MEDICAL IMAGE SEGMENTATION UNDER MULTIPLE REAL-WORLD CONSTRAINTS

CHANGTAI LI

Cover:
A GPU-embedded AI eye projects light onto floating brain MRI slices, highlighting a precisely segmented lesion, symbolising intelligent medical image analysis, computational vision, and accurate lesion detection in a cinematic black scene, generated by ChatGPT.

Medical Image Segmentation under Multiple Real-World Constraints

Changtai Li
PhD Thesis

The research for this dissertation was conducted at:

the Scientific Visualization and Computer Graphics (SVCG) research group, part of the Bernoulli Institute (BI) and the Faculty of Science and Engineering (FSE) at the University of Groningen, the Netherlands

and

the Artificial Intelligence and 3D Visualization (AI3D) research group, part of the School of Intelligence Science and Technology at the University of Science and Technology Beijing, China.

university of groningen

Medical Image Segmentation under Multiple Real-World Constraints

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Thursday 18 June 2026 at 9.00 hours

by

Changtai Li

born on 1 April 1999

Supervisors
Prof. J. Kosinka
Prof. A.C. Telea
Prof. X. Ban

Co-supervisors
Dr. S.D. Frey

Assessment committee
Prof. K. Bunte
Prof. P.M.A. van Ooijen
Prof. A.P.J.M. Siebes

The only way to do good work is to love what you do.
— Steve Jobs

ABSTRACT

Medical image segmentation (MIS) is fundamental to clinical diagnosis, treatment planning, and longitudinal monitoring. However, its translation to routine clinical practice is frequently impeded by three pervasive challenges in data curation and deployment. First, annotations are often produced through a two-stage workflow in which non-experts create coarse masks, and clinicians subsequently refine them, leading to *label-quality imbalance* and systematic bias in the available supervision. Second, obtaining dense expert annotations for volumetric imaging is expensive and time-consuming, resulting in *label scarcity* that limits supervised training, particularly for 3D tumour segmentation. Third, segmentation performance can depend strongly on informative imaging modalities (e.g., PET or MRI), yet such modalities may be unavailable at inference due to acquisition cost, accessibility, or patient burden, creating *modality-information asymmetry*. This dissertation addresses these challenges by developing learning frameworks that compensate for unreliable labels, limited annotations, and missing modalities while preserving clinically relevant accuracy.

First, we consider the common two-stage annotation workflow in which non-experts provide coarse masks that are subsequently refined by clinicians. We propose a recurrent refinement learning framework that iteratively transforms non-expert delineations into expert-level segmentations. By explicitly modelling the refinement process and leveraging discrepancy-aware supervision across iterations, the method learns latent correction patterns from paired annotations, thereby improving robustness to label-quality imbalance without discarding imperfect labels.

Second, we study label-efficient 3D tumour segmentation under limited expert annotations. We introduce a self-supervised rotation learning strategy tailored to consecutively sliced MRI volumes, in which the encoder is trained to capture 3D spatial and geometric regularities through a rotation-based pretext task. The learned representations transfer effectively to downstream voxel-wise segmentation, yielding improved performance when only a small fraction of labelled volumes is available.

Third, we address modality-information asymmetry, where weak modalities such as non-contrast CT provide limited lesion contrast, whereas stronger modalities (e.g., PET or MRI) are costly or unavailable at inference. We propose a two-stage cross-modal knowledge transfer framework: an asymmetric relationship modelling module constructs high-confidence fused representations from multi-modal inputs, and a diffusion-inspired progressive transfer mechanism distils these repre-

sentations into a weak-modality encoder. This enables accurate segmentation using only the weak modality at inference time, substantially narrowing the performance gap induced by missing modalities.

For all our contributions, extensive experiments on multiple public and clinical datasets demonstrate consistent improvements in robustness, label efficiency, and generalisation under realistic constraints. Beyond achieving higher accuracy, the proposed methods offer a principled path towards more scalable and deployable segmentation systems: they reduce dependence on dense expert annotation, leverage imperfect supervision more effectively, and mitigate reliance on costly modalities at inference. Collectively, this work advances practical learning paradigms for medical image analysis in clinical settings where supervision and modality availability are inherently constrained.

SAMENVATTING

Medische beeldsegmentatie is essentieel voor klinische diagnostiek, behandelplanning en longitudinale monitoring. De vertaling naar de routinematige klinische praktijk wordt echter vaak belemmerd door drie wijdverbreide uitdagingen in datacuratie en implementatie. Ten eerste worden annotaties vaak verkregen via een tweestapsworkflow waarbij niet-experts grove maskers maken die vervolgens door clinici worden verfijnd, wat leidt tot *onbalans in labelkwaliteit* en systematische vertekening in de beschikbare supervisie. Ten tweede is het verkrijgen van dichte expertannotaties voor volumetrische beeldvorming kostbaar en tijdrovend, wat resulteert in *schaarste aan labels* die supervised training beperkt, met name voor 3D-tumoursegmentatie. Ten derde kan segmentatieprestatie sterk afhankelijk zijn van informatieve beeldmodaliteiten (bijv. PET of MRI), terwijl dergelijke modaliteiten bij inferentie mogelijk niet beschikbaar zijn vanwege acquisitiekosten, toegankelijkheid of belasting voor de patiënt, wat *modaliteits-informatie-asymmetrie* veroorzaakt. Dit proefschrift pakt deze uitdagingen aan door leerframeworks te ontwikkelen die compenseren voor onbetrouwbare labels, beperkte annotaties en ontbrekende modaliteiten, terwijl klinisch relevante nauwkeurigheid behouden blijft.

Ten eerste beschouwen wij de gangbare tweestapsannotatieworkflow waarin niet-experts grove maskers leveren die vervolgens door clinici worden verfijnd. Wij stellen een recurrent verfijningsleerframework voor dat niet-expertafbakening iteratief omzet in expert-niveau segmentaties. Door het verfijningsproces expliciet te modelleren en discrepantie-bewuste supervisie over iteraties te benutten, leert de methode latente correctiepatronen uit gepaarde annotaties, waardoor de robuustheid tegen onbalans in labelkwaliteit toeneemt zonder imperfecte labels te verwerpen.

Ten tweede bestuderen wij label-efficiënte 3D-tumoursegmentatie onder beperkte expertannotaties. Wij introduceren een zelfgesuperviseerde rotatieleerstrategie die is toegespitst op opeenvolgende MRI-slices, waarbij de encoder wordt getraind om 3D-ruimtelijke en geometrische regulariteiten te vangen via een rotatiegebaseerde pretexttaak. De aangeleerde representaties dragen effectief over naar downstream voxelgewijze segmentatie, wat leidt tot betere prestaties wanneer slechts een klein deel van de gelabelde volumes beschikbaar is.

Ten derde behandelen wij modaliteits-informatie-asymmetrie, waarbij zwakkere modaliteiten zoals non-contrast CT een beperkt laesiecontrast bieden, terwijl sterkere modaliteiten (bijv. PET of MRI) kostbaar zijn of bij inferentie niet beschikbaar. Wij stellen een tweestaps crossmodale kennisoverdrachtsframework voor: een module voor asymme-

trische relatie-modellering construeert hoogbetrouwbare gefuseerde representaties uit multimodale input, en een diffusie-geïnspireerd mechanisme voor progressieve overdracht distilleert deze representaties naar een encoder voor de zwakke modaliteit. Dit maakt nauwkeurige segmentatie mogelijk met uitsluitend de zwakke modaliteit tijdens de testfase, en verkleint substantieel de prestatiekloof die ontstaat door ontbrekende modaliteiten.

Uitgebreide experimenten op meerdere publieke en klinische datasets tonen consistente verbeteringen in robuustheid, labelefficiëntie en generaliseerbaarheid onder realistische beperkingen. Naast hogere nauwkeurigheid bieden de voorgestelde methoden een principieel pad naar beter schaalbare en beter inzetbare segmentatiesystemen: zij verminderen de afhankelijkheid van dichte expertannotatie, benutten imperfecte supervisie effectiever en beperken de afhankelijkheid van kostbare modaliteiten tijdens inferentie. Gezamenlijk bevordert dit werk praktische leerparadigma's voor medische beeldsegmentatie in klinische omgevingen waar supervisie en modaliteitsbeschikbaarheid inherent beperkt zijn.

PUBLICATIONS

This thesis is the result of the following publications:

- R. Jiang†, **C. Li**†, X. Ban*, S. Yin, C. Yao, Y. Guo, and M. S. Obaidat. From non-expert to expert: Recurrent refined learning for medical image segmentation. *In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2086–2093. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10821757

- **C. Li**†, R. Jiang†, S. Yin, J. Yang, and X. Ban*. Self-supervised rotation learning for 3D segmentation on nasopharyngeal carcinoma MRI images. *In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3529–3534. IEEE, 2023. doi: 10.1109/BIBM58861.2023.10385483

- **C. Li**, C. Yao*, W. Liu, X. Wang, J. Kosinka, S. Frey, A. C. Telea, and X. Ban. Weak-to-strong: Empowering non-contrast CT for accurate lesion segmentation via cross-modal knowledge transfer. To be submitted.

Other selected publications during the PhD study are listed below:

- **C. Li**†, Y. Zhang†, Y. Jia, Y. Guo, C. Yao*, X. Ban, and Y. He*. HRTEM-GAN: Structure-preserving restoration of low-quality atomic-scale HRTEM images. *Nano Research*. SciOpen, 2026. doi: 10.26599/NR.2026.94908715

- **C. Li**†, X. Han†, C. Yao, Y. Guo, Z. Li, L. Jiang, W. Liu, H. Huang, H. Fu*, and X. Ban*. A novel training-free approach to efficiently extracting material microstructures via visual large model. *Acta Materialia*, Volume 290, 120962. Elsevier, 2025. doi: 10.1016/j.actamat.2025.120962

- **C. Li**†, R. Jiang†, H. Wang, W. Xue, and Y. Guo, and X. Ban*. DeepMMP: Efficient 3D perception of microstructures from serial section microscopic images. *Computational Materials Science*, Volume 235, 112826. Elsevier, 2024. doi: 10.1016/j.commatsci.2024.112826

† Equal contribution
* Corresponding author

CONTENTS

LIST OF ABBREVIATIONS AND NOTATIONS

General and Imaging

MIA	medical image analysis
MIS	medical image segmentation
CT	computed tomography
NCCT	non-contrast computed tomography
MRI	magnetic resonance imaging
PET	positron emission tomography
ADC	apparent diffusion coefficient (MRI)
HU	Hounsfield unit
US	ultrasound
GTV	gross tumour volume
GTVnx	gross tumour volume of nasopharyngeal primary

Architectures and Components

FCN	fully convolutional network
CNN	convolutional neural network
ViT	vision transformer
MHA	multi-head attention
SSM	structured state space model
SS2D	selective scan 2D
MLP	multi-layer perceptron
LN	layer normalization
DWConv	depthwise convolution
ReLU	rectified linear unit

Loss Functions and Metrics

BCE	binary cross-entropy
CE	cross-entropy
WCE	weighted cross-entropy
MSE	mean squared error
KL	Kullback–Leibler divergence

DSC	Dice similarity coefficient
IoU	intersection over union
HD	Hausdorff distance
HD95	95th-percentile Hausdorff distance
ASSD	average symmetric surface distance
RVE	relative volume error
TP	true positive
TN	true negative
FP	false positive
FN	false negative
Prec	precision
Recall	recall (sensitivity)

Methods

SSL	self-supervised learning
FT	fine-tuning
ReReNet	recurrent refined network
NPC	nasopharyngeal carcinoma
3DRotNPC	3D rotation learning for NPC
ResUNet3D	3D residual U-Net backbone
ARM	asymmetric relationship modeling
PKT	progressive knowledge transfer
MFEM	mutual feature enhancement module
CFAM	complementary feature aggregation module

1 INTRODUCTION

1.1 PRELIMINARIES

Medical image analysis (MIA) has become a central component of contemporary clinical workflows [91, 126], supporting diagnosis, treatment planning, and longitudinal monitoring through quantitative assessment of patient-specific anatomy and pathology (see Fig. 1.1).

As the fundamental technique of MIA, medical imaging can be broadly defined as the acquisition and computational reconstruction of in vivo representations of human anatomy, physiology, and molecular processes using physical signals—such as X-ray attenuation, magnetic resonance, acoustic scattering, or radiotracer emission [185]. These representations are encoded as digital data, typically as 2D images composed of pixels (e.g., radiographs, individual slices) or as 3D/4D volumes composed of voxels (e.g., computed tomography (CT)/magnetic resonance imaging (MRI) volumes and dynamic sequences), enabling quantitative interrogation of structure and function across space and time. Medical imaging supports clinical decision-making and biomedical research. %

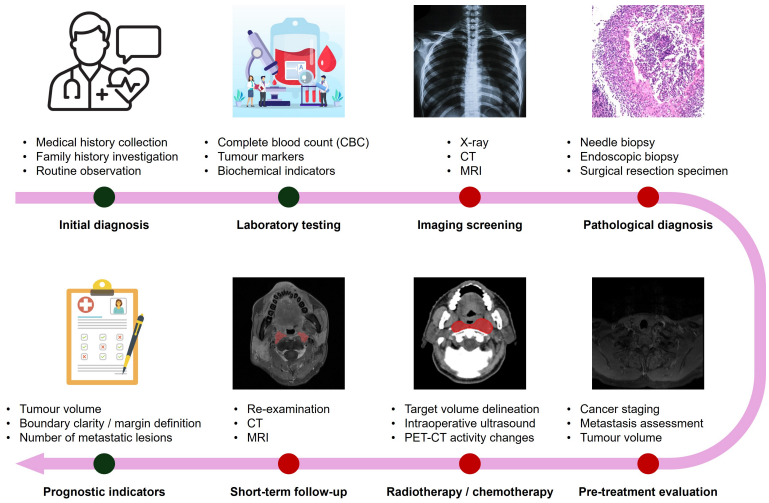


Figure 1.1: The comprehensive patient journey in clinical practice begins with initial consultation and laboratory screening, proceeds through pathological confirmation, and concludes with pre-treatment evaluation, therapy, and prognostic monitoring. Medical image analysis plays a central role throughout this workflow.

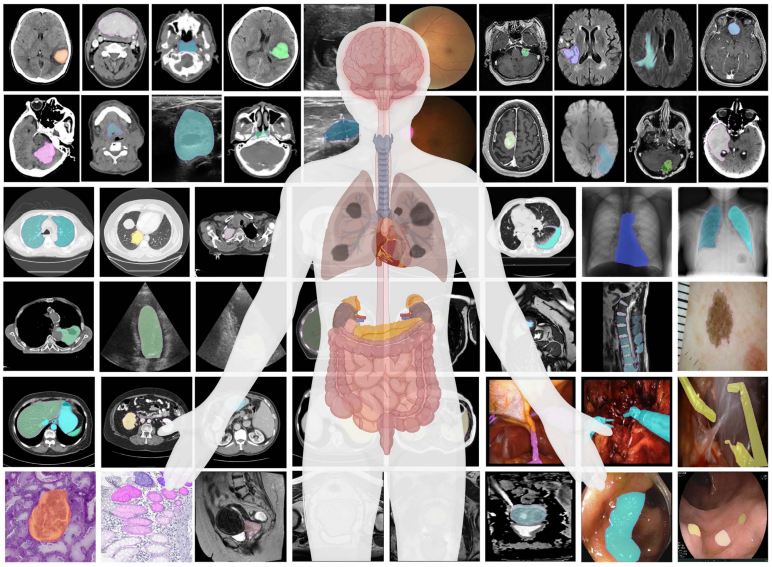


Figure 1.2: Segmentation tasks in medical image analysis. Image taken from [103].

Among the spectrum of tasks in medical image computing, image segmentation occupies a particularly critical role [5, 112] because it provides structured, spatially resolved delineations of organs and lesions that enable downstream measurements, risk stratification, and intervention guidance (see Fig. 1.2). In practice, accurate segmentation can reduce inter-observer variability in reporting [96], facilitate reproducible biomarker extraction, and improve the efficiency of routine clinical assessment, especially in scenarios where manual delineation is time-consuming and subject to subjective interpretation [65, 120].

Despite substantial progress driven by deep learning [77, 91], segmentation systems that perform well under benchmark conditions often degrade when deployed in realistic settings. The dominant bottlenecks are rarely architectural, but instead arise from constraints that are intrinsic to clinical data acquisition and annotation. To clarify, we use *annotation* to assign semantic classes (e.g., benign or malignant) to image elements (e.g., voxels), whereas *delineation* refers more specifically to tracing the spatial extent of a structure by placing its boundary (i.e., drawing contours) to form a segmentation mask. Specifically, the available supervision is frequently imperfect in **quality** due to heterogeneous annotator expertise and systematic boundary inconsistencies [71, 128], limited in **quantity** because expert voxel-wise labelling is costly [33, 127], and restricted in **information availability** at inference when highly informative modalities are absent or impractical to acquire [7, 130]. These constraints define a pragmatic problem setting: developing segmentation

methods that remain reliable and deployable under quality-, quantity-, and modality-limited supervision, while retaining clinically meaningful accuracy.

1.2 KEY CHALLENGES IN PRACTICAL MEDICAL IMAGE SEGMENTATION

Practical medical image segmentation (MIS) is embedded in the clinical workflow illustrated in Fig. 1.1, where image acquisition, delineation, treatment planning, and follow-up are tightly coupled stages. In routine practice, segmentation quality is not determined by algorithmic design alone, but by constraints arising along the patient journey, including imaging protocols, reporting requirements, and annotation workflows. In this thesis, we focus on three recurring limitations that most directly determine whether a segmentation model can be trained reliably and deployed robustly: (i) heterogeneous **annotation quality**, reflecting inter-observer variability during pathological confirmation and treatment planning; (ii) restricted **annotation quantity**, due to the time-intensive nature of expert delineation in pre-treatment evaluation and follow-up; and (iii) an **information gap** between strong and weak imaging modalities at inference, since advanced modalities (e.g., PET or MRI) may not be available at all clinical stages.. These factors are structurally linked to different phases of the patient journey. Limited annotation budgets often amplify label noise, while modality constraints during routine screening or follow-up reduce lesion conspicuity, increasing both inter-observer variability and the difficulty of learning stable decision boundaries. Together, they define the practical operating conditions under which MIS systems must function.

Challenge 1: Annotation quality (heterogeneous and imperfect supervision). High-quality voxel-wise annotation in the case of 3D images is labour-intensive and requires specialised expertise, leading to supervision that is frequently heterogeneous in practice [71, 119].

Labels, which indicate if an image area is either inside or outside a segmented region, may be produced by annotators with different levels of training, under time constraints, or using varying institutional guidelines, resulting in systematic discrepancies in target extent and boundary placement [96]. This is particularly pronounced for pathologies with weak contrast, infiltrative growth patterns, or ambiguous margins, where even expert delineations can vary [74]. Such variability is hard to model well as independent random noise; instead, it often exhibits structured patterns, including consistent under-segmentation of low-contrast regions, omission of small satellites, and over-smoothing of irregular boundaries.

From a machine learning perspective, low-quality annotations degrade the fidelity of the supervisory signal and can bias both the es-

timated shape prior and the calibration of predicted probabilities [108]. Over-penalising uncertain boundary regions may encourage conservative predictions, whereas training on systematically over-inclusive labels may inflate false positives and reduce clinical utility [172]. Importantly, naïvely mixing various styles of labels (e.g., conservative and inclusive) can cause the model to fit to annotator-specific idiosyncrasies rather than to imaging evidence [128]. A practical segmentation system must therefore be robust to label imperfections while still improving boundary consistency and lesion completeness in a manner aligned with clinical expectations.

Challenge 2: Annotation quantity (label scarcity and data efficiency). Beyond label quality, the amount of annotated data available for supervised training is typically limited [127]. This constraint is especially acute for volumetric imaging, where voxel-wise delineation across a full 3D scan can require substantial expert time [33]. Consequently, many datasets contain only a modest number of annotated volumes, with substantial class imbalance between foreground structures (the target anatomy or pathology to be segmented) and background (all remaining outside the target region). Limited annotated sample size and skewed class distributions hinder generalisation across scanners, acquisition settings, and patient populations, and can lead to unstable optimisation and poor robustness to distribution shift.

The quantity limitation is not merely a matter of collecting more cases. Clinical cohorts are often constrained by prevalence, inclusion criteria, and data governance, and annotator availability scales poorly. As a result, the effective training regime is frequently characterised by low-label budgets, partial annotations, or selective labelling of representative slices [171]. Under these conditions, a central methodological requirement is data efficiency: the ability to exploit unlabelled or weakly labelled images to learn transferable representations [68, 76], and to translate those representations into consistent improvements in downstream segmentation with minimal additional annotation effort. In 3D settings, this also entails leveraging volumetric continuity and anatomical coherence to avoid learning slice-wise shortcuts that do not persist across the volume [184, 188].

Challenge 3: Strong–weak modality gap (inference-time information asymmetry). A third constraint concerns the information available at deployment [154]. In many clinical scenarios, "strong" modalities (i.e., imaging sources that offer higher lesion-to-background contrast and richer tissue characterisation than the deployment modality) provide clearer lesion conspicuity and more reliable boundary cues, for example, due to functional contrast, contrast enhancement, or improved tissue differentiation [149]. However, these modalities may be unavailable or impractical in routine deployment because of

cost, scanner availability, examination time, contraindications (e.g., contrast agents), or protocol variability across institutions [7]. In contrast, "weak" modalities may be more available and faster to acquire; yet, these provide limited discriminative evidence for the target structure. This mismatch creates an information asymmetry: strong modalities may be accessible during method development or in subsets of training data, yet the deployed system must operate using weak modalities alone [157].

This setting imposes a strict constraint on model design. Methods that rely on multi-modal fusion at inference are not deployable when strong inputs are absent, while approaches that attempt to synthesise missing modalities introduce an additional source of error that can propagate into segmentation outputs [106]. A robust solution must therefore retain the benefits of strong modalities during training while maintaining a unimodal inference pathway [87]. Concretely, the challenge is to transfer clinically meaningful cues from strong to weak modalities in a way that improves delineation and reduces ambiguity on weak inputs, without embedding an implicit dependency on unavailable information at deployment time.

Problem setting derived from the challenges. Taken together, these challenges define a pragmatic problem formulation for MIS: learning methods that (i) remain stable under heterogeneous and imperfect labels, (ii) achieve strong performance with limited annotated volumes by exploiting unlabelled data, and (iii) support deployment under weak-modality-only inference while leveraging stronger modalities when they are available during training.

1.3 CONTRIBUTIONS

This thesis investigates MIS under three practical constraints that commonly arise in clinical deployment: heterogeneous **annotation quality** (corresponding to Challenge 1), limited **annotation quantity** (corresponding to Challenge 2), and **inference-time information asymmetry** (corresponding to Challenge 3) between strong and weak modalities. Correspondingly, the thesis makes three methodological contributions, each targeting one constraint while maintaining a deployment-oriented perspective.

Contribution 1 (C-1): Learning under heterogeneous annotation quality. We address the impact of imperfect and heterogeneous supervision by introducing a learning formulation that explicitly leverages the structure of label corrections observed in realistic annotation workflows. Rather than treating annotation errors as independent noise, the proposed approach exploits refinement-related signals to improve robustness to systematic label bias and to stabilise boundary delineation. This contribution targets improved segmentation reliability when train-

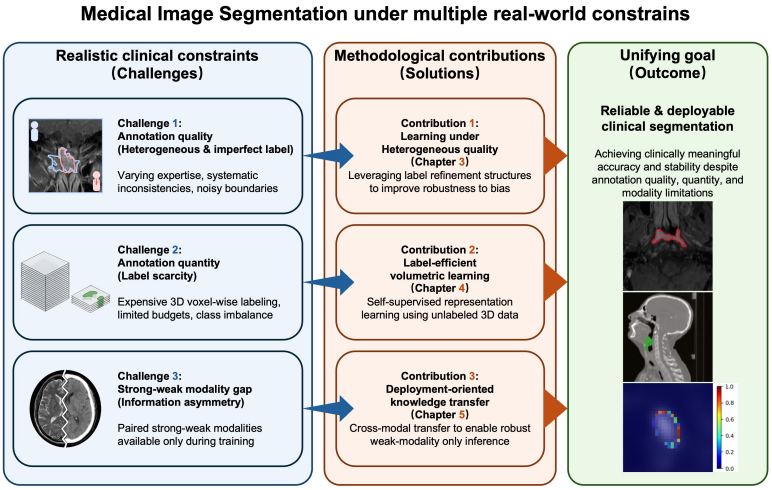


Figure 1.3: Challenges and contributions of the thesis.

ing labels are produced under varying expertise levels and delineation conventions.

Contribution 2 (C-2): Label-efficient learning under limited annotation quantity. To reduce reliance on large collections of expert voxel-wise annotations, we develop a label-efficient training strategy for volumetric segmentation that leverages unlabelled 3D data through self-supervised representation learning. By constructing a pretraining objective that exploits volumetric continuity and spatial regularities, the learned representations provide a strong initialisation for downstream segmentation and improve generalisation in low-label regimes. This contribution is designed for practical settings where annotation budgets are constrained, and foreground structures are scarce.

Contribution 3 (C-3): Deployment-oriented segmentation under strong–weak modality asymmetry. We consider scenarios in which strong modalities provide valuable delineation cues during development but are unavailable at deployment. To bridge this gap, we propose a cross-modal knowledge transfer strategy that exploits strong-modality information during training while using a weak-modality-only inference pathway. The resulting models retain part of the strong-modality benefit without introducing dependency on unavailable inputs, supporting robust deployment in weak-modality settings.

Unifying perspective. Together, these contributions provide a coherent framework for improving segmentation performance under quality-, quantity-, and modality-limited supervision. Rather than be-

ing three independent threads, they can be understood as addressing complementary constraints that typically arise at different stages of a practical ML engineering pipeline for MIS. Concretely, ReReNet (C-1) improves robustness to heterogeneous label quality by exploiting refinement structure when coarse and expert annotations co-exist; the self-supervised pre-training strategy (C-2) then improves data efficiency by leveraging large unlabelled volumetric archives to reduce dependence on dense expert delineations; finally, the cross-modal transfer component (C-3) exploits strong modalities during training to produce a deployable model that remains reliable when only a weak modality is available at inference. This sequential view explains the *unifying* goal: the thesis promotes principled use of auxiliary signals available during data creation or training (refinement structure, unlabelled volumes, and strong modalities) to progressively enhance reliability and deployability in the intended clinical setting. At the same time, these contributions do not need to be combined into a single end-to-end system; depending on dataset availability and deployment constraints, each component can also be adopted independently. The challenges and methodological contributions are summarised in Fig. 1.3.

1.4 THESIS ORGANISATION

This thesis is organised around the three already mentioned practical constraints for MIS—annotation quality, annotation quantity, and strong–weak modality asymmetry—and presents one methodological contribution for each.

Chapter 2 introduces the background required for the remainder of the thesis. It summarises the fundamental concepts in MIS, including common problem formulations, architectures, training objectives, and evaluation metrics, and provides a structured review of related literature aligned with the three constraint settings considered in this work.

Chapter 3 addresses the challenge of heterogeneous **annotation quality**. It presents the proposed learning formulation that leverages refinement-related signals to improve robustness under imperfect supervision, followed by experimental evaluation and analysis. This chapter develops **C-1**.

Chapter 4 focuses on **annotation quantity** and label efficiency. It introduces the proposed self-supervised representation learning strategy for volumetric data and demonstrates its effectiveness for 3D segmentation under limited labelled data. This chapter develops **C-2**.

Chapter 5 considers the **strong–weak modality** setting, where strong modalities may be available during training but not at deployment. It presents the proposed cross-modal knowledge transfer approach for weak-modality-only inference and evaluates its performance in clinically motivated scenarios. This chapter develops **C-3**.

Finally, Chapter 6 summarises the main findings, discusses limitations, and outlines directions for future research.

2 BACKGROUND

This chapter establishes the technical foundations and reviews prior work relevant to the three methodological contributions of this thesis. We begin by summarising the core concepts of MIS, including commonly used problem formulations, network architectures, training objectives, and evaluation metrics (Section 2.1). The subsequent sections provide targeted literature reviews aligned with each of the three constraint settings addressed in this thesis. Section 2.2 discusses existing strategies for handling heterogeneous annotation quality and learning under imperfect supervision, setting the stage for the recurrent refinement approach presented in Chapter 3. Section 2.3 reviews self-supervised and label-efficient learning methods for volumetric medical imaging, motivating the rotation-based pretraining strategy introduced in Chapter 4. Finally, Section 2.4 surveys multi-modal fusion and cross-modal knowledge transfer techniques, providing context for the strong-to-weak modality transfer framework developed in Chapter 5. We then formalise the problem and the three constraints that motivate the thesis contributions (Section 2.5). A brief summary (Section 2.6) concludes the chapter.

2.1 FUNDAMENTALS OF MEDICAL IMAGE SEGMENTATION

Image segmentation is a fundamental computer vision task that aims to decompose an image into a set of disjoint, coherent regions according to predefined criteria, such as appearance, geometry, or semantic meaning. Depending on the target granularity, segmentation can range from low-level grouping of pixels with similar intensity or texture to high-level delineation of objects of interest by assigning each pixel to a specific class or label [148].

Segmentation problems are commonly categorised according to the prediction target (see Fig. 2.1).

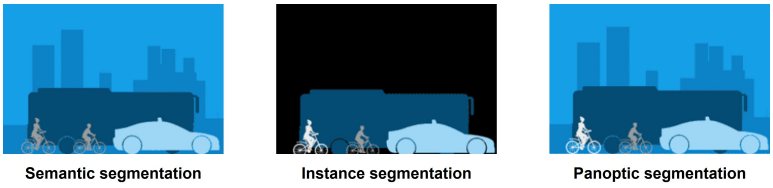


Figure 2.1: Three types of segmentation problems. Image adapted from [75].

Semantic segmentation assigns a category label to every pixel/voxel, producing a dense map that partitions the image into labelled regions. **Instance segmentation** extends this setting by additionally separating individual object instances within the same category, yielding a distinct mask for each instance. **Panoptic segmentation** unifies these formulations by jointly predicting per-pixel category labels and instance identities for countable objects ("things") while also labelling amorphous regions ("stuff") in a single representation [112].

MIS, a major application domain of image segmentation, aims to partition an image into meaningful anatomical structures or pathological regions (e.g., organs, lesions, blood vessels) from background and is a core component of many clinical practices [148]. Within medical image analysis pipelines, semantic segmentation is often used to localise task-relevant anatomy or lesions and to suppress irrelevant structures, thereby supporting more accurate boundary delineation and downstream decision-making (e.g., organ or lesion contouring) [5].

Developing algorithms that accurately delineate organs or lesions typically requires task-specific imaging data together with reference annotations that serve as ground truth. Such data are acquired using a range of clinical modalities, including X-ray, positron emission tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US) [154]. Prior to the widespread adoption of deep learning [77], MIS was primarily addressed with hand-crafted or rule-based techniques, such as edge detection, template matching, region growing, graph-cut formulations, active contours, and classical machine-learning methods [148].

This section provides a concise overview of fundamental concepts and commonly adopted technical paradigms in MIS, with an emphasis on deep learning approaches that have become predominant in recent years. Specifically, we summarise typical problem formulations, representative neural network architectures, widely used training objectives (loss functions), and standard evaluation metrics. Classical non-deep-learning pipelines (e.g., graph-based methods, thresholding, or hand-crafted feature engineering) are not discussed in detail here, as contemporary research and most state-of-the-art systems increasingly rely on deep neural networks for segmentation. The material in this section establishes the conceptual and technical background required to follow the methodological chapters that follow.

2.1.1 *Common Network Architectures*

Deep learning approaches to MIS are predominantly built upon fully convolutional, encoder–decoder paradigms that map an input image (or volume) to a dense, pixel-/voxel-wise prediction (see Fig. 2.2). Early fully convolutional networks (FCNs) [99] established the principle of learning a hierarchical representation via successive down-sampling

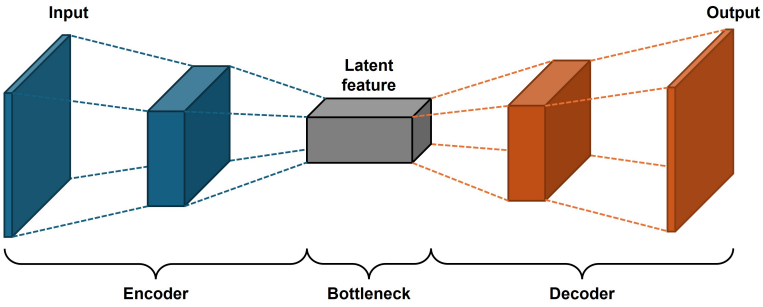


Figure 2.2: Encoder–Decoder architecture.

and restoring spatial resolution through up-sampling to obtain fine-grained segmentation outputs. Within this family, U-Net [120] has emerged as the most widely adopted backbone in medical imaging due to its symmetric design and skip connections, which effectively fuse high-resolution localisation cues with low-resolution semantic context and perform well even under limited annotation regimes. Beyond the canonical U-Net, common architectural lines include (i) alternative semantic segmentation backbones (e.g., DeepLab-style atrous/dilated designs [27] and encoder–decoder variants such as SegNet [8]), (ii) modifications to the encoder backbone (e.g., residual or dense blocks, 3D convolutions or separable 3D operators for volumetric data), and (iii) modules that enhance feature fusion and representation, such as attention mechanisms, multi-scale context aggregation (e.g., pyramid or atrous spatial pooling), and multi-modal fusion with multiple encoders [5, 148].

More recently, transformer-based components have been integrated either as complements to convolutional U-shaped networks or as standalone U-shaped transformer backbones, aiming to better capture long-range dependencies while maintaining localisation through hierarchical decoding and skip connections [6, 81, 125]. Probabilistic extensions have also been explored to model ambiguity and uncertainty in predictions [74, 108]. These architectural directions provide the context for the following sections, where we first detail the standard U-Net and then review representative classes of U-Net-centric improvements.

Fully Convolutional Networks (FCNs). FCNs were among the first deep architectures to enable end-to-end *dense* prediction for semantic segmentation by removing the fixed-size constraint of conventional image classifiers [99]. The central idea is to reinterpret a classification ConvNet as a per-pixel predictor by replacing the terminal fully connected layers with convolutional layers (equivalently, viewing a fully connected layer as a convolution whose kernel covers the entire input support).

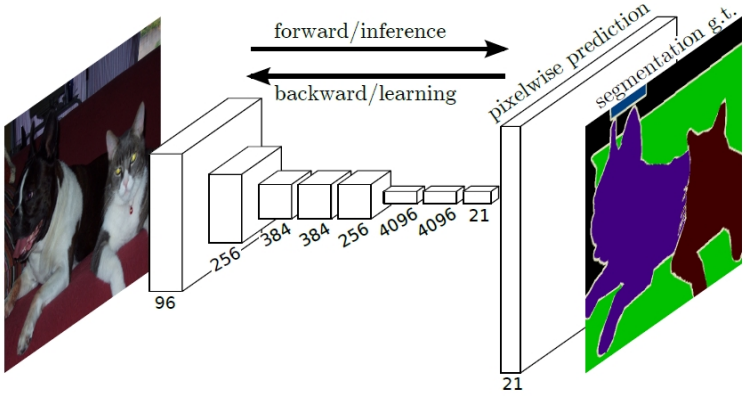


Figure 2.3: Fully Convolutional Networks (FCNs). Image taken from [99].

This "convolutionalisation" yields a network that can accept an input of arbitrary spatial size and produce a correspondingly sized output score map, with inference and learning performed efficiently in a whole-image manner rather than patch-by-patch (see Fig. 2.3).

Because typical ConvNet encoders employ pooling and strided convolutions, the raw output score maps are spatially coarse (e.g., with an effective stride of 32 pixels in VGG-style backbones). FCNs address this by introducing *in-network upsampling* using learnable transposed convolution (deconvolution) layers that convert low-resolution score maps into dense predictions at the original input resolution.

However, naïve single-step upsampling from the coarsest resolution tends to produce overly smooth boundaries. To mitigate the inherent "what-versus-where" tension in semantic segmentation, FCNs further introduce *skip connections* that fuse deep, semantically rich but low-resolution features with shallower, high-resolution features that preserve local appearance cues (see Fig. 2.4). This multi-scale fusion leads

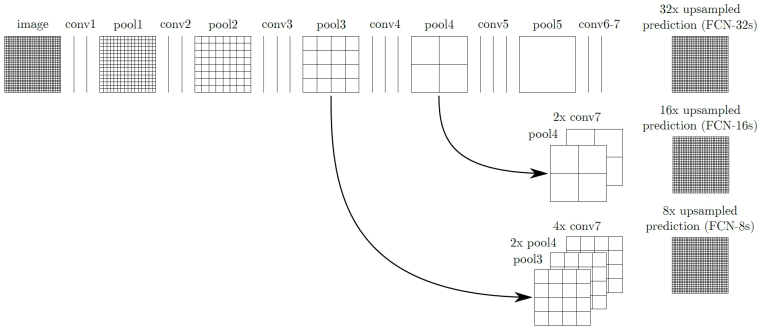


Figure 2.4: Illustration of skip connections in FCNs. Image take from [99].

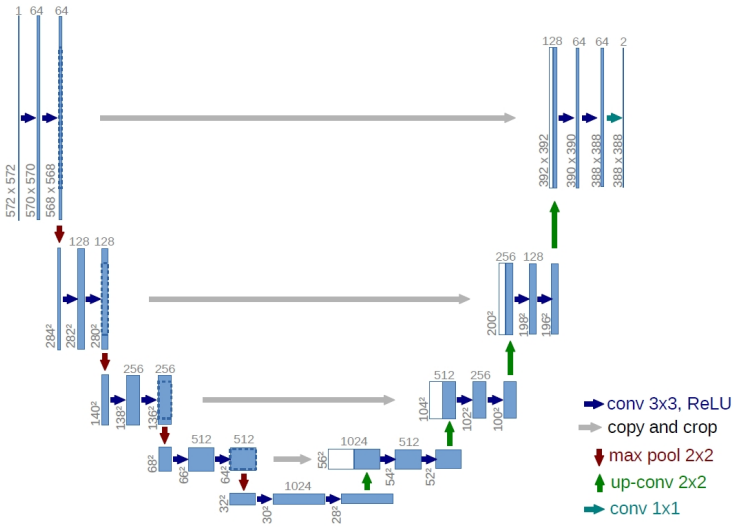


Figure 2.5: U-Net architecture. Image taken from [120].

to the well-known FCN-32s/FCN-16s/FCN-8s variants, where intermediate predictions (e.g., from `pool4` and `pool3`) are combined with the final-layer prediction to progressively refine spatial detail.

Conceptually, FCNs establish the encoder–decoder paradigm for segmentation: an encoder produces hierarchical representations with increasing receptive field, while a decoder restores spatial resolution to obtain pixelwise predictions. This design principle underpins many subsequent medical segmentation architectures, most notably U-Net, which generalises skip-based feature fusion within a symmetric encoder–decoder topology.

2.1.2 *U-Net and Its Extensions*

U-Net [120] is a seminal encoder–decoder architecture that has become the de facto baseline for MIS. Its defining characteristic is the symmetric "U-shaped" topology, where a contracting path progressively aggregates contextual information via down-sampling, and an expanding path restores spatial resolution via up-sampling. Crucially, U-Net introduces skip connections that directly propagate high-resolution encoder features to the corresponding decoder stages, as shown in Fig. 2.5. This design mitigates the loss of fine spatial details induced by pooling, enabling accurate localisation and boundary delineation—properties that are particularly important in medical imaging, where targets often exhibit low contrast, ambiguous margins, and substantial inter-patient variability. Owing to its favourable accuracy–efficiency trade-off and its

compatibility with diverse convolutional and attention/state-space extensions, U-Net has served as a standard backbone across a wide range of modalities and anatomies. In this thesis, U-Net provides the core architectural template: all subsequent segmentation models are built upon, or are direct extensions of, the U-Net paradigm, allowing methodological contributions to be isolated and compared under a consistent backbone.

The original U-Net builds on a fully convolutional encoder–decoder design and was explicitly motivated by biomedical settings where annotated data are limited [120]. To improve generalisation under scarce supervision, it combines standard training-time augmentations (e.g., rotations, intensity perturbations) with elastic deformations that mimic plausible tissue variability. The network comprises three main components. First, the contracting (encoder) path captures contextual information through a sequence of blocks, each typically formed by two 3×3 convolutions with ReLU activations followed by 2×2 max-pooling. The pooling operations progressively enlarge the receptive field while controlling computational cost. Second, a bottleneck stage at the lowest resolution applies additional 3×3 convolutions to form high-level semantic representations. Third, the expanding (decoder) path restores spatial resolution by repeated up-sampling (commonly via 2×2 transposed convolutions) followed by 3×3 convolutions and ReLU activations to refine predictions.

A key challenge in such encoder–decoder pipelines is that deep representations tend to sacrifice spatial precision, which is critical for delineating anatomical boundaries. U-Net addresses this by introducing skip connections between encoder and decoder feature maps at corresponding scales. These connections concatenate high-resolution encoder features with decoder features, injecting localisation cues and enabling the decoder to recover fine details. Finally, a 1×1 convolution maps the fused representation to the target label space. For inference on large images, U-Net also adopts an overlap-tile strategy to reduce border artefacts that can arise from missing context near image boundaries. The original U-Net achieved strong performance and efficient inference on many biomedical benchmarks, establishing the U-shaped architecture and skip-connection design as a standard backbone in MIS. We next review the extensions of U-Net that have been proposed in the literature from different architectural perspectives.

2.1.2.1 *CNN-Based Extensions*

Skip-connection redesign and enrichment. Beyond the original one-to-one skips, later models either increase skip density or explicitly process skip features to reduce the encoder–decoder semantic gap and suppress irrelevant activations. UNet++ [187] introduces nested and dense skip pathways to enable multi-scale aggregation across interme-

diate nodes. UNet 3+ [61] connects each decoder stage to all encoder stages (and earlier decoder stages) via explicit up/down-sampling and per-skip convolutions to realise full-scale fusion. BiO-Net adds backwards (decoder-to-encoder) skip paths and a recursive traversal to iteratively refine representations. Attention U-Net [116] inserts attention gates on skip features to emphasise salient regions. BCDU-Net employs bi-directional ConvLSTM during skip fusion to model non-linear dependencies between encoder and decoder features.

Backbone modernisation with stronger CNN blocks. A natural line of improvement replaces the vanilla U-Net convolutional blocks with more powerful modules that enhance optimisation stability, receptive field coverage, or parameter efficiency. Residual connections (ResNet-style) [54] facilitate gradient flow through deeper encoder–decoder stacks [180] and have become standard in volumetric variants such as V-Net [111]. Dense connectivity (DenseNet-style [60]) promotes feature reuse and implicit deep supervision, as exemplified by H-DenseUNet [83] for hybrid 2D/3D liver segmentation. Multi-resolution blocks, such as those in MultiResUNet [64], stack 3×3 convolutions in an inception-like arrangement to capture multi-scale cues without explicitly enlarging the kernel size. Additional variants include deformable convolutions for adaptive spatial sampling [40], separable 3D convolutions for reduced memory footprint [30], and recurrent convolutional blocks for iterative feature refinement [114]. At the bottleneck, where spatial resolution is lowest, attention-style modules capture long-range dependencies within compressed feature maps, while atrous spatial pyramid pooling (ASPP) aggregates multi-scale context without a proportional increase in parameters.

2.1.2.2 *Transformer-Based Extensions*

The success of Transformers [141] in natural language processing has motivated their adoption in (medical) image segmentation [41, 52], where self-attention mechanisms can capture long-range spatial dependencies that are difficult for local convolutions to model. The Vision Transformer (ViT) [41] tokenises images into patches and applies a Transformer encoder to capture global context, while leveraging CNN skip features for precise localisation (see Fig. 2.6).

At the heart of Transformer lies the *self-attention* mechanism, which computes pairwise interactions among all tokens in a sequence. Given an input sequence $\mathbf{X} \in \mathbb{R}^{N \times d}$ of N tokens each with embedding dimension d, self-attention first projects $\mathbf{X}$ into queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{2.1}$$

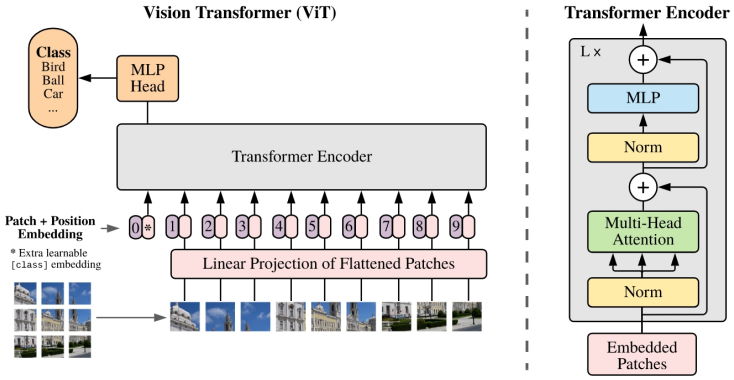


Figure 2.6: ViT architecture. Image taken from [41].

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices. The attention weights are computed by measuring the compatibility between queries and keys via scaled dot-product, followed by a softmax normalisation:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right), \tag{2.2}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the attention weight matrix and each row sums to one. The scaling factor $\sqrt{d_k}$ prevents the dot products from growing large in magnitude, which would push the softmax into regions of extremely small gradients. The self-attention output is then computed as a weighted sum of the value vectors:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{A}\mathbf{V}. \tag{2.3}$$

To capture diverse relationships, Transformers employ *multi-head attention* (MHA), which runs h parallel attention heads with independent projections and concatenates their outputs:

$$\begin{aligned} \text{MHA}(\mathbf{X}) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_O, \\ \text{head}_i &= \text{Attention}(\mathbf{X}\mathbf{W}_Q^i, \mathbf{X}\mathbf{W}_K^i, \mathbf{X}\mathbf{W}_V^i), \end{aligned} \tag{2.4}$$

where $\mathbf{W}_O \in \mathbb{R}^{hd_k \times d}$ is a learnable output projection matrix that maps the concatenated h attention heads back to the model embedding dimension d. Multi-head attention enables the model to jointly attend to information from different representation subspaces at different positions. However, the pairwise computation has $\mathcal{O}(N^2)$ complexity, which becomes prohibitive for high-resolution images where N can reach tens of thousands of tokens.

Two main integration strategies have emerged for adapting Transformers to MIS: hybrid CNN–Transformer architectures and standalone Transformer backbones.

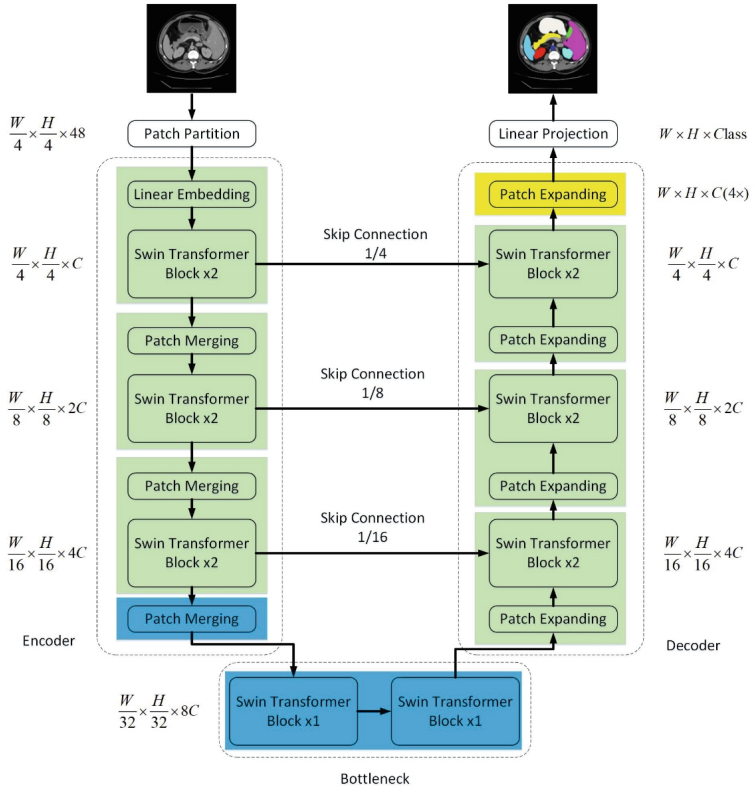


Figure 2.7: Swin-Unet architecture. Image taken from [19].

Hybrid CNN–Transformer U-Nets. These architectures retain CNN encoders and decoders for local detail extraction and boundary recovery, while inserting Transformer modules to model long-range dependencies and cross-scale interactions. TransUNet [25] tokenises CNN feature maps and applies a ViT-style encoder to inject global context, while leveraging CNN skip features for precise localisation. TransBTS [149] inserts Transformer blocks at the bottleneck of a 3D CNN encoder–decoder to balance volumetric locality with global modelling under memory constraints. CoTr [162] employs deformable attention to reduce the quadratic cost of standard self-attention and to facilitate multi-scale fusion. UCTransNet [144] introduces a channel-wise Transformer module to fuse multi-scale encoder features and mitigate scale-wise semantic gaps before decoding.

Standalone Transformer U-shaped backbones. An alternative approach replaces most convolutional processing with hierarchical Transformer blocks while preserving a U-shaped multi-stage design.

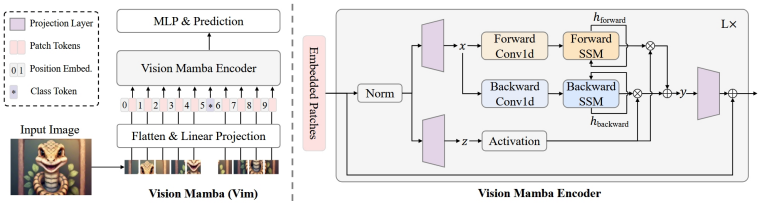


Figure 2.8: Vision Mamba architecture. Image taken from [189].

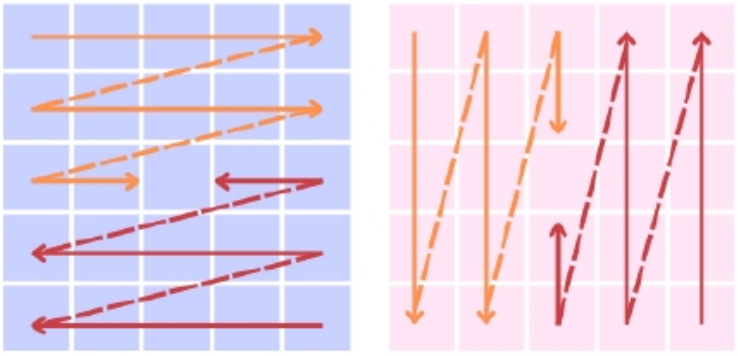
MedT adopts gated axial-attention to improve data efficiency under limited supervision. Swin-Unet [19] leverages shifted-window self-attention with patch merging and expanding operations to construct a scalable hierarchical encoder–decoder (see Fig. 2.7). MISSFormer combines overlapping patch embeddings with locality-enhanced feed-forward components (e.g., depth-wise convolutions) and Transformer-based context bridges to stabilise local continuity and multi-scale fusion.

2.1.2.3 *Mamba-Based Extensions*

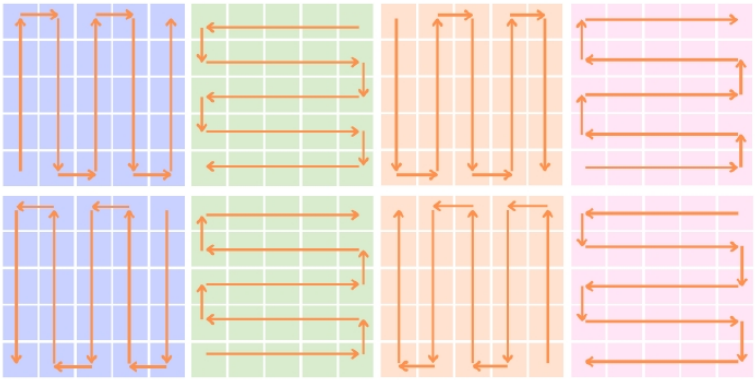
Recent advances in structured state space models (SSMs) have introduced *Mamba*, a selective SSM that performs sequence modelling via input-dependent gating over a state-space recurrence (i.e., a linear-time "scan") rather than pairwise token interactions, with computational and memory costs that scale linearly with sequence length [10].

As a pioneering work in dense prediction, Vision Mamba [189] extends Mamba to two-dimensional feature maps via bidirectional state space models (see Fig. 2.8). Later Vision Mamba variants extend SSMs to two-dimensional feature maps via various scanning mechanisms, including selective scan 2D (SS2D) [98], and Zigzag [59], enabling long-range dependency modelling on images without the quadratic cost of Transformers. This property is particularly attractive in MIS, where high-resolution 2D slices and volumetric 3D data often make attention-based designs expensive.

Integration patterns in U-Net. For MIS, Mamba blocks are introduced into U-Net-like encoder–decoder pipelines: (i) inserting Mamba blocks before the first encoder stage to enrich early representations; (ii) augmenting or replacing encoder stages with Vision Mamba blocks (e.g., VSS/SS2D-based blocks) to strengthen global context modelling; (iii) applying Mamba blocks in the bottleneck to increase receptive field while preserving the convolutional inductive bias elsewhere; and (iv) incorporating Mamba modules along skip pathways to refine the fusion of encoder and decoder features. These variants retain the multi-scale hi-



(a) SS2D operator.



(b) Zigzag operator.

Figure 2.9: Two representative 2D scanning strategies for adapting Mamba to image data. Images taken from [10].

erarchy and skip fusion of U-Net while substituting parts of the feature transformation with SSM-based mixers.

Scanning mechanisms for 2D/3D segmentation. A central architectural component of Vision Mamba is the *scan* that converts a 2D (or 3D) tensor into sequences processed by SSMs. Beyond early bidirectional scans [189] (i.e., scanning the feature map in both forward and reverse order and fusing the two passes so each token receives context from both directions), SS2D [98] performs directional scanning (top-to-bottom, left-to-right, and reverse directions) and merges the resulting representations to emulate the effect of global context aggregation (see Fig. 2.9a).

Subsequent variants aim to mitigate patch discontinuities and improve locality/globality trade-offs, including continuous 2D scans and zigzag scans [59] (see Fig. 2.9b) to better preserve neighbourhood conti-

nuity; local scans to restrict scanning within windows; multi-head and multi-path scans to diversify routes and fuse them adaptively; omnidirectional scan to combine refined local features (e.g., via depthwise convolution) with complementary global scan streams; hierarchical scan to aggregate information at multiple granularities; and segmentation-oriented designs such as tri-orientated spatial scanning (e.g., scanning over height, width, and channels); and slice-aware scans for volumetric inputs. These scanning operators form the implementation basis for many Mamba-based segmentation backbones.

Representative Mamba-based segmentation backbones. Building on these scanning operators, a rapidly expanding family of Mamba–U-Net variants has emerged for both 2D and 3D MIS. VM-UNet [122] tokenises images via patch embedding and uses stacked Vision Selective Scan (VSS) blocks with patch merging in the encoder and patch expansion in the decoder, maintaining U-Net-style skip connections while using SSM-based mixing within stages. Mamba-UNet [153] similarly follows an encoder–decoder structure but converts feature maps into sequences for VSS processing and reconstructs spatial resolutions via upsampling in the decoder, with skips connecting corresponding scales. LKM-UNet [147] combines depthwise convolution with large-kernel Vision Mamba blocks that integrate pixel-wise SSM processing and patch-level modelling, using bidirectional Mamba blocks to enhance context aggregation while decoding with convolutional upsampling and residual connections. H-Vmunet [159] introduces high-order VSS modules by modifying SS2D into an H-SS2D design and employs attention-like bridges (channel and spatial) for encoder–decoder information fusion. SegMamba [163] targets volumetric tumour analysis with a tri-orientated spatial Mamba block that scans along multiple spatial/channel orientations and directions, explicitly adapting scanning to 3D contexts.

Hybrid and weak/semi-supervised extensions. Several architectures explicitly hybridise Mamba with convolutional and/or Transformer components. Weak-Mamba-UNet [152] combines a CNN U-Net branch for local structure, a Transformer branch for global context, and a Visual Mamba branch for efficient long-range dependency modelling, and is evaluated under scribble-based weak supervision with losses defined over pseudo-labels and scribble constraints. Semi-Mamba-UNet [102] adopts a semi-supervised setting that combines supervised losses (e.g., Dice and cross-entropy) with pseudo-label-based semi-supervised terms and pixel-level contrastive objectives. Other hybrid designs include Swin-UMamba [93], which couples pretrained Mamba-style encoders with U-Net decoding components, and adapter-style integrations where Mamba blocks are used to adapt large prompt-

able models (e.g., adding Mamba blocks atop Transformer encoders and using 3D convolutional decoders).

Efficiency considerations. Mamba-based segmentation models can offer favourable accuracy–efficiency trade-offs relative to Transformer-heavy backbones. LightM-UNet [88], for instance, achieves competitive segmentation performance with substantially fewer parameters and GFLOPs than Swin UNETR [51] on representative datasets, while Mamba-UNet [153] has been shown to outperform Swin-UNet [19] on multi-organ benchmarks (e.g., ACDC and Synapse CT) in both Dice and Hausdorff distance (defined next in Section 2.1.4). In such settings, SSM-based mixers can thus serve as a practical alternative to attention mechanisms when computational budgets or input resolutions make self-attention costly.

2.1.3 *Common Loss Functions*

The choice of loss function plays a critical role in training segmentation networks, as it directly shapes the optimisation landscape and influences how the model balances different error types. In MIS, loss functions must address several recurring challenges: severe class imbalance (foreground lesions often occupy a small fraction of the image), ambiguous or fuzzy boundaries in the input images that we want to segment, and considerable variability in target size and shape of the segments to be created across patients. Here, "class imbalance" refers to voxel-/pixel-level imbalance (i.e., the foreground occupies far fewer pixels/voxels than the background within each image), rather than imbalance in the number of object instances across classes.

Below, we review the main families of losses commonly employed. Let $\Omega = \{1, \dots, H\} \times \{1, \dots, W\}$ denote the discrete spatial grid of pixel locations (and $\Omega = \{1, \dots, H\} \times \{1, \dots, W\} \times \{1, \dots, D\}$ for 3D volumes), with $|\Omega| = H \times W$ (or $H \times W \times D$). Throughout, let $y \in \{0, 1\}^{|\Omega|}$ denote the ground-truth binary mask (that we train the model on) and $\hat{p} \in [0, 1]^{|\Omega|}$ the predicted foreground probability map over the spatial domain Ω; for multi-class settings with C classes, $y_{i,c}$ and $\hat{p}_{i,c}$ represent the one-hot label and predicted probability for class $0 \le c \le C - 1$ at location i, respectively.

Cross-entropy losses. Cross-entropy is the most straightforward choice, treating segmentation as independent per-pixel classification. Binary cross-entropy (BCE) penalises deviations between predicted probabilities and binary labels:

$$\mathcal{L}_{\mathrm{BCE}}(y, \hat{p}) = -\sum_{i \in \Omega} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big]. \tag{2.5}$$

Its multi-class counterpart, categorical cross-entropy (CE), sums over all classes:

$$\mathcal{L}_{\mathrm{CE}}(y, \hat{p}) = -\sum_{i \in \Omega} \sum_{c=1}^{C} y_{i,c} \log \hat{p}_{i,c}. \tag{2.6}$$

When foreground pixels are rare, the loss is dominated by the abundant background class. Weighted cross-entropy alleviates this by assigning higher weights α_c to under-represented classes:

$$\mathcal{L}_{\mathrm{WCE}}(y, \hat{p}) = -\sum_{i \in \Omega} \sum_{c=1}^{C} \alpha_c \, y_{i,c} \log \hat{p}_{i,c}. \tag{2.7}$$

Overlap-based losses. Rather than aggregating pixel-wise errors, overlap-based losses directly optimise set-similarity measures between predicted and ground-truth regions, making them naturally robust to class imbalance. The soft Dice loss, derived from the Dice similarity coefficient (detailed in Section 2.1.4), is perhaps the most widely used in medical imaging:

$$\mathcal{L}_{\mathrm{Dice}}(y, \hat{p}) = 1 - \frac{2 \sum_{i \in \Omega} y_i \hat{p}_i}{\sum_{i \in \Omega} y_i + \sum_{i \in \Omega} \hat{p}_i}. \tag{2.8}$$

For multi-class tasks, a common practice is to average per-class Dice losses so that each class contributes equally regardless of its prevalence. The soft Jaccard (IoU) loss follows a similar rationale but uses the intersection-over-union formulation:

$$\mathcal{L}_{\mathrm{IoU}}(y, \hat{p}) = 1 - \frac{\sum_{i \in \Omega} y_i \hat{p}_i}{\sum_{i \in \Omega} y_i + \sum_{i \in \Omega} \hat{p}_i - \sum_{i \in \Omega} y_i \hat{p}_i}. \tag{2.9}$$

Tversky loss generalises Dice by introducing asymmetric penalties for false positives (FP) and false negatives (FN):

$$\mathcal{L}_{\mathrm{Tv}}(y, \hat{p}) = 1 - \frac{\sum_i y_i \hat{p}_i}{\sum_i y_i \hat{p}_i + \alpha \, \mathrm{FP} + \beta \, \mathrm{FN}}, \tag{2.10}$$

where $\mathrm{FP} = \sum_i (1 - y_i) \hat{p}_i$ and $\mathrm{FN} = \sum_i y_i (1 - \hat{p}_i)$. Setting $\alpha = \beta = 0.5$ recovers Dice; increasing β penalises missed detections more heavily, which is often desirable in clinical applications where under-segmentation can have serious consequences.

Focal and hard-example mining losses. In highly imbalanced settings, even weighted cross-entropy may be overwhelmed by easy-to-classify background pixels. Focal loss addresses this by down-weighting well-classified examples through a modulating factor $(1 - \hat{p})^{\gamma}$:

$$\mathcal{L}_{\mathrm{Focal}}(y, \hat{p}) = -\sum_{i \in \Omega} \Big[\alpha \, y_i (1 - \hat{p}_i)^{\gamma} \log \hat{p}_i + (1 - \alpha)(1 - y_i) \hat{p}_i^{\gamma} \log(1 - \hat{p}_i) \Big].$$

(2.11)

When $\gamma > 0$, confident predictions contribute less to the loss, focusing learning on hard, ambiguous pixels. Focal Tversky loss applies the same principle to overlap-based objectives by raising the Tversky loss to a power γ, combining the benefits of asymmetric FP/FN control with hard-example emphasis.

Boundary- and distance-aware losses. Region-based losses treat all misclassified pixels equally, potentially under-penalising errors near object boundaries where clinical accuracy matters most. Boundary-aware losses incorporate geometric information by weighting errors according to their distance from the ground-truth contour. A representative example is the boundary loss, which integrates the predicted probability against a signed distance transform $d(i)$ of the ground truth contour:

$$\mathcal{L}_{\text{Boundary}}(y, \hat{p}) = \sum_{i \in \Omega} \hat{p}_i \, d(i), \tag{2.12}$$

where $d(i) = \phi_y(i)$ denotes the signed distance transform (SDF) computed from the ground-truth mask y (positive outside the object and negative inside, with $d(i) = 0$ on ∂Y). Because this formulation is linear in $\hat{p}$ and may be unstable alone, it is typically combined with a region loss (e.g., Dice or CE) using a training-schedule for the loss weights (e.g., gradually increasing the boundary-term coefficient over epochs). Hausdorff-distance-based losses build on distance maps but replace the non-smooth set operations (and the discrete boundary extraction from predictions) with smooth relaxations so that gradients can be propagated to the network outputs.

Compound objectives. In actual workflows, practitioners often combine losses from different families to leverage complementary strengths. A widely adopted combination is Dice plus cross-entropy:

$$\mathcal{L}(y, \hat{p}) = \lambda \, \mathcal{L}_{\text{CE}}(y, \hat{p}) + (1 - \lambda) \, \mathcal{L}_{\text{Dice}}(y, \hat{p}), \tag{2.13}$$

where Dice encourages region overlap, and CE provides well-calibrated per-pixel gradients. Other common variants include BCE+Dice for binary tasks, CE+IoU, and Dice combined with boundary losses to sharpen contour predictions.

2.1.4 *Evaluation Metrics*

Quantitative evaluation is essential for comparing segmentation algorithms and assessing their clinical utility in absolute or relative terms. These evaluation criteria are conceptually related to the loss functions

used for training: both quantify discrepancies between predictions and references, and ideally, the optimisation objective should align with the evaluation measures.

Unlike natural-image segmentation benchmarks where a single accuracy number may suffice, MIS demands a multi-faceted evaluation: clinicians care not only about overall region overlap but also about boundary precision, boundary smoothness, volumetric accuracy, and whether lesions are detected at all. This subsection surveys the evaluation metrics most commonly reported in the MIS literature. Throughout, let Y and $\hat{Y}$ denote the sets of ground-truth and predicted foreground voxels, respectively, and let ∂Y, $\partial \hat{Y}$ be their respective boundaries.

Region-overlap metrics. The simplest way to assess segmentation quality is to measure how well the predicted region overlaps with the ground truth. The **Dice similarity coefficient (DSC)**, also known as the F_1 score, is by far the most widely reported metric in medical imaging:

$$\mathrm{DSC}(Y, \hat{Y}) = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}. \tag{2.14}$$

DSC ranges from 0 (no overlap) to 1 (perfect agreement) and can be interpreted as the harmonic mean of precision and recall. A closely related measure is the **Jaccard index** (intersection-over-union, IoU):

$$\mathrm{IoU}(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}. \tag{2.15}$$

Although DSC and IoU are monotonically related, DSC tends to yield higher numerical values for the same prediction, which is worth noting when comparing results across papers. Both metrics are inherently size-dependent: since DSC/IoU are normalised set-overlap measures, their scores change much more when the involved sets (Y and $\hat{Y}$) are small—missing or adding only a few voxels can cause a large relative drop—whereas the same absolute voxel error produces only a minor change when Y and $\hat{Y}$ are large.

Sensitivity, specificity, and precision. Overlap metrics conflate over-segmentation and under-segmentation into a single number. To disentangle these failure modes, it is useful to examine **sensitivity** (recall) and **precision** separately. Sensitivity measures the fraction of true foreground voxels that are correctly detected:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2.16}$$

where TP and FN denote true positives and false negatives, respectively. Low sensitivity indicates under-segmentation—the model misses parts

of the target. Precision, conversely, measures the fraction of predicted foreground voxels that are correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.17}$$

Low precision signals over-segmentation—the model includes spurious regions. **Specificity** (true negative rate) is sometimes reported but is less informative when the background dominates, as is typical in lesion segmentation.

Boundary and distance metrics. Overlap-based scores can be high even when predicted boundaries deviate noticeably from the ground truth, particularly for large, blob-like targets. For applications such as radiotherapy planning, where contour accuracy directly affects treatment margins, boundary-aware metrics are indispensable. In practice, however, evaluation metrics such as HD/HD95 are often non-smooth or expensive to optimise directly, so differentiable surrogates (e.g., Dice- or boundary-weighted losses) are used while these metrics are reserved for evaluation. The **Hausdorff distance (HD)** quantifies the worst-case boundary error:

$$\text{HD}(\partial Y, \partial \hat{Y}) = \max\Big\{ \max_{a \in \partial Y} d(a, \partial \hat{Y}),\ \max_{b \in \partial \hat{Y}} d(b, \partial Y) \Big\}, \tag{2.18}$$

where $d(a, B) = \min_{b \in B} \|a - b\|$ is the point-to-set Euclidean distance (i.e., the Euclidean distance transform of the set B evaluated at a). Because a single outlier can dominate HD, the **95th-percentile Hausdorff distance (HD95)** is often preferred:

$$\text{HD95}(\partial Y, \partial \hat{Y}) = \max\Big\{ P_{95}\big(\{d(a, \partial \hat{Y})\}_{a \in \partial Y}\big),\ P_{95}\big(\{d(b, \partial Y)\}_{b \in \partial \hat{Y}}\big) \Big\}. \tag{2.19}$$

For a more global view of boundary accuracy, and also to remove outlier effects, the **average symmetric surface distance (ASSD)** averages all point-to-surface distances:

$$\text{ASSD}(\partial Y, \partial \hat{Y}) = \frac{1}{|\partial Y| + |\partial \hat{Y}|} \Big(\sum_{a \in \partial Y} d(a, \partial \hat{Y}) + \sum_{b \in \partial \hat{Y}} d(b, \partial Y) \Big). \tag{2.20}$$

Lower values indicate tighter boundary agreement; ASSD is particularly informative when clinical tolerance is specified in millimetres, as it quantifies surface deviation in physical space rather than voxel units.

Volume-based metrics. In longitudinal studies or treatment-response assessment, the total lesion volume may be more clinically

relevant than voxelwise overlap. The **relative volume error (RVE)** captures systematic over- or under-estimation:

$$\mathrm{RVE}(Y, \hat{Y}) = \frac{|\hat{Y}| - |Y|}{|Y|} \times 100\%. \tag{2.21}$$

A positive RVE indicates over-segmentation; a negative value indicates under-segmentation. When the sign is unimportant, the absolute relative volume error |RVE| is reported instead.

Case-level detection. For screening applications, the primary question is whether a lesion is present at all, rather than how precisely it is delineated. Case-level metrics treat each scan (or patient) as a single sample: a case is positive if the ground-truth lesion volume exceeds zero and predicted positive if the model outputs any foreground (or a connected component above a size threshold). Sensitivity, specificity, and precision can then be computed over cases rather than voxels, providing a detection-oriented perspective that complements voxelwise evaluation.

In practice, no single metric tells the whole story, and a comprehensive evaluation typically combines several complementary measures. Notably, most metrics are unambiguous at their extremes: DSC/IoU reaching 1 (or distance-based errors reaching 0) implies an exact match to the reference. Away from these perfect values, different metrics capture different error modes, so introducing multiple measures helps interpret how two methods differ. Overlap scores such as DSC or IoU are nearly universal and serve as the primary indicator of segmentation quality, but reporting at least one boundary metric (HD95 or ASSD) is strongly recommended, especially for thin structures or small lesions where overlap alone can be misleadingly high. Sensitivity and precision help diagnose whether errors are predominantly false negatives or false positives, thereby guiding targeted model refinement. When volumetric accuracy matters clinically, RVE or absolute volume error should be included to capture systematic over- or under-estimation. Finally, statistical significance tests or confidence intervals over the test set add rigour to method comparisons and support reproducible conclusions.

2.2 LEARNING FROM IMPERFECT ANNOTATIONS

2.2.1 *Sources of Annotation Imperfection in Medical Image Segmentation*

In practical medical image segmentation, the supervision signal is typically an observed label map $\tilde{y}$ produced under time, expertise, and protocol constraints, whereas the target of interest is an implicit (often unobserved) reference y^* that approximates the underlying anatomy or pathology. The mismatch $\tilde{y} \neq y^*$ is rarely random; it is structured by

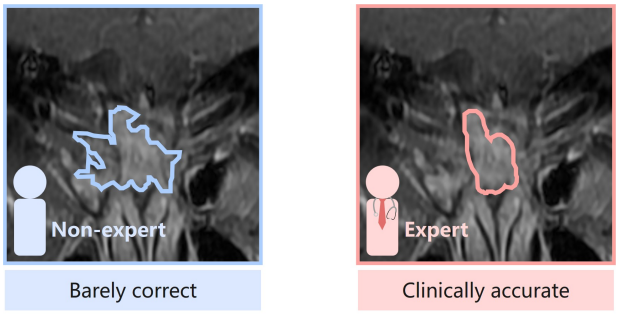


Figure 2.10: Illustration of annotation quality differences in medical image segmentation. **Left**: A non-expert annotation (blue contour) that is "barely correct"—it roughly captures the region of interest but exhibits substantial over-segmentation and imprecise boundaries. **Right**: An expert annotation of the same input image (pink contour) that is "clinically accurate"—produced by a trained clinician with domain knowledge, resulting in precise boundary delineation suitable for clinical decision-making. The discrepancy between non-expert and expert labels exemplifies the structured noise that arises from annotator expertise differences.

imaging physics, clinical conventions, and human perception, and it often concentrates around object boundaries and small or low-contrast lesions [71, 128]. Figure 2.10 illustrates this phenomenon.

A major source of imperfection is **intrinsic ambiguity** in the image evidence. Partial-volume effects [70], limited spatial resolution, motion artefacts, and low signal-to-noise ratios can render boundaries ill-defined, so multiple plausible delineations can be clinically acceptable [4]. This phenomenon is exacerbated in volumetric data when slice thickness is large relative to in-plane resolution, leading to anisotropy and reduced z-axis boundary sharpness, which increases uncertainty around superior–inferior boundaries [32, 124].

A second source is **inter- and intra-observer variability**. Even when image quality is high, clinicians may apply different implicit rules for what constitutes tumour extent, necrosis, oedema, or infiltrative margins, and a single annotator may exhibit inconsistency across sessions [4, 96]. Such variability is not well captured by independent and identically distributed (i.i.d.) label-noise assumptions; instead, it correlates with anatomy, contrast, lesion size, and local texture statistics [96].

Third, **protocol- and workflow-induced bias** can introduce systematic errors. Differences in institutional guidelines, annotation tools, and time budgets can yield consistent over- or under-segmentation patterns [71]. In multi-centre datasets, domain shifts in acquisition parameters may interact with these biases, making certain structures harder to label reliably [155].

Finally, **heterogeneous annotation quality** arises when labels are produced by annotators with varying expertise, or when coarse labels are later refined by experts [128]. This setting is common in scalable dataset curation and motivates methods that explicitly model a refinement process from weak to strong supervision, rather than treating all labels as exchangeable noisy observations.

2.2.2 *Learning Strategies: Robust Losses, Uncertainty Modelling, and Refinement/Transfer Viewpoints*

Learning with imperfect annotations can be framed as minimising an empirical risk under corrupted supervision. Let x denote the image and $\tilde{y}$ the observed label; a generic objective is

$$\min_{\theta} \; \mathbb{E}_{(x,\tilde{y})} \left[\ell\big(f_\theta(x), \tilde{y}\big) \right], \tag{2.22}$$

where $\ell(\cdot)$ is typically a Dice- or cross-entropy-based segmentation loss, and f_θ denotes the segmentation model (parameterised by θ) that maps an input image x to a pixel-/voxel-wise prediction. When $\tilde{y}$ contains structured errors, naïvely optimising this objective may encourage the model to fit annotation artefacts, particularly in boundary regions and for minority classes [71, 128, 172].

Robust losses aim to reduce sensitivity to corrupted labels by modifying the training objective or the contribution of unreliable samples/voxels. A common approach is to reweight losses using confidence or consistency signals, thereby down-weighting regions likely to be mislabeled [71, 128]. Sample selection methods, exemplified by co-teaching, train two networks that exchange small-loss samples to mitigate memorisation of noise [50]. Extensions such as stochastic co-teaching improve robustness when the noise rate is unknown and have been evaluated in medical imaging contexts [36]. More recent formulations learn to bootstrap robust supervision by optimising sample weights or targets, providing a principled bridge between loss reweighting and self-training [186]. In segmentation, robust learning is often implemented at the voxel or region level to reflect the spatially localised nature of annotation errors [169, 172].

Uncertainty modelling treats label imperfection as a consequence of ambiguity rather than purely adversarial noise. If multiple annotations are available, one can represent supervision as a distribution over labels (or boundary bands) instead of a single hard mask, and optimise a likelihood that accounts for annotator disagreement [4, 96]. When only a single mask is provided, uncertainty can be approximated through Bayesian and non-Bayesian approaches; these uncertainty estimates can guide selective learning, calibration, and quality control [124]. In radiotherapy and other settings with limited data, probabilistic segmentation models can provide voxel-wise uncertainty estimates that

reflect both ambiguous boundaries and data scarcity [32]. Importantly, uncertainty-aware formulations align with clinical practice, where ambiguity is expected in low-contrast regions and should not be over-penalised [4, 124].

Refinement/transfer viewpoints interpret imperfect annotations as weaker supervision that can be progressively improved. Teacher–student training (i.e., a stronger teacher model provides soft targets or feature guidance that a student model is trained to mimic), pseudo-labelling, and early-learning strategies aim to transfer knowledge from a stronger signal (e.g., a model trained on cleaner data, an ensemble, or an expert subset) to a student model [71, 94, 128].

Such approaches exploit the observation that deep networks tend to fit clean patterns earlier than noise, allowing the training process to be guided towards reliable regions [94]. In paired-quality settings, where labels of different fidelity coexist, the mapping from $\tilde{y}$ to y^* can be modelled as a correction process, enabling iterative refinement and discrepancy-driven supervision [128, 155]. This perspective is particularly suitable for scalable medical annotation workflows, where coarse labels are produced cheaply and then selectively refined, and it provides a conceptual bridge to non-expert-to-expert learning in which refinement dynamics are learned rather than assumed.

In Chapter 3, we build directly on this background by framing non-expert annotations as a weak and spatially structured supervision signal. Rather than treating label noise as random corruption, we distinguish an (unobserved) latent expert-level mask from the observed imperfect annotation and learn an explicit refinement process that progressively transforms the latter into the former.

2.3 SELF-SUPERVISED LEARNING FOR VOLUMETRIC MEDICAL IMAGING

This section introduces the general workflow of self-supervised learning (SSL) for volumetric medical images and highlights properties that motivate geometry-aware proxy tasks. The discussion is intentionally method-agnostic and serves as background for Chapter 4, where we instantiate these principles for label-efficient tumour segmentation [68, 135].

2.3.1 *Spatial Characteristics of Volumetric Medical Imaging*

Unlike natural images captured by a camera in a single exposure, volumetric medical scans are typically acquired as volumetric measurements and then reconstructed slice-by-slice along an anatomical axis—for example, an MRI scanner excites tissue in sequential planes from

head to foot, producing a stack of 2D images that together form a 3D representation of the patient's anatomy (see Fig. 2.11).

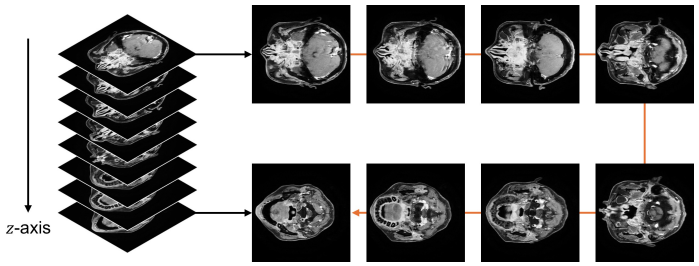


Figure 2.11: Illustration of a partial volumetric MRI scan of the brain. Slice images taken from [48].

Formally, let $x \in \mathbb{R}^{H \times W \times D}$ denote such a 3D scan defined on a voxel grid $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\} \times \{1, \ldots, D\}$, where H and W are the in-plane height and width, and D indexes the acquisition direction (often referred to as the z-axis or slice direction).

This acquisition process endows volumetric medical data with two distinctive properties that differentiate them from 2D generic images. First, *inter-slice continuity*: because anatomical structures extend smoothly through space, adjacent slices depict gradually varying cross-sections of the same organs, vessels, and lesions—a tumour visible in slice d will typically appear in slices $d - 1$ and $d + 1$ with only minor changes in shape and intensity [158]. Second, *anatomical coherence*: the spatial arrangement of structures follows predictable patterns dictated by human anatomy, so that the relative positions of organs (e.g., the heart lies anterior to the spine) remain consistent across patients and provide strong contextual cues for localisation [135]. Third, *positional constraints*: clinical acquisition protocols impose strong pose and field-of-view priors, so major structures typically appear in constrained regions of the image and are approximately aligned with the scanner axes, rather than occurring at arbitrary locations or orientations [45].

In practice, this continuity is modulated by the physical voxel spacing $\mathbf{s} = (s_x, s_y, s_z)$. Clinical protocols often favour high in-plane resolution at the expense of slice thickness, resulting in *anisotropic* volumes where $s_z \gg s_x \approx s_y$—for instance, a typical abdominal CT may have in-plane spacing of 0.7 mm but slice thickness of 5 mm [65]. Such anisotropy affects the effective receptive field along the acquisition axis and can introduce discontinuities that complicate 3D reasoning. Nonetheless, the underlying anatomical coherence persists: even with coarse z-resolution, the ordering and relative positioning of structures remain informative. These properties suggest that a representation encoder should capture not only local appearance within each slice but also consistent 3D morphology and relative spatial arrangement across slices. In practice, however, many medical pipelines still operate on 2D representations due to

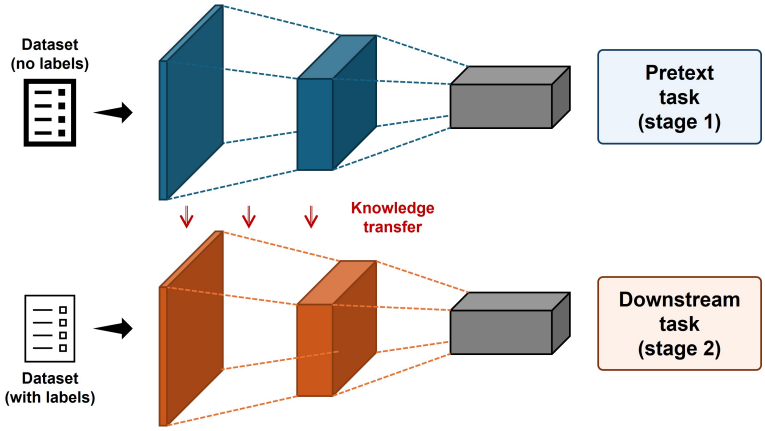


Figure 2.12: Illustration of the two stages of self-supervised learning (SSL): *pre-training* for pretext (or proxy) task learning and *fine-tuning* for downstream task learning.

memory and computational constraints and due to anisotropic voxel spacing, and even in 3D settings, it is often convenient to exploit the acquisition ordering to construct self-supervision signals.

A useful abstraction is to view a volume as an ordered sequence of slices $\{x^{(d)}\}_{d=1}^{D}$ with $x^{(d)} \in \mathbb{R}^{H\times W}$. Under this view, meaningful self-supervision can be constructed from constraints that couple neighbouring slices—for example, predicting the relative position of two slices, detecting whether slices have been reordered, or inferring geometric transformations applied to a sub-volume [68]. Such constraints are particularly valuable in label-limited regimes common to medical imaging, where expert annotations are scarce and expensive, yet large archives of unlabelled scans are readily available.

2.3.2 *Self-Supervised Pre-training Framework*

The central motivation behind SSL is to exploit large amounts of unlabelled data—which are abundant in clinical archives—to learn transferable representations before fine-tuning on scarce expert annotations (see Fig. 2.12). This leads to a two-stage paradigm that has become standard in label-efficient MIA. In the first stage, an encoder f_θ is pre-trained on an unlabelled dataset $\mathcal{D}_u = \{x_j\}_{j=1}^{m}$ (where each x_j denotes an input image/volume) by optimising a *proxy task* that derives supervision from the data itself, without any manual labels. Typical proxy tasks include reconstruction-based objectives (e.g., autoencoder-style reconstruction of the input, or inpainting masked regions), transformation prediction (e.g., rotation or relative-position prediction), and contrastive learning that matches representations across augmented views of the same sam-

ple [68]. The objective is to learn representations that capture semantically meaningful structure:

$$\theta^\star \;=\; \arg\min_{\theta}\; \mathbb{E}_{x\sim\mathcal{D}_u}\Big[\mathcal{L}_{\text{ssl}}\big(f_\theta(x)\big)\Big]. \tag{2.23}$$

In the second stage, the pre-trained encoder is transferred to a task-specific model g_ϕ, which is fine-tuned on a small labelled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{n}$ (typically $n \ll m$):

$$\phi^\star \;=\; \arg\min_{\phi}\; \mathbb{E}_{(x,y)\sim\mathcal{D}_l}\Big[\mathcal{L}_{\text{task}}\big(g_\phi(x), y\big)\Big]. \tag{2.24}$$

In practice, the transfer is realised by initialising the encoder weights of g_ϕ from $\theta^\star$ and learning a task-specific head or decoder on top. The key design choice lies in the SSL objective $\mathcal{L}_{\text{ssl}}$, which is defined by a proxy task that generates supervision from the input x alone—via transformations, masking, or structural constraints—without requiring ground-truth labels [68].

Recent medical SSL work has emphasised that transferability depends on which information is preserved during pre-training [76]. For example, a unified visual information preservation perspective combines pixel-level restoration on corrupted inputs with feature-level comparison across views, and employs multi-scale representations to retain fine-grained cues while promoting invariances useful for downstream medical analysis [184]. Such formulations motivate two general design principles for volumetric settings: (a) reconstruction-style objectives can preserve local detail that is critical for medical interpretation [188], and (b) multi-scale features are beneficial when downstream targets vary substantially in size [121]. In addition, large-scale 3D pre-training has incorporated explicit geometric context priors by constructing supervisory signals from spatial relationships between crops and enforcing consistency across data sources via teacher–student learning [56]. These examples illustrate complementary ways to inject 3D inductive bias into $\mathcal{L}_{\text{ssl}}$ without relying on annotations.

In parallel, related research has addressed essentially the same challenge (**CII**)—learning under label scarcity in biomedical imaging—but in a different (and comparatively easier) problem setting, namely 2D image classification [11, 13, 15]. Benato et al. propose Deep Feature Annotation (DeepFA) and its variants [15], where deep features (e.g., from VGG-style encoders) are projected to 2D for interactive or semi-automatic label propagation [12, 17], often coupled with confidence-based sampling and iterative pseudo-labelling to reduce expert effort [14, 16], and more recently enhanced via contrastive learning to improve latent-space separability and thus pseudo-label quality [11]. These works demonstrate that exploiting unlabelled data through representation learning and guided pseudo-labelling can substantially improve performance

with minimal supervision. However, they focus on 2D image classification rather than dense 3D voxel-wise segmentation, where supervision is spatially dense, errors are boundary-localised, and the semantic and geometric consistency must hold across slices in anisotropic volumes [65, 135, 158]. This makes the label-scarcity setting substantially more challenging and motivates volumetric SSL proxy tasks that explicitly encode inter-slice continuity and 3D geometric regularities, as detailed next.

2.3.3 *Representation Learning via Spatial Transformations*

A common strategy for volumetric SSL is to define proxy tasks using spatial transformations that must be inferred from the observed volume, thereby encouraging the encoder to encode 3D structure rather than superficial appearance [135, 158]. Formally, let $\mathcal{T}$ denote a family of geometric transformations, and sample a random transform t from $\mathcal{T}$ (denoted $t \sim \mathcal{T}$), e.g., a rotation, translation, or permutation along z. Applying t to an input volume (or a sub-volume) yields $\tilde{x} = t(x)$. Transformation-based SSL typically trains an encoder (and possibly a prediction head) to solve:

$$\theta^\star = \arg\min_\theta \mathbb{E}_{x\sim\mathcal{D}_u} \mathbb{E}_{t\sim\mathcal{T}} \Big[\mathcal{L}_{\text{proxy}}\big(h(f_\theta(\tilde{x})),\, \pi(t)\big) \Big], \tag{2.25}$$

where $\pi(t)$ denotes the supervisory signal induced by t (e.g., a discrete rotation class, continuous parameters, or relative position labels), and $h(\cdot)$ is a lightweight predictor. The key assumption is that predicting $\pi(t)$ requires recognising anatomical structures and their spatial relationships. This assumption aligns with volumetric data: inter-slice coherence provides additional constraints (e.g., consecutive slices should remain anatomically plausible under the predicted transform), while anisotropic spacing limits which transformations are physically meaningful.

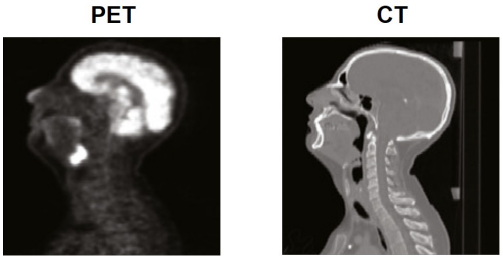
Geometry-aware objectives can also be realised through crop-based context modelling, where spatial overlap or relative position between sub-volumes is used as supervision [158]. In particular, geometric context priors may define continuous labels derived from overlap proportions between crops, turning spatial layout into a learnable signal that is tightly coupled to volumetric structure [135]. Compared with purely instance-discrimination objectives, such geometric supervision is directly tied to 3D organisation and is therefore well matched to medical volumes in which anatomical distributions are spatially structured.

In Chapter 4, we operationalise these ideas in a label-efficient tumour segmentation setting by designing a proxy task that exploits inter-slice continuity and geometric regularities in 3D MRI, and by adopting the pre-training–fine-tuning protocol in Eqs. (2.23)–(2.24).

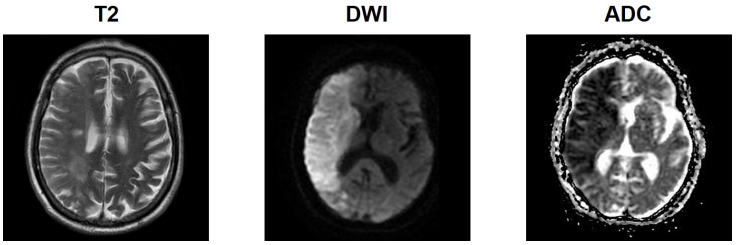
2.4 MULTI-MODAL LEARNING AND KNOWLEDGE TRANSFER

2.4.1 *Differences between Medical Imaging Modalities*

Medical imaging modalities differ in the physical principles used to form images, which in turn determines the type and reliability of the information available for segmentation. X-ray CT measures X-ray attenuation and provides high spatial resolution with relatively stable intensity statistics across scanners after standard calibration, making it well-suited for delineating dense structures and contrast-enhanced vasculature [154].



(a) PET and CT (sagittal view of head and neck).



(b) MRI sequences: T2-weighted, diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) map (axial view of brain).

Figure 2.13: Comparison of different medical imaging modalities. (a) PET provides functional information (e.g., metabolic activity) with lower spatial resolution and higher noise, whereas CT offers high-resolution anatomical detail with clear bone and soft-tissue boundaries. (b) Different MRI sequences capture complementary tissue properties: T2-weighted images highlight fluid and edema, DWI is sensitive to restricted water diffusion (e.g., acute ischemia), and ADC maps quantify diffusion coefficients to distinguish true restriction from T2 shine-through. PET and CT images taken from [117], and MRI images taken from [57].

MRI acquires signals governed by proton density and relaxation properties, and therefore offers rich soft-tissue contrast through multiple sequences (e.g., T1-, T2-weighted, diffusion, or contrast-enhanced variants), at the cost of more complex intensity distributions and protocol-

dependent appearance [7]. Nuclear medicine modalities such as PET and SPECT measure radiotracer uptake and provide functional information related to metabolism or perfusion, but typically exhibit lower spatial resolution and higher noise, often requiring anatomical guidance from CT or MRI for precise boundary localisation [1, 46]. Ultrasound forms images from acoustic scattering and is characterised by operator dependence and speckle artefacts, yielding substantial appearance variability across acquisitions [154].

As illustrated in Fig. 2.13, these modalities provide fundamentally different views of the same anatomy. PET highlights regions of elevated metabolic activity but lacks the spatial precision to delineate anatomical boundaries, whereas CT provides sharp structural detail but limited soft-tissue contrast. MRI sequences such as T2-weighted, DWI, and ADC each emphasise different tissue properties, enabling multiparametric characterisation of pathology—yet their intensity distributions are sequence-dependent and can vary substantially across scanners and protocols.

From a machine learning perspective, such modality-specific characteristics pose distinct challenges for segmentation algorithms. The same anatomical structure may be conspicuous in one modality while poorly contrasted in another; conversely, modality-specific artefacts (e.g., MRI intensity non-uniformity, CT beam hardening, PET partial volume effects) can distort both local textures and global intensity statistics [29, 39]. In addition, multimodal datasets frequently exhibit imperfect spatial correspondence due to patient motion, differing slice thicknesses, and inconsistent fields of view, which introduces misregistration and resolution mismatch [7, 165]. As a result, multimodal learning for medical image segmentation must handle both semantic complementarity (different cues across modalities) and statistical heterogeneity (different intensity and noise characteristics), while remaining robust to missing or unreliable modalities at deployment [7, 130].

2.4.2 *Multi-Modal Learning for Medical Image Segmentation*

Multi-modal learning aims to exploit complementary information from multiple imaging sources to improve segmentation accuracy and robustness. A common formulation assumes a set of modality observations $\{x^{(m)}\}_{m=1}^{M}$ for a given subject (i.e., $x^{(m)}$ denotes the image/volume of that subject in modality m, not a modality-specific dataset), and seeks to predict a segmentation mask y using a model f_θ that conditions on all or a subset of modalities, i.e., $\hat{y} = f_\theta(x^{(1)}, \ldots, x^{(M)})$ [165]. In practice, training and deployment conditions may differ: all modalities can be available during training while only a subset is available during inference, or modality availability may vary across sites and patients [7, 72]. This motivates architectures and objectives that both leverage complementary information and tolerate missing inputs.

Architecturally, fusion strategies are often categorised by the stage at which modalities interact. Let $\phi^{(m)}$ denote the encoder for modality m, producing feature maps $z^{(m)} = \phi^{(m)}(x^{(m)})$, and let ψ denote the decoder that produces the final segmentation. *Early fusion* concatenates modalities at the input level and learns shared low-level features:

$$\hat{y} = \psi\big(\phi([x^{(1)}; x^{(2)}; \ldots; x^{(M)}])\big), \tag{2.26}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. This approach can be sensitive to intensity scale differences and misregistration [130, 160]. *Intermediate fusion* employs modality-specific encoders followed by feature interaction modules:

$$\hat{y} = \psi\big(\mathcal{F}(z^{(1)}, z^{(2)}, \ldots, z^{(M)})\big), \tag{2.27}$$

where $\mathcal{F}$ is a fusion operator (e.g., concatenation, element-wise addition, or attention-based aggregation) that combines modality-specific representations [1, 46]. *Late fusion* combines modality-specific predictions:

$$\hat{y} = \mathcal{G}\big(\psi^{(1)}(z^{(1)}), \psi^{(2)}(z^{(2)}), \ldots, \psi^{(M)}(z^{(M)})\big), \tag{2.28}$$

where $\psi^{(m)}$ denotes a modality-specific decoder (or prediction head) producing an output from $z^{(m)}$, and $\mathcal{G}$ aggregates per-modality outputs (e.g., via averaging or learned weighting), which can improve robustness but may underutilise cross-modality complementarity when boundaries are ambiguous in the weak modality [130]. Recent approaches increasingly use attention-based interactions to condition features from one modality on another. For instance, cross-attention is computed as:

$$\tilde{z}^{(1)} = \mathrm{softmax}\left(\frac{Q^{(1)} K^{(2)\top}}{\sqrt{d}}\right) V^{(2)}, \tag{2.29}$$

where $Q^{(1)} = z^{(1)}\mathbf{W}_Q$, $K^{(2)} = z^{(2)}\mathbf{W}_K$, and $V^{(2)} = z^{(2)}\mathbf{W}_V$ are linear projections of the modality-specific feature maps $z^{(1)}$ and $z^{(2)}$ into the query, key, and value embedding subspaces, respectively.. This mechanism allows one modality to query and aggregate relevant information from another, focusing transfer on anatomically consistent regions [18, 38, 143].

A central practical issue is learning with missing modalities. Approaches that assume complete modality sets at inference are often unsuitable in constrained settings where certain acquisitions are expensive, time-consuming, or contraindicated [7]. To address this, multimodal segmentation models commonly incorporate modality dropout or stochastic masking during training to improve robustness, or use mixture-of-experts and routing mechanisms to adapt computation to

available modalities [72, 182]. An alternative line of work attempts modality completion by synthesising the missing modality or its latent representation, followed by standard multimodal fusion [167]. While completion can recover complementary cues, its downstream utility depends on the fidelity of the generated signal and can be degraded by domain shift or anatomical outliers [43, 167]. These considerations motivate knowledge-transfer approaches that exploit strong modalities during training while producing models that operate reliably using only weak modalities at deployment [105].

2.4.3 *Knowledge Transfer for Medical Image Segmentation*

Knowledge transfer provides a principled mechanism to reuse information learned in a source setting to improve performance in a target setting, particularly when the target is constrained by limited supervision or limited modality availability. In MIS, transfer commonly appears as (i) representation transfer via pre-training followed by task-specific fine-tuning, (ii) teacher–student knowledge distillation, and (iii) self-training with pseudo-labels [105, 127] (also see Section 2.2.2. Although these paradigms differ in implementation, they share the goal of injecting privileged information available during training into a deployable model that must operate under stricter constraints [105].

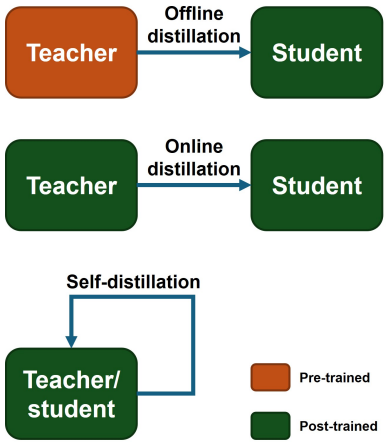


Figure 2.14: Three paradigms of knowledge distillation. **Offline distillation**: a pre-trained (frozen) teacher transfers knowledge to a trainable student. **Online distillation**: both teacher and student are trained jointly, with knowledge flowing from teacher to student during training. **Self-distillation**: a single network acts as both teacher and student, transferring knowledge from deeper layers or earlier training epochs to itself.

Knowledge distillation is particularly relevant when a strong source model has access to richer inputs or stronger supervision than the target model. As illustrated in Fig. 2.14, knowledge distillation can be categorised into three paradigms based on the training dynamics between teacher and student. In *offline distillation*, the teacher is pre-trained and frozen; the student learns to mimic the teacher's outputs or intermediate representations without affecting the teacher's parameters. In *online distillation*, both teacher and student are trained simultaneously, allowing bidirectional or mutual learning where the teacher can also benefit from the student's gradients. In *self-distillation*, a single network serves as both teacher and student, typically by transferring knowledge from deeper to shallower layers, from larger to smaller sub-networks, or from earlier to later training iterations [105].

Orthogonal to the choice of training paradigm is the question of *what* knowledge is transferred and *where* in the network hierarchy the transfer occurs. Let f_T denote a teacher and f_S a student, with intermediate features $z_T^{(l)}$ and $z_S^{(l)}$ at layer l, and output logits p_T and p_S. Depending on the level of abstraction, distillation objectives can be broadly grouped into three categories, as follows [105].

Output-level distillation encourages the student prediction to match the teacher prediction through a divergence term. A common choice is the Kullback–Leibler (KL) divergence on temperature-softened logits:

$$\begin{aligned} \mathcal{L}_{\text{out}} &= \tau^2 \cdot \mathrm{KL}\big(\sigma(p_T/\tau) \,\|\, \sigma(p_S/\tau)\big), \\ \mathrm{KL}(q \,\|\, r) &= \sum_{c=1}^{C} q_c \log \frac{q_c}{r_c}, \end{aligned} \tag{2.30}$$

where $q = \sigma(p_T/\tau)$ and $r = \sigma(p_S/\tau)$, $\sigma(\cdot)$ is the softmax function and $\tau > 1$ is a temperature hyperparameter that smooths the probability distribution to reveal inter-class relationships [87, 105].

Feature-level distillation aligns intermediate representations at one or multiple scales:

$$\mathcal{L}_{\text{feat}} = \sum_{l \in \mathcal{S}} \big\| g_l(z_S^{(l)}) - z_T^{(l)} \big\|_2^2, \tag{2.31}$$

where $\mathcal{S}$ is the set of layers at which alignment is enforced and g_l is an optional projection head that matches the student's feature dimension to the teacher's. This is beneficial for dense prediction where localisation and boundary cues are distributed across layers [24, 87].

Relational distillation transfers pairwise similarities or structural constraints between features:

$$\mathcal{L}_{\text{rel}} = \sum_{i,j} \big| \langle z_T^{(i)}, z_T^{(j)} \rangle - \langle z_S^{(i)}, z_S^{(j)} \rangle \big|, \tag{2.32}$$

where $\langle \cdot, \cdot \rangle$ denotes a similarity measure (e.g., cosine similarity or Euclidean distance). This aims to preserve geometric or anatomical relations that may be difficult to learn from the weak input alone [105].

In practice, the overall training objective combines these terms with a task-specific supervised loss:

$$\mathcal{L} = \mathcal{L}_{\text{seg}}(y, \hat{y}_S) + \lambda_1 \mathcal{L}_{\text{out}} + \lambda_2 \mathcal{L}_{\text{feat}} + \lambda_3 \mathcal{L}_{\text{rel}}, \tag{2.33}$$

where $\mathcal{L}_{\text{seg}}$ is typically a combination of cross-entropy and Dice loss (see earlier Section 2.1.3, and $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters controlling the contribution of each distillation term. In segmentation, these objectives are frequently augmented with uncertainty-aware weighting or boundary-focused constraints to prevent propagating teacher errors into the student [87].

Cross-modal knowledge transfer addresses the scenario where a strong modality $x^{(s)}$ is available during training but a weak modality $x^{(w)}$ is used at inference. It can also be viewed through the perspective of *Learning Using Privileged Information* (LUPI), introduced by Vapnik and Vashist [133, 140]. In the LUPI setting, privileged information is available only during training to facilitate learning, but is unavailable at inference time. While LUPI was originally introduced in the context of support vector machine (SVM)-based learning, modern deep learning approaches commonly instantiate LUPI through teacher–student distillation [107], where privileged information is accessible only to the teacher during training. In medical imaging, this setting naturally arises when a strong modality $x^{(s)}$ (e.g., contrast-enhanced MRI or PET) is accessible during training, whereas only a weak modality $x^{(w)}$ is available at deployment due to cost, acquisition time, or clinical constraints. A common approach trains a teacher f_T using both modalities (or their fused representations) and distils the resulting knowledge into a student f_S that observes only the weak modality [166, 178]:

$$\mathcal{L}_{\text{cross}} = \mathcal{L}_{\text{seg}}\big(y, f_S(x^{(w)})\big) + \lambda \cdot d\big(f_T(x^{(w)}, x^{(s)}), f_S(x^{(w)})\big), \tag{2.34}$$

where $d(\cdot, \cdot)$ measures the discrepancy between teacher and student outputs (or features), and the teacher leverages the strong modality to provide more accurate guidance. A typical choice is an ℓ_2 distance (or KL divergence) between their output logits/probability maps for output-level distillation. For feature-level distillation, d is often instantiated as an MSE term that aligns intermediate representations across selected layers, optionally via a projection head to match feature dimensions.

The effectiveness of this transfer depends on (i) *where* in the network/representation hierarchy knowledge is transferred (e.g., at the output logits, at intermediate multi-scale feature layers, or at both), (ii) *what* is transferred (semantic features, boundary information, uncertainty, or lesion-centric cues), and (iii) *how* the transfer is scheduled over training (i.e., via a curriculum that gradually introduces/strengthens the distillation objective, rather than enforcing one-shot direct matching from the start) [66, 105]. Progressive transfer is often preferred when the representation gap between modalities is large, as direct feature re-

gression can be unstable under misregistration and heterogeneous appearance distributions [157, 175]. Overall, knowledge transfer provides a unifying lens for constrained medical image segmentation: it exploits privileged information at training time while preserving feasibility and robustness at deployment [135, 158, 184].

2.5 PROBLEM STATEMENT

This thesis studies **lesion semantic segmentation**, where each pixel (or voxel) in a medical image is assigned a clinically meaningful label, typically distinguishing lesion tissue from normal tissue. Given a 2D or 3D medical image (e.g., CT or MRI), the goal is to produce a label mask of the same spatial size via dense prediction. The task is commonly formulated as binary segmentation ("lesion" versus "non-lesion") and can be extended to multi-class settings. It is clinically relevant to workflows such as treatment planning, computer-aided diagnosis, and longitudinal monitoring.

Formulation. Let $x \in \mathbb{R}^{H\times W\times D\times M}$ denote a medical image and $y \in \{0, 1, \ldots, C-1\}^{H\times W\times D}$ the corresponding label mask, where C is the number of classes. A segmentation model f_θ predicts

$$\hat{y} = f_\theta(x), \qquad \hat{y} \in \{0, 1, \ldots, C-1\}^{H\times W\times D}. \tag{2.35}$$

In the binary case ($C = 2$), $\hat{y}$ can be interpreted as the lesion mask.

Practical constraints and thesis focus. In clinical workflows, supervision signals are limited and heterogeneous, and modality availability often differs across sites and times. This thesis addresses three interrelated constraints that motivate the methodological contributions in Chapter 3, Chapter 4, and Chapter 5, as follows.

(1) Heterogeneous annotation quality. Annotation quality often varies: coarse masks from non-experts are commonly refined by experts, yielding paired non-expert and expert labels. The challenge is to exploit these correction patterns rather than discarding coarse annotations. Chapter 3 introduces a refinement-based framework that learns from paired non-expert and expert labels, modelling the coarse-to-fine correction process and leveraging discrepancy-aware optimisation to improve segmentation.

(2) Label scarcity and volumetric structure. Expert labelling is expensive, so only a small fraction of images (especially 3D volumes) are fully annotated, while large collections remain unlabeled. This raises the need for label-efficient learning strategies that leverage unlabelled data and the inherent spatial continuity of volumetric imaging. Chapter 4 focuses on this setting and proposes a self-supervised pre-training strategy tailored to 3D volumes (e.g., rotation prediction over consecutive slices) to improve segmentation under limited annotations.

(3) Asymmetry between strong and weak modalities. In many clinical settings, *strong* modalities (e.g., PET, MRI) provide rich functional or soft-tissue contrast and make lesion regions more pronounced, but they are not always available at inference due to cost, equipment, or protocol constraints. *Weak* modalities (e.g., non-contrast CT) are more widely available but offer subtler lesion contrast, making accurate segmentation harder when they are used alone. Standard multi-modal fusion requires all modalities at inference and therefore does not address the case where only the weak modality is available at test time. Chapter 5 addresses this gap with a two-stage framework that transfers lesion-specific knowledge from strong to weak modalities, enabling accurate segmentation using the weak modality alone at inference.

Accordingly, the three methodology chapters that follow each target one of these constraints and develop dedicated solutions grounded in the foundations and gaps identified in this chapter.

2.6 CHAPTER SUMMARY

This chapter has established the technical foundations and surveyed prior work relevant to the three methodological contributions of this thesis. We summarise the key observations and the research gaps that motivate the subsequent chapters.

Fundamentals of medical image segmentation. We reviewed the core building blocks of modern segmentation systems, including encoder–decoder architectures (e.g., U-Net and its variants), commonly used loss functions (cross-entropy, Dice-based losses), and standard evaluation metrics (Dice coefficient, IoU, Hausdorff distance). These components provide the technical vocabulary and baseline infrastructure upon which the contributions of this thesis are built. Note that we do not aim to advance these fundamental architectures, losses, or metrics; instead, they serve as our shared methodological foundation, upon which we develop higher-level contributions to address the thesis challenges under imperfect and limited supervision.

Learning under heterogeneous annotation quality. The literature review in Section 2.2 revealed that most existing approaches handle imperfect supervision by adopting robust objectives and reweighting schemes, selecting or filtering training signals, modelling uncertainty or annotator disagreement with probabilistic formulations, and/or using teacher–student transfer (self-training, pseudo-labelling, distillation) to bootstrap cleaner supervision. However, in realistic clinical annotation workflows, label imperfections often exhibit structured patterns—particularly when initial annotations by non-specialised personnel are subsequently corrected by expert physicians. This refinement structure remains underexploited. *Research gap:* Methods that explicitly model

and leverage the latent correction patterns encoded in multi-stage annotation workflows are lacking. Chapter 3 addresses this gap by proposing a recurrent refinement framework that learns to transform non-expert annotations into expert-level segmentations.

Label-efficient learning for volumetric data. Section 2.3 surveyed self-supervised strategies designed to reduce annotation burden. While contrastive learning and reconstruction-based pretext tasks have shown promise in natural image domains, their adaptation to 3D medical imaging remains limited, particularly for exploiting the inherent spatial continuity of volumetric data. *Research gap:* Self-supervised proxy tasks tailored to the geometric and anatomical structure of consecutively sliced medical volumes are underexplored. Chapter 4 addresses this gap by introducing a rotation-based pretraining strategy that captures 3D spatial regularities.

Bridging the strong–weak modality gap. Section 2.4 examined multi-modal fusion techniques and knowledge distillation methods. Existing fusion approaches typically require all modalities at inference, limiting their applicability when strong modalities (e.g., PET, MRI) are unavailable in routine clinical practice. Meanwhile, standard knowledge distillation methods often apply one-shot alignment, which may be unstable when the representational gap between teacher and student is large. *Research gap:* Effective mechanisms for transferring lesion-specific knowledge from strong to weak modalities—while maintaining a weak-modality-only inference pathway—remain underdeveloped. Chapter 5 addresses this gap through a two-stage framework comprising asymmetric relationship modelling and progressive knowledge transfer.

3

RECURRENT REFINED NETWORK FOR MEDICAL IMAGE SEGMENTATION UNDER LABEL-QUALITY IMBALANCE

This chapter presents the first methodological contribution of this thesis, targeting the challenge of heterogeneous annotation quality introduced in Chapter 1. As discussed, practical medical image segmentation is frequently constrained by imperfect labels produced under varying expertise levels and delineation conventions. Rather than treating annotation errors as independent noise, the approach developed here exploits the structure of label corrections observed in realistic annotation workflows—where initial annotations by non-specialised personnel are subsequently refined by expert physicians. This refinement process encodes latent correction patterns that, if leveraged appropriately, can improve both boundary consistency and lesion completeness.

To this end, we propose ReReNet, a recurrent refined network for lesion segmentation that learns from non-expert to expert. It achieves progressively refined segmentation results through multiple iterations with tailored discrepancy-aware supervision. During training, as the iterative process perceives discrepancies between refining and expert labels, the model gradually grasps the knowledge of turning the barely correct into the clinically accurate. We validate ReReNet's capability on three medical image segmentation (MIS) datasets, including magnetic resonance (MR) and computed tomography (CT) modalities. Comparison results indicate that the proposed approach achieves superior performance by introducing the designed recurrent mechanism and outperforms mainstream methods, demonstrating the effectiveness of mining hidden correction patterns by utilizing non-expert information[1].

3.1 INTRODUCTION

Medical image segmentation (MIS), identifying and isolating critical regions like lesions, blood vessels, and tissues [126], plays a major role in early diagnosis, staging, treatment planning, and predicting prognosis [148]. Imaging techniques like computed tomography (CT), magnetic resonance (MR), etc., offer non-invasive ways to get detailed information about lesions, including their relationships to surrounding

1 Parts of this chapter were published in: R. Jiang†, **C. Li**†, X. Ban*, S. Yin, C. Yao, Y. Guo, and M. S. Obaidat. From non-expert to expert: Recurrent refined learning for medical image segmentation. *In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2086–2093. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10821757

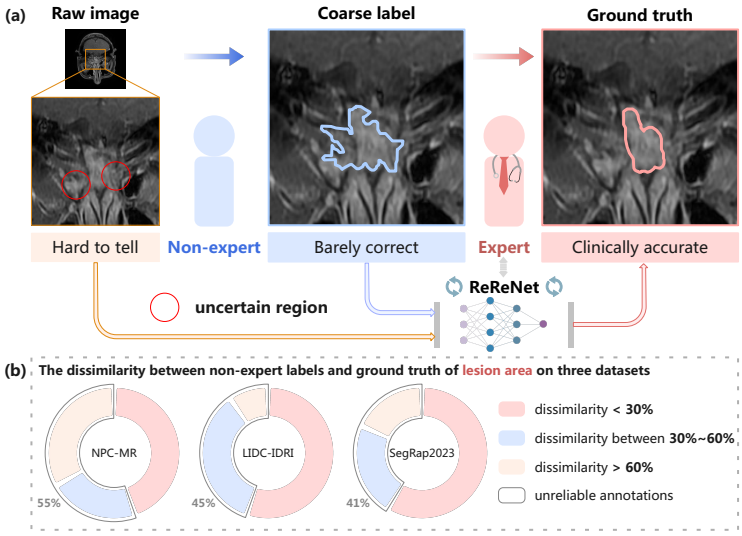


Figure 3.1: The motivation of ReReNet. (a) A common way to generate ground truth is for a non-expert to delineate the lesion coarsely, followed by refinement by an expert. ReReNet aims to mimic the refining process by recurrently pushing the model to output an expert-like mask. (b) On the three datasets used in our evaluation, nearly half of the lesions have a dissimilarity of more than 30% in lesion area between coarse labels and ground truth.

structures [100, 151]. To automate the delineation process, which requires time and expertise, numerous techniques have been developed and applied to real-world scenarios [20, 97], among which prevailing deep-learning-based ones are recognised to identify complex patterns from massive data [126, 183].

However, training a capable model needs abundant high-quality annotations [80, 84]. For MIS, annotating presents distinct challenges compared to natural scenes. The inherent complexity of medical images requires extensive domain expertise, hindering accurate labelling by untrained individuals [86]. Conversely, annotating in natural scenes for image segmentation or object detection can be delegated to non-experts, enabling rapid dataset curation. The difficulty in recognising regular targets (buildings, persons, lane lines, etc.) is minimal. In contrast, medical MR and CT imaging necessitate careful judgment and confirmation for seasoned technicians to identify critical regions [22, 179], especially in complex scenarios with blurred boundaries and intermingled tissues.

A prevalent solution in practice for lesion annotation adopts a two-step approach [119], as shown in Fig. 3.1 (a). Firstly, non-expert annotators with brief training perform coarse labelling, capturing diagnostically relevant information like location and size. Secondly, experts, typ-

ically senior physicians, review and correct these coarse annotations. This strategy balances annotation quality with resource utilisation, significantly reducing expert workload and overall costs. However, these non-expert labels are typically overlooked during the training of the lesion segmentation model. Indeed, mislabelling occurs more frequently among non-experts due to the uncertainty of blurred boundaries and tumour infiltration. Fig. 3.1 (b) depicts the discrepancies between non-expert labels and ground truth. We argue that, regardless of their accuracy, modelling the correction process from unreliable (non-expert) to reliable (expert) labels can guide the model to explicitly discover refinement patterns and leverage helpful information in non-expert labels, thereby achieving more precise and robust lesion segmentation.

To this end, we propose ReReNet, a recurrent refined framework for lesion segmentation. To the best of our knowledge, this is the first study of using non-expert labels to assist in MIS tasks. Specifically, we treat the non-expert label as the initial coarse label. The output mask feeds into the model recurrently as the coarse label of the next refinement stage. Those iterative processes continually enhance the accuracy of the segmentation. Further, to address the issue encountered when naïvely learning with the blunt repetition that the performance unexpectedly decreases instead of increases, we design a discrepancy-aware optimisation strategy to calculate the discrepancy loss that explicitly considers label differences and guides the model toward better learning from coarse labels.

We assess the performance and validate the effectiveness of ReReNet on three datasets, NPC-MR (in-house), LIDC-IDRI [3], and SegRap2023 [101]. Compared with other mainstream MIS methods, ReReNet performs better by mimicking the correction process conducted by expert physicians from gradually refined coarse labels. Taking a naïve U-Net [120] as the cornerstone, the modified version (ReReNet) improves the DSC by 4.36%, 1.17%, and 1.11%, respectively. Furthermore, we conduct segmentation experiments at different label ratios during training on the NPC-MR dataset. With only 10% of labels, ReReNet achieves the performance of the baseline model trained on 50% of labels. This substantial reduction in reliance on high-standard annotations highlights ReReNet's potential in MIS.

The main contributions of our work are:

- We propose ReReNet, a novel learning framework to harness non-expert labels that may have been previously overlooked and to refine segmentation results recurrently.
- We develop a discrepancy-aware optimisation strategy, perceiving the differences to expert labels as the process iterates. This strategy of calculating the loss guides the network model to focus on these discrepancies.

- We conduct extensive experiments on three lesion segmentation datasets, of which the numerical and visual results illustrate that the proposed ReReNet delivers higher segmentation accuracy and alleviates the label demand compared to typical methods. Ablation studies also prove the efficacy of the recurrent mechanism and the discrepancy-aware supervision.

3.2 RELATED WORK

In the past decade, the boosted computational capacity encouraged the development of numerous deep learning-based automatic MIS algorithms. After fully convolutional networks (FCNs) [99], the emergence of the simple yet effective U-Net [120] established the U-shaped architecture as a de facto model in medical image analysis in either 2D [65, 116] or 3D [33, 110]. Recently, transformer has been introduced into MIS, such as UNETR [52], TransUNet [26], and Swin-Unet [19], etc.

The challenges addressed in this chapter are rooted in the broader setting of learning from imperfect annotations, where the discrepancy between an observed label map and an implicit reference often arises from intrinsic ambiguity, observer variability, and protocol-induced bias, and tends to concentrate around boundaries and small or low-contrast lesions (see Chapter 2 Section 2.2). Building on that background, we briefly highlight here only the aspects most relevant to non-expert-to-expert refinement, and keep the discussion focused on the literature closest to our formulation.

Several essential phenomena in Medical images that raise the difficulty of accurately segmenting the tumour lesion are blurred lesion boundaries and noise [86]. To overcome the issue, some methods adopt the coarse-to-fine paradigm [55, 92, 104, 161]. The first stage generates an approximate region, and the second stage refines this region using more sophisticated models. CFU-Net [170] introduces an additional decoding path at the feature level, using the coarse part to guide the decoding of the refined part. [85] utilises semantic information from feature maps of different layers to refine the boundaries of coarse masks, thereby enhancing segmentation performance.

However, most coarse-to-fine pipelines above treat “coarse” as an intermediate prediction (or a low-resolution/ROI cue) generated by a first-stage model, rather than as an observed supervision signal provided by non-expert annotators. As a result, they rarely exploit paired heterogeneous-quality annotations (non-expert vs. expert) to explicitly learn a refinement mapping that corrects structured boundary errors in blurred and intermingled lesion regions. Using coarse labels, we innovatively formulate the recurrent refined network to model the coarse-to-fine pattern from non-expert labels to expert labels, thereby realising accurate lesion segmentation.

3.3 METHOD

We present a novel learning framework to harness abundant coarse label information previously underutilised and introduce a discrepancy-aware optimisation strategy to enhance this learning paradigm. We first formulate the problem of segmenting the lesion or organ region and clarify the mathematical notations. Next, we delve into the specific concept and methodology of the framework. Based on that, we describe the dedicated optimisation strategy tailored to this scenario. Fig. 3.2 demonstrates the details of the proposed ReReNet and the Discrepancy loss.

3.3.1 *ReReNet Framework*

Inspired by the coarse-to-fine approach [55, 55, 85], we design a recurrent and progressive learning architecture that utilises coarse labels to assist in segmentation. Non-expert labels are treated as the initial coarse labels. The output masks are fed into the model recurrently as the coarse labels of the next refinement stage, continually enhancing the segmented results.

Considering the input space $\mathcal{X}$ (i.e., the space of all 2D input images to be segmented), each image is a set of N pixels, where $N = H \times W$. Each pixel stores an intensity value (single-channel, e.g., CT/MRI) or a vector of channel values (multi-channel, e.g., multi-sequence or multi-modal inputs). In a standard semantic segmentation setup, given an image $x \in \mathcal{X}$, our objective is to learn a mapping that assigns a label $y_i \in \mathcal{Y}$ to each pixel of x, representing its semantic category, i.e., foreground (lesion) or background. The mapping is achieved by a learnable encoder-decoder model f_θ, predicting the probability of mapping from the image space $\mathcal{X}$ to the label space $\mathcal{Y}$. The output segmentation mask is $\hat{y} = \{\arg\max_{c \in \mathcal{Y}} p_i^c\}_{i=1}^N$, where p_i^c is the model's predicted probability of pixel x_i belonging to category c. Hence, $\hat{y} = f_\theta(x)$.

In the training and inference phase of ReReNet, the final segmentation mask is obtained through multiple stages, as shown in Fig. 3.2 (a). We divide this process into K stages. During training, at each stage $1 \leq k \leq K$, the output mask $\hat{y}^{k-1}$ from the previous stage is input into the network, aiding in producing a more refined segmentation outcome, which is denoted as:

$$\hat{y}^k = f_\theta^{k-1}(x, \hat{y}^{k-1}). \tag{3.1}$$

The parameters of the model f_θ^{k-1} are updated by calculating the loss between $\hat{y}^k$ and y to obtain f_θ^k for the subsequent stage. Upon the completion of training, we have the final model f_θ^* after the last stage. It is noted that the recurrent mechanism is also employed in inference, where the mask from the previous stage is concatenated with the raw image to input into the model. After iterating K times, the whole process is finished and generates the final segmentation result $\hat{y}^k = f_\theta^*(x, \hat{y}^{k-1})$.

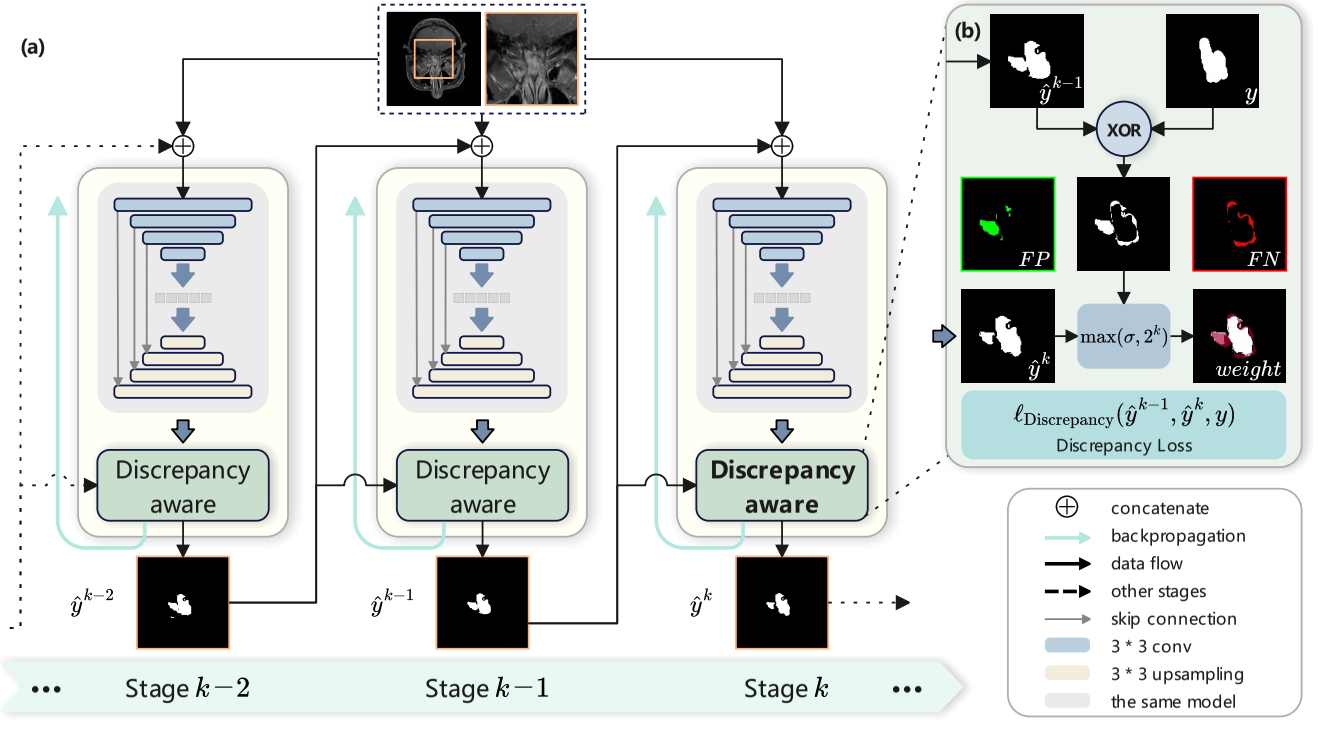


Figure 3.2: The scheme of the proposed ReReNet. (a) The overall data flow of the recurrent process from stage $k-2$ to stage k is depicted. The output mask $\hat{y}$ of one stage is reused as the input of the next stage. (b) The discrepancy-aware optimisation strategy is demonstrated. The discrepancy is obtained by the coarse label and ground truth to weight the loss.

For a given image x, during the segmentation process, the image x and the initial coarse label (non-expert label) $\hat{y}^0$ are concatenated and merged along the feature channel to serve as the input into the model. ReReNet outputs the segmentation mask $\hat{y}^1$. Subsequently, the discrepancy between $\hat{y}^1$ and the ground truth y is obtained for the loss calculation, and the network's parameters are updated through gradient descent. This step is defined as stage 1, whose output $\hat{y}^1$ is utilised as the coarse label for the next stage. The process is repeated K times to complete the training of one batch, and the inference process follows the same procedure.

3.3.2 *Discrepancy-Aware Optimisation Strategy*

Upon the ReReNet, we also propose a discrepancy-aware optimisation strategy to optimise the model's training. Mainstream approaches commonly use the Dice loss (introduced in Chapter 2 Section 2.1.3) for network optimisation. Due to the varying sizes of target regions, it is more robust against class imbalance issues and may offer greater stability during training. For ease of reading, we repeat the general form of the Dice loss:

$$\mathcal{L}_{\text{Dice}}(\hat{y}, y) = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}. \tag{3.2}$$

However, solely using the Dice loss overlooks the valuable information on the coarse label. To effectively utilise this information, we introduce a discrepancy-aware optimisation strategy, i.e., a loss function, enhancing the network's perception of the discrepancy.

The following describes the difference-aware optimisation strategy and its implementation, as shown in the discrepancy-aware module in Fig. 3.2 (b). As understood from the ReReNet framework paradigm, the inputs to the network are $x \oplus \hat{y}^{k-1}$, where x represents the original input image, $\hat{y}^{k-1}$ is the coarse label from the previous stage, and $\oplus$ denotes the concatenate operation. The network output is $\hat{y}^k$, the current stage's output mask. The discrepancy is defined as $\text{XOR}(\hat{y}^{k-1}, y)$, representing the portions incorrectly classified by the model in the previous stage, containing the false positive (FP) and false negative (FN). For these incorrectly segmented regions, a corresponding weight map is built for parameter optimisation, manifested as a loss function. The weight w is used for calculating the loss, where λ is the weight factor, and is defined as

$$w = \lambda(\hat{y}^{k-1} \oplus y) + 1. \tag{3.3}$$

Based on the weight map, the discrepancy loss can be defined as:

$$\ell_{\text{Discr}}(\hat{y}^{k-1}, \hat{y}^k, y) = -\sum_{i}^{N} (w_i \cdot (y_i \cdot \log(\hat{y}_i^k) + (1 - y_i) \cdot \log(1 - \hat{y}_i^k))).$$

(3.4)

Using both ℓ_{Dice} and $\ell_{\text{Discrepancy}}$, the overall network optimisation loss for the ReReNet can be expressed as:

$$\ell(\hat{y}^{k-1}, \hat{y}^{k}, y) = \alpha \cdot \ell_{\text{Dice}}(\hat{y}, y) + (1-\alpha) \cdot \ell_{\text{Discr}}(\hat{y}^{k-1}, \hat{y}^{k}, y), \quad (3.5)$$

where α is the hyperparameter that balances the two losses. To address the unexpected drop in performance observed with naïve cyclic learning concerning the different stages of k, we assign varying weights to the different stages. With the progression of recurrent iterations, the model incrementally intensifies the penalty for errors, expecting better results, as given by:

$$\ell(\hat{y}^{k-1}, \hat{y}^{k}, y)^{k} = \max(\sigma, 2^{k}) * \ell(\hat{y}^{k-1}, \hat{y}^{k}, y), \quad (3.6)$$

where σ is the initial factor. We optimise the entire model across all its stages based on $\ell(\hat{y}^{k-1}, \hat{y}^{k}, y)^{k}$.

3.4 EXPERIMENTS

3.4.1 *Datasets*

We next evaluate the proposed method on three datasets: NPC-MR, LIDC-IDRI, and SegRap2023, as shown in Fig. 3.3.

NPC-MR. This is an in-house dataset collected at the Second Affiliated Hospital of Guangxi Medical University. It comprises MR scans from 362 patients with nasopharyngeal carcinoma (NPC), a malignancy with a high incidence in East and Southeast Asia, particularly in Southern China [31]. Following a two-stage clinical annotation protocol, 10 briefly trained annotators produced coarse lesion delineations using ITK-SNAP [173]; these were subsequently reviewed and corrected by five experienced medical professionals to obtain clinically accurate labels treated as expert ground truth. All slices were cropped to 512×512 pixels. After annotation, quality control was performed to rectify potential connectivity and inter-slice continuity errors, ensuring dataset integrity and usability. The average Dice similarity coefficient (DSC) between the paired non-expert and expert masks is 68.5%, indicating a substantial gap between coarse and clinically corrected annotations.

LIDC-IDRI. This public CT dataset comprises 1,018 cases with multi-observer lung nodule annotations [3]. Scans annotated by at least four physicians were selected, and images were cropped to 256×256 pixels. To derive paired masks as a proxy for expert/non-expert supervision from the multi-observer annotations, pixels with an agreement rate exceeding 50% were treated as the reference (consensus) labels, while the

Table 3.1: Performance evaluation of ReReNet ($\mathcal{K} = 5$) compared with other typical segmentation models whose implementations are publicly available.

Dataset	Method	DSC (%)	IoU (%)	ASSD ↓	HD95 ↓	Precision (%)	Recall (%)
NPC-MR	U-Net	72.19	59.29	5.88	15.70	75.84	74.43
	Attention U-Net	73.18	60.85	7.22	18.35	73.95	78.43
	DeepLabv3	69.90	56.23	6.33	**15.03 (-0.67)**	73.24	72.08
	TransUNet	68.67	55.58	7.60	20.81	68.93	74.43
	Swin-Unet	67.23	53.12	7.08	19.12	69.86	70.97
	ReReNet(ours)	**77.54 (+4.36)**	**66.22 (+5.37)**	**5.36 (-0.52)**	16.39	**78.71 (+2.87)**	**81.39 (+2.96)**
LIDC-IDRI[3]	U-Net	81.28	71.17	3.09	7.05	82.72	84.20
	Attention U-Net	78.63	68.11	4.02	10.75	80.22	82.04
	DeepLabv3	41.15	33.75	51.97	103.55	42.01	43.07
	TransUNet	82.38	71.40	2.18	6.49	82.45	85.16
	Swin-Unet	78.46	67.49	5.79	12.78	79.76	81.04
	ReReNet(ours)	**83.55 (+1.17)**	**73.40 (+2.00)**	**1.13 (-1.05)**	**2.60 (-3.89)**	**84.58 (+1.86)**	**88.39 (+3.23)**
SegRap2023[101]	U-Net	75.45	62.15	3.86	6.89	75.40	81.62
	Attention U-Net	77.34	64.88	3.51	6.15	76.28	83.31
	DeepLabv3	68.92	54.88	4.94	8.67	70.37	74.84
	TransUNet	74.83	62.00	5.96	9.46	71.35	83.76
	Swin-Unet	67.37	53.72	10.47	20.35	67.30	73.94
	ReReNet(ours)	**78.45 (+1.11)**	**65.96 (+1.08)**	**3.34 (-0.17)**	**5.03 (-1.12)**	**77.28 (+1.00)**	**85.52 (+1.76)**

Bold marks the best result; underline marks the second best.

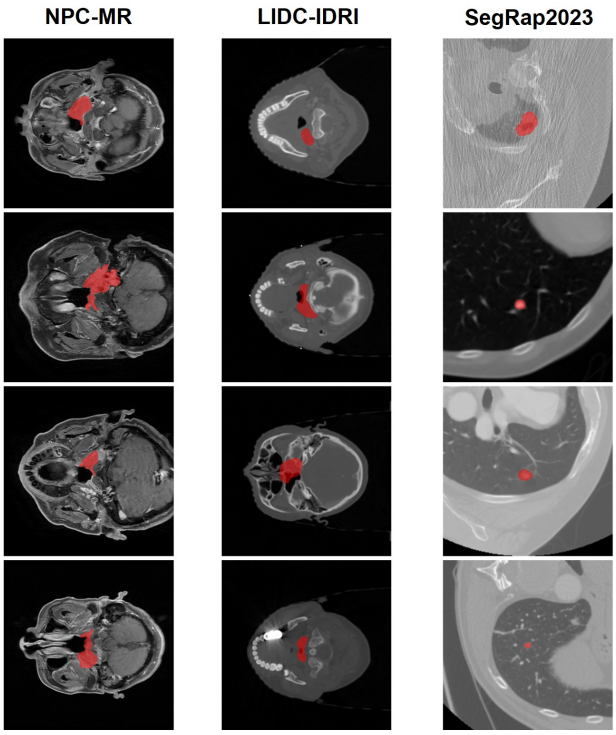


Figure 3.3: Example images from the experimental datasets. Red regions indicate the ground-truth lesion masks.

single annotation with the largest deviation from this consensus was designated as the coarse label. The average DSC between the resulting non-expert and expert pairs is 66.61%.

SegRap2023. This public CT dataset contains gross tumour volume annotations for the primary nasopharyngeal lesion (GTVnx) from 200 NPC patients [101]. Because this dataset does not provide non-expert annotations or multi-observer annotations, non-expert labels are simulated by applying random perspective transformations to the expert ground-truth masks. The average DSC between the simulated non-expert labels and expert ground truth is 60.63%.

3.4.2 *Implementation Details*

Expert labels are treated as ground truth, and non-expert labels are treated as the initial coarse labels for ReReNet. To ensure a comparable distribution of grayscale values across images, normalisation is per-

formed using the mean and standard deviation of the entire training set with z-score standardisation.

Other models, except ReReNet, take a single medical image as input because these models are not designed to incorporate additional label information; they are trained with the Dice loss.

ReReNet was trained using medical images and coarse labels. During the first 20 epochs, only the Dice loss was used to avoid potential instability from introducing the discrepancy loss in the early learning phase.

Models were trained for 100 epochs on the training set, and the best-performing model on the validation set was selected for final metric evaluation on the test set. The Adam optimiser was used for parameter updates, with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}.

For all experiments, hyperparameters are configured as follows, determined via grid search: $\lambda = 10$ in (3.3), $\alpha = 0.8$ in (3.5), and $\sigma = 1$ in (3.6). All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU for training and testing. The operating system is Ubuntu 20.04 LTS, and we use PyTorch (1.13.1) to implement our network and training.

3.4.3 *Experimental Results*

We use multiple evaluation metrics to assess model performance comprehensively. For details and formulas, see Chapter 2 Section 2.1.4. The Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) are the primary metrics for segmentation accuracy, reflecting overlap between predicted and ground-truth segmentations. Boundary differences are assessed using the average symmetric surface distance (ASSD) and the 95th Percentile Hausdorff Distance (HD95). Precision measures the accuracy of positive predictions, while Recall assesses the ability to identify relevant instances.

3.4.3.1 *Segmentation Performance*

We assessed ReReNet by comparing its results with several mainstream segmentation models on the NPC-MR, LIDC-IDRI [3], and SegRap2023 [101] datasets. Of note, unlike semi-supervised learning, which generates pseudo-labels for unlabelled samples and utilises them with certain strategies to improve performance [171], our solution is *fully supervised*; the non-expert labels are provided by non-professionals and have one-to-one corresponding ground truth, which are processed by humans. Therefore, we selected representative supervised segmentation architectures for comparison with ReReNet, as follows. First, we select U-Net as the cornerstone model due to its widespread use and well-established performance in MIS tasks. We also conducted compar-

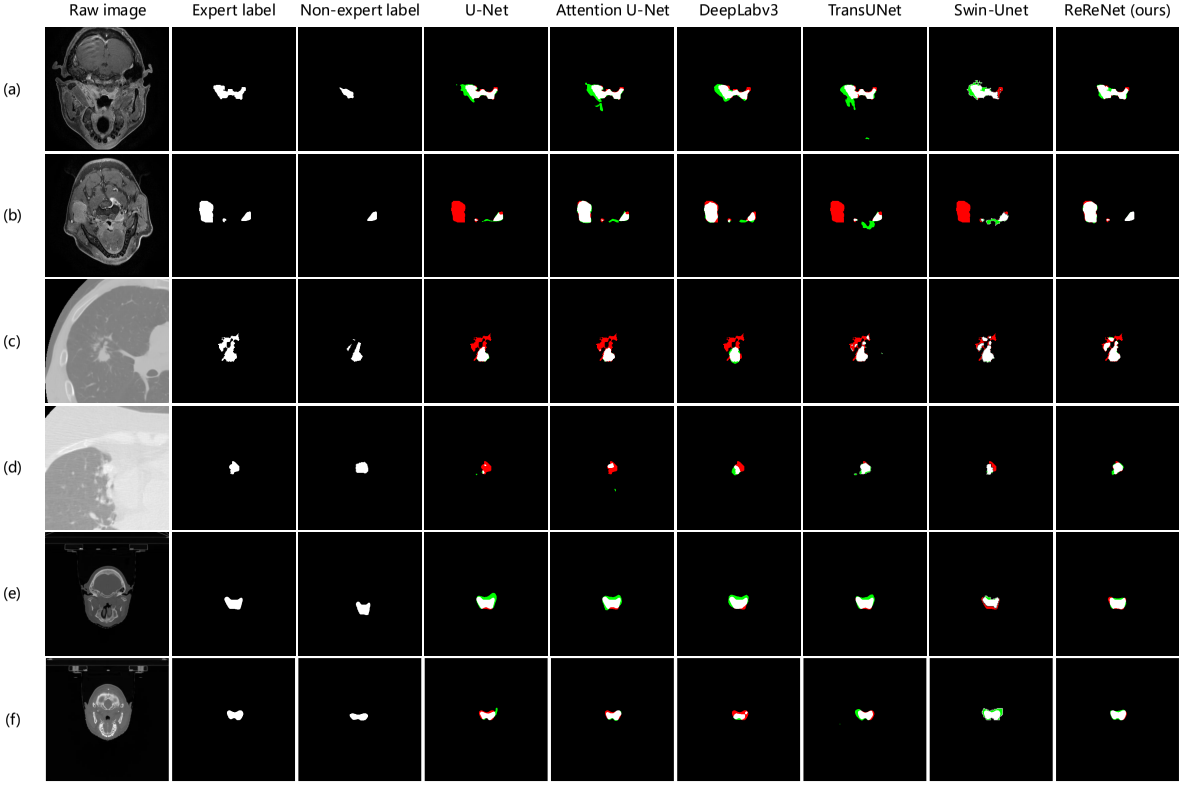


Figure 3.4: Qualitative results of ReReNet comparing other approaches. (a) and (b) belong to NPC-MR, (c) and (d) belong to LIDC-IDRI, and (e) and (f) belong to SegRap2023. White denotes correct segmentation, red denotes under-segmentation, and green denotes over-segmentation.

ative experiments with other prominent architectures, including Attention U-Net [116], DeepLabv3 [28], and the Transformer-based models TransUNet [26] and Swin-Unet [19].

As Table 3.1 shows, ReReNet demonstrates higher performance than the other tested models across all three datasets. On NPC-MR, ReReNet achieves the most substantial improvements, with DSC and IoU increasing by 4.36% and 5.37%, respectively, relative to the second-best method (Attention U-Net), indicating higher segmentation accuracy. The improvement in Precision (2.87%) and Recall (2.96%) demonstrates that ReReNet achieves a better balance between avoiding false positives and capturing true positives. Notably, ReReNet achieves the best ASSD (5.36 mm), reducing boundary errors by 0.52 mm compared to U-Net, which suggests superior boundary delineation capability. While DeepLabv3 shows a marginal advantage in HD95 (0.67 mm reduction), ReReNet maintains competitive boundary accuracy while having very good overall segmentation metrics.

On LIDC-IDRI, ReReNet achieves consistent improvements across all metrics, with DSC and IoU increasing by 1.17% and 2.00%, respectively, compared to TransUNet (the second-best method). The boundary accuracy improvements are notable: ReReNet reduces ASSD by 1.05 mm (from 2.18 mm to 1.13 mm) and HD95 by 3.89 mm (from 6.49 mm to 2.60 mm), representing substantial enhancements in boundary precision. The Recall improvement of 3.23% (from 85.16% to 88.39%) indicates that ReReNet is particularly effective at identifying lung nodules that might be missed by other methods, which is crucial for clinical applications where false negatives can have serious consequences.

On SegRap2023, ReReNet maintains its performance advantage with improvements of 1.11% in DSC and 1.08% in IoU over Attention U-Net. The boundary metrics show consistent improvements, with ASSD reduced by 0.17 mm and HD95 reduced by 1.12 mm. The Recall improvement of 1.76% (from 83.76% to 85.52%) demonstrates ReReNet's ability to capture more complete lesion boundaries, which is essential for accurate gross tumour volume delineation in radiotherapy planning.

Across all three datasets, ReReNet outperforms all CNN-based architectures (U-Net, Attention U-Net, DeepLabv3) and Transformer-based models (TransUNet, Swin-Unet). This cross-architecture superiority suggests that the benefit of incorporating non-expert labels through discrepancy-aware supervision is architecture-agnostic and could be generalised to different model designs. The consistent performance gains across datasets with different characteristics (MR vs CT, different anatomical regions, varying annotation quality) further validate the robustness and generalizability of the proposed approach.

Fig. 3.4 presents a visual qualitative comparison of the binary masks generated by ReReNet and other models. ReReNet yields more accurate segmentation results with fewer over-segmentation and under-segmentation errors. For instance, in Fig. 3.4 (b), left-side lesions that

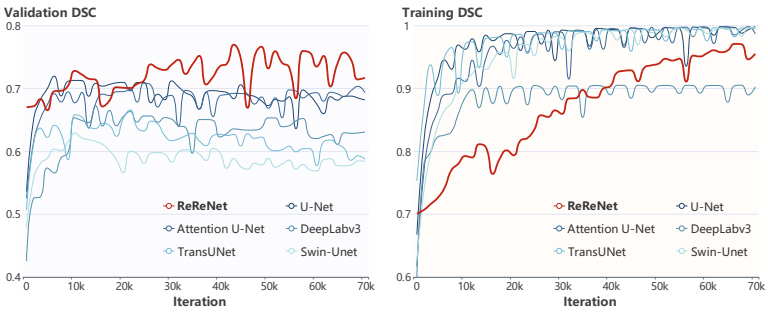


Figure 3.5: Trend of the DSC with training iteration on the NPC-MR dataset.

are incorrectly identified or completely missed by other models are correctly recognised by ReReNet. Moreover, guided by the non-expert label, ReReNet achieves more accurate segmentation of the right-side lesions, demonstrating its ability to leverage coarse annotations to refine boundary delineation. The visualisations across different datasets consistently show that ReReNet produces segmentation masks that are closer to the ground truth with better preservation of lesion boundaries and reduced spurious detections.

Fig. 3.5 illustrates the DSC variation on the validation set during training on the NPC-MR dataset. As training progresses, other methods gradually exhibit signs of overfitting after 10,000 iterations, with their validation DSC plateauing or even declining. This overfitting behaviour is particularly evident in Transformer-based models (TransUNet and Swin-Unet), which show more pronounced performance degradation in later training stages. In contrast, with discrepancy-aware supervision, ReReNet exhibits improved training effectiveness and stability, particularly after 30,000 iterations. While fluctuations are observed, the validation DSC does not show a distinct downward trend, suggesting no clear overfitting. Note that our comparisons use the best validation checkpoint for each method (equivalently, allowing early stopping). Hence, even after removing the effect of late-iteration overfitting via model selection, ReReNet remains consistently better than the baselines in terms of peak DSC. The discrepancy-aware loss mechanism helps ReReNet maintain a steady learning trajectory by adaptively weighting the supervision signal based on the reliability of non-expert labels, thereby mitigating overfitting and enabling continued performance improvement throughout training. This training stability is crucial for practical applications where robust convergence behaviour is essential.

Overall, the experimental results demonstrate the strong performance of ReReNet in the MIS task, with significant improvements across the evaluation metrics. These findings indicate that the proposed ReReNet can effectively utilise coarse labels to enhance segmentation

Table 3.2: Performance gains of ReReNet compared with the baseline model at different label ratios on NPC-MR dataset.

Metric	Method	Label ratio					Average
		10%	**30%**	**50%**	**70%**	**100%**	
DSC (%)	U-Net	60.34	65.72	69.41	71.21	72.19	67.77
	ReReNet	**68.17** (+7.83)	**69.81** (+4.09)	**70.23** (+0.82)	**74.54** (+3.33)	**77.54** (+5.35)	**72.06** (+4.29)
IoU (%)	U-Net	46.65	51.64	56.01	58.20	59.29	54.35
	ReReNet	**56.39** (+9.74)	**57.83** (+6.19)	**58.18** (+2.17)	**62.23** (+4.03)	**66.22** (+6.93)	**60.17** (+5.82)
HD95 ↓	U-Net	43.22	28.00	**19.30** (-6.63)	**15.38** (-1.74)	**15.70** (-0.69)	24.32
	ReReNet	**27.49** (-15.73)	**25.56** (-2.44)	25.93	17.12	16.39	**22.50** (-1.82)
Precision (%)	U-Net	57.40	61.20	73.85	74.18	75.84	68.50
	ReReNet	**76.08** (+18.68)	**73.44** (+12.24)	**74.18** (+0.33)	**74.56** (+0.38)	**78.71** (+2.87)	**75.39** (+6.89)

performance, achieving superior results through discrepancy-aware supervision and iterative refinement.

3.4.3.2 *Performance at Different Label Ratios*

To assess ReReNet under varying label budgets (on NPC-MR), we construct five training subsets: 10%, 30%, 50%, 70%, and 100% of the available paired samples (each with both non-expert and expert masks). This setting directly reflects **CII (annotation quantity)** by evaluating performance when only a small fraction of paired annotations can be obtained. Segmentation results of ReReNet and the baseline model are compared across these label ratios, as illustrated in Table 3.2. As the ratio of labels increases, the performance metrics consistently improve, confirming the well-established principle that larger training datasets enhance model performance. Moreover, the baseline model improves more markedly when the label ratio increases from 10% to 50%, whereas the improvement from 50% to 100% exhibits diminishing marginal returns. This reflects the common trade-off between marginal accuracy gains and the effort required to acquire additional labels.

In these experiments, ReReNet reduces reliance on labels while maintaining accurate segmentation. Across label ratios, ReReNet generally outperforms the baseline segmentation model. With only 10% labels, ReReNet attains better performance (IoU 56.39%) than the baseline model trained with 50% labels (IoU 56.01%). With 70% labels, ReReNet (DSC 74.54%) surpasses the baseline model trained with 100% labels (DSC 72.19%).

These results indicate that ReReNet can achieve strong segmentation performance with limited label resources, corresponding to **CII (annotation quantity)** of this thesis: when only a modest number of expert-annotated cases are available due to the high cost and poor scalability of voxel-wise delineation. By exploiting coarse-to-expert refinement signals rather than requiring additional fully labelled volumes, ReReNet improves data efficiency and thus offers a practical solution to the high cost of labelling in MIS.

3.4.3.3 *Visualisation of Intermediate Output*

Figure 3.6 visualises the output mask at each step during the iterative inference of ReReNet ($K = 5$). In the first step, the output segmentation contains many inaccurate regions. As iterations proceed, the model gradually corrects these regions and ultimately produces more accurate segmentation results. These visualisations demonstrate ReReNet's ability to reduce segmentation errors by iteratively updating and refining its outputs.

Table 3.3: Ablation study of introducing non-expert labels, recurrent mechanism, and discrepancy-aware optimization strategy.

Non-Expert	Expert	Recurrent	Discrepancy	DSC (%)
✓	✗	✗	✗	65.60 (-6.59)
✗	✓	✗	✗	72.19
✓	✓	✓	✗	76.71 (+4.52)
✓	✓	✓	✓	**77.54 (+5.35)**

Table 3.4: Study on hyperparameter K.

Total stages	$K = 1$	$K = 3$	$K = 5$	$K = 7$	$K = 9$
DSC (%)	77.15	75.36	**77.54**	72.28	72.27

3.4.4 *Effectiveness of Components of ReReNet*

Table 3.3 illustrates the unreliability of non-expert labels for the NPC-MR dataset. When supervised only by such non-expert labels, the trained UNet performs worse than those supervised by expert labels, as anticipated. We next evaluate the effectiveness of the recurrent mechanism and the discrepancy-aware optimisation strategy. ReReNet ($K = 5$), equipped solely with the recurrent framework, achieves a DSC of 76.71%, exceeding the baseline by 4.52%. Incorporating the discrepancy-aware module further boosts DSC by 0.83%, reaching 77.54%. This improvement underscores the efficacy of the recurrent mechanism and the discrepancy-aware module.

3.4.5 *Impact of Varying the Number of Total Stages*

A study on the hyperparameter K (total stages of ReReNet) is conducted on NPC-MR, with results displayed in Table 3.4 and Fig. 3.6. Here, K controls the number of recurrent refinement stages during inference for the refinement depth of the overall pipeline. At $K = 1$, ReReNet already achieves a strong DSC of 77.15%, supporting the benefit of incorporating coarse labels. As K increases, performance drops initially, peaks at $K = 5$ (DSC= 77.54%), and then degrades substantially with further stages. This indicates that moderate recurrence facilitates refinement, whereas excessive iterations increase loop complexity and error accumulation. Accordingly, we adopt $K = 5$ as the default setting. A study on the hyperparameter K (total stages of ReReNet) is conducted on NPC-MR, with results displayed in Table 3.4 and Fig. 3.6.

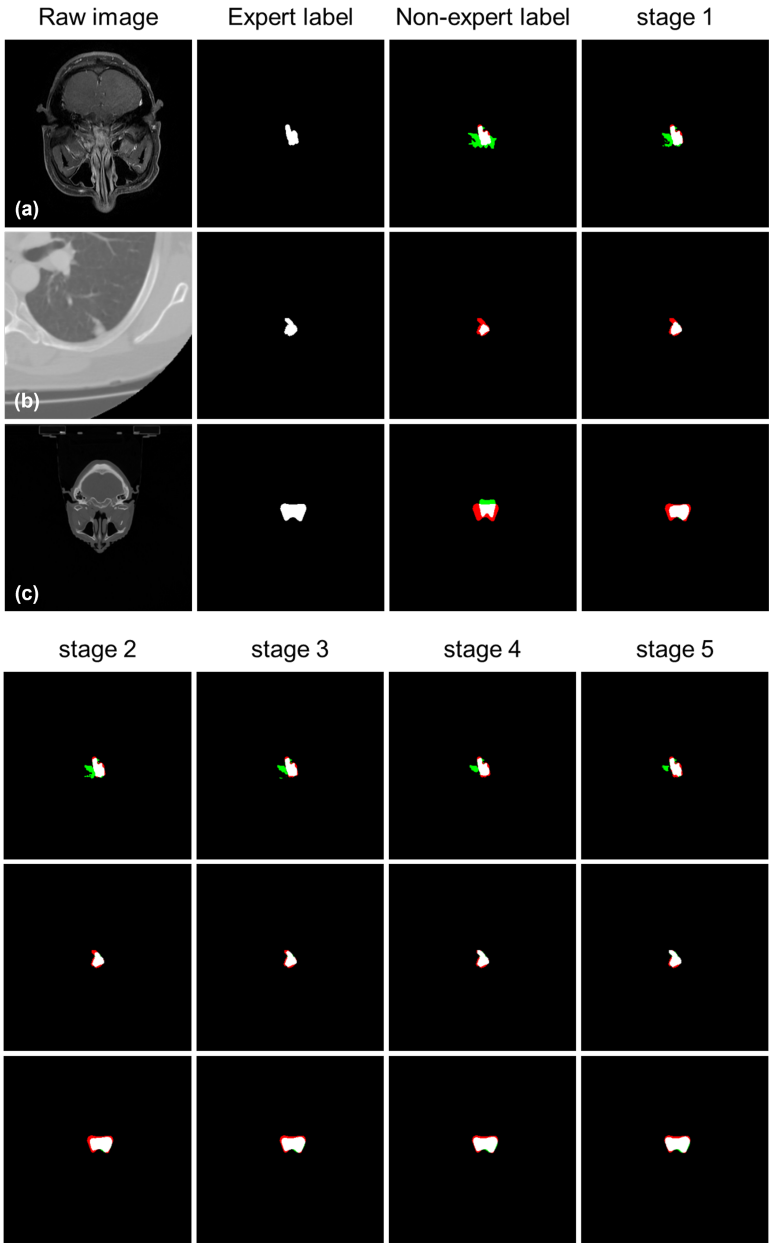


Figure 3.6: Intermediate output masks of ReReNet from stage 1 to 5 when $K = 5$, white denotes correct segmentation, red denotes under-segmentation, and green denotes over-segmentation. Best viewed in colour. Panels (a)–(c) show three examples from NPC-MR, LIDC-IDRI, and SegRap2023, respectively. As refinement progresses, the segmentation masks gradually improve.

3.5 CHAPTER SUMMARY

In this chapter, we have presented the Recurrent Refined Network (ReReNet), a novel framework that exploits the two-stage annotation workflow—where non-experts provide coarse masks subsequently refined by clinicians—to learn latent correction patterns for medical image segmentation. By applying a discrepancy-aware optimisation strategy, ReReNet internalises the expert refinement process and progressively transforms approximate delineations into clinically accurate segmentations.

Extensive experiments on three datasets spanning MR and CT modalities demonstrate that ReReNet consistently outperforms mainstream architectures (used in a supervised mode) in lesion segmentation. The results confirm that mining hidden correction patterns from imperfect labels provides a practical path toward robust segmentation: the model improves both boundary consistency and lesion completeness without discarding valuable non-expert annotations.

Several directions remain for future work. First, outputs from rule-based algorithms or foundation models could replace non-expert labels at inference, broadening applicability to scenarios where paired annotations are unavailable. Second, extending the framework to native 3D architectures would better capture the volumetric context inherent to medical imaging data.

The next chapter refines the label-limited experimental regime already considered here. Whereas this chapter focuses on learning from heterogeneous annotation quality via non-expert-to-expert refinement, Chapter 4 makes limited annotation quantity the primary constraint and addresses it through self-supervised representation learning on unlabelled volumetric data.

4

SELF-SUPERVISED ROTATION LEARNING FOR MEDICAL IMAGE SEGMENTATION UNDER LABEL-QUANTITY LIMITATION

The preceding chapter addressed the challenge of heterogeneous annotation quality, demonstrating how refinement-based signals can improve segmentation robustness under imperfect supervision. This chapter turns to a complementary constraint that frequently co-occurs in practice: limited annotation quantity. In volumetric medical imaging, the cost of expert voxel-wise delineation is substantial, and available labelled data often constitute only a small fraction of the acquired scans. As introduced in Chapter 1, this thesis addresses annotation quantity limitations through label-efficient learning strategies that exploit unlabelled volumes via self-supervised representation learning.

To this end, we propose 3DRotNPC, a framework based on a tailored self-supervised learning (SSL) strategy to accurately segment tumour regions under the circumstance of label limitation. Segmenting the tumours of nasopharyngeal carcinoma (NPC) in Magnetic Resonance Imaging (MRI) images serves as the clinical application scenario, where the precise identification and location of lesion regions typically require plenty of labels. To learn the rich 3D spatial and geometric nature of MRI images in a self-supervised manner, we design a proxy task of randomly selecting and rotating images in consecutively sliced data. After the SSL pre-training stage, the learned parameters of the Convolutional Neural Networks (CNNs) based model are transferred to adapt to the downstream segmentation task. We verify the capability of 3DRotNPC on an NPC tumour dataset collected and curated from clinical treatment in a representative hospital. Extensive experiments demonstrate that our approach delivers considerable gains in downstream 3D voxel segmentation[1].

4.1 INTRODUCTION

Nasopharyngeal carcinoma (NPC) is one of the highly prevalent malignant tumours in East and Southeast Asia. Early detection and early treatment can significantly improve prognosis [31]. Several imaging techniques are commonly utilised for staging and radiotherapy, such as magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET/CT). Clinically, radiotherapy is rel-

1 Parts of this chapter were published in: **C. Li**†, C. Yao, X. Ban*, S. Yin, and M. S. Obaidat. Self-Supervised Rotation Learning for 3D Segmentation on Nasopharyngeal Carcinoma MRI Images. *In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3529–3534. IEEE, 2023. doi: 10.1109/BIBM58861.2023.10385483

atively effective in the treatment of NPC [150]. It is of vital importance to achieve precise identification of the tumour [109], which is the basis of the following diagnosis and treatment.

Many methods have been proposed to segment the regions of NPC tumours [9, 97, 136]. As the boundary of the tumour is usually unclear and easily intermingles with the surrounding tissue, rule-based methods are barely effective under these circumstances [37, 137]. Deep learning-based methods bring promising performances on the NPC segmentation task [63, 90, 97]. Architectures of CNNs and Transformers have been around to model the tumour region and background relationships.

Although the former methods have attained considerable results [136, 138], we believe that there are still two main challenges to accurately and efficiently obtaining the regions of the tumour. On the one hand, 3D information has yet to be leveraged [78]. There are numerous works focusing on the 2D image segmentation of MRI or CT images of NPC [82]. However, the characteristics hidden in the sliced 3D volume data are not fully utilised [73], i.e., the continuity and similarity between every two adjacent slices. On the other hand, a notorious problem of medical image analysis is that data labelling often requires expert knowledge and costs substantial human resources. The laborious labelling work hinders the progress of precisely and efficiently detecting the tumour regions.

To address the above issues, we propose 3DRotNPC, a tailored self-supervised learning (SSL) strategy to learn the spatial and geometrical natures of the consecutive slices of MRI images. Specifically, we design a new proxy task for the consecutively sliced MRI images. A continuous fragment of slices can be seen as a 3D volume. Before feeding into the neural network, several slices in the fragment are randomly selected to be rotated. The CNN model takes the transformed continuous slices as input and encodes the volume to a low-dimensional latent feature representation. After that, a modified classification head is set to output the prediction of the rotation angle of each slice. After the unsupervised pre-training stage, the downstream tumour segmentation task is conducted to deliver the binary mask of tumour regions. For simplicity, we employ a regular but effective 3D voxel segmentation model, ResUNet3D [79], in order to validate the efficiency and effectiveness of the proposed 3DRotNPC strategy. We have carried out several experiments on proxy feature representation learning tasks and downstream transferring tasks. The results show that models trained based on the 3DRotNPC strategy deliver better 3D voxel segmentation results than models without 3DRotNPC, especially for 10% to 30% of the data labels.

To our best knowledge, 3DRotNPC is the first work to accomplish 3D tumour segmentation of NPC with the SSL paradigm. In general, our contribution in this direction is threefold:

- We propose 3DRotNPC, a framework to realise segmenting the tumour regions of NPC MRI images effectively, mainly comprising self-supervised pre-training and label-efficient 3D voxel segmentation.
- We design a tailored self-supervised Rotation Learning proxy task that models the rich 3D spatial and geometrical nature of consecutively sliced MRI images, benefiting the concerned downstream tumour segmentation task.
- We validate the 3D voxel segmentation performance of 3DRotNPC through abundant experiments and conclude that the proposed strategy benefits knowledge transfer from the pre-trained model to the downstream task.

4.2 RELATED WORK

With the rapidly increasing demand for automated identification of lesion areas in medical images [65], a large number of works have emerged in the field of MIS [67, 132]. Nonetheless, for nasopharyngeal carcinoma, the number of reported solutions to segment the regions of tumours is not particularly high. One of the main reasons is that it has a prominent geographical distribution, distinctly different from other epithelial head and neck tumours [31, 134]. In recent years, a number of related methods have been proposed. A multi-stage rendering approach adopts a progressive strategy to address the background dominant problem [85]. Li et al. [86] put forward NPCNet, a milestone work that detects primary tumours and metastatic lymph nodes jointly. Considering the data form of the MRI or CT images, the 3D spatial information contained in consecutive slices has not been mined. There exist several works to formulate and solve the 3D voxel segmentation in medical image analysis [168].

A key problem of the aforementioned Deep Learning based methods is the extensive and costly data labelling. SSL [174], another type of learning paradigm, aims to learn robust and generable feature representations by predicting the transformation properties of the data themselves [115], drastically alleviating the demand for the number of labels. As detailed in Chapter 2 Section 2.3, volumetric medical imaging is acquired slice-by-slice and exhibits inter-slice continuity and anatomical coherence, which naturally motivates geometry-aware proxy tasks that exploit 3D structure.

Several works have tried to obtain spatial and geometrical properties through SSL. Properties such as relative location [2], sequence order [164], and motion statistics [146] have been explored. Jing et al. [69] make the CNN learn spatiotemporal features by predicting the rotation of the whole video. [135] randomly crop and augment the input CT images, and use inpainting and rotation prediction as proxy tasks for

learning contextual representations. However, despite these advances, SSL remains underexplored for NPC segmentation, particularly in settings that must exploit the intrinsic 3D continuity of MRI volumes. Most existing proxy tasks are formulated for generic 2D/temporal data and do not explicitly encourage slice-consistent geometric understanding in consecutively acquired scans. This motivates our approach: we design an SSL pre-training task that explicitly models 3D inter-slice continuity and geometric consistency, enabling label-efficient volumetric segmentation under limited expert annotation.

4.3 METHODOLOGY

4.3.1 *Overview*

The overall pipeline of our 3DRotNPC is illustrated in Fig. 4.1. The preprocessing and transformation of a fragment of consecutive slices are performed first, as shown at the top of the figure. Meanwhile, the classification labels are generated. Then, self-supervised pre-training is conducted, shown in the blue part of the figure. Finally, the 3D voxel segmentation task is carried out with the parameters pre-trained on the proxy task, including the encoder and bottleneck part of the model, which are to be fine-tuned, depicted in the red part. We employ a modified ResUNet3D architecture to learn the 3D spatial and geometric features. In our setting, the input volume has a depth of 16, where each z-axis layer corresponds to one slice in a fragment of consecutive slices of MRI images.

4.3.2 *Self-Supervised Rotation Learning*

4.3.2.1 *Motivation*

A proper pre-training strategy, i.e., finding beneficial proxy tasks for effective supervision, plays a fundamental role in SSL. We have observed the continuity and similarity between slice-by-slice contents. In the actual practice of diagnosing, experienced experts usually look through the context of one particular slice to judge the existence of the tumour. In addition, geometric features such as symmetry and the approximate location of the interested tumour regions are mainly considered. Motivated by that prior knowledge, we seek to let the model imitate those behaviours by implicitly learning the spatial and geometric nature of the raw data.

4.3.2.2 *Randomized Rotation Transformation*

Inspired by 3DRotNet[69] and Swin UNETR[135], we modify and improve the slices rotation strategy. Firstly, for a fragment of consecutive

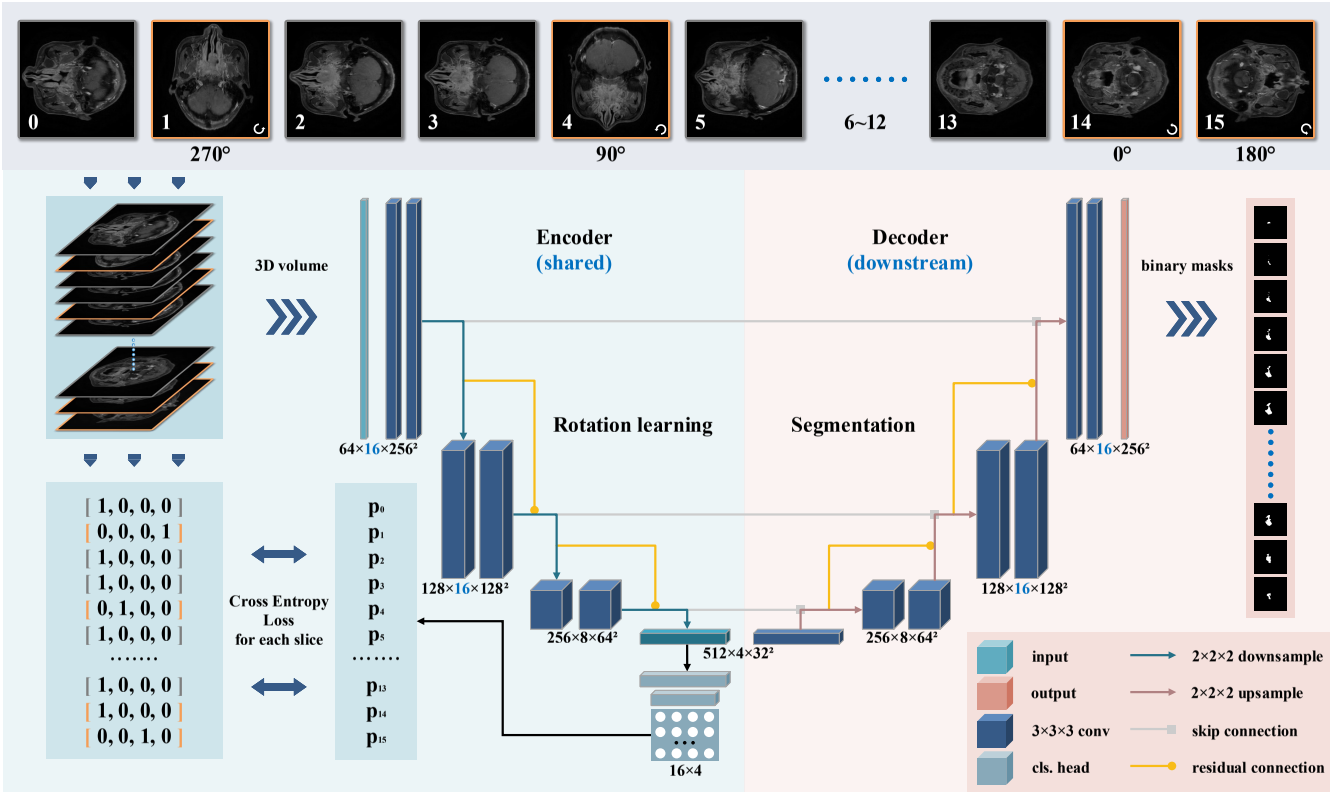


Figure 4.1: The overview pipeline of 3DRotNPC includes data preparation, a self-supervised proxy task, and a downstream 3D voxel segmentation task.

slices, 4 slices (one quarter of the number of slices of one sample) are randomly selected to be rotated. For the sake of achieving the perception of geometric characteristics, we randomly rotate the selected 4 slices, with rotation angles of 90, 180, and 270 degrees; the rationale for choosing four slices is provided in Section 4.4.3.2 It should be noted that each slice also has a probability (25%) of not being rotated. Degrees of 0, 90, 180, and 270 correspond to the classes (0, 1, 2, 3), respectively. The transformation is:

$$T(x \mid \mathbf{s}, \mathbf{y}) = \text{Rot}\left(x_{s \in \mathbf{s}}, y \in \mathbf{y}\right), \tag{4.1}$$

where x is a multi-layer volume, s is the vector of indices for selected slices, and y is the vector of rotation classes for them.

Secondly, once the initial volume is transformed following the rotation protocol, it is prepared to be fed into the encoder part of the modified ResUNet3D $R(\cdot \mid \theta_R)$. The encoder part of it, $E(\cdot \mid \theta_E)$, encodes the input data. After multiple consecutive convolutions and corresponding downsampling operations, the original data is compressed into a low-dimensional feature representation.

Thirdly, we design a classification head $H(\cdot \mid \theta_H)$, after $E(\cdot \mid \theta_E)$, to predict the rotation angle of each input slice. We reshape the output probability map to (batch size, 16, 4). The second term represents the depth of the volume, and the third term is the predicted class scores (normalised to a probability distribution) over the four rotation angles for each slice after the sigmoid activation. To get the loss of rotation angle prediction, we adopt the Weighted Cross Entropy (WCE) method to calculate the loss between the angle classification and the generated true classification for each slice separately, and weigh it based on the number of rotations that happened or not, namely:

$$\mathcal{L}_{\text{slice}}\ (x_s \mid \theta_C) = -\frac{1}{M} \sum_{y=1}^{M} w_y \log\left(C\left(T\left(x_s \mid s, y\right) \mid \theta_C\right)\right). \tag{4.2}$$

Among them, $C(\cdot \mid \theta_C)$ is the combination of the encoder part and the classification head, x_s in a one-layer slice, w_y is a class-dependent weight that re-scales the cross-entropy contribution of rotation class y to compensate for class imbalance (in particular, the higher prevalence of the 0° class) and to discourage the trivial solution of predicting all slices as unrotated. We compute w_y per mini-batch using inverse-frequency weighting, i.e., $w_y = 1/p_y^{(b)}$, where $p_y^{(b)}$ is the empirical proportion of class y within the current mini-batch. At last, all the losses are summed up and averaged to get the overall loss:

$$\mathcal{L}\left(x \mid \theta_C\right) = \min_{\theta_C} \frac{1}{N} \sum_{i=1}^{N} l_{\text{slice}}\ (x_i \mid \theta_C), \tag{4.3}$$

where N is the total number of slices in one single volume. To summarise, when the model finishes the 3DRotNPC proxy task, we consider

it capable of perceiving the adequate spatial and geometrical nature of the data, and that perception is expected to empower the downstream 3D voxel segmentation.

4.3.3 *3D Voxel Segmentation*

In most SSL methods, the architecture of the proxy and downstream tasks is identical. Here, we are concerned with the extent to which the designed self-supervision affects the downstream segmentation. For simplicity and following the conventions of SSL, we use ResUNet3D in the two stages: we copy the parameters of the pre-trained encoder and bottleneck parts and paste them into the segmentation model correspondingly as initialisation parameters, and then fine-tune the model. The decoder part of the ResUNet3D is utilised to generate the segmentation binary masks, illustrated on the right of Fig. 4.1. After pre-training with 3DRotNPC, the model is expected to be integrated with 3D spatial and geometric information. In the next part, our experiments verify the performance of 3D voxel segmentation with the number of labels as a variable.

4.4 RESULTS

4.4.1 *Data Collection and Curation*

We conduct experiments on the collected 1000+ samples of MRI images of NPC from 422 patients, with 856 of these samples containing NPC tumours. For each sample, the spatial resolution of the axial slices ranges from 0.25 mm to 1.29 mm, with a slice interval of 5.2–5.7 mm. Each axial slice has 512×512 pixels, and the number of slices acquired per scan is from 24 to 52. Among these samples, 166 have the NPC tumour regions labelled by skilled medical experts. These precise manual labels serve as the golden standard for the tumour segmentation task. Partial samples are shown in Fig.4.2.

For the dataset preparation, in the self-supervised pre-training task, we employ the complete dataset of NPC MRI images without labels, and in the supervised segmentation task, we build our dataset based on individual patients, and each patient's data encompasses varying numbers of slices. Out of the 166 patients, 135 samples are assigned to the training set, 15 to the validation set, and 16 to the test set. Additionally, to enhance computational efficiency, we crop out regions outside the nasal cavity. The input for the model is scaled down to the shape of (16, 256, 256). The image normalisation, label alignment, and other preprocessing are performed to ensure data consistency and usability.

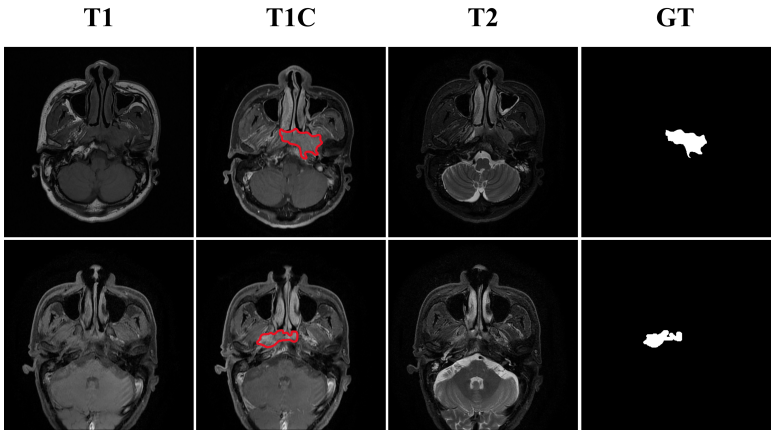


Figure 4.2: The MRI image examples include sequences such as T1-weighted (T1), contrast-enhanced T1-weighted (T1C), and T2-weighted (T2). It can be seen that the T1C delivers relatively clear effects with appropriate contrast.

4.4.2 *Implementation Details*

The model training and inference are performed on a Deep Learning server with one NVIDIA GeForce RTX 3090 24GB GPU, implemented using the PyTorch library. We adopt the Adam optimiser to update parameters for 100 epochs, with an initial learning rate of 1×10^{-4}. Regarding the self-supervised pre-training task, the model with the best performance in the proxy classification task is chosen as the pre-trained model. In segmentation, the Dice and IoU coefficients are calculated to evaluate the performance. The model exhibiting the best performance on the validation set is chosen as the final inference model. Considering the sparsity along the z-axis, we modify the ResUNet3D to preserve continuous spatial information. Specifically, the first pooling does not involve the operation along the z-axis and is exclusively performed in 2D.

4.4.3 *Experimental Results on Single-Centre Setting*

We initially performed a selection of base models to ensure that the selected one is appropriate for our study. Experiments of supervised 3D voxel segmentation of NPC tumours are conducted employing the typical CNN-based 3D-UNet [33], and ResUNet3D [79], and self-attention-based UNETR [52]. Based on empirical results (shown in Table 4.1), ResUNet3D has superior performance compared to other architectures. For the uncompetitiveness of UNETR, our speculation is that available sample-label pairs are inadequate for exploiting its advantages in full,

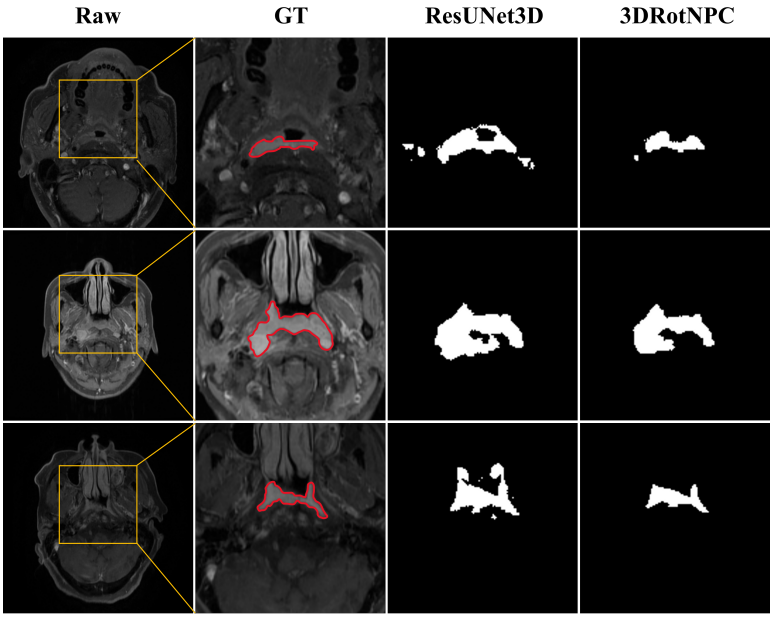


Figure 4.3: Visualization comparisons of segmented binary masks among ResUNet3D and 3DRotNPC at the label ratio of 30%.

which corroborates the viewpoint from [131]. Consequently, we opt to use ResUNet3D as our primary model.

4.4.3.1 *Segmentation Performance*

We conducted a series of experiments, comparing the performances among 3D voxel segmentation models trained with and without rotation prediction pre-training, to validate the effectiveness of incorporating the SSL strategy. To assess model performances across varying scales of the dataset, we arrange the data into four levels of the number of labels: 10%, 30%, 50%, and 100%, as presented in Table 4.1. By comparing the results of the modified ResUNet3D (trained from scratch) and the 3DRotNPC at different label ratios, we observe that as the label ratio increases, the segmentation results measured by two evaluation metrics also improve, which confirms the validity of adding more labels to train on.

On average, the improvement in accuracy from a 10% to 30% label ratio is more substantial than the improvement from a 30% to 100% ratio, suggesting diminishing marginal returns in accuracy gain with increasing label quantities. This presents a practical problem of the trade-off between investing more label effort for marginal accuracy improvements.

Table 4.1: Comparisons of 3D Voxel Segmentation Performance Between 3DRotNPC and Other Models under Different Label Ratios

Metric	Mode\Ratio	5%	10%	30%	50%	100%	Avg
Dice (%)	UNETR	31.19	33.65	42.53	46.69	52.23	41.26
	3D-UNet	34.35	44.72	58.61	58.32	57.65	50.73
	ResUNet3D	54.59	61.94	69.23	69.94	71.91	65.52
	3DRotNPC	**54.63 (+0.04)**	**63.88 (+1.93)**	**70.31 (+1.08)**	**71.04 (+1.10)**	**72.21 (+0.30)**	**66.41 (+0.89)**
IoU (%)	UNETR	19.79	22.16	29.64	33.22	37.83	28.53
	3D-UNet	22.31	31.72	44.30	44.53	43.92	37.36
	ResUNet3D	38.93	45.84	53.71	54.88	57.17	50.11
	3DRotNPC	**39.12 (+0.18)**	**47.86 (+2.02)**	**55.04 (+1.33)**	**55.86 (+0.98)**	**57.35 (+0.17)**	**51.04 (+0.94)**

Table 4.2: Ablation Studies of 3DRotNPC at K=2, 4, and 6

Metric	Mode\Ratio	Avg
Dice (%)	ResUNet3D	65.59
	Rot-2	66.55 (+0.96)
	Rot-4	**67.09 (+1.50)**
	Rot-6	66.14 (+0.56)
IoU (%)	ResUNet3D	49.77
	Rot-2	50.77 (+1.00)
	Rot-4	**51.45 (+1.68)**
	Rot-6	50.67 (+0.90)

In our experiments, 3DRotNPC substantially mitigates the model's dependency on labels, achieving accurate segmentation with minimal labels. Specifically, across different label ratios, 3DRotNPC consistently outperforms non-pre-trained segmentation models. The largest improvement can be found for the 10% label ratio for 3DRotNPC. Seeing the Dice coefficient at the 10% label ratio, the proxy task yields a 1.93% improvement over the plain ResUNet3D's result of 61.94%.

Importantly, the experimental results demonstrate that our self-supervised pre-training, with only 50% of the labels available, achieves 71.04%. 3DRotNPC elevates 3D voxel segmentation performance to a level close to that achieved with nearly full labels (71.91%), indicating the efficiency of our method. Visualisations of generated segmentation masks of ResUNet3D and 3DRotNPC are demonstrated in Fig. 4.3. Tumour regions of NPC are delineated in red. Our method produces more precise, clearer binary masks than ResUNet3D.

Notably, the gain from 3DRotNPC in the downstream segmentation task is less pronounced at label ratios above 50% than at lower ratios. Using the Dice coefficient to elaborate, the improvement is only 0.3% at the 100% label ratio. When training samples are as few as six (5%), there is also barely an improvement using the rotation strategy. We argue that in the scenario with sufficiently many samples, the generalisation brought by the designed proxy task is not that beneficial for the downstream task because of the dominance of large-scale data and supervised training. On the other hand, when the number of labels is extremely low, the model is not able to learn useful features from monotonous data, not to mention the effectiveness of transferred knowledge.

These results and findings strongly indicate that the proposed self-supervised rotation learning strategy exerts a positive influence on the downstream segmentation task with a limited quantity of labels.

4.4.3.2 *Ablation Analysis*

To verify the effectiveness of the chosen number of slices to be rotated, i.e., $k = 4$. we conduct ablation studies on the variable of the number of slices (2, 4, and 6) to be rotated. The results are given in Table 4.2, reporting the average results of performances under 10% and 30% label ratios. With 4 slices rotated, the segmentation performance is the best. The average Dice and IoU coefficients are improved by 1.5 and 1.68, respectively. We conjecture that when there are relatively more or fewer slices to be rotated, the difficulty of predicting the proper rotation classes is either too hard or too simple. That is to say, the model does not capture the 3D characteristics appropriately, resulting in no comparable gains in the downstream segmentation task.

4.4.4 *Experimental Results on Multi-Centre Setting*

As shown in the previous section, even with an SSL strategy, the sparsity of the data distribution under very low sample sizes still significantly limits the transfer efficiency of learned representations; dependence on data quantity remains one of the major challenges in medical image segmentation. Single-centre data are inherently limited, as each institution serves a finite patient population. Thus, how to effectively integrate multi-centre medical data, build datasets of sufficient scale and representativeness, and extract generalisable features from them becomes an important problem for improving model performance.

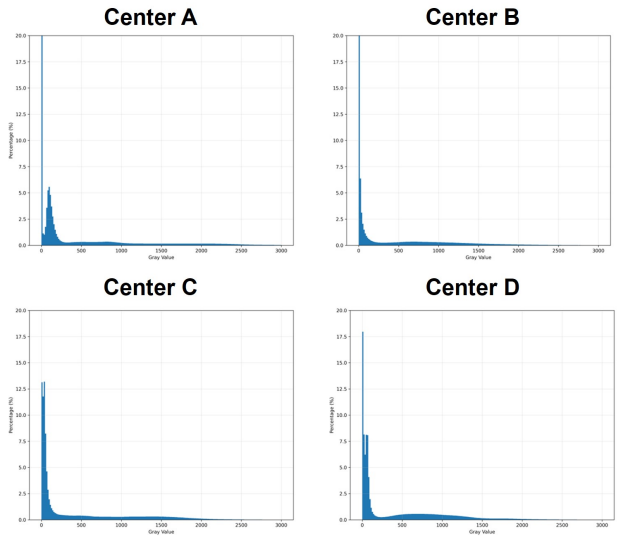


Figure 4.4: Intensity statistics of the multi-centre datasets (Centre A, Centre B, Centre C, NPC-MR3D).

Multi-centre medical image fusion, however, faces serious technical challenges due to heterogeneity. This heterogeneity may stem from multiple factors, including but not limited to the brand (e.g., GE, Siemens, Philips) and model of imaging equipment, the parameter settings of imaging protocols (such as resolution, contrast, and scan time), the characteristics of patient populations (e.g., ethnicity, age, and sex distribution), and the skill level of operators. These differences can significantly affect the generalisation ability of algorithms, especially in medical image segmentation tasks.

To validate the feature extraction and generalisation capability of the proposed method on cross-centre data, generalisation experiments are conducted on four datasets: the three anonymised centres (Centre A, Centre B, Centre C) from the publicly available multi-centre data [145], and NPC-MR3D. Sample examples from the four datasets are shown in Fig. 4.4. All four datasets consist of nasopharyngeal MRI examinations, with the target lesions being the primary nasopharyngeal carcinoma and lymph node metastases.

Table 4.3: Statistical comparison of the four multi-center datasets (sample size, intensity statistics).

Statistics	**Center A**	**Center B**	**Center C**	**NPC-MR3D**
Number of samples	50	50	60	300
Annotated samples	50	50	60	166
Average intensity	718.33	641.91	504.34	514.59
Standard deviation	872.42	679.80	660.09	598.64
Median intensity	217	448	104	179

As shown in Table 4.3, the four datasets (Centre A, Centre B, Centre C, NPC-MR3D) exhibit clear statistical differences in both sample distribution and imaging characteristics, including the number of samples, mean intensity, standard deviation of intensity, and median intensity. Centre A, Centre B, and Centre C contain 50, 50, and 60 cases, respectively, all with full annotations; the NPC-MR3D set is limited to 300 cases (of which 166 are annotated) to avoid an excessive imbalance in sample size across centres. The mean intensities of the four datasets are 718.33, 641.91, 504.34, and 514.59, respectively. The mean intensity of NPC-MR3D is close to that of Centre C, but its intensity standard deviation is lower than the other three; Centre A has the highest mean and standard deviation, and Centre C has the lowest median intensity. Fig. 4.4 illustrates the intensity distribution across centres; it can be seen that different centres exhibit pronounced differences in intensity distribution.

4.4.4.1 *Cross-Centre Generalization Analysis*

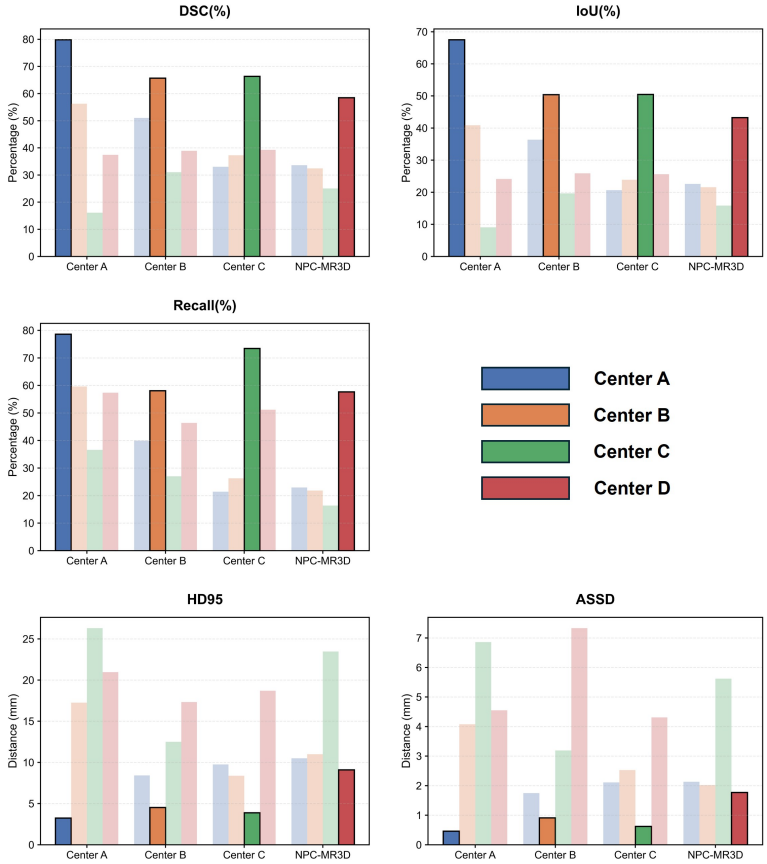


Figure 4.5: Cross-centre generalisation performance. Models are independently trained on four different centres (NPC-MR3D, A, B, C) and directly tested on all centres. The bars represent DSC (%) for each train-test combination.

We first independently train models on datasets from each of the four centres and directly test them on all centres to evaluate the cross-centre transferability. The experimental results are shown in Fig. 4.5. Observing each group of experiments, the model achieves the best performance on the data from its training centre. For instance, the model trained on NPC-MR3D achieves a DSC of 58.50% when tested on NPC-MR3D, significantly higher than its performance on the datasets from centres A, B, and C (33.62%, 32.44%, and 25.06%, respectively). Similarly, the models trained on centres A, B, and C also achieve the highest DSC on their respective datasets, reaching 79.82%, 65.70%, and 66.35%.

The experimental results reveal that while models perform well on data from their training centre, the performance drops significantly when tested on data from other centres. This phenomenon fully exposes the generalisation limitations of models trained on single-centre data, which struggle to adapt to the data distribution differences across different medical centres. Furthermore, this experiment validates an important conclusion: due to the influence of factors such as imaging equipment, acquisition parameters, patient populations, and annotation standards on medical imaging data, domain-specific knowledge is difficult to transfer directly to other datasets, indicating a significant domain shift problem. This observation further emphasises the insufficiency of single-centre training methods in multi-centre scenarios and provides research directions for subsequent studies on how to improve the generalisation capability of models.

4.4.4.2 *Self-Supervised Pre-Training on Multi-Centre Data*

Leveraging the 3DRotNPC paradigm, we apply the SSL strategy to multi-centre data by incorporating the rotation prediction task. During the proxy task training, the model is pre-trained in an unsupervised manner using all unlabelled medical imaging data from four different medical centres. Through the learning process of the proxy task, the model can learn discriminative visual feature representations from large amounts of unlabeled data, which can overcome the domain shift effect caused by multi-centre data distribution differences to a certain extent, decouple domain-related features, and extract universal representations. After pre-training, the encoder network parameters learned from the proxy task are used as the initialisation weights for the downstream task. This parameter transfer approach fully utilises the universal feature representation capability obtained through self-supervised learning.

Table 4.4 presents the performance comparison between 3DRotNPC and the vanilla model (without pre-training) on different datasets. The model after self-supervised pre-training shows improvements in DSC and IoU across all datasets, indicating enhanced accuracy in target region prediction. For example, on the Centre A dataset, DSC improves from 79.82% to 82.21%, and HD95 decreases by 1.04, demonstrating that the model can more precisely delineate lesion regions. Furthermore, on Centre B and Centre C datasets, DSC improves by 6.01% and 7.99% , respectively, indicating improved model performance on multi-centre data. The experimental results show that self-supervised pre-training effectively enhances the segmentation performance of the model, manifested as improvements in segmentation accuracy, boundary prediction capability, and generalisation performance.

Fig. 4.6 shows the 2D slice visualisation results on the four datasets, where (a, b), (c, d), (e, f), and (g, h) are sample examples from Centre

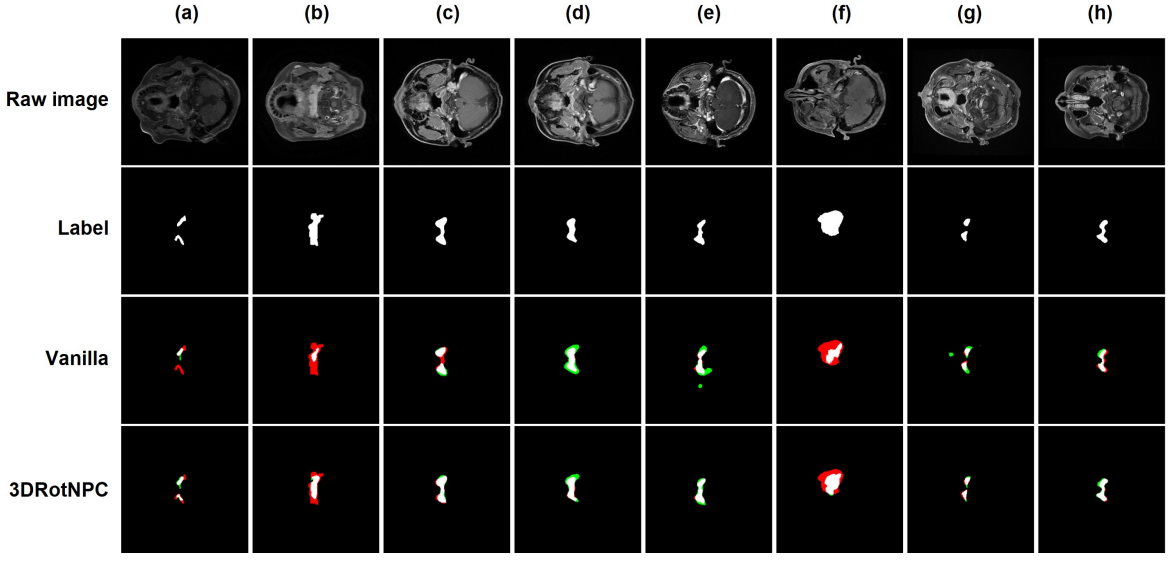


Figure 4.6: Multi-centre data 2D slice visualisation results. (a, b), (c, d), (e, f), and (g, h) are sample examples from Centre A, Centre B, Centre C, and NPC-MR3D, respectively. 3DRotNPC achieves better results in subjective visual comparison.

Table 4.4: Multi-Center Training Comparative Experiment

Metric	Method	A	B	C	NPC-MR3D
DSC (%)	vanilla	79.82	65.70	66.35	58.50
	3DRotNPC	**82.21**	**71.71**	**74.34**	**59.27**
IoU (%)	vanilla	67.50	50.42	50.47	43.27
	3DRotNPC	**70.95**	**56.63**	**59.64**	**43.58**
ASSD ↓	vanilla	0.46	0.91	0.62	1.77
	3DRotNPC	**0.34**	**0.58**	**0.61**	**1.54**
HD95 ↓	vanilla	3.25	4.53	3.89	9.10
	3DRotNPC	**2.21**	**3.47**	**3.26**	**7.74**
Prec (%)	vanilla	84.73	82.90	71.14	67.31
	3DRotNPC	**85.03**	**86.70**	**77.46**	**67.76**
Recall (%)	vanilla	78.61	58.10	73.46	**57.67**
	3DRotNPC	**81.99**	**64.29**	**75.58**	56.79

A, Centre B, Centre C, and NPC-MR3D, respectively. It can be observed that 3DRotNPC achieves better results in subjective visual comparison.

Although 3DRotNPC brings significant performance improvements of 2.39%, 6.01%, and 7.99% on Centre A, Centre B, and Centre C, respectively, the improvement on the NPC-MR3D dataset is relatively small, with DSC only increasing from 58.50% to 59.27% and IoU improving by only 0.31%. In contrast, the DSC improvements on Centre A, B, and C datasets are more substantial. This difference may stem from the sufficiency of NPC-MR3D training data. The training sample size of the NPC-MR3D dataset is much larger than that of Centres A, B, and C, allowing the model to fully learn the distribution characteristics of this data during the vanilla training phase, thus limiting the additional gains from self-supervised pre-training.

In contrast, the number of samples of Centre A, B, and C are relatively small, and the model may face data insufficiency issues during direct training, while self-supervised learning can effectively compensate for this deficiency, thereby bringing more significant performance improvements. This result indicates that in medical scenarios, especially in cases of small-sample hospital data, self-supervised learning can effectively alleviate data insufficiency problems and improve the generalisation capability of models. Therefore, self-supervised learning has important application value for resource-constrained medical institutions and can serve as an effective means to enhance model robustness and cross-centre adaptability.

4.5 CHAPTER SUMMARY

In this chapter, we have presented 3DRotNPC, a self-supervised learning framework that exploits the 3D spatial and geometric structure of volumetric MRI data to reduce reliance on voxel-wise annotations. By designing a rotation prediction proxy task on consecutively sliced data, the encoder learns anatomically meaningful representations without requiring labels. Transferring the pre-trained parameters to the downstream segmentation task yields substantial performance gains under label-limited regimes.

Extensive experiments on nasopharyngeal carcinoma MRI data demonstrate that 3DRotNPC delivers considerable improvements when only 10% to 50% of the labelled data are available. The results confirm that self-supervised pre-training provides a cost-effective path toward accurate volumetric segmentation: with relatively fewer labels, the pre-trained model achieves performance comparable to fully supervised baselines trained on larger annotated datasets.

Several directions remain for future work. First, exploring complementary proxy tasks—such as masked autoencoding (MAE) and contrastive learning—may yield more expressive representations and further improve downstream performance. Second, combining multiple pretext objectives in a curriculum or multi-task framework could enhance generalisation across different imaging protocols and anatomical regions.

Having addressed the first two constraints—label-quality imbalance and label-quantity limitation—the next chapter considers the third constraint that arises at deployment: the information asymmetry between strong and weak imaging modalities.

5

CROSS-MODAL KNOWLEDGE TRANSFER FOR MEDICAL IMAGE SEGMENTATION UNDER MODALITY-INFORMATION ASYMMETRY

The preceding chapters addressed two constraints intrinsic to the annotation process: heterogeneous label quality (Chapter 3) and limited label quantity (Chapter 4). This chapter turns to a third constraint that arises at deployment: the information asymmetry between strong and weak imaging modalities. As introduced in Chapter 1, strong modalities such as positron emission tomography (PET) or contrast-enhanced magnetic resonance imaging (MRI) provide enhanced lesion conspicuity, i.e., lesions stand out more clearly from surrounding tissue due to higher contrast and clearer boundaries, making them easier to detect and delineate. However, they may be unavailable in routine clinical practice due to cost, acquisition time, or patient contraindications. In contrast, weak modalities such as non-contrast computed tomography (NCCT) are widely available but offer limited discriminative evidence for subtle lesions. The challenge is therefore to leverage strong-modality information during training while maintaining a weak-modality-only inference pathway.

To this end, we propose a novel two-stage framework that enables accurate segmentation using only the weak modality (NCCT) at inference. Rather than simple fusion, our framework transfers knowledge from fused features to the weak-modality encoder to strengthen its representational capability. Specifically, we first introduce a dual-branch feature-mining mechanism that dynamically captures and integrates asymmetric cross-modal lesion features at the patch level. We next devise a diffusion-inspired knowledge transfer strategy that systematically guides the weak-modality encoder to assimilate integrated high-confidence semantic representations. Unlike conventional one-shot alignment, our mechanism bridges the representation gap in a stable, step-by-step manner, ensuring that the weak-modality encoder progressively perceives and internalises decisive features. Extensive experiments on three public datasets show that our method achieves competitive performance in multi-modal settings and considerably reduces the performance gap between strong and weak modalities in NCCT-only scenarios[1].

1 Parts of this chapter are in the following submission: **C. Li**, C. Yao*, W. Liu, X. Wang, J. Kosinka, S. Frey, A. C. Telea, and X. Ban. Weak-to-Strong: Empowering Non-Contrast CT for Accurate Lesion Segmentation via Cross-Modal Knowledge Transfer. *IEEE Transactions on Multimedia*, Under Review.

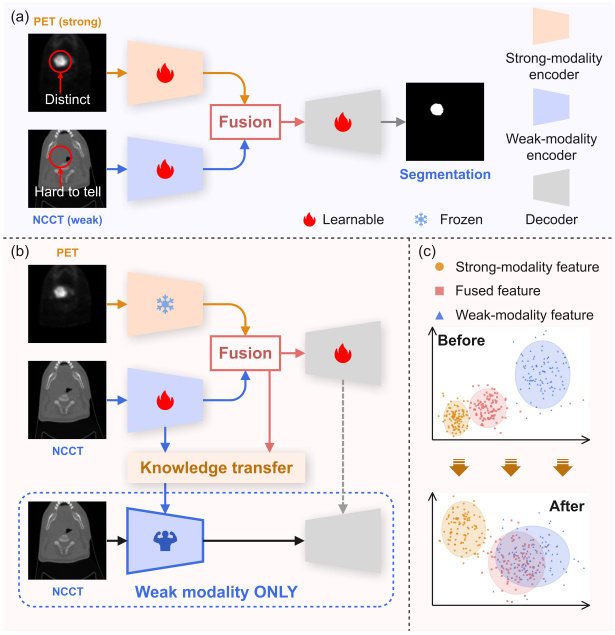


Figure 5.1: Operation of our proposed framework. The strong modality (PET) clearly highlights lesion regions; the weak modality (NCCT) captures subtle intensity variations. (a) Existing methods generally focus on fusing strong and weak modalities, requiring *both* inputs at inference. (b) We aim to enhance the weak-modality encoder's ability to capture subtle yet important lesion features, enabling accurate segmentation using NCCT *alone* at inference. (c) Comparison of feature distributions before and after knowledge transfer. After transfer, the output of the weak-modality encoder aligns more closely with the fused high-confidence features.

5.1 INTRODUCTION

Lesion segmentation has long had a crucial role in medical image analysis, serving as a fundamental step in disease diagnosis, treatment, and prognosis [5, 34, 129, 139]. Segmentation tasks in medical imaging are uniquely challenging due to low contrast, blurred boundaries, and the need for precise delineation of subtle pathological regions [45, 86]. Recent years have witnessed a surge in the development of such methods with a focus on leveraging deep learning techniques to enhance performance when using various imaging modalities [47, 48, 117, 156].

For non-contrast computed tomography (NCCT) scans, they mainly deliver structural and anatomical information without functional or metabolic context. Lesions often exhibit subtle intensity differences compared to surrounding tissues, making them hard to separate [123],

as illustrated in Fig. 5.1 (a). Hence, accurate segmentation on NCCT scans relies on non-obvious yet important features that are hard to explicitly define or supervise. Compared with NCCT, positron emission tomography (PET) and magnetic resonance imaging (MRI) are often referred to as *strong* modalities since they provide rich functional or soft-tissue contrast information, making lesion regions more pronounced [21, 53]. For instance, PET imaging detects metabolic activity in lesions, allowing for precise localization [183], while MRI provides detailed anatomical information and soft-tissue contrast, enabling accurate delineation of lesion boundaries [44]. However, PET and MRI also have limitations, such as the need for more specialised equipment, longer acquisition times, and higher costs, which limit their availability and accessibility [20]. This motivates the challenge of achieving accurate segmentation performance when using the *weak* modality (NCCT) alone, particularly at the inference stage.

The key question we address in this chapter is: *how to enhance the weak-modality encoder to capture non-obvious yet important features for accurate segmentation using NCCT alone.* Most related work uses multi-modal fusion to leverage complementary information for enhanced segmentation performance; this does not cover the case where the strong modality is unavailable at inference [1, 46, 143, 166, 181] (see Fig. 5.1 (a)). Other studies address the missing modality problem through feature enhancement or generation [72, 167, 182]. However, their focus is mainly on intra-modal variations within MRI sequences rather than distinct multi-modal imaging. A key question, not addressed by these studies, is how to make the weak-modality encoder capture the "hidden" features present in NCCT whose lesion-specific subtlety is difficult to perceive effectively. To this end, we devise a two-stage framework which transfers knowledge from the strong to the weak modality via fused features, progressively guiding the weak-modality encoder to identify subtle yet crucial features and thereby enabling precise lesion segmentation using NCCT alone (see Fig. 5.1 (b)).

To facilitate effective fusion and subsequent knowledge transfer, one must account for the fine-grained spatial characteristics of lesions [45]. In the dual-modality medical imaging setting, since the strong modality typically offers distinct lesion information, while the weak modality contains subtle and structural evidence, the two modalities should not be fused symmetrically. This motivates a fusion mechanism that uses stronger representations to guide weak-feature enhancement while still preserving complementary information. For this, we propose a dual-branch architecture, termed Asymmetric Relationship Modelling (ARM), as the first stage of our framework. Using Vision Mamba [98, 189] as the core component of feature interaction and integration, ARM has linear complexity and long-range dependency modelling, which helps capture patch-level semantic commonalities while explicitly modelling asymmetric cross-modal relationships through mutual feature

enhancement and complementary feature aggregation. Taken together, these aspects provide the needed informative fused representations for subsequent transfer.

Despite the asymmetric relationships learned in the first stage of our framework, a representational gap remains between the unimodal weak-modality encoder and the enhanced multi-modal features (see "Before" part of Fig. 5.1 (c)). To bridge the gap, our framework's second stage introduces a progressive, noise-aware knowledge transfer strategy inspired by diffusion principles, termed Progressive Knowledge Transfer (PKT), which enables smooth and robust transfer of high-confidence semantic knowledge. By gradually guiding the learning process rather than enforcing direct alignment, PKT strengthens the representational capability of the weak-modality encoder. This enables the encoder to capture decisive lesion features—aligning its feature distribution closely with the fused features—and ultimately achieve accurate lesion segmentation using the weak modality alone (see "After" part of Fig. 5.1 (c)).

In summary, our main contributions are as follows:

- We propose a novel two-stage framework designed to boost lesion segmentation performance using the weak modality (NCCT) alone. By effectively transferring knowledge from multi- to weak-modality, we achieve accurate lesion segmentation without auxiliary guidance at inference.
- We introduce the Asymmetric Relationship Modelling (ARM) approach, a feature interaction and integration module powered by Vision Mamba. ARM leverages long-range dependency modelling to effectively extract shared decisive semantic commonalities hidden beneath modality-specific differences.
- We devise a Progressive Knowledge Transfer (PKT) mechanism inspired by diffusion processes. By smoothly distilling high-confidence, decisive semantic features from the fused representation, PKT bridges the representational gap between fused and weak spaces.
- Extensive experiments on three public datasets show that our approach outperforms existing methods in both multi- and weak-modality scenarios. It exceeds the best-performing baseline by an average of 2.2% Dice in multi-modal settings and improves NCCT-only segmentation performance by an average of 6.1%.

5.2 RELATED WORK

5.2.1 *Lesion Segmentation across Imaging Modalities*

Lesion segmentation performance in medical imaging is inherently constrained by the imaging modality. Different modalities exhibit fundamentally different lesion saliency and contrast characteristics [20, 47, 48, 117, 123]. Even with identical segmentation backbones, segmentation performance is chiefly bounded by the imaging modality itself. As discussed in Chapter 2 Section 2.4.1, such modality-specific constraints originate from distinct image formation principles and lead to pronounced statistical heterogeneity and semantic complementarity across modalities (e.g., CT for anatomical detail versus PET/MRI for functional or soft-tissue cues).

Recent studies have addressed the scenario where specific modalities are absent at inference [7, 72, 167, 182]. A unified hyper-network framework [167] was proposed to dynamically aggregate information for dual-level completion. Similarly, MMCFormer [72], a co-training framework that utilises Multi-Scale Contextual Agreement (MSCA) modules to distil knowledge from a full-modality network to its missing-modality counterpart. However, most existing missing-modality approaches address modality absence within MRI sequences only, designed to handle incomplete modality inputs by reducing representation discrepancies through completion or alignment strategies. In contrast, our work explicitly targets the substantial semantic gap between NCCT and strong modalities (like PET and MRI), and aims to maintain stable segmentation performance when the strong modality is entirely unavailable at inference.

5.2.2 *Fusion-Based Lesion Segmentation*

Integrating complementary modalities [154], such as PET/CT or MRI/CT, has been extensively explored to enhance lesion segmentation [1, 46, 143, 166, 181]. Typical approaches employ data-level, feature-level, or decision-level fusion to leverage cross-modal information [44]. For instance, AATSN [1] utilises a fusion-attention decoder to efficiently integrate anatomical CT spatial features with metabolic PET information. MFCNet [143] introduces mutual calibration blocks to recalibrate semantic representations via attention mechanisms, while MMCA-Net [181] employs a transformer-integrated Y-Net to capture deep cross-modal dependencies. Despite performance improvements, these methods rely on the availability of *all* modalities at inference and tend to degrade severely when the strong modality is absent. This limits deployment in clinical scenarios where PET or MRI scans are unavailable due to high cost, long acquisition times, or safety concerns. Consequently, there is a growing need for methods that transfer cross-modal

knowledge to enhance NCCT segmentation, which is the focus of our work.

5.2.3 *Cross-Modal Knowledge Transfer*

Cross-modal learning enables models to exploit complementary information across different imaging modalities, thereby improving robustness and performance in numerous tasks [105]. In practice, knowledge distillation (KD) is widely adopted to transfer informative representations between different representational spaces [56]. KD is commonly instantiated through feature-level or output-level distillation, with extensions to relational and hierarchical formulations [23, 129].

Beyond a direct or one-shot approach, recent studies have explored progressive and diffusion-based [58] distillation strategies to stabilise knowledge transfer by gradually narrowing the representation gap between teacher and student models [89, 113, 118]. For example, DiffKD [62] models student features as noisy observations of teacher representations and uses explicit denoising processes to recover clean features to improve students' effective learning capability, primarily in same-modality teacher–student settings. In medical image segmentation, generative diffusion models have been used to address data heterogeneity and domain shift. Cascaded diffusion models [176] are used to progressively normalise non-IID distributions across multi-centre datasets, while DiffuSeg [177] synthesises domain-consistent images via conditional diffusion to facilitate segmentation under annotation-scarce target domains.

While these methods demonstrate that progressive optimisation and noise-aware supervision can stabilise representation learning, they are typically built upon explicit diffusion models and are mainly designed for *same-modality denoising* or data synthesis scenarios. In contrast, we adopt a diffusion-inspired progressive transfer strategy that does not rely on generative diffusion models, but instead gradually distils high-confidence fused representations to bridge the substantial semantic gap between asymmetric modalities and empower the weak-modality encoder. This setting aligns with the practical constraint in Chapter 2 Section 2.4.3: strong modalities provide privileged information during training but may be unavailable at deployment, requiring the student to inherit clinically useful cues while operating solely on NCCT.

5.3 METHODOLOGY

5.3.1 *Framework Overview*

Our proposed framework comprises two training stages designed to enhance lesion segmentation using the weak modality (NCCT). In the

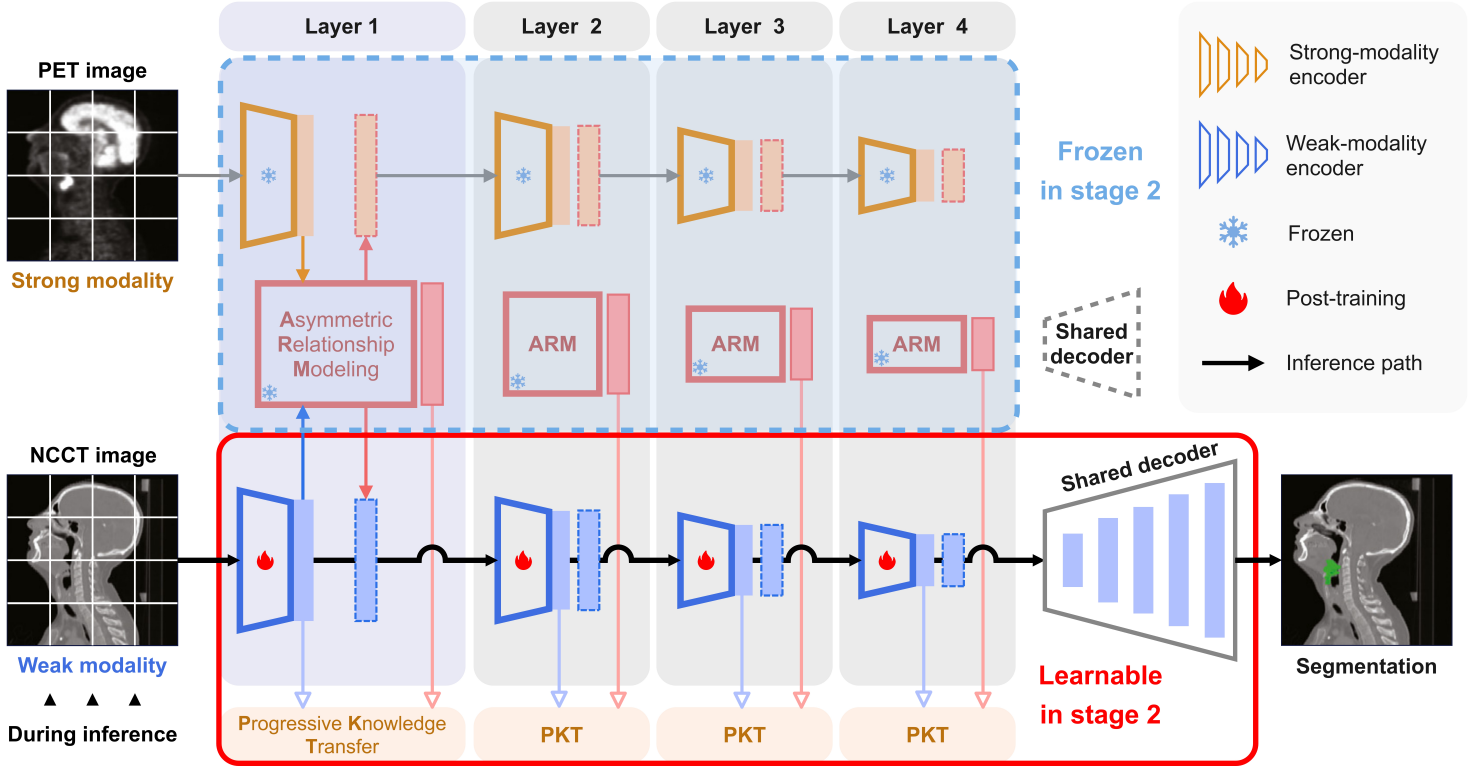


Figure 5.2: Schematic of our proposed framework when transferring knowledge from multi- to weak-modality (the second stage). After multi-modal training, the parameters of the weak-modality branch are updated through Progressive Knowledge Transfer (PKT). At inference, the strong-modality branch, the intermediate branch (ARMs), and the PKTs are not used. ARM and PKT are described in detail in Fig. 5.3.

first stage, taking paired strong and weak modalities as input, the ARM (Section 5.3.2) captures asymmetric dependencies to generate enhanced fused representations and produce a multi-modal segmentation prediction supervised by ground truth. In the second stage, the PKT mechanism (Section 5.3.3) distils these high-confidence fused representations into the NCCT encoder, with the strong-modality encoder and ARM frozen. Crucially, at inference, the network requires only the weak modality as input to predict the final lesion segmentation map. An overview of our framework is shown in Fig. 5.2, with emphasis on the second stage: at each encoder layer, the ARM-enhanced feature and the corresponding weak-modality feature are jointly fed into PKT to progressively empower the weak-modality encoder.

5.3.2 *Asymmetric Relationship Modelling*

It is essential to account for the heterogeneous sizes and dispersed spatial distribution of lesion regions in both the strong and weak modalities, for which patch-level semantic commonalities underlying modality-specific differences must be effectively modelled. For this, we introduce a tailored Vision Mamba-based dual-branch architecture to capture the shared, yet asymmetric, relationships between the strong and weak modalities (see Fig. 5.3). Our proposed ARM comprises two components, namely the Mutual Feature Enhancement Module (MFEM) and the Complementary Feature Aggregation Module (CFAM), described below.

Vision Mamba Preliminaries: Modern Structured State Space Models (SSMs), notably Mamba [35, 189] (more details provided in Chapter 2 Section 2.1.2.3), are typically built around an input-dependent selective state space layer for sequence mixing, together with lightweight linear projections, gating, and normalisation components. This architecture offers an efficient alternative to Transformers by bridging continuous-time formulations with discrete deep learning and achieving linear computational complexity with respect to sequence length.

Recent advances have extended Mamba to computer vision [95] and medical image analysis [10] through two-dimensional selective scan operators (SS2D) [98, 122] which perform directional and spatially-aware scanning over 2D feature maps to model long-range dependencies with linear complexity and reduced computational overhead, an advantage particularly relevant for medical imaging scenarios. Motivated by these developments, we incorporate SS2D blocks into our proposed ARM to conduct feature interaction and integration, providing a stronger foundation for cross-modal feature enhancement than conventional convolutions or self-attention mechanisms [41].

Mutual Feature Enhancement Module (MFEM): Taking strong- and weak-modality features as input, each branch independently encodes

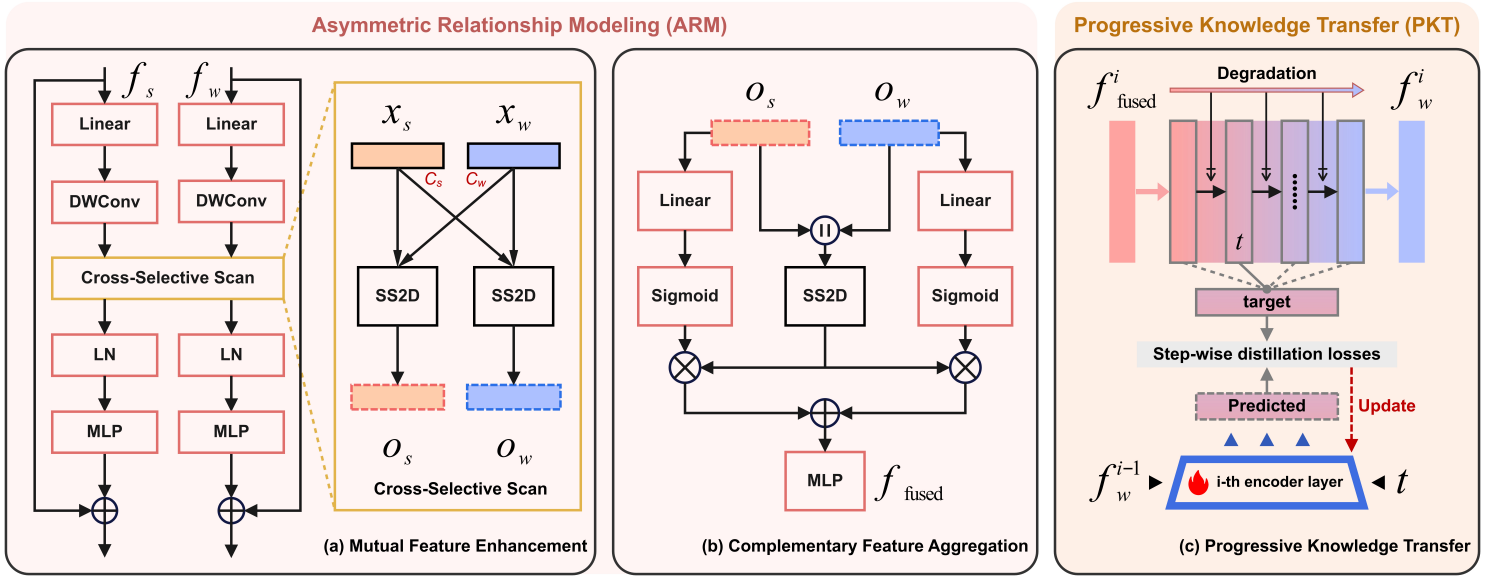


Figure 5.3: Details of our proposed two-stage framework. ARM, used in the first stage, consists of the Mutual Feature Enhancement Module (MFEM) and the Complementary Feature Aggregation Module (CFAM); PKT is used in the second stage. (a) MFEM models asymmetric cross-modal interactions via cross-selective scan. (b) CFAM aggregates and redistributes shared contextual information to refine lesion-specific features. (c) PKT progressively transfers the ARM-fused knowledge to the weak-modality encoder through a step-wise degradation process. SS2D denotes a 2D Selective Scan operator.

modality-specific information to extract discriminative features. To capture lesion-specific discrepancies while enabling effective cross-modal interaction, we propose the MFEM (see Fig. 5.3 (a)). Initially, the local structures of the strong (f_s) and weak (f_w) modality features are strengthened via a linear projection and a depthwise convolution (DWConv) given by

$$x_k = \mathrm{DWConv}(\mathrm{Linear}(f_k)), \quad k \in \{s, w\}. \tag{5.1}$$

To facilitate deep semantic interaction, we adopt the cross-selective scan mechanism [142] to enforce a cross-modal information exchange. For each modality $k \in \{s, w\}$, the input-dependent parameters $(\mathbf{B}_k, \mathbf{C}_k, \mathbf{\Delta}_k)$ are dynamically generated from the input features x_k, enabling adaptive modulation of the state transitions conditioned on modality-specific context. Here, $\mathbf{\Delta}_k$ controls the discretisation timescale, while $\mathbf{B}_k$ and $\mathbf{C}_k$ respectively govern the input injection and output projection in the state space model. The hidden states $\mathbf{h}_k$ are updated according to the standard selective scan rule

$$\mathbf{h}_k^t = \bar{\mathbf{A}}_k \mathbf{h}_k^{t-1} + \bar{\mathbf{B}}_k x_k^t, \quad k \in \{s, w\}, \tag{5.2}$$

where $\bar{\mathbf{A}}_k$ and $\bar{\mathbf{B}}_k$ denote the discretised state transition and input matrices obtained via zero-order hold discretization [49].

Crucially, to inject complementary context from the weak modality into the strong modality (and vice versa), we exchange the output projection matrices $\mathbf{C}_k$ during the readout phase. The cross-enhanced outputs y_s and y_w are formulated as

$$\begin{aligned} y_s^t &= \mathbf{C}_w \mathbf{h}_s^t + \mathbf{D}_s x_s^t, \\ y_w^t &= \mathbf{C}_s \mathbf{h}_w^t + \mathbf{D}_w x_w^t, \end{aligned} \tag{5.3}$$

where $\mathbf{D}_k$ is the residual connection parameter. This mechanism allows the strong modality to guide the reconstruction of the weak modality using its own contextual dynamics, thereby enhancing semantic consistency. The final outputs o_s and o_w of MFEM are obtained via a residual connection and a Multi-Layer Perceptron (MLP) as

$$o_k = f_k + \mathrm{MLP}(\mathrm{LN}(y_k)), \quad k \in \{s, w\}, \tag{5.4}$$

where LN denotes layer normalization.

Complementary Feature Aggregation Module (CFAM): While MFEM promotes mutual enhancement, the feature representations remain in separate streams. We design the CFAM module to merge these streams into a unified representation and subsequently redistribute the refined global context (see Fig. 5.3 (b)). The enhanced features o_s and o_w are first concatenated and fused through an SS2D block to capture long-range spatial dependencies via

$$g_{\mathrm{sh}} = \mathrm{SS2D}(\mathrm{Concat}(o_s, o_w)). \tag{5.5}$$

Considering the inherent asymmetry between the strong and weak modalities, we use a gated fusion mechanism to selectively filter information. A sigmoid activation function $\sigma(\cdot)$ generates modality-specific adaptive gates, i.e., $\alpha_k = \sigma(\text{Linear}(o_k))$. The final refined features f_{fused} are computed as

$$\begin{aligned} f_k^{\text{out}} &= \text{Linear}(o_k + \alpha_k \odot g_{\text{sh}}), \\ f_{\text{fused}} &= f_s^{\text{out}} + f_w^{\text{out}}. \end{aligned} \tag{5.6}$$

This design ensures that the strong modality preserves its detailed anatomical information while selectively absorbing complementary information from the shared global context g_{sh}. Together, MFEM and CFAM constitute the ARM stage, which effectively models the asymmetric relationship prior to the Progressive Knowledge Transfer (PKT).

5.3.2.1 *Loss Function*

During the first stage, we train our framework using an equally weighted hybrid segmentation loss combining Dice loss ($\mathcal{L}_{\text{Dice}}$) and cross-entropy loss ($\mathcal{L}_{\text{CE}}$) given by

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}. \tag{5.7}$$

5.3.3 *Progressive Knowledge Transfer*

In the second stage of the framework, we introduce PKT to systematically distil the integrated knowledge captured by ARM into the weak-modality encoder (see Fig. 5.3 (c)). Instead of enforcing one-shot feature alignment, PKT adopts a diffusion-inspired progressive scheme with controlled degradation, enabling stable and effective knowledge assimilation. Note that during PKT, the ARMs and strong-modality encoders are fixed.

Let f_w^i denote a weak-modality feature and f_{fused}^i the corresponding fused feature produced by ARM at the i-th encoder stage. We design a step-wise degradation trajectory that progressively drifts from the fused feature toward the weak-modality feature. Specifically, we define a deterministic interpolated state at diffusion step t as

$$\bar{f}^{i,t} = \rho_t \, f_{\text{fused}}^i + (1 - \rho_t) f_w^i, \quad t = 1, \ldots, T, \tag{5.8}$$

where $\rho_t \in [0, 1]$ is a monotonically decreasing schedule (e.g., $\rho_1 \approx 1$ and $\rho_T \approx 0$).

To simulate progressive degradation, we further inject Gaussian noise at each step to obtain the noisy target

$$f_{\text{tar}}^{i,t} = \sqrt{\alpha_t} \, \bar{f}^{i,t} + \sqrt{1 - \alpha_t} \, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{5.9}$$

where $\alpha_t \in (0, 1]$ controls the signal-to-noise ratio at step t.

The weak-modality encoder is trained to progressively absorb the distilled knowledge by supervising each encoder layer with time-conditioned targets. Specifically, the i^{th} encoder layer predicts the degraded target feature at step t from the $i-1^{th}$ layer representation as

$$\hat{f}_w^{i,t} = \Phi_w^i(f_w^{i-1}, t), \tag{5.10}$$

where step t is only used during training (not required at inference).

The progressive distillation loss consists of two components. The first term enforces similarity between the predicted feature and the degraded target as

$$\mathcal{L}_{\text{mse}}^i = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_w^{i,t} - f_{\text{tar}}^{i,t}\right\|_2^2. \tag{5.11}$$

To stabilise the degradation trajectory and prevent abrupt representation shifts across diffusion steps, we further impose an inter-step smoothness regularisation on the predicted feature sequence given by

$$\mathcal{L}_{\text{smooth}}^i = \frac{1}{T-1}\sum_{t=2}^{T}\left\|\hat{f}_w^{i,t} - \hat{f}_w^{i,t-1}\right\|_2^2. \tag{5.12}$$

The PKT loss at the i^{th} stage is thus formulated as

$$\mathcal{L}_{\text{PKT}}^i = \mathcal{L}_{\text{mse}}^i + \gamma\,\mathcal{L}_{\text{smooth}}^i, \tag{5.13}$$

where γ controls the smoothness regularisation.

The overall training objective combines the progressive knowledge transfer loss with the segmentation loss as

$$\mathcal{L} = \lambda\sum_i \mathcal{L}_{\text{PKT}}^i + \mathcal{L}_{\text{seg}}, \tag{5.14}$$

where λ controls the contribution of PKT. Through this interpolated and regularised progressive scheme, the weak-modality encoder gradually internalises high-confidence semantic knowledge from the fused features.

Algorithm 5.1 outlines the PKT training procedure. By progressively degrading fused representations toward weak-modality representations and enforcing step-wise smoothness, PKT enables stable knowledge transfer that strengthens the weak-modality encoder for NCCT-only inference.

Algorithm 5.1: Progressive Knowledge Transfer (PKT).

for each training batch (x_w, x_s, y) **do**
  Extract weak-modality features $\{f_w^i\}$ and obtain fused features $\{f_{\text{fused}}^i\}$ via ARM
  for each encoder stage $i = 1, \ldots, L$ **do**
    for $t = 1, \ldots, T$ **do**
      Construct the degraded target via interpolation and noise injection (Eqs. (5.8) and (5.9))
      Predict weak-modality feature (Eq. (5.10))
    end for
    Compute $\mathcal{L}_{\text{PKT}}^i$ and step-wise smoothness
    Update weak-modality encoder block i
  end for
  Update network with segmentation loss $\mathcal{L}_{\text{seg}}$
end for

5.4 RESULTS

5.4.1 *Experimental Setup*

5.4.1.1 *Baselines*

We evaluate our proposed framework by comparing it against a diverse set of methods, categorised into representative medical image segmentation backbones and related advanced fusion-based approaches. Firstly, to explore the intrinsic performance disparity between weak and strong modalities, we use three widely adopted architectures: UNet [120], a standard CNN-based baseline; TransUNet [25], which integrates Transformers to capture global context; and VM-UNet [122], a recently introduced Mamba-based network known for efficient long-range dependency modelling. Secondly, to assess the superiority of our framework over existing fusion-based approaches, we compare it with leading methods in the field of medical multi-modal segmentation for PET/NCCT and MRI/NCCT, namely MSAM [46], MMCL [166], AATSN [1], MFCNet [143], Mirror U-Net [106], and MMCA-Net [181]. Comparing with these methods allows us to validate the effectiveness of our approach in both multi-modal and NCCT-only inference scenarios.

5.4.1.2 *Datasets*

We conducted extensive experiments on three publicly available dual-modality medical image datasets: HECKTOR 2022e [117], AUTOPET [47], and APIS [48]. For the 2D image segmentation implementation, we extracted axial slices from the 3D volumes, filtering out background slices that contained no anatomical structures or lesions. Note that due to multi-centre acquisition and different scanners/protocols,

the original scan resolution is not identical across cases; a unified pre-processing is applied as detailed below.

HECKTOR (head and neck tumour segmentation). This dataset originates from the MICCAI 2022 HECKTOR challenge and is a standard benchmark for multi-modal PET/NCCT lesion segmentation. It contains 524 cases collected from seven clinical centres, with standardised pre-processing including resampling and region cropping. In our experiments, PET is the strong modality, and NCCT is the weak modality.

AUTOPET (automated lesion segmentation in whole-body FDG-PET/CT). This dataset, associated with the MICCAI 2022 AutoPET challenge, comprises 1014 whole-body FDG-PET/CT studies, from which we selected a subset containing 210 NCCT cases.

APIS (A Paired CT-MRI Dataset for Ischemic Stroke Segmentation). Released as part of the ISBI 2023 challenge, APIS comprises paired NCCT and ADC (MRI) images from 96 ischemic stroke cases. The ADC maps provide clear lesion delineation and are treated as the strong modality.

5.4.1.3 *Implementation Details*

To ensure data consistency and optimal network convergence, we apply modality-specific pre-processing and normalisation techniques across all datasets. For NCCT, we first perform intensity clipping to remove outliers and focus on relevant anatomical structures. Specifically, the clipping ranges are set to $[-1024, 1024]$ Hounsfield Units (HU) for the HECKTOR and AUTOPET datasets, and $[-450, 1050]$ HU for the APIS dataset, after which the intensities are linearly mapped to the range $[-1, 1]$. Regarding the strong modalities, PET images undergo z-score normalisation to standardise the distribution of metabolic intensities; the MRI (ADC) maps in the APIS dataset are similarly normalised to align feature distributions. All 2D slices are uniformly resized to 256×256 pixels using bilinear interpolation for image intensities and nearest-neighbour interpolation for the corresponding masks. Furthermore, to formulate a specific binary lesion segmentation problem, we explicitly filter ground-truth annotations to include only malignant tumour classes (for HECKTOR and AUTOPET) or ischemic lesions (for APIS) as the foreground. All other classes, including benign findings or physiological high-uptake regions, are treated as the background class.

All experiments were implemented in PyTorch and conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB memory). The networks were trained using the AdamW optimiser with an initial learning rate of 1×10^{-3} for the first stage and 1×10^{-4} for the second stage. A cosine annealing learning rate schedule was adopted to ensure stable convergence. The batch size was set to 16 for all experiments, and models were trained for 200 epochs. For the PKT stage, the diffusion

steps were set to $T = 10$, and the balancing coefficient λ was empirically set to 1.0 across all datasets.

5.4.2 *Evaluation of Different Modality Settings*

We first evaluate lesion segmentation performance under different modality training and inference settings. As shown in the upper part of Table 5.1, models trained and tested on the strong modality (PET for HECKTOR/AUTOPET and MRI for APIS) consistently outperform their NCCT-based counterparts, reflecting the superior lesion sensitivity of strong modalities. For instance, VM-UNet achieves a Dice of 0.673 on PET versus 0.554 on NCCT in HECKTOR, 0.693 versus 0.580 in AUTOPET, and 0.630 on MRI versus 0.394 on NCCT in APIS. In contrast, NCCT-only settings yield the lowest performance across all architectures, with markedly lower Dice/IoU and higher HD95 values. Moreover, the relative improvements from NCCT to PET/MRI are consistent across datasets, suggesting that the modality gap is inherent and not architecture-dependent. Quantitatively, the performance gap between strong and weak modalities averages approximately 12% across the three representative architectures.

The lower part of Table 5.1 details the feasibility of transferring knowledge from paired modalities to the weak modality. Directly applying models trained on paired data (e.g., VM-UNet with early or concatenation fusion) to the weak modality results in catastrophic performance degradation, with Dice scores dropping to as low as 0.055–0.166 on HECKTOR and 0.049–0.152 on AUTOPET. Although further fine-tuning the weak-modality encoder on NCCT data significantly recovers performance, it still lags behind the strong modality's upper bound by a large margin. In contrast, our proposed framework achieves a Dice score of 0.623 on HECKTOR and 0.609 on AUTOPET using only NCCT at inference, approaching the performance of the strong modality. This shows that our framework effectively bridges the performance gap between the weak and strong modalities and surpasses standard fine-tuning strategies.

5.4.3 *Effectiveness of Asymmetric Relationship Modelling*

5.4.3.1 *Quantitative Results*

We comprehensively evaluate the segmentation performance of the proposed framework against baseline approaches on the three datasets. The quantitative results are summarised in Table 5.2 and discussed below.

On the HECKTOR dataset, our method shows superior performance compared to all competing fusion mechanisms. Specifically, the configuration without PKT, which leverages the ARM with dual-modality inputs, achieves a Dice score of 0.794 and an HD95 of 7.5. This perfor-

Table 5.1: Segmentation performance across different methods, modalities, and knowledge transfer strategies on three datasets

Method	Training Modality	Inference Modality	Post-training	HECKTOR (PET/NCCT)				AUTOPET (PET/NCCT)				APIS (MRI/NCCT)			
				Dice↑	IoU↑	HD95↓	Recall↑	Dice↑	IoU↑	HD95↓	Recall↑	Dice↑	IoU↑	HD95↓	Recall↑
UNet [120]	Weak	Weak	-	0.531	0.392	12.5	0.603	0.542	0.384	15.2	0.611	0.336	0.237	26.5	0.398
	Strong	**Strong**	-	**0.646**	**0.476**	**9.2**	**0.704**	**0.657**	**0.487**	**11.5**	**0.716**	**0.541**	**0.403**	**15.8**	**0.612**
TransUNet [25]	Weak	Weak	-	0.516	0.372	11.6	0.616	0.558	0.402	14.1	0.638	0.372	0.281	25.2	0.436
	Strong	**Strong**	-	**0.639**	**0.470**	**8.8**	**0.723**	**0.683**	**0.504**	**10.5**	**0.754**	**0.577**	**0.440**	**14.0**	**0.657**
VM-UNet [122]	Weak	Weak	-	0.554	0.413	11.0	0.640	0.580	0.421	13.5	0.663	0.394	0.302	23.8	0.473
	Strong	**Strong**	-	**0.673**	**0.504**	**8.2**	**0.764**	**0.693**	**0.513**	**10.2**	**0.779**	**0.630**	**0.473**	**13.2**	**0.702**
VM-UNet (early)	Paired	Weak	-	0.055	0.029	41.0	0.077	0.049	0.027	45.2	0.068	0.031	0.022	58.0	0.041
	Paired	Weak	FT	0.514	0.363	12.3	0.584	0.530	0.371	16.0	0.601	0.351	0.252	26.5	0.411
VM-UNet (concat)	Paired	Weak	-	0.136	0.078	36.5	0.179	0.122	0.073	39.8	0.145	0.064	0.032	52.0	0.093
	Paired	Weak	FT	0.540	0.380	11.5	0.632	0.573	0.412	14.2	0.654	0.378	0.290	24.8	0.451
Ours	Paired	Weak	-	0.166	0.095	33.5	0.219	0.152	0.088	36.5	0.190	0.083	0.041	49.0	0.114
	Paired	Weak	PKT	0.623	0.454	10.3	0.691	0.609	0.460	11.0	0.707	0.484	0.362	21.5	0.544

"Strong" refers to PET for HECKTOR/AUTOPET and MRI for APIS; "Weak" refers to NCCT for all datasets. "Paired" indicates training with both strong- and weak-modality images. "FT" indicates fine-tuning, and "PKT" indicates the proposed Progressive Knowledge Transfer. The best results within each representative architecture are shown in **bold**.

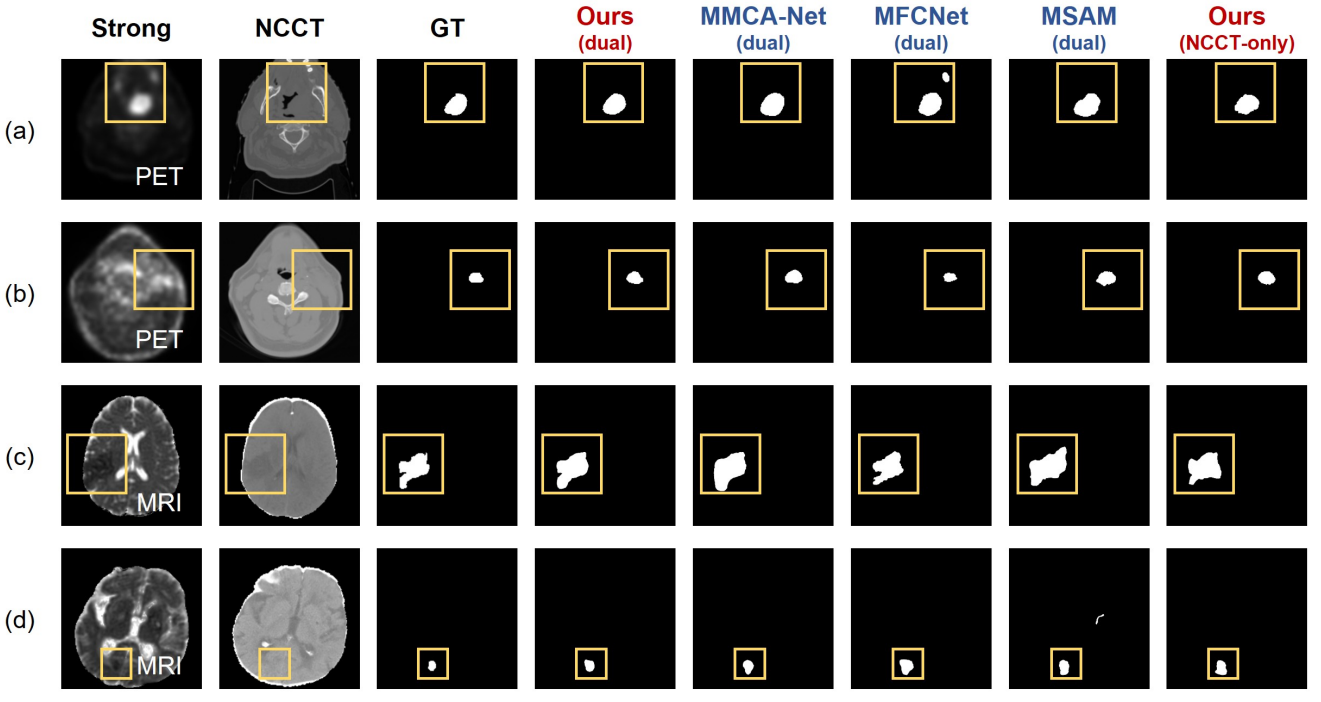


Figure 5.4: Qualitative comparison of lesion segmentation results on representative cases. Rows (a) and (b) correspond to PET/NCCT pairs, while rows (c) and (d) show MRI/NCCT pairs. In the dual-modality setting (the fourth column), our approach yields more accurate lesion delineation and more effective suppression of false positives than competing methods, particularly in cases with blurred lesion boundaries or low-intensity contrast in NCCT. Notably, even when operating in the NCCT-only inference setting (the last column), our method preserves segmentation quality comparable to the dual-modality configuration, demonstrating the effectiveness and robustness of the proposed framework in empowering the weak modality.

Table 5.2: Quantitative comparison of multi-modal segmentation performance on three datasets

Method	HECKTOR				AUTOPET				APIS			
	Dice↑	**IoU↑**	**HD95↓**	**Recall↑**	**Dice↑**	**IoU↑**	**HD95↓**	**Recall↑**	**Dice↑**	**IoU↑**	**HD95↓**	**Recall↑**
VM-UNet (early)	0.688	0.529	9.8	0.747	0.701	0.542	10.5	0.759	0.432	0.273	23.5	0.504
VM-UNet (concat)	0.719	0.557	9.2	0.778	0.732	0.581	9.8	0.794	0.450	0.291	22.8	0.533
VM-UNet (add)	0.704	0.543	9.5	0.762	0.712	0.551	10.1	0.770	0.439	0.280	23.2	0.521
VM-UNet (multiply)	0.703	0.543	9.3	0.771	0.720	0.562	9.9	0.784	0.457	0.296	22.5	0.540
MSAM [46]	0.733	0.580	8.9	0.792	0.740	0.586	9.5	0.796	0.482	0.321	21.0	0.559
MMCL [166]	0.645	0.477	10.5	0.716	0.681	0.510	11.2	0.739	0.392	0.244	25.0	0.461
AATSN [1]	0.721	0.560	8.7	0.781	0.728	0.569	9.2	0.790	0.503	0.332	20.0	0.583
MFCNet [143]	0.762	0.614	8.2	0.811	0.770	0.631	8.8	0.819	<u>0.603</u>	<u>0.432</u>	<u>16.0</u>	<u>0.681</u>
Mirror U-Net [106]	0.705	0.547	9.0	0.766	0.720	0.563	9.6	0.781	0.511	0.344	19.0	0.592
MMCA-Net [181]	<u>0.774</u>	<u>0.635</u>	<u>7.9</u>	<u>0.822</u>	<u>0.781</u>	<u>0.641</u>	<u>8.5</u>	<u>0.834</u>	0.554	0.381	17.4	0.630
Ours (w/o PKT)	**0.794**	**0.663**	**7.5**	**0.850**	**0.810**	**0.682**	**8.0**	**0.871**	**0.619**	**0.448**	**15.6**	**0.704**

All results are reported using dual-modality input at inference. The best and second-best results are shown in **bold** and <u>underlined</u>, respectively.

mance surpasses recent advanced methods, including MMCA-Net (Dice = 0.774, HD95 = 7.9) and MFCNet (Dice = 0.762, HD95 = 8.2), validating the effectiveness of the proposed state-space modelling in capturing cross-modal features. For the AUTOPET dataset, our framework similarly establishes a new leading method. The framework (w/o PKT) setting attains a Dice score of 0.810 and an HD95 of 8.0, outperforming the second-best method, MMCA-Net, which achieves a Dice score of 0.781 and an HD95 of 8.5. This consistent improvement highlights the robustness of the ARM module in handling metabolic feature discrepancies across different PET-NCCT distributions. On the APIS dataset, which presents significant challenges due to the subtle intensity differences of ischemic stroke lesions, our method still maintains its leading position: Our framework (w/o PKT) yields a Dice score of 0.619 and HD95 of 15.6, effectively outperforming MFCNet (Dice = 0.603, HD95 = 16.0) and MMCA-Net (Dice = 0.554, HD95 = 17.4).

These results confirm that, prior to the knowledge transfer stage, our dual-branch architecture explicitly models the asymmetric relationship between modalities, resulting in more accurate lesion delineation than existing fusion baselines.

5.4.3.2 *Qualitative Results*

Figure 5.4 compares our approach (in both dual-modality and NCCT-only settings) with the competing multi-modal fusion methods on representative cases from the HECKTOR and APIS datasets. Visually, the strong modalities exhibit superior lesion conspicuity compared with NCCT, characterised by hyper-intense (bright) regions on PET and hypo-intense (dark) areas on MRI. In the dual-modality setting (without PKT), our approach delivers visibly more accurate lesion delineations (see the fourth column) than competing multi-modal fusion methods across both PET/NCCT (rows (a) and (b)) and MRI/NCCT (rows (c) and (d)) cases. The segmentation masks are highly consistent with the ground-truth masks, with smooth and complete boundaries even for lesions exhibiting blurred contours and heterogeneous uptake in the strong modality. In conclusion, the proposed ARM module is effective in capturing the asymmetric relationship between modalities and integrating the complementary information from both modalities. Other compared methods tend to over-segment the lesions or generate erroneous shapes.

5.4.4 *Effectiveness of Progressive Knowledge Transfer*

5.4.4.1 *Comparison under NCCT-Only Inference*

Firstly, we assess the fine-tuning strategy across different methods on the NCCT of the HECKTOR and APIS datasets, as detailed in Table 5.3.

Table 5.3: Comparison of knowledge transfer strategies (post-training) across different methods on HECKTOR and APIS datasets when using only NCCT at inference

Method	Post-Training	HECKTOR		APIS	
		Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
VM-UNet (Single)	N/A	0.554	11.0	0.394	23.8
VM-UNet (Early)	FT	0.514	12.3	0.351	26.5
VM-UNet (Concat)	FT	0.540	11.5	0.378	24.8
MSAM [46]	FT	0.530	11.8	0.374	25.0
MMCL [166]	FT	0.408	15.5	0.355	26.0
AATSN [1]	FT	0.543	11.4	0.394	24.5
MFCNet [143]	FT	<u>0.571</u>	<u>10.9</u>	0.411	23.0
Mirror U-Net [106]	FT	0.442	14.8	0.381	24.2
MMCA-NET [181]	FT	0.563	11.1	<u>0.414</u>	<u>22.9</u>
Ours	FT	0.523	12.0	0.419	22.5
	MSE	0.558	11.2	0.430	22.3
	PKT (w/o noise)	0.592	10.8	0.463	21.8
	PKT (w/ noise)	**0.623**	**10.3**	**0.484**	**21.5**

"FT" represents fine-tuning the weak-modality encoder without PKT. The "Single" baseline denotes training solely on NCCT. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

Table 5.4: Ablation study on the number of PKT layers

PKT Layers	HECKTOR		APIS	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
N/A	0.554	11.0	0.394	23.8
L_4	0.591	10.7	0.407	23.0
L_3, L_4	0.593	10.6	0.452	22.0
L_2, L_3, L_4	0.600	10.5	0.458	21.9
L_1, L_2, L_3, L_4	**0.623**	**10.3**	**0.484**	**21.5**

The impact of progressively applying PKT from the deepest layer (L_4) to the shallowest layer (L_1) is investigated. "Layers" denotes the specific encoder stages where PKT is active. The best results are shown in **bold**.

For methods requiring dual-modality inputs, we duplicate the NCCT modality as the strong modality. It is evident that simply fine-tuning multi-modal models on NCCT often underperforms the single-modality baseline. Similarly, our framework without PKT (Ours (FT)) struggles (0.523 Dice on HECKTOR), confirming that, without effective knowledge transfer, the structural discrepancy leads to model degradation when the strong modality is absent.

Secondly, ablation results on aspects of PKT further highlight its necessity. As shown in the bottom part of Table 5.3, replacing fine-tuning with a simple MSE-based feature regression brings only limited gains (0.430 vs. 0.419 on APIS), suggesting that naïve feature alignment is insufficient. Further, introducing progressive distillation (PKT (w/o noise) leads to a notable improvement (+3.4% Dice), which is then enhanced by noise-conditioned targets, achieving the best results on both HECKTOR (0.623 Dice) and APIS (0.484 Dice).

In the last column of Fig. 5.4, it can be observed that even when only NCCT is available at inference, our method yields plausible results. This confirms that progressive, diffusion-inspired distillation is highly effective in transferring informative knowledge to the weak-modality encoder.

5.4.4.2 *Ablation on the Position of PKT*

To evaluate the effect of PKT at different feature abstraction levels, we conduct an ablation study by progressively extending PKT from the deepest layer (L_4) to the shallowest layer (L_1), as detailed in Table 5.4. The results show a clear monotonic improvement in segmentation performance as PKT is applied to more layers. Applying PKT only at the

deepest layer (L_4) already yields noticeable gains over the baseline, indicating that transferring high-level semantic information benefits coarse lesion localisation. However, restricting PKT to deep features limits boundary refinement and results in suboptimal performance. Incorporating shallower layers (L_3–L_1) leads to consistent additional improvements, highlighting the importance of low-level feature transfer for capturing fine-grained structural details in NCCT images. Overall, applying PKT across all layers (L_1–L_4) achieves the best performance, confirming that comprehensive multi-scale knowledge transfer is essential for effectively bridging the strong-to-weak modality gap.

5.4.4.3 *Interpretability Analysis*

We further investigate how PKT enhances lesion sensitivity in the weak modality using the IBIX framework [42], which visualises mask-level sensitivity responses by measuring output variations under localised input perturbations. Figure 5.5 compares the prediction and sensitivity maps produced by the strong- and weak-modality branches. As shown in the first row, the PET's output exhibits highly concentrated activations along the ground-truth lesion boundary, reflecting its strong metabolic contrast and reliable lesion localisation. In contrast, the NCCT branch before PKT produces diffuse and noisy sensitivity responses, with activations spreading beyond the lesion region and inducing false positives. After applying PKT, the NCCT model demonstrates markedly improved localisation, with sensitivity responses becoming more focused around the true lesion and substantially reduced spurious activations.

This comparison also highlights the complementary characteristics of PET and NCCT. PET provides strong functional information enabling robust coarse lesion localisation, but its responses tend to be spatially smoothed due to limited resolution and partial-volume effects [70], resulting in less detailed lesion shapes. In contrast, NCCT predictions are constrained by anatomical structures and intensity gradients, which discourage unrealistic extension into adjacent cavities. By transferring PET-derived localisation cues through PKT, the NCCT branch is guided to suppress structurally inconsistent activations while preserving anatomically plausible boundaries, thereby achieving accurate and focused lesion sensitivity.

5.5 CHAPTER SUMMARY

In this chapter, we have presented a cross-modal knowledge transfer framework that addresses the modality-information asymmetry problem in medical image segmentation. The proposed Asymmetric Relationship Modelling (ARM) module constructs high-confidence fused representations from multi-modal inputs by modelling patch-level

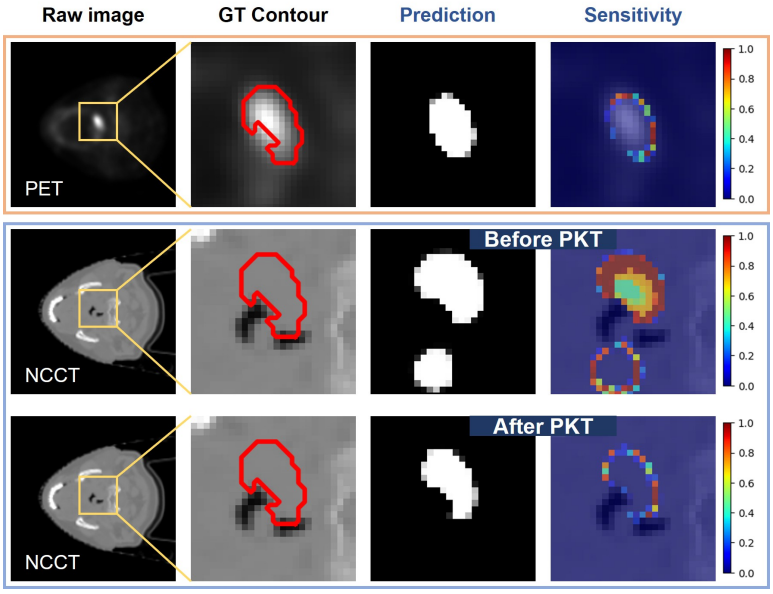


Figure 5.5: Visualisation of lesion predictions and sensitivity maps using the IBIX framework. From top to bottom, rows correspond to the PET output, the NCCT output before PKT, and the NCCT output after PKT. After applying PKT, the NCCT activations become more localised and better aligned with the true lesion regions, in contrast to the diffuse and noisy responses observed before PKT.

cross-modal relationships with state-space operators. The Progressive Knowledge Transfer (PKT) mechanism then distils these enhanced representations into a weak-modality encoder through a diffusion-inspired step-wise scheme, avoiding one-shot overload and stabilising optimisation.

Extensive experiments on three public datasets demonstrate that our framework attains superior segmentation accuracy, outperforming the second-best method by an average of 2.2% in Dice score in multi-modal settings. Crucially, using the weak modality alone at inference, our method achieves around 6.1% Dice improvement in single-modality segmentation. Interpretation analysis further confirms that PKT improves boundary fidelity and suppresses false positives, providing a practical path toward reliable NCCT-only lesion segmentation in resource-constrained clinical scenarios.

Several directions remain for future work. First, region-adaptive knowledge transfer strategies could be explored by disentangling localisation and shape modelling across modalities. Second, extending AR-M/PKT to 3D volumetric encoders and hybrid 2D–3D schemes would better leverage spatial context. Third, broader validation across diseases

and centres with prospective clinical studies would provide stronger evidence for generalizability and clinical utility.

The next chapter concludes this thesis by summarising the main contributions and discussing future research directions.

6 CONCLUSION

This thesis has addressed three practical constraints that commonly limit the deployment of deep-learning-based medical image segmentation systems: heterogeneous annotation quality, limited annotation quantity, and inference-time modality asymmetry. Each constraint was tackled through a dedicated methodological contribution, collectively forming a coherent framework for improving segmentation reliability under realistic clinical conditions. This chapter summarises the main findings, discusses limitations, and outlines directions for future research.

6.1 SUMMARY OF CONTRIBUTIONS

The central motivation of this thesis stems from a pragmatic observation: segmentation methods that excel on curated benchmarks often degrade when confronted with the imperfect, scarce, or incomplete supervision that characterises real-world clinical data. Rather than pursuing architectural novelty, this work focused on developing training strategies and knowledge transfer mechanisms that make principled use of auxiliary signals available during annotation or data acquisition. The three contributions are summarised below.

Contribution 1: Learning under heterogeneous annotation quality (Chapter 3). We introduced **ReReNet**, a recurrent refined network that learns to transform coarse, non-expert annotations into clinically accurate delineations by modelling the latent correction patterns observed in realistic annotation workflows. The key insight is that label refinement—where initial annotations by non-specialised personnel are subsequently corrected by expert physicians—encodes structured information that, if leveraged appropriately, can improve both boundary consistency and lesion completeness. ReReNet achieves progressively refined segmentation through multiple iterations with discrepancy-aware supervision, enabling the model to internalise the process of "turning the barely correct into the clinically accurate". Experiments on three medical image segmentation datasets spanning MR and CT modalities demonstrated that ReReNet consistently outperforms mainstream architectures, validating the effectiveness of mining hidden correction patterns from imperfect labels.

Contribution 2: Label-efficient learning under limited annotation quantity (Chapter 4). We developed **3DRotNPC**, a self-supervised learning framework that exploits the 3D spatial and geomet-

ric structure of volumetric MRI data to reduce reliance on voxel-wise annotations. The core idea is to design a proxy task—predicting the rotation applied to randomly selected slices within consecutive volumes—that encourages the encoder to learn anatomically meaningful representations without requiring labels. After pre-training, the learned parameters are transferred to the downstream segmentation task, yielding substantial performance gains under label-limited regimes. Extensive experiments on nasopharyngeal carcinoma MRI data showed that 3DRotNPC delivers considerable improvements when only 10% to 50% of the labelled data are available, demonstrating that self-supervised pre-training provides a cost-effective path toward accurate volumetric segmentation.

Contribution 3: Deployment-oriented segmentation under strong–weak modality asymmetry (Chapter 5). We proposed a two-stage framework comprising **Asymmetric Relationship Modelling (ARM)** and **Progressive Knowledge Transfer (PKT)** to bridge the performance gap between strong modalities (e.g., PET or contrast-enhanced MRI) and weak modalities (e.g., non-contrast CT) at inference. ARM captures patch-level cross-modal relationships using state-space operators, yielding enhanced, lesion-sensitive fused representations. PKT then progressively distils these representations into the weak-modality encoder through a diffusion-inspired, stepwise scheme, avoiding one-shot overload and stabilising optimisation. Experiments on three public datasets demonstrated that our framework achieves an average improvement of 2.2% Dice in multi-modal settings and, crucially, narrows the strong–weak performance gap by approximately 6.1% Dice when only the weak modality is available at inference. Interpretability analysis confirmed that PKT improves boundary fidelity and suppresses false positives, providing a practical path toward reliable weak-modality-only lesion segmentation.

Unifying perspective. Together, these contributions address complementary facets of a common problem: how to train robust segmentation models when the available supervision is imperfect in quality, limited in quantity, or constrained by modality availability at deployment. The three methods share a design philosophy of exploiting auxiliary signals—refinement structure, unlabelled volumes, and strong-modality information—that are often available during data creation or model development but are rarely utilised effectively. Moreover, these contributions could be combined within a single pipeline when multiple constraints co-occur. For example, one could pre-train an encoder with SSL (CII), transfer strong-modality knowledge during training (CIII), and then apply recurrent refinement to exploit coarse-to-expert supervision (CI), yielding a unified system. By making principled use of these signals, the proposed approaches improve deployability and clinical util-

ity without requiring prohibitively large annotation budgets or multi-modal inputs at inference.

6.2 LIMITATIONS AND FUTURE WORK

While the proposed methods have demonstrated promising results, several limitations warrant further investigation. These limitations also point to opportunities for future research.

Extension to 3D architectures. The methods developed in Chapter 3 and Chapter 5 were evaluated in a slice-wise 2D setting, where training and inference operate on individual axial slices. Although 2D processing is computationally efficient and widely adopted in clinical workflows, volumetric scans inherently encode rich 3D contextual information that slice-wise processing cannot fully exploit, such as through-plane boundary continuity and coherent 3D shape constraints. A natural extension is to implement ReReNet and the ARM/PKT framework on native 3D backbones (e.g., 3D U-Net/UNETR-style encoders) or on hybrid 2.5D/3D schemes that incorporate neighbouring slices while respecting anisotropic voxel spacing. This direction is particularly relevant for lesions with complex 3D morphology or anisotropic growth patterns.

Generalisation across clinical scenarios. The experiments in this thesis focused on specific clinical applications, including nasopharyngeal carcinoma segmentation, head-and-neck tumour delineation, and ischemic stroke lesion detection. While these scenarios are representative, broader validation across a wider range of diseases, anatomical regions, imaging protocols, and clinical centres would strengthen the evidence for generalisability. Prospective studies with external validation cohorts are needed to assess real-world deployment performance.

Initialisation strategies for recurrent refinement. In ReReNet, the initial coarse segmentation during inference is derived from non-expert labels, which may not always be available in practice. Alternative initialisation strategies—such as outputs from rule-based algorithms, atlas-based priors, or foundation models—could extend the applicability of the recurrent refinement paradigm to settings where non-expert annotations are absent.

Richer self-supervised proxy tasks. The rotation prediction task in 3DRotNPC represents one of many possible pretext tasks for self-supervised learning on volumetric data. Exploring complementary objectives, such as masked autoencoding, contrastive learning, or multi-task pretraining, may yield more expressive representations and further improve downstream segmentation performance. Combining multiple

proxy tasks in a curriculum or multi-objective framework is a promising avenue.

Region-adaptive knowledge transfer. The current PKT mechanism applies uniform distillation across the entire feature map. Introducing region-adaptive transfer strategies—for example, by disentangling localisation cues and shape modelling across modalities—could allow more targeted knowledge transfer and reduce the risk of propagating modality-specific artefacts into the weak-modality encoder.

Integration with foundation models. Recent advances in vision-language foundation models (e.g., SAM, MedSAM) offer powerful zero-shot and few-shot segmentation capabilities. Investigating how the methods developed in this thesis can be integrated with or benefit from such foundation models—for example, using foundation-model outputs as initialisation for recurrent refinement or as pseudo-labels for self-supervised pretraining—represents an exciting direction for future work.

6.3 CLOSING REMARKS

Medical image segmentation remains a cornerstone of computer-aided diagnosis and treatment planning. This thesis has shown that practical constraints—imperfect labels, scarce annotations, and modality limitations—need not be insurmountable obstacles. By designing methods that exploit the structure of annotation workflows, the continuity of volumetric data, and the complementary information across modalities, it is possible to train segmentation models that are more robust, more data-efficient, and more deployable in resource-constrained clinical settings. We hope that the contributions presented here will inspire further research toward bridging the gap between benchmark performance and real-world clinical utility.

BIBLIOGRAPHY

[1] I. Ahmad, Y. Xia, H. Cui, and Z. U. Islam. AATSN: Anatomy aware tumor segmentation network for PET-CT volumes and images using a lightweight fusion-attention mechanism. *Computers in Biology and Medicine*, 157:106748, 2023.

[2] U. Ahsan, R. Madhok, and I. Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189, 2019.

[3] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011.

[4] G. Athanasiou, J. L. Arcos, and J. Cerquides. Enhancing medical image segmentation: Ground truth optimization through evaluating uncertainty in expert annotations. *Mathematics*, 11(17):3771, 2023.

[5] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof. Medical image segmentation review: The success of U-Net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10076–10095, 2024.

[6] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof. Advances in medical image analysis with vision Transformers: A comprehensive review. *Medical Image Analysis*, 91:103000, 2024.

[7] R. Azad, M. Dehghanmanshadi, N. Khosravi, J. Cohen-Adad, and D. Merhof. Addressing missing modality challenges in MRI im-

ages: A comprehensive review. *Computational Visual Media*, 11 (2):241–268, 2025.

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

[9] X. Bai, Y. Hu, G. Gong, Y. Yin, and Y. Xia. A deep learning approach to segmentation of nasopharyngeal carcinoma using computed tomography. *Biomedical Signal Processing and Control*, 64: 102246, 2021.

[10] S. Bansal, S. A, M. P. J, M. S, S. Madisetty, M. Z. U. Rehman, C. S. Raghaw, G. Duggal, and N. Kumar. A comprehensive survey of Mamba architectures for medical image analysis: Classification, segmentation, restoration and beyond, 2024. URL https://arxiv.org/abs/2410.02362. arXiv:2410.02362.

[11] B. Benato, A. Falcão, and A. C. Telea. Linking data separation, visual separation, and classifier performance using pseudo-labeling by contrastive learning. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 315–324, 2023.

[12] B. C. Benato, J. F. Gomes, A. C. Telea, and A. X. Falcão. Semi-automatic data annotation guided by feature space projection. *Pattern Recognition*, 109:107612, 2021.

[13] B. C. Benato, J. F. Gomes, A. C. Telea, and A. X. Falcão. Semi-supervised deep learning based on label propagation in a 2D embedded space. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, May 10–13, 2021, Revised Selected Papers*, pages 371–381, 2021.

[14] B. C. Benato, A. C. Telea, and A. X. Falcão. Iterative pseudo-labeling with deep feature annotation and confidence-based sampling, 2021.

[15] B. C. Benato, A. C. Telea, and A. X. Falcão. Deep feature annotation by iterative meta-pseudo-labeling on 2D projections. *Pattern Recognition*, 141:109649, 2023.

[16] B. C. Benato, C. Grosu, A. X. Falcão, and A. C. Telea. Human-in-the-loop: Using classifier decision boundary maps to improve pseudo labels. *Computers & Graphics*, 124:104062, 2024.

[17] B. C. Benato, A. C. Telea, and A. X. Falcão. Semi-supervised learning with interactive label propagation guided by feature space projections. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 392–399, 2018.

[18] L. Bi, M. Fulham, S. Song, D. D. Feng, and J. Kim. Hyper-connected transformer network for multi-modality PET-CT segmentation. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4, 2023.

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 205–218, 2022.

[20] K. Cao, Y. Xia, J. Yao, X. Han, L. Lambert, T. Zhang, W. Tang, G. Jin, H. Jiang, X. Fang, I. Nogues, X. Li, W. Guo, Y. Wang, W. Fang, M. Qiu, Y. Hou, T. Kovarnik, M. Vocka, Y. Lu, Y. Chen, X. Chen, Z. Liu, J. Zhou, C. Xie, R. Zhang, H. Lu, G. D. Hager, A. L. Yuille, L. Lu, C. Shao, Y. Shi, Q. Zhang, T. Liang, L. Zhang, and J. Lu. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nature Medicine*, 29(12):3033–3043, 2023.

[21] J. A. Chalela, C. S. Kidwell, L. M. Nentwich, M. Luby, J. A. Butman, A. M. Demchuk, M. D. Hill, N. Patronas, L. Latour, and S. Warach. Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. *The Lancet*, 369(9558):293–298, 2007.

[22] F. Chen, H. Han, P. Wan, L. Chen, W. Kong, H. Liao, B. Wen, C. Liu, and D. Zhang. Do as sonographers think: Contrast-enhanced ultrasound for thyroid nodules diagnosis via microvascular infiltrative awareness. *IEEE Transactions on Medical Imaging*, 43(11): 3881–3894, 2024.

[23] H. Chen, F. Shao, W. Jing, H. Wang, and Q. Jiang. Cross-modal hierarchical knowledge distillation for image aesthetics assessment. *IEEE Transactions on Multimedia*, 27:2556–2569, 2025.

[24] H. Chen, H. Ding, Y. Jiang, J. Lan, K. C. Li, G. W. Y. Cheng, N. C. Ng, Y. Pu, J. Cai, L. t. Lin, and J. S. Yoo. REACT-KD: Region-aware cross-modal topological knowledge distillation for interpretable medical image classification, 2025. URL https://arxiv.org/abs/2508.02104. arXiv:2508.02104.

[25] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and

Y. Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

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

[27] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848, 2018.

[28] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 833–851, 2018.

[29] Q. Chen, J. Zhang, R. Meng, L. Zhou, Z. Li, Q. Feng, and D. Shen. Modality-specific information disentanglement from multi-parametric MRI for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 43(5):1958–1971, 2024.

[30] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao. S3D-UNet: Separable 3D U-Net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 358–368, 2019.

[31] Y. P. Chen, A. T. C. Chan, Q. T. Le, P. Blanchard, Y. Sun, and J. Ma. Nasopharyngeal carcinoma. *The Lancet*, 394(10192):64–80, 2019.

[32] P. Chlap, H. Min, J. Dowling, M. Field, K. Cloak, T. Leong, M. Lee, J. Chu, J. Tan, P. Tran, T. Kron, M. Sidhom, K. Wiltshire, S. Keats, A. Kneebone, A. Haworth, M. A. Ebert, S. K. Vinod, and L. Holloway. Uncertainty estimation using a 3D probabilistic U-Net for segmentation with small radiotherapy clinical trial datasets. *Computerized Medical Imaging and Graphics*, 116:102403, 2024.

[33] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 9901, pages 424–432, 2016.

[34] L. Dai, B. Sheng, T. Chen, Q. Wu, R. Liu, C. Cai, L. Wu, D. Yang, H. Hamzah, Y. Liu, X. Wang, Z. Guan, S. Yu, T. Li, Z. Tang, A. Ran, H. Che, H. Chen, Y. Zheng, J. Shu, S. Huang, C. Wu, S. Lin, D. Liu, J. Li, Z. Wang, Z. Meng, J. Shen, X. Hou, C. Deng, L. Ruan, F. Lu, M. Chee, T. C. Quek, R. Srinivasan, R. Raman, X. Sun, Y. X. Wang, J. Wu, H. Jin, R. Dai, D. Shen, X. Yang, M. Guo, C. Zhang, C. Y. Cheung, G. S. W. Tan, Y. C. Tham, C. Y. Cheng, H. Li, T. Y. Wong, and W. Jia. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*, 2024.

[35] T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[36] B. D. de Vos, G. E. Jansen, and I. Išgum. Stochastic co-teaching for training neural networks with unknown levels of label noise. *Scientific Reports*, 13:16875, 2023.

[37] Y. Deng, C. Li, X. Lv, W. Xia, L. Shen, B. Jing, B. Li, X. Guo, Y. Sun, C. Xie, and L. Ke. The contrast-enhanced MRI can be substituted by unenhanced MRI in identifying and automatically segmenting primary nasopharyngeal carcinoma with the aid of deep learning models: An exploratory study in large-scale population of endemic area. *Computer Methods and Programs in Biomedicine*, 217: 106702, 2022.

[38] Z. Diao, H. Jiang, and T. Shi. A spatial squeeze and multimodal feature fusion attention network for multiple tumor segmentation from PET–CT volumes. *Engineering Applications of Artificial Intelligence*, 121:105955, 2023.

[39] Z. Diao, H. Jiang, T. Shi, and Y. D. Yao. Siamese semi-disentanglement network for robust PET-CT segmentation. *Expert Systems with Applications*, 223:119855, 2023.

[40] S. Dong, J. Zhao, M. Zhang, Z. Shi, J. Deng, Y. Shi, M. Tian, and C. Zhuo. DeU-Net: Deformable U-Net for 3D cardiac MRI video segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 98–107, 2020.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[42] M. Espadoto, S. B. Martins, W. Branderhorst, and A. Telea. Machine learning—basic unsupervised methods (cluster analysis methods, t-SNE). In *Clinical Applications of Artificial Intelligence in Real-World Data*, pages 141–159. Springer International Publishing, Cham, 2023.

[43] L. Fan, Y. Ou, C. Zheng, P. Dai, T. Kamishima, M. Ikebe, K. Suzuki, and X. Gong. MDA: An interpretable multi-modal fusion with missing modalities and intrinsic noise, 2024. URL https://arxiv.org/abs/2406.10569. arXiv:2406.10569.

[44] C. M. Feng, Y. Yan, G. Chen, Y. Xu, Y. Hu, L. Shao, and H. Fu. Multimodal transformer for accelerated MR imaging. *IEEE Transactions on Medical Imaging*, 42(10):2804–2816, 2023.

[45] Y. Feng, Y. Wang, H. Li, M. Qu, and J. Yang. Learning what and where to segment: A new perspective on medical image few-shot segmentation. *Medical Image Analysis*, 87:102834, 2023.

[46] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim. Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3507–3516, 2021.

[47] S. Gatidis, T. Hepp, M. Früh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Küstner, C. Cyran, and D. Rubin. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

[48] S. Gómez, E. Rangel, D. Mantilla, A. Ortiz, P. Camacho, E. de la Rosa, J. Seia, J. S. Kirschke, Y. Li, M. El Habib Daho, and F. Martínez. APIS: a paired CT-MRI dataset for ischemic stroke segmentation - methods and challenges. *Scientific Reports*, 14(1): 20543, 2024.

[49] A. Gu, K. Goel, and C. Re. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[50] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8536–8546, 2018.

[51] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, 2022.

[52] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1748–1758, 2022.

[53] D. He, W. Li, G. Wang, Y. Huang, and S. Liu. DM-FNet: Unified multimodal medical image fusion via diffusion process-trained encoder-decoder. *IEEE Transactions on Multimedia*, 27:9415–9428, 2025.

[54] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[55] S. He, Y. Ji, Y. Zhang, A. Zeng, D. Pan, J. Lin, and X. Zhang. CFNet: A coarse-to-fine framework for coronary artery segmentation. In *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 431–442, 2023.

[56] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531. arXiv:1503.02531.

[57] R. Hird. The basics of mri interpretation, 2020. URL https://geekymedics.com/the-basics-of-mri-interpretation/.

[58] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020.

[59] V. T. Hu, S. A. Baumann, M. Gui, O. Grebenkova, P. Ma, J. Fischer, and B. Ommer. ZigMa: A DiT-style zigzag Mamba diffusion model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 148–166, 2024.

[60] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[61] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. W. Chen, and J. Wu. UNet 3+: A full-scale connected UNet for medical image segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059, 2020.

[62] T. Huang, Y. Zhang, M. Zheng, S. You, F. Wang, C. Qian, and C. Xu. Knowledge diffusion for distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 65299–65316, 2023.

[63] Z. Huang, S. Tang, Z. Chen, G. Wang, H. Shen, Y. Zhou, H. Wang, W. Fan, D. Liang, Y. Hu, and Z. Hu. TG-Net: Combining transformer and GAN for nasopharyngeal carcinoma tumor segmentation based on total-body uEXPLORER PET/CT scanner. *Computers in Biology and Medicine*, 148:105869, 2022.

[64] N. Ibtehaz and M. S. Rahman. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, 2020.

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

[66] W. Ji and A. C. S. Chung. Unsupervised domain adaptation for medical image segmentation using Transformer with meta attention. *IEEE Transactions on Medical Imaging*, 43(2):820–831, 2024.

[67] H. Jiang, P. Cao, M. Xu, J. Yang, and O. Zaiane. Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction. *Computers in Biology and Medicine*, 127:104096, 2020.

[68] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2021.

[69] L. Jing, X. Yang, J. Liu, and Y. Tian. Self-supervised spatiotemporal feature learning via video rotation prediction, 2018. URL https://arxiv.org/abs/1811.11387. arXiv:1811.11387.

[70] H. Jomaa, R. Mabrouk, and N. Khlifa. Post-reconstruction-based partial volume correction methods: A comprehensive review. *Biomedical Signal Processing and Control*, 46:131–144, 2018.

[71] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[72] S. Karimijafarbigloo, R. Azad, A. Kazerouni, S. Ebadollahi, and D. Merhof. MMCFormer: Missing modality compensation transformer for brain tumor segmentation. In *Medical Imaging with Deep Learning*, pages 1144–1162, 2024.

[73] T. Klinghoffer, P. Morales, Y. G. Park, N. Evans, K. Chung, and L. J. Brattain. Self-supervised feature extraction for 3D axon segmentation. In *CVPRW*, pages 4213–4219, 2020.

[74] S. A. A. Kohl, B. Romera-Paredes, C. Meyer, J. D. Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. A. Eslami, D. J. Rezende, and O. Ronneberger. A probabilistic U-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6965–6975, 2018.

[75] V. Kookna. Semantic vs. instance vs. panoptic segmentation, 2022. URL https://www.pyimagesearch.com/2022/02/07/semantic-vs-instance-vs-panoptic-segmentation/.

[76] R. Krishnan, P. Rajpurkar, and E. J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6 (12):1346–1352, 2022.

[77] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

[78] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman. 3D UX-Net: A large kernel volumetric ConvNet modernizing hierarchical Transformer for medical image segmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[79] K. Lee, J. Zung, P. Li, V. Jain, and H. S. Seung. Superhuman accuracy on the SNEMI3D connectomics challenge, 2017. URL https://arxiv.org/abs/1706.00120. arXiv:1706.00120.

[80] C. Li, R. Jiang, S. Yin, J. Yang, and X. Ban. Self-supervised rotation learning for 3D segmentation on nasopharyngeal carcinoma MRI images. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3529–3534, 2023.

[81] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou. Transforming medical imaging with Transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, 85:102762, 2023.

[82] S. Li, Y. Q. Deng, H. L. Hua, S. L. Li, X. X. Chen, B. J. Xie, Z. Zhu, R. Liu, J. Huang, and Z. Z. Tao. Deep learning for locally advanced nasopharyngeal carcinoma prognostication based on pre- and post-treatment MRI. *Computer Methods and Programs in Biomedicine*, 219:106785, 2022.

[83] X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

[84] X. Li, W. Tan, P. Liu, Q. Zhou, and J. Yang. Classification of COVID-19 chest CT images based on ensemble deep learning. *Journal of Healthcare Engineering*, 2021(1), 2021.

[85] Y. Li, H. Peng, T. Dan, Y. Hu, G. Tao, and H. Cai. Coarse-to-fine nasopharyngeal carcinoma segmentation in MRI via multi-stage rendering. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 623–628, 2020.

[86] Y. Li, T. Dan, H. Li, J. Chen, H. Peng, L. Liu, and H. Cai. NPC-Net: Jointly segment primary nasopharyngeal carcinoma tumors and metastatic lymph nodes in MR images. *IEEE Transactions on Medical Imaging*, 41(7):1639–1650, 2022.

[87] P. Liang, J. Chen, Q. Chang, and L. Yao. RSKD: Enhanced medical image segmentation via multi-layer, rank-sensitive knowledge distillation in vision Transformer models. *Knowledge-Based Systems*, 293:111664, 2024.

[88] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, and L. Ma. LightM-UNet: Mamba assists in lightweight UNet for medical image segmentation, 2024. URL https://arxiv.org/abs/2403.05246. arXiv:2403.05246.

[89] Q. Lin, K. He, Y. Zhu, F. Xu, E. Cambria, and M. Feng. Cross-modal knowledge diffusion-based generation for difference-aware medical VQA. *IEEE Transactions on Image Processing*, 34:2421–2434, 2025.

[90] Z. Ling, G. Tao, Y. Li, and H. Cai. NPCFORMER: Automatic nasopharyngeal carcinoma segmentation based on boundary attention and global position context attention. In *ICIP*, pages 1981–1985, 2022.

[91] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[92] G. Liu, Y. Jiang, D. Liu, B. Chang, L. Ru, and M. Li. A coarse-to-fine segmentation frame for polyp segmentation via deep and classification features. *Expert Systems with Applications*, 214:118975, 2023.

[93] J. Liu, H. Yang, H. Y. Zhou, L. Yu, Y. Liang, Y. Yu, S. Zhang, H. Zheng, and S. Wang. Swin-UMamba†: Adapting Mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(10):3898–3908, 2025.

[94] S. Liu, M. Zou, N. Liu, Y. Li, and W. Zheng. A teacher-guided early-learning method for medical image segmentation from noisy labels. *Complex & Intelligent Systems*, 10(6):8011–8026, 2024.

[95] X. Liu, C. Zhang, F. Huang, S. Xia, G. Wang, and L. Zhang. Vision Mamba: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2025.

[96] X. Liu, F. Xing, T. Marin, G. E. Fakhri, and J. Woo. Variational inference for quantifying inter-observer variability in segmentation of anatomical structures. In *Medical Imaging 2022: Image Processing*, volume 12032, page 120321M, 2022.

[97] Y. Liu, X. Yuan, X. Jiang, P. Wang, J. Kou, H. Wang, and M. Liu. Dilated adversarial U-Net network for automatic gross tumor volume segmentation of nasopharyngeal carcinoma. *Applied Soft Computing*, 111:107722, 2021.

[98] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 103031–103063, 2024.

[99] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[100] Z. Lu, Y. Liu, M. Jin, X. Luo, H. Yue, Z. Wang, S. Zuo, Y. Zeng, J. Fan, Y. Pang, J. Wu, J. Yang, and Q. Dai. Virtual-scanning light-field microscopy for robust snapshot high-resolution volumetric imaging. *Nature Methods*, 20(5):735–746, 2023.

[101] X. Luo, J. Fu, Y. Zhong, S. Liu, B. Han, M. Astaraki, S. Bendazzoli, I. Toma-Dasu, Y. Ye, Z. Chen, Y. Xia, Y. Su, J. Ye, J. He, Z. Xing, H. Wang, L. Zhu, K. Yang, X. Fang, Z. Wang, C. W. Lee, S. J. Park, J. Chun, C. Ulrich, K. H. Maier-Hein, N. Ndipenoch, A. Miron, Y. Li, Y. Zhang, Y. Chen, L. Bai, J. Huang, C. An, L. Wang, K. Huang, Y. Gu, T. Zhou, M. Zhou, S. Zhang, W. Liao, G. Wang, and S. Zhang. SegRap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Medical Image Analysis*, 101:103447, 2025.

[102] C. Ma and Z. Wang. Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 300:112203, 2024.

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

[104] Q. Ma, C. Zu, X. Wu, J. Zhou, and Y. Wang. Coarse-to-fine segmentation of organs at risk in nasopharyngeal carcinoma radiotherapy. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 358–368, 2021.

[105] A. M. Mansourian, R. Ahmadi, M. Ghafouri, A. M. Babaei, E. B. Golezani, Z. yasamani Ghamchi, V. Ramezanian, A. Taherian, K. Dinashi, A. Miri, and S. Kasaei. A comprehensive survey on knowledge distillation. *Transactions on Machine Learning Research*, 2025.

[106] Z. Marinov, S. Reiß, D. Kersting, J. Kleesiek, and R. Stiefelhagen. Mirror U-Net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2283–2293, 2023.

[107] M. Martínez-García, J. Vadillo, M. Pedersoli, I. Inza, and J. A. Lozano. Privileged learning via a multi-task distilled approach. *Pattern Recognition*, 178:113389, October 2026.

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

[109] M. Meng, B. Gu, L. Bi, S. Song, D. D. Feng, and J. Kim. DeepMTS: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT. *IEEE Journal of Biomedical and Health Informatics*, 26 (9):4497–4507, 2022.

[110] F. Milletari, N. Navab, and S. A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[111] F. Milletari, N. Navab, and S. A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation, 2016. URL https://arxiv.org/abs/1606.04797. arXiv:1606.04797.

[112] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[113] M. Moniruzzaman and Z. Yin. Progressive knowledge distillation from different levels of teachers for online action detection. *IEEE Transactions on Multimedia*, 27:1526–1537, 2025.

[114] M. Mubashar, H. Ali, C. Grönlund, and S. Azmat. R2U++: a multiscale recurrent residual U-Net with dense skip connections for medical image segmentation. *Neural Computing and Applications*, 34(20):17723–17739, 2022.

[115] B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[116] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention U-Net: Learning where to look for the pancreas. In *Proceedings of the Medical Image Deep Learning (MIDL)*, 2022.

[117] V. Oreiller, V. Andrearczyk, M. Jreige, S. Boughdad, H. Elhalawani, J. Castelli, M. Vallières, S. Zhu, J. Xie, Y. Peng, A. Iantsen, M. Hatt, Y. Yuan, J. Ma, X. Yang, C. Rao, S. Pai, K. Ghimire, X. Feng, M. A. Naser, C. D. Fuller, F. Yousefirizi, A. Rahmim, H. Chen, L. Wang, J. O. Prior, and A. Depeursinge. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Medical Image Analysis*, 77:102336, 2022.

[118] Y. Qing, S. Liu, H. Wang, and Y. Wang. DiffUIE: Learning latent global priors in diffusion models for underwater image enhancement. *IEEE Transactions on Multimedia*, 27:2516–2529, 2025.

[119] R. Raumanns, G. Schouten, M. Joosten, J. Pluim, and V. Cheplygina. ENHANCE (enriching health data by annotations of crowd and experts): A case study for skin lesion classification. *Journal of Machine Learning for Biomedical Imaging*, 1:1–26, 2021.

[120] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[121] H. R. Roth, C. Shen, H. Oda, T. Sugino, M. Oda, Y. Hayashi, K. Misawa, and K. Mori. A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 417–425, 2018.

[122] J. Ruan, J. Li, and S. Xiang. VM-UNet: Vision mamba UNet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, page 3767748, 2025.

[123] M. Salehjahromi, T. V. Karpinets, S. J. Sujit, M. Qayati, P. Chen, M. Aminu, M. B. Saad, R. Bandyopadhyay, L. Hong, A. Sheshadri, J. Lin, M. B. Antonoff, B. Sepesi, E. J. Ostrin, I. Toumazis, P. Huang, C. Cheng, T. Cascone, N. I. Vokes, C. Behrens, J. H. Siewerdsen, J. D. Hazle, J. Y. Chang, J. Zhang, Y. Lu, M. C. Godoy, C. Chung, D. Jaffray, I. Wistuba, J. J. Lee, A. A. Vaporciyan, D. L. Gibbons, G. Gladish, J. V. Heymach, C. C. Wu, J. Zhang, and J. Wu. Synthetic PET from CT improves diagnosis and prognosis for lung cancer: Proof of concept. *Cell Reports Medicine*, 5(3):101463, 2024.

[124] E. Scalco, S. Pozzi, G. Rizzo, and E. Lanzarone. Uncertainty quantification in multi-class segmentation: Comparison between Bayesian and non-Bayesian approaches in a clinical perspective. *Medical Physics*, 51(9):6090–6102, 2024.

[125] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, 2023.

[126] D. Shen, G. Wu, and H. I. Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

[127] W. Shen, Z. Peng, X. Wang, H. Wang, J. Cen, D. Jiang, L. Xie, X. Yang, and Q. Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9284–9305, 2023.

[128] J. Shi, K. Zhang, C. Guo, Y. Yang, Y. Xu, and J. Wu. A survey of label-noise deep learning for medical image analysis. *Medical Image Analysis*, 95:103166, 2024.

[129] Y. Shu, H. Li, B. Xiao, X. Bi, and W. Li. Cross-mix monitoring for medical image segmentation with limited supervision. *IEEE Transactions on Multimedia*, 25:1700–1712, 2023.

[130] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

[131] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your ViT? data, augmentation, and regularization in vision Transformers. *Transactions on Machine Learning Research*, 2022.

[132] W. Tan, P. Huang, X. Li, G. Ren, Y. Chen, and J. Yang. Analysis of segmentation of lung parenchyma based on deep learning methods. *Journal of X-Ray Science and Technology*, 29(6):945–959, 2021.

[133] J. Tang, Q. Gou, S. Fu, K. Zhao, T. Dong, and Y. Tian. Efficient and robust kernel learning with class-wise privileged information for pattern classification. *Pattern Recognition*, 179:113522, November 2026.

[134] P. Tang, C. Zu, M. Hong, R. Yan, X. Peng, J. Xiao, X. Wu, J. Zhou, L. Zhou, and Y. Wang. DA-DSUnet: Dual attention-based dense SU-net for automatic head-and-neck tumor segmentation in MRI images. *Neurocomputing*, 435:103–113, 2021.

[135] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of Swin Transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20698–20708, 2022.

[136] G. Tao, H. Li, L. Liu, and H. Cai. Detection-and-excitation neural network achieves accurate nasopharyngeal carcinoma segmentation in multi-modality MR images. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1063–1068, 2021.

[137] G. Tao, H. Li, J. Huang, C. Han, J. Chen, G. Ruan, W. Huang, Y. Hu, T. Dan, B. Zhang, S. He, L. Liu, and H. Cai. SeqSeg: A sequential method to achieve nasopharyngeal carcinoma segmentation free from background dominance. *Medical Image Analysis*, 78:102381, 2022.

[138] G. Tao, H. Li, D. Lu, Z. Ling, L. Liu, and H. Cai. Reler: Relearning controversial regions to accurately segment nasopharyngeal carcinoma. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1207–1212, 2022.

[139] Z. Tu, Z. Zhu, Y. Duan, B. Jiang, Q. Wang, and C. Zhang. A spatial-temporal progressive fusion network for breast lesion segmentation in ultrasound videos. *IEEE Transactions on Multimedia*, 27: 7470–7482, 2025.

[140] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, July 2009.

[141] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6000–6010, 2017.

[142] Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, and Y. Xie. Sigma: Siamese Mamba network for multi-modal semantic

segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1734–1744, 2025.

[143] F. Wang, C. Cheng, W. Cao, Z. Wu, H. Wang, W. Wei, Z. Yan, and Z. Liu. MFCNet: A multi-modal fusion and calibration networks for 3D pancreas tumor segmentation on pet-ct images. *Computers in Biology and Medicine*, 155:106657, 2023.

[144] H. Wang, P. Cao, J. Wang, and O. R. Zaiane. UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2441–2449, 2022.

[145] H. Wang, J. Chen, S. Zhang, Y. He, J. Xu, M. Wu, J. He, W. Liao, and X. Luo. Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. *IEEE Transactions on Medical Imaging*, 43(12): 4078–4090, 2024.

[146] J. Wang and J. Jiao. Self-supervised video representation learning by uncovering spatio-temporal statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 2022.

[147] J. Wang, J. Chen, D. Chen, and J. Wu. LKM-UNet: Large kernel vision mamba UNet for medical image segmentation. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, editors, *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 360–370, 2024.

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

[149] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. TransBTS: Multimodal brain tumor segmentation using Transformer. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 109–119, 2021.

[150] Z. Wang, M. Fang, J. Zhang, L. Tang, L. Zhong, H. Li, R. Cao, X. Zhao, S. Liu, R. Zhang, X. Xie, H. Mai, S. Qiu, J. Tian, and D. Dong. Radiomics and deep learning in nasopharyngeal carcinoma: A review. *IEEE Reviews in Biomedical Engineering*, pages 1–18, 2023, in press.

[151] Z. Wang, M. Fang, J. Zhang, L. Tang, L. Zhong, H. Li, R. Cao, X. Zhao, S. Liu, R. Zhang, X. Xie, H. Mai, S. Qiu, J. Tian, and D. Dong. Radiomics and deep learning in nasopharyngeal carcinoma: A review. *IEEE Reviews in Biomedical Engineering*, 17: 118–135, 2024.

[152] Z. Wang and C. Ma. Weak-mamba-unet: Visual mamba makes cnn and vit work better for scribble-based medical image segmentation, 2024. URL https://arxiv.org/abs/2402.10887. arXiv:2402.10887.

[153] Z. Wang, J. Q. Zheng, Y. Zhang, G. Cui, and L. Li. Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation, 2024. URL https://arxiv.org/abs/2402.05079. arXiv:2402.05079.

[154] E. Warner, J. Lee, W. Hsu, T. Syeda-Mahmood, C. E. Kahn Jr., O. Gevaert, and A. Rao. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *International Journal of Computer Vision*, 132(9):3753–3769, 2024.

[155] Y. Wei, Y. Deng, C. Sun, M. Lin, H. Jiang, and Y. Peng. Deep learning with noisy labels in medical prediction problems: a scoping review. *Journal of the American Medical Informatics Association*, 31(7):1596–1607, 2024.

[156] J. Wen, F. Qin, J. Du, M. Fang, X. Wei, C. L. P. Chen, and P. Li. Msgfusion: Medical semantic guided two-branch network for multimodal brain image fusion. *IEEE Transactions on Multimedia*, 26: 944–957, 2024.

[157] J. Wu, D. Guo, G. Wang, Q. Yue, H. Yu, K. Li, and S. Zhang. FPL+: Filtered pseudo label-based unsupervised cross-modality adaptation for 3D medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3098–3109, 2024.

[158] L. Wu, J. Zhuang, and H. Chen. Large-scale 3D medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[159] R. Wu, Y. Liu, P. Liang, and Q. Chang. H-vmunet: High-order vision Mamba UNet for medical image segmentation. *Neurocomputing*, 624:129447, 2025.

[160] L. Xiang, Y. Chen, W. Chang, Y. Zhan, W. Lin, Q. Wang, and D. Shen. Deep-learning-based multi-modal fusion for fast MR reconstruction. *IEEE Transactions on Biomedical Engineering*, 66 (7):2105–2114, 2019.

[161] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E. K. Fishman, and A. L. Yuille. Recurrent saliency transformation network for tiny target segmentation in abdominal CT scans. *IEEE Transactions on Medical Imaging*, 39(2):514–525, 2020.

[162] Y. Xie, J. Zhang, C. Shen, and Y. Xia. CoTr: Efficiently bridging CNN and Transformer for 3D medical image segmentation. In

Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), pages 171–180, 2021.

[163] Z. Xing, T. Ye, Y. Yang, D. Cai, B. Gai, X. J. Wu, F. Gao, and L. Zhu. SegMamba-V2: Long-range sequential modeling mamba for general 3-D medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(1):4–15, 2026.

[164] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10326–10335, 2019.

[165] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with Transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[166] Z. Xue, P. Li, L. Zhang, X. Lu, G. Zhu, P. Shen, S. A. Ali Shah, and M. Bennamoun. Multi-modal co-learning for liver lesion segmentation on PET-CT images. *IEEE Transactions on Medical Imaging*, 40(12):3531–3542, 2021.

[167] H. Yang, J. Sun, and Z. Xu. Learning unified hyper-network for multi-modal MR image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 42(12): 3678–3689, 2023.

[168] J. Yang, B. Wu, L. Li, P. Cao, and O. Zaiane. MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT. *Computerized Medical Imaging and Graphics*, 92:101957, 2021.

[169] S. Ye, Y. Xu, D. Chen, S. Han, and J. Liao. Learning a single network for robust medical image segmentation with noisy labels. *IEEE Transactions on Medical Imaging*, 43(9):3188–3199, 2024.

[170] H. Yin and Y. Shao. CFU-Net: A coarse–fine U-Net with multilevel attention for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 72:5020412, 2023.

[171] L. Yu, S. Wang, X. Li, C. W. Fu, and P. A. Heng. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 605–613, 2019.

[172] S. Yu, M. Chen, E. Zhang, J. Wu, H. Yu, Z. Yang, L. Ma, X. Gu, and W. Lu. Robustness study of noisy annotation in deep learning based medical image segmentation. *Physics in Medicine & Biology*, 65(17):175007, 2020.

[173] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[174] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4L: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019.

[175] G. Zhang, X. Qi, J. Wu, B. Yan, and G. Wang. IPLC+: SAM-guided iterative pseudo label correction for source-free domain adaptation in medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 29(12):9060–9072, 2025.

[176] H. Zhang, M. Chen, Y. Liu, G. Luo, and Y. Zhu. Non-IID medical image segmentation based on cascaded diffusion model for diverse multi-center scenarios. *IEEE Journal of Biomedical and Health Informatics*, 29(7):5042–5055, 2025.

[177] L. Zhang, F. Wu, K. Bronik, and B. W. Papiez. DiffuSeg: Domain-driven diffusion for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3619–3631, 2025.

[178] Y. Zhang, J. Yang, J. Tian, Z. Shi, C. Zhong, Y. Zhang, and Z. He. Modality-aware mutual learning for multi-modal medical image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 589–599, 2021.

[179] Y. Zhang, R. Xi, L. Zeng, D. Towey, R. Bai, R. Higashita, and J. Liu. Structural priors guided network for the corneal endothelial cell segmentation. *IEEE Transactions on Medical Imaging*, 43(1):309–320, 2024.

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

[181] W. Zhao, Z. Huang, S. Tang, W. Li, Y. Gao, Y. Hu, W. Fan, C. Cheng, Y. Yang, H. Zheng, D. Liang, and Z. Hu. MMCA-NET: A multi-modal cross attention transformer network for nasopharyngeal carcinoma tumor segmentation based on a total-body PET/CT system. *IEEE Journal of Biomedical and Health Informatics*, 28 (9):5447–5458, 2024.

[182] Z. Zhao, H. Yang, and J. Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 183–192, 2022.

[183] Y. Zhong, C. Cai, T. Chen, H. Gui, J. Deng, M. Yang, B. Yu, Y. Song, T. Wang, X. Sun, J. Shi, Y. Chen, D. Xie, C. Chen, and Y. She. PET/CT based cross-modal deep learning signature to predict occult nodal metastasis in lung cancer. *Nature Communications*, 14 (1):7513, 2023.

[184] H. Y. Zhou, C. Lu, C. Chen, S. Yang, and Y. Yu. A unified visual information preservation framework for self-supervised pretraining in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8020–8035, 2023.

[185] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers. 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, 2021.

[186] Y. Zhou, X. Li, F. Liu, Q. Wei, X. Chen, L. Yu, C. Xie, M. P. Lungren, and L. Xing. L2b: Learning to bootstrap robust models for combating label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23523–23533, 2024.

[187] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. 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, 2018.

[188] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021.

[189] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

ACKNOWLEDGMENTS

A thesis is never the work of one person alone, and I am fortunate to have been accompanied by many who made this journey both possible and meaningful.

I owe my first and deepest thanks to my supervisors at the University of Groningen: Prof. Jiří Kosinka, Prof. Alexandru C. Telea, and Dr. Steffen Frey. Jiří steered my study within the SVCG group with quiet steadiness from the very first day; it is no exaggeration to say that this thesis took shape under his care. Alexandru had a gift for seeing the heart of a problem and reflecting it back to me with clarity—every conversation sharpened both my writing and my thinking. Steffen brought a refreshing directness to our meetings, and the pointed questions he raised during the development of the papers did much to elevate the quality of this work.

I am equally grateful to my supervisors in China: Prof. Xiaojuan Ban, Prof. Chao Yao, and Prof. Xiaokun Wang. Xiaojuan gave me room to explore while always being there when I needed guidance; her patience and trust allowed me to grow at my own pace. Chao Yao and Xiaokun Wang offered thoughtful advice on writing and revising the papers that form part of this thesis, and their careful eye for detail saved me from many missteps.

Warm thanks go to the members of the SVCG group, with whom I shared not only an office but also ideas, laughter, and the occasional frustration. It has been a genuine pleasure to work alongside such generous and good-humoured colleagues, and I will carry those memories with me well beyond Groningen.

I gratefully acknowledge the financial support of the China Scholarship Council, without which this research would not have been possible. I also thank the Faculty of Science and Engineering and the Bernoulli Institute at the University of Groningen for providing the facilities and a welcoming working environment, and the administrative staff for the many small kindnesses that kept daily life running smoothly.

Finally, to my parents: thank you for believing in me long before I had reason to believe in myself. Whatever merit this work may have, it belongs as much to you as it does to me.

SHORT RÉSUMÉ

Changtai Li

PERSONAL DATA

Name:	Changtai Li
Email:	lichangtai17@gmail.com
Orcid:	0000-0001-7985-9704

SCIENTIFIC EDUCATION

Current | 06/2025	PhD candidate in Computer Science (Joint-PhD program) University of Groningen, Groningen, the Netherlands Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence
09/2021	University of Science and Technology Beijing, Beijing, China School of Intelligence Science and Technology

07/2021 | 09/2017	Bachelor of Engineering in Computer Science University of Science and Technology Beijing, Beijing, China School of Computer and Communication Engineering

SELECTED PUBLICATIONS

C. Li†, Y. Zhang†, Y. Jia, Y. Guo, C. Yao*, X. Ban, and Y. He*. HRTEM-GAN: Structure-preserving restoration of low-quality atomic-scale HRTEM images. *Nano Research*. SciOpen, 2026. doi: 10.26599/NR.2026.94908715

C. Li†, X. Han†, C. Yao, Y. Guo, Z. Li, L. Jiang, W. Liu, H. Huang, H. Fu*, and X. Ban*. A novel training-free approach to efficiently extracting material microstructures via visual large model. *Acta Materialia*, Volume 290, 120962. Elsevier, 2025. doi: 10.1016/j.actamat.2025.120962

C. Li†, R. Jiang†, H. Wang, W. Xue, Y. Guo, and X. Ban*. DeepMMP: Efficient 3D perception of microstructures from serial section microscopic images. *Computational Materials Science*, Volume 235, 112826. Elsevier, 2024. doi: 10.1016/j.commatsci.2024.112826

R. Jiang†, **C. Li**†, X. Ban*, S. Yin, C. Yao, Y. Guo, and M. S. Obaidat. From non-expert to expert: Recurrent refined learning for medical image segmentation. *In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2086–2093. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10821757

C. Li†, R. Jiang†, S. Yin, J. Yang, and X. Ban*. Self-supervised rotation learning for 3D segmentation on nasopharyngeal carcinoma MRI images. *In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3529–3534. IEEE, 2023. doi: 10.1109/BIBM58861.2023.10385483

B. Zhang, Q. Zhu, C. Xu, **C. Li**, Y. Ma, Z. Ma, S. Liu, R. Shao, Y. Xu, B. Jiang, L. Gao, X. Pang, Y. He*, G. Chen*, and L. Qiao*. Atomic-scale insights on hydrogen trapping and exclusion at incoherent interfaces of nanoprecipitates in martensitic steels. *Nature Communications*, Volume 13, 3858. Nature Publishing Group, 2022. doi: 10.1038/s41467-022-31665-x

† Equal contribution
* Corresponding author

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Final Version as of May 22, 2026 (`classicthesis`).