radiologists. The images were acquired using either a 1.5T Philips Ingenia or a 1.5T Siemens Aera scanner. The typical spatial resolution was $0.89 \times 0.89 \times 10$ mm with no slice gap for the Philips scans, and $1.7 \times 1.7 \times 8$

Algorithm 6 Pseudo-Labeling of unlabelled data x_j^u by TFFS-MedSAM2*.

Input: Unlabelled data x_j^u and labelled dataset $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^M$

Output: Pseudo-label $\hat{y}_j^u$ for x_j^u

- 1: $\hat{p}_j^u \leftarrow 0$ ▷ Zero Initialisation
 - 2: **for** $i \leftarrow 1$ to M **do**
 - 3: $\{x_i^c, y_i^l\} \leftarrow$ Evenly interleave the slices of the smaller volume into the slices of the larger volume.
 - 4: $\hat{p}_j^u \leftarrow \hat{p}_j^u + \text{Sigmoid}(f_\theta(x_j^u \mid \{x_i^c, y_i^l\}))$
 - 5: **end for**
 - 6: $\hat{p}_j^u \leftarrow \hat{p}_j^u / M$ ▷ Get Average
 - 7: $\hat{y}_j^u \leftarrow \text{threshold}(\hat{p}_j^u)$ ▷ Get Binary Mask
-

mm with a 2 mm inter-slice gap for the Siemens scans.

To evaluate the few-shot segmentation capability of TFFS-MedSAM2, an episodic evaluation strategy is adopted, which is commonly used in few-shot learning to simulate real-world scenarios where only a limited number of annotated samples are available for each task. In this setting, each episode is treated as an independent segmentation task composed of a support set and a query (test) set. The support set provides a small number of labelled examples to guide the model, while the query set contains unseen samples used for evaluation.

Specifically, to ensure statistical robustness of the results, 600 episodes for the **Cardiac** dataset are randomly generated. In the 1-shot setting, each episode includes one labelled image in the support set, whereas in the 5-shot setting, the support set consists of five labelled images. In both cases, the query set contains 15 unlabelled images for performance evaluation. All competing methods are evaluated using the same set of episodes to ensure fairness and reproducibility. This episodic design allows for consistent comparison across methods and provides a reliable estimate of model generalisation under extremely limited supervision.

$R = 3$ is set in TFFS-MedSAM2, and keep other settings the same. In ad-

dition to comparing the original TFFS-MedSAM2 with the SOTA method UniverSeg (Butoi* et al., 2023), a comparison with a variant, denoted as TFFS-MedSAM2* and implemented as Algorithm 6, was further included. This variant introduces a more efficient pseudo-labelling strategy aimed at improving computational efficiency. In the original TFFS-MedSAM2, pseudo masks for unlabelled data are generated by randomly inserting a fully labelled 3D volume into the unlabelled volume at several random slice locations along the z-axis. In contrast, TFFS-MedSAM2* replaces this random insertion process with a deterministic, evenly spaced insertion scheme. Specifically, the slices from the volume with fewer slices are evenly interleaved into the larger volume while preserving the original slice order within each volume.

Same with TFFS-MedSAM2, all input frames were resized to 512×512 in TFFS-MedSAM2*. However, the input frames were resized to 128×128 in UniverSeg due to its default setting. In model evaluation, the predicted masks were resized back to the original image resolution and compared to the ground truth masks for metrics calculation. The mean Dice coefficient (DC) and average symmetric surface distance (ASSD) across all episodes are reported. The Wilcoxon Signed Rank Test (WSRT) is used to assess the statistical significance between the results of the two methods. In addition, the average model inference time and GPU memory usage for pseudo mask generation of one request data are shown in Table 6.5.

Results and Discussion

The quantitative results of the few-shot learning experiments are summarised in Table 6.4, which reports the mean Dice and ASSD for all evaluated methods. As expected, the 5-shot setting consistently achieves better

Table 6.4: Mean Dice and HD of the few-shot learning methods on the Cardiac dataset. Both 1-shot and 5-shot settings are evaluated. **Bold** ones have the best mean performance. † indicates statistical significance compared to the bold results ($p < 0.01$ measured by WSRT). TFFS and TFFS* indicate TFFS-MedSAM2 and TFFS-MedSAM2*, respectively.

	Dice			HD		
	LVBP	LVM	RVBP	LVBP	LVM	RVBP
	1 shot					
TFFS	82.30	68.73	63.11	1.51†	1.12	3.10
TFFS*	80.60†	67.76†	60.98†	1.50	1.23†	3.73†
UniverSeg	44.43†	50.18†	44.31†	4.38†	4.94†	4.48†
	5 shot					
	LVBP	LVM	RVBP	LVBP	LVM	RVBP
	5 shot					
TFFS	86.31	71.08	69.55	0.73	0.87	1.68
TFFS*	84.74†	70.02†	68.75†	0.81†	0.94†	1.88†
UniverSeg	60.55†	65.43†	60.70†	2.21†	2.55†	2.22†

segmentation performance than the 1-shot setting across most cases, confirming that additional annotated examples improve the model’s generalisation and stability.

Among the compared approaches, the original TFFS-MedSAM2 and its variant outperform the SOTA UniverSeg method across all segmentation classes in terms of both Dice and ASSD, with statistical significance measured by WSRT on the local **Cardiac** dataset, demonstrating the superior few-shot generalisation capability of the MedSAM2-based approach. Notably, while TFFS-MedSAM2* exhibits slightly lower accuracy than the original design, the performance gap between the two remains minimal.

Table 6.5 demonstrates that TFFS-MedSAM2* achieves comparable segmentation accuracy while substantially reducing inference time by almost R -fold, where R is the number of random insertion trials used in the original algorithm (see Algorithm. 5). This improvement reflects the efficiency of the deterministic slice interleaving strategy adopted in TFFS-MedSAM2*, which eliminates the need for multiple random insertions and averaging.

Moreover, although UniverSeg demonstrates a shorter overall model de-

Table 6.5: Average model inference time (MIT, in seconds) and GPU memory usage (GM, in GB) of generating the pseudo mask for one request volume. TFFS and TFFS* indicate TFFS-MedSAM2 and TFFS-MedSAM2*, respectively.

	TFFS	TFFS*	UniverSeg	TFFS	TFFS*	UniverSeg
	MIT			GM		
1 shot	20.68	8.42	0.69	0.18	0.18	0.39
5 shot	113.91	46.35	1.28	0.18	0.18	2.11

ployment time, its inference process requires substantially more GPU memory. In addition, since all support volumes are concatenated and processed jointly within the network, as the number of support images increases, both the input dimensionality and computational workload of UniverSeg increase proportionally, leading to higher GPU memory usage during inference.

In contrast, TFFS-MedSAM2 and TFFS-MedSAM2* perform pseudo-label generation independently for each support volume. Each support image is combined with the request volume in a separate forward pass to produce individual predictions, which are subsequently averaged to obtain the final pseudo-mask. Because these predictions are computed sequentially rather than simultaneously, the overall GPU memory consumption remains nearly constant regardless of the number of support images provided.

Overall, although TFFS-MedSAM2 achieves slightly better performance than TFFS-MedSAM2*, from a practical perspective, the efficiency gain makes TFFS-MedSAM2* a preferable choice in clinical or large-scale semi-supervised applications of SSL-MedSAM2, where computational resources or inference time are limited. Furthermore, since both variants produce pseudo labels of comparable quality, the subsequent iterative refinement by nnU-Net in the SSL pipeline could effectively correct minor inconsistencies, resulting in final segmentation masks of similar accuracy and robustness.

6.4 Discussion and Conclusions

This work proposed SSL-MedSAM2, a unified semi-supervised medical image segmentation framework that integrates a large foundation model (MedSAM2) with a self-adaptive segmentation backbone (nnU-Net) through a novel training-free few-shot learning branch (TFFS-MedSAM2). The framework is designed to address the challenge of limited labelled data in clinical imaging by effectively leveraging both annotated and unannotated volumes.

Comprehensive experiments were conducted to evaluate SSL-MedSAM2 under both semi-supervised and few-shot learning regimes. In the semi-supervised setting, the framework achieved SOTA performance on the binary-class MICCAI 2025 CARE-LiSeg dataset, demonstrating the generalisability and robustness of the proposed method even when the labelled dataset and the unlabelled dataset are in different phases (e.g, GED4 and T1). In the few-shot setting, TFFS-MedSAM2 was evaluated independently on the multi-class local Cardiac dataset under 1-shot and 5-shot conditions. The results revealed that the proposed few-shot branch not only surpasses the performance of the existing SOTA method UniverSeg, but also maintains high accuracy even under extreme data scarcity. Moreover, the modified variant TFFS-MedSAM2* produced comparable segmentation results while substantially reducing inference time, highlighting its potential for efficient deployment in time-critical clinical environments.

Despite its strong performance, the proposed framework has several limitations. First, the pseudo-labelling process in TFFS-MedSAM2, although effective, still relies on the structural priors captured by the pretrained MedSAM2 model, which may limit its adaptability to unseen modalities or imaging protocols with drastically different intensity distributions. More-

over, when the data distribution of the support labelled data and query unlabelled data differs too much (e.g., GED4 and T2WI in **CARE** dataset), TFFS-MedSAM2 tends to generate low-quality pseudo masks. In addition, the model inference time of TFFS-MedSAM2 remains relatively long. This computational overhead can hinder real-time or high-throughput deployment in clinical workflows.

In future work, several directions will be explored to further enhance the generalisability and efficiency of the proposed framework. Firstly, further modifications to TFFS-MedSAM2 will be investigated, aiming for better and more robust performance on 1-shot/5-shot segmentation. Thus, SSL-MedSAM2 can face more challenging clinical applications when very few manual annotations are available. Another promising direction is to develop cross-domain adaptive pseudo-labelling strategies that can automatically assess and refine pseudo-mask quality across heterogeneous imaging sources. Potential remedies include sequence-specific adapters, contrast-normalisation, or cross-sequence consistency training. Furthermore, efforts will be directed toward incorporating modality-invariant representation learning to enhance the robustness of MedSAM2-based pseudo-mask generation for unseen modalities. Last, even the efficiency of TFFS-MedSAM2 is largely constrained by MedSAM2, potential strategies include parallelising the support–query inference, or introducing efficient sampling schemes to reduce redundant support slices will be explored. Thus, the clinical applicability can be broadened, and a more general-purpose, data- and time-efficient medical segmentation framework can be established.

Chapter 7

Concluding Remarks and Future Work

Medical image segmentation lies at the heart of computer-assisted diagnosis and treatment planning, serving as the foundation for quantitative analysis, surgical navigation, and disease monitoring. Yet despite the transformative success of deep learning in natural image analysis, its adaptation to medical imaging remains constrained by three intertwined challenges: (1) limited and heterogeneous annotated datasets; (2) high computational demands of complex neural architectures, particularly Transformers; and (3) the prohibitive cost and effort of expert annotation. This thesis was therefore motivated by a fundamental question articulated in Chapter 1: How can deep learning models be designed to simultaneously achieve high efficiency, strong generalisability, and minimal dependence on large annotated datasets for robust medical image segmentation in diverse and evolving clinical environments?

To address this question, the thesis followed a coherent research path that progressively tackled these challenges through four major methodological

innovations—CTranS, MO-CTranS, CRFTrans, and SSL-MedSAM2—each building upon the limitations and insights of the previous stage. These models were not conceived as isolated algorithms but rather as sequential steps toward a unified goal: constructing efficient and generalisable deep learning frameworks that can function robustly in realistic, resource-constrained clinical environments. The research progression follows a logical evolution of capability: starting from the design of an effective hybrid architecture as the baseline (CTranS), moving to multi-organ and cross-dataset generalisation (MO-CTranS), then optimising architectural efficiency for deployment (CRFTrans), and finally addressing annotation scarcity through semi-supervised learning and few-shot learning settings with foundation models (SSL-MedSAM2).

Across these four main contributions, the thesis systematically advances the field from accuracy-oriented architecture design to efficient and generalisable learning frameworks, culminating in annotation-efficient adaptation through foundation models. Each contribution addresses a distinct but interconnected aspect of the central problem, collectively forming an end-to-end research pathway that reconciles theoretical innovation with clinical practicality. The results presented throughout the thesis demonstrate that through principled architectural design and task-aware learning, deep segmentation models can become not only more accurate and efficient but also more generalisable and resource-conscious, thus paving the way for their eventual translation from the research lab to clinical routine.

7.1 Contributions

This section summarises the principal contributions of the thesis and explains how they collectively form an end-to-end framework for accurate, efficient, and generalised medical image segmentation.

CTranS – A Multi-Resolution Convolution–Transformer Network for Medical Image Segmentation

The first contribution, CTranS, establishes the methodological foundation for the entire thesis. CTranS was conceived to overcome the intrinsic trade-off between local precision and global contextual awareness that characterises traditional convolutional and Transformer models.

CTranS introduces a hybrid encoder–decoder architecture that fuses a Convolutional Neural Network (CNN) based encoder and a Transformer-based decoder through a Pyramid Feature Fusion (PFF) module. The CNN encoder efficiently captures fine-grained structural information at multiple scales, while the Transformer decoder models long-range dependencies across regions and slices. The PFF mechanism ensures that information from multiple spatial resolutions is preserved and hierarchically aggregated, mitigating the feature degradation often observed in purely convolutional designs. Unlike prior hybrid networks that fuse features only at a single scale, CTranS performs multi-resolution alignment between the encoder and decoder, ensuring a balanced representation of local detail and global structure.

Comprehensive experiments on five public datasets demonstrated that CTranS consistently outperformed the conventional state-of-the-art (SOTA) CNNs (e.g., ResUNet (Zhang et al., 2018)) and SOTA Transformer-based hybrids

(e.g., FAT-Net (Wu et al., 2022), TransUNet (Chen et al., 2021b)) in Dice coefficient and surface distance metrics. CTranS also required fewer parameters and shorter training time per epoch, providing a strong argument for its clinical feasibility.

CTranS contributes a general blueprint for synergistic feature learning, showing that the complementary roles of convolutional locality and transformer globality can be unified into a single, stable training pipeline. The insights derived here, particularly the role of hierarchical feature fusion, form the architectural backbone for the subsequent models proposed in the thesis.

MO-CTranS – A Unified Multi-Organ Segmentation Model Learning from Multiple Heterogeneously Labelled Datasets

Building upon the CTranS baseline, the second contribution extends the scope of segmentation from a single, fully labelled dataset to multiple partially annotated datasets, a realistic but challenging scenario in clinical research.

The proposed MO-CTranS (Multi-Organ CTranS) introduces task-specific tokens within the Transformer decoder. Each token acts as a latent query that modulates the attention map according to the current organ or dataset, thereby enabling the network to infer context-aware representations even when label sets differ across datasets. This mechanism resolves label conflicts—for example, when an organ annotated in one dataset appears as background in another—and permits simultaneous training on a heterogeneous mixture of data sources without complex multi-stage pseudo-labelling pipelines or dataset-specific decoders.

Evaluations across several multi-organ Magnetic Resonance Imaging (MRI) segmentation benchmarks confirmed that MO-CTranS not only maintains the accuracy of models trained on individual datasets but also achieves superior cross-dataset generalisation. The network successfully learned a shared representation space where common anatomical features are reinforced and dataset-specific biases are suppressed. Compared with the conventional multi-decoder architectures, MO-CTranS significantly reduced the number of learnable parameters. Compared to the baseline CTranS, no extra learnable parameters are introduced in MO-CTranS.

This study provides the first demonstration within this thesis that generalisability can be embedded as an architectural property rather than achieved through extensive data augmentation or post-hoc domain adaptation. The use of task-specific tokens inspires later work in large multimodal models where tokenised prompts guide task specialisation, thus connecting this research to the broader evolution of foundation-model design.

CRFTrans – A Recursive Vision Transformer Layer Inspired by Conditional Random Fields

The third contribution, CRFTrans, addresses the critical limitation of efficiency in Transformer-based segmentation models, which traditionally rely on deeply stacked self-attention layers and exhibit quadratic computational complexity.

CRFTrans reinterprets the attention mechanism through the operations of Conditional Random Field (CRF) inference. Instead of stacking multiple independent Transformer blocks, a single recurrent self-attention layer is iteratively applied, emulating mean-field message passing among image tokens. This design allows the model to progressively refine predictions across

iterations using shared parameters, effectively transforming depth-wise hierarchy into temporal recurrence. The approach maintains contextual reasoning capability while drastically reducing parameter redundancy.

Empirical evaluation demonstrated that replacing conventional Transformer stacks with CRFTrans significantly reduced the parameter count with less memory consumption and fast training speed, while preserving (and in some cases improving) the segmentation accuracy. Statistical tests confirmed the absence of significant performance degradation compared to deeper architectures. This verifies that recurrence in depth can substitute for layer stacking without compromising representational capacity.

CRFTrans unifies two historically distinct paradigms—probabilistic graphical models and attention-based deep learning—within a single framework. This conceptual bridge provides new interpretability to Transformer operations, viewing each attention iteration as an inference step toward spatial label consistency. Consequently, CRFTrans contributes not only computational efficiency but also theoretical insight into how contextual dependencies can be recursively optimised in segmentation networks.

SSL-MedSAM2 – A Semi-Supervised Framework Powered by Few-Shot Learning of Foundation Models

The final contribution, SSL-MedSAM2, unifies the key ideas developed in this thesis and builds upon the recent shift toward foundation-model-based learning. While earlier models in this thesis rely on fully annotated datasets, SSL-MedSAM2 tackles the annotation bottleneck that limits its practical scalability.

SSL-MedSAM2 integrates two complementary components, which are a

Training-Free Few-Shot Segmentation (TFFS) branch based on the foundation model MedSAM2 (Ma et al., 2025), which automatically generates pseudo-labels for unannotated data by exploiting the memory-attention mechanism of SAM2 (Ravi et al., 2024); and a fully supervised refinement branch built upon adaptable backbones (e.g., nnU-Net (Isensee et al., 2021) or MO-CTranS), which iteratively refines these pseudo-labels through self-training. This hybrid framework couples the zero-shot generalisation of MedSAM2 with the task-specific adaptability of conventional segmentation models, achieving a virtuous cycle where foundation-model knowledge is progressively distilled into a compact medical network.

SSL-MedSAM2 was extensively validated on the MICCAI CARE-LiSeg 2025 challenge dataset (Liu et al., 2025; Gao et al., 2023; Wu and Zhuang, 2023) under the semi-supervised learning setting, and the TFFS branch alone was evaluated on a multi-class local clinical dataset under the few-shot learning setting. SSL-MedSAM2 achieved top-3 leaderboard ranking on both in-distribution and out-of-distribution test sets, confirming its robustness to vendor variation and limited labelling. Additionally, TFFS-MedSAM2 consistently outperforms the SOTA method UniverSeg, demonstrating the effectiveness of the few-shot learning capacity on the multi-class segmentation task.

This contribution illustrates the effective integration of foundation models with semi-supervised learning and few-shot learning. By leveraging SAM2’s few-shot capacity in a training-free manner, SSL-MedSAM2 achieves annotation efficiency without costly fine-tuning, opening a pathway toward “plug-and-play” segmentation systems that can be rapidly deployed across institutions and modalities.

Together, CTranS, MO-CTranS, CRFTrans, and SSL-MedSAM2 constitute

a coherent research framework that evolves from conventional fully supervised modelling toward intelligent, adaptive, and data-efficient paradigms. Each step not only addresses a specific technical challenge but also informs the next, forming a continuous progression toward clinically deployable deep learning systems.

7.2 Limitations

While the models proposed in this thesis collectively advance medical image segmentation in terms of efficiency, generalisability, and annotation efficiency, several limitations still remain. Recognising these limitations is essential for identifying directions for future research.

The principal limitation of CTranS lies in its 2D architectural formulation. All Transformer operations and the Pyramid Feature Fusion (PFF) mechanism are designed on 2D image grids, which means that volumetric spatial continuity is not explicitly modelled. When applied to 3D medical data such as Computed Tomography (CT) or MRI scans, the current implementation approximates 3D context by stacking three adjacent slices together to form a pseudo-RGB input and predicting on the middle slice. Although this “2.5D” strategy can partially utilise inter-slice information, it cannot fully capture the long-range dependencies and volumetric consistency inherent in 3D anatomy. Due to the quadratic complexity of self-attention scaling with volumetric tokens in the current Transformer layer, the 3D architectural formulation for CTranS is not feasible for clinical application.

This limitation is not only relevant to CTranS but also propagates to subsequent models—MO-CTranS and CRFTrans—since they inherit the same 2D Transformer backbone. In clinical workflows, however, most imaging

data are volumetric, and accurate delineation of organs or lesions relies heavily on spatial coherence across slices. Consequently, the 2D-based design restricts the direct applicability of the current models to real-world 3D segmentation tasks. Extending the Transformer blocks and the pyramid fusion operations into true 3D space, therefore represents a necessary next step toward clinical deployment.

The MO-CTranS framework successfully enables unified learning from multiple partially labelled datasets through the introduction of dataset-specific tokens. However, its current tokenisation strategy lacks a principled definition of dataset categorisation. In the experiments presented, the different datasets used originated from the same institution and shared similar acquisition conditions. The question of how to assign dataset tokens becomes non-trivial when dealing with multi-centre data or images acquired under heterogeneous scanner settings.

For example, if two datasets contain the same anatomical classes (e.g., liver, kidney) but are obtained from different hospitals or scanner vendors, it is unclear whether they should share the same token (treating them as one domain) or use distinct tokens (treating them as separate domains). The current formulation does not provide an explicit guideline or adaptive mechanism to determine this. As a result, the generalisation capability of MO-CTranS to cross-domain settings, where domain shift rather than label heterogeneity dominates, remains untested. Future exploration on how dataset tokens can encode not only label semantics but also domain attributes (e.g., acquisition protocol, modality type, or vendor) is essential for transforming MO-CTranS from a multi-task learner into a truly domain-adaptive segmentation framework.

Furthermore, the main limitation of CRFTrans lies in the sensitivity of

its performance to the number of iterative steps used during the recursive attention process. Although the model demonstrates the effectiveness of iterative message passing inspired by CRF inference, the optimal number of iterations has not been theoretically or empirically standardised. Too few iterations may result in under-refinement of contextual features and incomplete spatial consistency, while excessive iterations can lead to gradient accumulation, unstable convergence, or oscillatory learning behaviour.

Currently, the method relies on empirical tuning of this hyperparameter, and no analytical framework exists to predict or adaptively determine the suitable number of iterations for a given dataset or network configuration. This sensitivity implies the need for adaptive recurrence control strategies, such as dynamic stopping criteria or learnable iteration schedules that balance convergence speed and stability. A more systematic exploration of this relationship is therefore required to enhance both the robustness and theoretical soundness of CRFTrans.

The key limitation of SSL-MedSAM2 arises from its high dependency on the MedSAM2 foundation model. While MedSAM2 provides robust few-shot segmentation capability, its generalisation is constrained by the diversity of its pre-training data. When SSL-MedSAM2 is applied to datasets whose imaging characteristics differ substantially from those seen during MedSAM2’s pre-training, such as novel contrast agents, rare modalities, or unusual tissue textures, the quality of pseudo-labels generated by TFFS-MedSAM2 can deteriorate sharply.

Moreover, as discussed in Chapter 6, SSL-MedSAM2 currently struggles with cross-sequence adaptation. The pseudo-labels produced on unseen sequences can be inconsistent or spatially biased. Overcoming this limitation will require either fine-tuning the foundation model on representative med-

ical sequences or designing adaptive modules capable of aligning feature distributions across sequences during the semi-supervised training process.

Nevertheless, the thesis has substantially achieved its intermediate objectives: it has provided (1) a solid hybrid baseline architecture, (2) a generalisable multi-organ segmentation strategy, (3) a computationally efficient Transformer alternative, and (4) a foundation-model-driven semi-supervised framework. These advances form a strong stepping stone toward the ultimate goal of a unified, efficient, and adaptive 3D segmentation system. The remaining limitations identified above, therefore directly motivate the future research directions outlined in the next section.

7.3 Future Work

Building upon the limitations discussed in the previous section, several research directions can be pursued to further strengthen and extend the methods proposed in this thesis. These directions form a structured plan that addresses each model’s key shortcomings while guiding the long-term evolution toward a unified, efficient, and continuously learning medical segmentation framework.

The most immediate research priority is to develop a 3D extension of CTranS, enabling native volumetric feature learning while maintaining computational feasibility. A naïve implementation of 3D self-attention would increase the number of image tokens cubically with spatial resolution, leading to prohibitive memory and time costs. Future work will therefore focus on designing computationally efficient 3D attention mechanisms—for example, by applying windowed or axial attention, multi-scale feature integration, or hybrid 2D plus 1D factorised attention to capture inter-slice

dependencies without performing full 3D attention (Liu et al., 2021; Shaker et al., 2024; Perera et al., 2024).

This 3D extension will also benefit MO-CTranS and CRFTrans, which share the same 2D backbone, providing a common 3D-aware foundation that can be fine-tuned for diverse modalities such as CT, MRI, or ultrasound. The key research goal is to achieve true volumetric reasoning with near-2D computational cost, ensuring that 3D CTranS remains deployable in real clinical environments.

A second research direction concerns the cross-domain adaptability of MO-CTranS. Future experiments will systematically evaluate the model across multi-centre datasets that differ in scanner vendor, acquisition protocol, or patient population. The goal is to determine how dataset tokens should be defined when two datasets share overlapping organ classes but differ in domain characteristics.

Several strategies will be investigated, including introducing domain-aware tokens that encode metadata such as scanner type or acquisition phase, incorporating domain adversarial learning to align feature distributions across tokens, and developing a token clustering mechanism that automatically merges or splits dataset tokens based on learned inter-dataset similarity (Xie et al., 2022; Feng et al., 2023). Through these experiments, MO-CTranS can evolve from a label-heterogeneity learner into a domain-adaptive multi-organ model, capable of robust generalisation to unseen hospitals and imaging protocols—an essential requirement for large-scale clinical translation.

For CRFTrans, future research will concentrate on improving training stability and developing a principled mechanism to automatically determine the optimal number of iterations. This can be achieved by exploring adap-

tive recurrence control strategies, such as learnable gating modules that predict a convergence score and terminate recursion when further refinement yields negligible improvement, or theoretical analysis of iteration dynamics to derive closed-form criteria for convergence stability. These studies will deepen understanding of how recursive attention interacts with gradient propagation and will transform CRFTrans from an empirically tuned algorithm into a theoretically grounded, self-regulated iterative inference framework. Such adaptability is crucial for deploying recursive Transformers across diverse datasets without manual hyperparameter tuning.

Additionally, future research will aim to enable cross-sequence learning within the SSL-MedSAM2 framework. To overcome its dependence on the MedSAM2 foundation model and limited adaptability across image sequences, potential research directions include incorporating sequence-aware adapters that learn contrast-specific feature transformations between sequences, introducing meta-learning or prompt-tuning strategies allowing the model to generalise to unseen sequences with minimal examples, and designing a cross-sequence consistency loss ensuring that anatomical structures are consistently segmented across different contrast settings (Bo et al., 2025; Zhu et al., 2025). These developments will extend SSL-MedSAM2 from a few-shot, single-sequence learner to a universal semi-supervised segmentation framework capable of handling the full spectrum of sequence and modality variations encountered in clinical practice.

In the longer term, the most promising and impactful direction of this research lies in the development of a continual learning framework that unifies efficiency, generalisability, and adaptability in real-world medical imaging workflows. Continual learning is an emerging paradigm in artificial intelligence that enables models to continuously learn from new data or tasks without forgetting previously acquired knowledge (Kumari et al., 2023).

Unlike conventional training paradigms, where a model is trained once and then frozen, continual learning mimics the way humans accumulate experience over time, dynamically expanding and refining their understanding as new information becomes available.

In the context of medical image segmentation, continual learning holds exceptional significance. Clinical data are constantly evolving: new imaging devices are introduced, scanning protocols are updated, and novel disease patterns emerge. A segmentation model deployed in a hospital must therefore be capable of adapting to new distributions, for instance, new MRI sequences, scanner vendors, or patient populations, without losing its prior segmentation capability. Continual learning provides a viable solution to this bottleneck. By enabling models to incrementally assimilate new knowledge, it allows long-term adaptation and stability, thus ensuring sustained performance as data evolves over time (Kumari et al., 2023). In medical segmentation, this capability could substantially reduce the need for manual re-annotation and re-training, leading to more practical and sustainable deployment in clinical workflows.

The few-shot learning strategy introduced in TFFS-MedSAM2 represents an initial step toward continual learning. By leveraging pre-trained knowledge from MedSAM2 and adapting it to new tasks with only a handful of annotated examples, TFFS-MedSAM2 demonstrates that knowledge can be transferred and reused in the same segmentation tasks on the same data source. However, this adaptation remains task-specific and static; once trained, the model does not continue to evolve when new data or tasks appear. To reach the level of true continual learning, a model must possess a more effective, generalisable, and robust knowledge-transfer capability, thus allowing persistent learning across organs, sequences, and modalities while maintaining stability and preventing catastrophic forgetting.

To achieve true continual learning, future research must focus on building more effective, generalisable, and robust knowledge transfer mechanisms that allow models to integrate information from diverse sources and evolve continuously across tasks, domains, and modalities. Future research will therefore focus on establishing effective and efficient knowledge transfer from large foundation models to downstream medical segmentation tasks as the cornerstone of continual learning.

Foundation models such as SAM2 have already demonstrated remarkable generalisation across domains due to their large-scale, diverse pre-training. However, transferring this broad knowledge to specialised medical tasks in a way that is both efficient and stable remains an open challenge. A key research focus will be developing strategies for cross-organ, cross-sequence/vendor, and cross-modality knowledge transfer.

Beyond designing knowledge transfer mechanisms, a critical and emerging research direction will involve quantifying the transferability of knowledge between tasks (Wu et al., 2025b). Understanding which knowledge can be transferred, how much can be transferred, and under what conditions transfer is beneficial is essential for efficient continual learning. This will require developing transferability metrics and analytical frameworks capable of measuring the relationship between a source task (or domain) and a target task. Such measures could predict how well a foundation model or pre-trained representation can support a new segmentation task before extensive fine-tuning is performed.

Lastly, future work should also investigate explainable AI and uncertainty quantification as essential components for clinical deployment. While the models developed in this thesis prioritise segmentation accuracy, efficiency, and generalisability, clinical practice also requires transparency and re-

liability. Explainable AI techniques, such as attention visualisation and feature attribution methods, can help clinicians understand which image regions contribute most to the predicted segmentation, enabling verification of anatomical plausibility and identification of potential failure cases. In parallel, uncertainty quantification methods—including Bayesian inference, Monte Carlo dropout, or ensemble-based approaches—can provide pixel-wise or structure-level confidence estimates, which are particularly valuable in low-contrast regions, rare pathologies, or under domain shift. Integrating explainability and uncertainty estimation therefore represents a necessary step toward improving user confidence and facilitating the translation of deep learning-based segmentation models into routine clinical workflows.

Ultimately, these advancements will lay the theoretical and methodological groundwork for a lifelong, knowledge-transferable medical segmentation system—one that continuously evolves, generalises across domains, and remains robust and efficient throughout its operational lifetime in clinical practice.

Bibliography

- Al Hasan, S., Mahim, S., Hossen, M. E., Hasan, M. O., Islam, M. K., Livreri, P., Khan, S. U., Alibakhshikenari, M., and Miah, M. S. (2025). Dsit unet a dual stream iterative transformer based unet architecture for segmenting brain tumors from flair mri images. *Scientific Reports*, 15(1):13815.
- Bai, Y., Yu, Q., Yun, B., Jin, D., Xia, Y., and Wang, Y. (2024). Fs-medsam2: Exploring the potential of sam2 for few-shot medical image segmentation without fine-tuning. *arXiv e-prints*, pages arXiv–2409.
- Bengio, Y., Goodfellow, I., and Courville, A. (2017). *Deep learning*, volume 1. MIT press Cambridge, MA, USA.
- Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. (2015). Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111.
- Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. (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.

- Bo, Y., Zhu, Y., Li, L., and Zhang, H. (2025). Famnet: Frequency-aware matching network for cross-domain few-shot medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1889–1897.
- Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., and De Bruijne, M. (2019). Semi-supervised medical image segmentation via learning consistency under transformations. In *International conference on medical image computing and computer-assisted intervention*, pages 810–818. Springer.
- Boykov, Y. and Jolly, M.-P. (2000). Interactive organ segmentation using graph cuts. In *International conference on medical image computing and computer-assisted intervention*, pages 276–286. Springer.
- Butoi*, V. I., Ortiz*, J. J. G., Ma, T., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2023). Universeg: Universal medical image segmentation. *International Conference on Computer Vision*.
- Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.
- Chen, B., Liu, Y., Zhang, Z., Lu, G., and Zhang, D. (2021a). Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *arXiv preprint arXiv:2107.05274*.
- Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al. (2024a). Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310.
- Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L.,

- and Zhou, Y. (2021b). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
- Chen, S., Lin, L., Cheng, P., and Tang, X. (2024b). Aslseg: Adapting sam in the loop for semi-supervised liver tumor segmentation.
- Chen, S., Ma, K., and Zheng, Y. (2019). Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*.
- Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.
- Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.
- Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, L. (2018). Universal transformers. *arXiv preprint arXiv:1807.03819*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dmitriev, K. and Kaufman, A. E. (2019). Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9501–9511.
- Dong, H., Gu, H., Chen, Y., Yang, J., Chen, Y., and Mazurowski, M. A. (2024). Segment anything model 2: an application to 2d and 3d medical images. *arXiv preprint arXiv:2408.00756*.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Feng, W., Ju, L., Wang, L., Song, K., Zhao, X., and Ge, Z. (2023). Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 623–631.
- Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.
- Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., and Rajpoot, N. (2019). Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European congress on digital pathology*, pages 11–19. Springer.
- Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X., and Wang, J. (2025). Medical image segmentation: A comprehensive review of deep learning-based methods. *Tomography*, 11(5):52.
- Gao, Y., Zhou, M., and Metaxas, D. N. (2021). Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer.
- Gao, Z., Liu, Y., Wu, F., Shi, N., Shi, Y., and Zhuang, X. (2023). A reliable and interpretable framework of multi-view learning for liver fibrosis staging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 178–188.

- Gee, J. C., Reivich, M., and Bajcsy, R. (1993). Elastically deforming a three-dimensional atlas to match anatomical brain images.
- Gehrig, M. and Scaramuzza, D. (2023). Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893.
- Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., and Dou, Q. (2024a). 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324.
- Gong, Z., French, A. P., Qiu, G., and Chen, X. (2024b). Ctrans: A multi-resolution convolution-transformer network for medical image segmentation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.
- Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2018). Densely connected convolutional networks.
- Huang, R., Zheng, Y., Hu, Z., Zhang, S., and Li, H. (2020). Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In *Medical Image Computing and Computer Assisted*

Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, pages 146–155. Springer.

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

Jafari, M., Auer, D., Francis, S., Garibaldi, J., and Chen, X. (2020). Dru-net: An efficient deep convolutional neural network for medical image segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1144–1148.

Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., and Luo, P. (2021). Multi-compound transformer for accurate biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 326–336. Springer.

Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., and Jin, C. (2024). Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 169:107840.

Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., and Ye, J. (2022). Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data*, 9(1):475.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331.

Kavur, A. E., Selver, M. A., Dicle, O., Barış, M., and Gezer, N. S. (2019). CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data.

- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.
- Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Kumari, P., Chauhan, J., Bozorgpour, A., Huang, B., Azad, R., and Merhof, D. (2023). Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. *arXiv preprint arXiv:2312.17004*.
- Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.
- Li, N., Xiong, L., Qiu, W., Pan, Y., Luo, Y., and Zhang, Y. (2023). Segment anything model for semi-supervised medical image segmentation via selecting reliable pseudo-labels. In *International Conference on Neural Information Processing*, pages 138–149. Springer.
- Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., and Goh, R. (2021a). Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*.

- Li, Y., Cai, W., Gao, Y., and Hu, X. (2021b). More than encoder: Introducing transformer decoder to upsample. *arXiv preprint arXiv:2106.10637*.
- Lian, S., Li, L., Luo, Z., Zhong, Z., Wang, B., and Li, S. (2023). Learning multi-organ segmentation via partial-and mutual-prior from single-organ datasets. *Biomedical Signal Processing and Control*, 80:104339.
- Lin, Y., Chen, Y., Cheng, K.-T., and Chen, H. (2023). Few shot medical image segmentation with cross attention transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 233–243. Springer.
- Liu, H., Xu, Z., Gao, R., Li, H., Wang, J., Chabin, G., Oguz, I., and Grbic, S. (2024). Cosst: Multi-organ segmentation with partially labeled datasets using comprehensive supervisions and self-training. *IEEE Transactions on Medical Imaging*.
- Liu, Y., Gao, Z., Shi, N., Wu, F., Shi, Y., Chen, Q., and Zhuang, X. (2025). Merit: Multi-view evidential learning for reliable and interpretable liver fibrosis staging. *Medical Image Analysis*, 102:103507.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.
- Luo, X., Chen, J., Song, T., and Wang, G. (2021). Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of*

- the AAAI conference on artificial intelligence*, volume 35, pages 8801–8809.
- Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1):654.
- Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., and Wang, B. (2025). Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.
- Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507.
- Miao, J., Chen, C., Zhang, K., Chuai, J., Li, Q., and Heng, P.-A. (2024). Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 167–177. Springer.
- Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE.
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

- Ouali, Y., Hudelot, C., and Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684.
- Perera, S., Navard, P., and Yilmaz, A. (2024). Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4988.
- Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., and Wang, S. (2019). X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P., and Khan, A. R., editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part III*, volume 11766 of *Lecture Notes in Computer Science*, pages 247–255. Springer.
- Ran, L., Li, Y., Liang, G., and Zhang, Y. (2024). Semi-supervised semantic segmentation based on pseudo-labels: A survey. *arXiv preprint arXiv:2403.01909*.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

- Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N., and Wachinger, C. (2020). ‘squeeze & excite’guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587.
- Sengupta, S., Chakrabarty, S., and Soni, R. (2025). Is sam 2 better than sam in medical image segmentation? In *Medical Imaging 2025: Image Processing*, volume 13406, pages 666–672. SPIE.
- Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.-H., and Khan, F. S. (2024). Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390.
- Shen, C., Li, W., Shi, Y., and Wang, X. (2024). Interactive 3d medical image segmentation with sam 2. *arXiv preprint arXiv:2408.02635*.
- Shen, Z., Liu, Z., and Xing, E. (2022a). Sliced recursive transformer. In *European Conference on Computer Vision*, pages 727–744. Springer.
- Shen, Z., Yang, H., Zhang, Z., and Zheng, S. (2022b). Automated kidney tumor segmentation with convolution and transformer network. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pages 1–12. Springer.
- Shi, Y., Zhang, Y., and Wang, S. (2023). Competitive ensembling teacher-student framework for semi-supervised left atrium mri segmentation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2231–2234. IEEE.
- Solatidehkordi, Z. and Zualkernan, I. (2022). Survey on recent trends in medical image classification using semi-supervised learning. *Applied Sciences*, 12(23):12094.

- Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). Seg-
menter: Transformer for semantic segmentation. In *Proceedings of the
IEEE/CVF international conference on computer vision*, pages 7262–
7272.
- Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., and Pais-
ley, J. (2022). Few-shot medical image segmentation using a global cor-
relation network with discriminative embedding. *Computers in biology
and medicine*, 140:105067.
- Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role mod-
els: Weight-averaged consistency targets improve semi-supervised deep
learning results. *Advances in neural information processing systems*,
30.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez,
A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
Advances in neural information processing systems, 30.
- Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su,
Y., Huang, Z., Shen, Y., et al. (2024). Sam-med3d: towards general-
purpose segmentation models for volumetric medical images. In *Euro-
pean Conference on Computer Vision*, pages 51–67. Springer.
- Wang, X., Yuan, Y., Guo, D., Huang, X., Cui, Y., Xia, M., Wang, Z.,
Bai, C., and Chen, S. (2022). Ssa-net: Spatial self-attention network
for covid-19 pneumonia infection segmentation with semi-supervised
few-shot learning. *Medical image analysis*, 79:102459.
- Wu, F. and Zhuang, X. (2023). Minimizing estimated risks on unlabeled
data: A new formulation for semi-supervised medical image segmen-
tation. *IEEE Transactions on Pattern Analysis and Machine Intelli-
gence*, 45(5):6021–6036.

- Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., and Wen, Z. (2022). Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76:102327.
- Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., and Jin, Y. (2025a). Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Wu, Y., Wang, J., Chen, X., Zhang, E., Tan, Y., Liu, H., and Li, Y. (2025b). Exploiting task relationships for continual learning using transferability-aware task embeddings. *arXiv preprint arXiv:2502.11609*.
- Xie, Q., Li, Y., He, N., Ning, M., Ma, K., Wang, G., Lian, Y., and Zheng, Y. (2022). Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. *IEEE Transactions on Medical Imaging*, 43(1):4–14.
- Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer.
- Xie, Y., Zhang, J., Xia, Y., and Shen, C. (2023). Learning from partially labeled data for multi-organ and tumor segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Xu, X., Sanford, T., Turkbey, B., Xu, S., Wood, B. J., and Yan, P. (2021a).

- Shadow-consistent semi-supervised learning for prostate ultrasound segmentation. *IEEE Transactions on Medical Imaging*, 41(6):1331–1345.
- Xu, Z., Zhang, W., Zhang, T., Yang, Z., and Li, J. (2021b). Efficient transformer for remote sensing image segmentation. *Remote Sensing*, 13(18):3585.
- Yao, H., Hu, X., and Li, X. (2022). Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3099–3107.
- Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., and Xie, Y. (2024). From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, 37(4):1529–1547.
- Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al. (2023). Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*.
- Yu, L., Wang, S., Li, X., Fu, C.-W., and Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 605–613. Springer.
- Zhang, G., Yang, Z., Huo, B., Chai, S., and Jiang, S. (2021a). Multiorgan segmentation from partially labeled datasets with conditional nnu-net. *Computers in Biology and Medicine*, 136:104658.
- Zhang, J., Xie, Y., Xia, Y., and Shen, C. (2021b). Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled

- datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204.
- Zhang, L., Deng, X., and Lu, Y. (2023a). Segment anything model (sam) for medical image segmentation: A preliminary review. In *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 4187–4194. IEEE.
- Zhang, Y., Higashita, R., Fu, H., Xu, Y., Zhang, Y., Liu, H., Zhang, J., and Liu, J. (2021c). A multi-branch hybrid transformer network for corneal endothelial cell segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–108. Springer.
- Zhang, Y., Lv, B., Xue, L., Zhang, W., Liu, Y., Fu, Y., Cheng, Y., and Qi, Y. (2025). Semisam+: Rethinking semi-supervised medical image segmentation in the era of foundation models.
- Zhang, Y. and Shen, Z. (2024). Unleashing the potential of sam2 for biomedical images and videos: A survey. *arXiv preprint arXiv:2408.12889*.
- Zhang, Y., Shen, Z., and Jiao, R. (2024a). Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238.
- Zhang, Y., Yang, J., Liu, Y., Cheng, Y., and Qi, Y. (2024b). Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3982–3986. IEEE.
- Zhang, Y. and Zhang, J. (2021). Dual-task mutual learning for semi-supervised medical image segmentation. In *Chinese Conference on*

Pattern Recognition and Computer Vision (PRCV), pages 548–559. Springer.

Zhang, Y., Zhou, T., Wang, S., Wu, Y., Gu, P., and Chen, D. Z. (2023b).

Samdsk: Combining segment anything model with domain-specific knowledge for semi-supervised learning in medical image segmentation. *arXiv preprint arXiv:2308.13759*.

Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., and Dalca, A. V. (2019). Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8553.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537.

Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., and Yu, Y. (2023). nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018).

Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer.

Zhu, Y., Li, M., Ye, Q., Wang, S., Xin, T., and Zhang, H. (2025). Robustemd: Domain robust matching for cross-domain few-shot medical image segmentation. *Artificial Intelligence in Medicine*, page 103197.

Zhu, Y., Wang, S., Xin, T., and Zhang, H. (2023). Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 271–280. Springer.

Zu, H., Ge, J., Xiao, H., Xie, J., Zhou, Z., Meng, Y., Ni, J., Niu, J., Zhang, L., Ni, L., et al. (2025). Rethinking few-shot medical image segmentation by sam2: A training-free framework with augmentative prompting and dynamic matching. *arXiv preprint arXiv:2503.04826*.