2026-04

A Baseline Analysis of Cross-Modal Liver Tumor Segmentation and the Role of Frozen Encoder

Iqbal, Md. Zafor

IUB

A Baseline Analysis of Cross-Modal Liver Tumor Segmentation and the Role of Frozen Encoder

April 2026

Prepared by:

Md. Zafor Iqbal

ID: 2111495

Department of Computer Science and Engineering

Independent University, Bangladesh

Supervised by:

Dr. Md Rashedur Rahman

Assistant Professor

Department of Computer Science and Engineering

Independent University, Bangladesh

In Partial Fulfillment of the Requirements for the Degree of Bachelor's of Computer Science & Engineering

Attestation

I are aware of the fact that plagiarism is strictly prohibited and it conflicts with my university's rules and regulations. I guarantee the authenticity of my work. I have used some copyrighted material and models of others in my project which has been properly cited following the international standards proper guidelines.

Author Name:

. . . . . .

Signature:

. . . . . .

Author Name:

. . . . . .

Signature:

. . . . . .

Author Name:

. . . . . .

Signature:

. . . . . .

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

I express my sincere gratitude to my supervisor, Dr. Md Rashedur Rahman, for his invaluable guidance throughout this project. I also thank Dr. Saadia Binte Alam and Dr. Ashraful Islam for their thoughtful feedback and guidance, which significantly improved the quality and rigor of my work.

I am grateful to the Center for Computational and Data Sciences (CCDS) & Independent University, Bangladesh, for providing computational resources that were essential to my research.

I would also like to acknowledge the technical support provided by Mr. Siam Tahsin Bhuiyan, Mr. Sefatul Wasi, Mr. Fatin Israq, Mr. Riyadul Islam, and Ms. Halima Khatun, whose contributions helped carry this project forward.

Finally, I used Large Language Models (LLM), specifically Claude Sonnet 4.5, Claude Opus 4.6, and Claude Haiku 4.5 from Anthropic, and Gemini 3 Pro from Google, to improve the clarity and language of my writing. These tools were used only for editorial purposes, and I confirm that the method, hypothesis, data analysis, result generation, and figures were produced entirely by the author.

List of Publications

M. Z. Iqbal, S. T. Bhuiyan, R. Islam, H. Khatun, S. K. Mazumder, R. Rahman, A. Islam and S. B. Alam, "A Two-Stage Deep Learning Framework for Liver and Hepatocellular Carcinoma Segmentation on MRI," *2025 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, Eastern University, Dhaka, Bangladesh, Nov. 29-30, 2025. (Presented)

Abstract

Accurate liver tumor segmentation is critical for surgical planning, volumetry, and treatment monitoring across diverse liver pathologies and clinical applications. Both Computed tomography (CT) and Magnetic resonance imaging (MRI) serve as essential modalities in clinical practice, yet segmentation models trained on one modality typically fail when applied to the other. This limitation reflects a fundamental gap in the understanding of why complex architectural innovations are necessary for cross-modal robustness. Most literature proposes sophisticated solutions without first establishing what simpler methods achieve or where they encounter irreducible obstacles.

This thesis provides a systematic baseline analysis of cross-modal liver and tumor segmentation, deliberately adopting simple approaches to characterize their capabilities and limitations. The study employs a frozen ResNet18 encoder combined with a U-Net decoder within a two-stage pipeline that first segments the liver, then detects lesions within the hepatic region. This straightforward architecture is evaluated across five public datasets spanning both modalities: LiverHCCSeg and CHAOS for MRI, and LiTS, 3D-IRCADb-01, and SLiver07 for CT.

Results reveal a clear performance hierarchy driven by training data properties. Within-modality transfer succeeds remarkably well when training data is sufficiently diverse. For liver segmentation, the model trained on LiTS generalized excellently to other CT benchmarks, achieving Dice Similarity Coefficient (DSC) of 0.976 on 3D-IRCADb-01 and 0.951 on SLiver07. In cross-modal evaluation, performance drops sharply. The LiTS-trained liver model reaches only DSC 0.196 on CHAOS MRI, and in tumor segmentation the contrast is even larger. Under the selected -20 to 400 Hounsfield Unit (HU) window, the LiTS tumor model achieved DSC 0.433 on the in-domain LiTS test set and DSC 0.467–0.470 on cross-domain CT-to-CT evaluation on 3D-IRCADb-01, but only DSC 0.090 on LiverHCCSeg MRI. This gap persisted even when ground-truth liver masks were used for Region of Interest (ROI) extraction, indicating that the primary limitation is feature mismatch in the frozen encoder rather than ROI quality or basic preprocessing.

The fundamental barrier is that CT and MRI rely on entirely different physical principles. CT measures X-ray attenuation, expressed in HU. MRI measures radiofrequency signal decay through T1 and T2 relaxation times. When processing signals from an unfamiliar modality, the encoder produces activations that follow patterns learned from natural images, fundamentally misaligned with medical tissue characterization.

These findings establish that simple baseline methods are effective within a single modality when training data is appropriately diverse. They also clarify precisely where these methods encounter fundamental limitations. By quantifying this boundary between tractable and intractable problems, this work provides the empirical foundation necessary to justify and guide future research into encoder adaptation and cross-modal architectural innovations.

Contents

List of Figures

List of Tables

Chapter-1: Introduction

Liver cancer remains one of the most lethal malignancies worldwide, with its incidence driven by a combination of chronic risk factors that have persisted for decades. A comprehensive analysis of the global cancer burden estimated approximately 906,000 new liver cancer cases and 830,000 deaths in 2020 alone, positioning it as the sixth most commonly diagnosed cancer and the third leading cause of cancer mortality globally [1]. These figures are not static. Projections indicate that by 2040, incidence may rise to over 1.4 million new cases annually, driven primarily by population growth and aging, as well as persistent exposure to hepatitis B virus (HBV) and hepatitis C virus (HCV), metabolic dysfunction-associated liver disease, and alcohol-related liver disease [1]. This trend indicates a deepening global health crisis that demands both preventive and diagnostic innovation.

Central to this crisis is the progressive deterioration of the liver through fibrosis and cirrhosis, which constitutes a dominant pathway to liver cancer. A global burden of disease analysis spanning 1990 to 2019 across 204 countries and territories documented trends in cirrhosis incidence and mortality and produced forecasts extending two decades into the future [2]. The study found that the age-standardized incidence rate of cirrhosis declined only marginally, from 25.7 per 100,000 in 1990 to 25.3 per 100,000 in 2019, yet the absolute number of cases continued to rise, driven primarily by population growth and aging rather than any true reduction in disease burden. Alcohol use emerged as the single largest contributor to cirrhosis-related disability-adjusted life years (DALY), accounting for 49.3% of DALYs and 48.4% of global deaths [2]. While developing regions carry a considerable burden due to endemic hepatitis B infection and limited screening infrastructure, the leading cause of cirrhosis globally shifted from hepatitis B to hepatitis C between 1990 and 2019 [2]. Projections from both Nordpred and BAPC models indicate that the absolute number of cirrhosis cases will continue rising through 2039, reinforcing the need to intercept the disease earlier in its course and to strengthen prevention efforts around the most common modifiable risk factors.

Among all primary liver malignancies, hepatocellular carcinoma (HCC) is by far the most prevalent, accounting for approximately 75–85% of cases, followed by intrahepatic cholangiocarcinoma. A detailed epidemiological review drawing on data from GLOBOCAN and national cancer registries elucidated the regional and etiological heterogeneity of liver cancer incidence and mortality [3]. In East Asia and Sub-Saharan Africa, chronic HBV infection remains the dominant risk factor; in North Africa and the Middle East, HCV predominates; while in Western Europe and North America, alcohol-related liver disease and the rapidly expanding epidemic of non-alcoholic fatty liver disease (NAFLD) are reshaping the etiological distribution [3]. Prevention strategies, including HBV vaccination, antiviral therapy, and alcohol reduction programs, have demonstrably reduced incidence

in some high-income settings, yet the overall global burden continues to climb.

The temporal trends in HCC epidemiology reveal a shifting landscape of disease. An analysis of global trends in HCC epidemiology, drawing on data from the global burden of disease study and GLOBOCAN registry, found that while age-standardized incidence rates have declined in historically high-burden regions such as East Asia following widespread HBV vaccination programs and agricultural policy reforms to reduce aflatoxin exposure, they continue to rise in Western nations, a divergence attributable to increasing metabolic dysfunction-associated steatotic liver disease (MASLD) and alcohol-related liver disease prevalence [4]. This review emphasized that the implications for screening, prevention, and therapeutic strategy differ substantially between regions and called for tailored public health approaches. Critically, most patients with HCC are still diagnosed at advanced stages of disease, largely owing to the suboptimal sensitivity of current surveillance tools and their underuse in clinical practice, with fewer than one in four at-risk patients undergoing surveillance globally [4]. This reality directly motivates the need for improved surveillance and early detection technologies, including emerging blood-based biomarker panels and abbreviated MRI protocols that are currently under prospective evaluation [4].

The biological mechanism linking chronic liver injury to malignant transformation is fibrosis, a process of pathological wound healing that drives disease progression across virtually all etiologies. A mechanistic review of fibrogenesis and its relationship to hepatocarcinogenesis detailed how iterative hepatic injury from viral, metabolic, or toxic sources activates hepatic stellate cells (HSCs), triggering their transdifferentiation into myofibroblasts under the influence of transforming growth factor-beta (TGF-β) and platelet-derived growth factor (PDGF) [5]. These activated HSCs deposit excessive extracellular matrix (ECM) components including collagen type I and III [5], progressively disrupting the liver parenchyma and replacing functional hepatocytes. Beyond structural disruption, the pro-fibrotic microenvironment promotes oncogenic signaling, as activated HSCs secrete cytokines and growth factors including hepatocyte growth factor (HGF), TGF-β, PDGF, interleukin-6 (IL-6), and Wnt ligands that act directly on tumor cells [5], while also suppressing adaptive immune responses through expression of PDL-1 and B7-H4 that induce T-cell exhaustion or anergy [5]. Additionally, HSCs promote an immunosuppressive environment by expanding regulatory T cells, collectively creating a microenvironment conducive to the survival and proliferation of pre-malignant cells. Consistent with this, fibrosis stage has been established as the single strongest independent predictor of liver-related mortality and time to development of severe liver disease in chronic liver disease patients, with up to 90% of HCC cases arising on the background of a cirrhotic liver.

The severity of fibrosis is not merely a histological curiosity but a direct determinant of patient mortality, as quantitatively demonstrated in a systematic review and meta-analysis of NAFLD outcomes [6]. By pooling data from five multinational cohort studies encompassing 1,495 NAFLD patients with biopsy-proven fibrosis staging and 17,452 patient-years of follow-up, this analysis established a monotonic relationship between fibrosis severity and both all-cause and liver-specific mortality. Using fibrosis stage 0 as the reference population, the pooled all-cause mortality rate ratios (MRR) increased progressively: stage 1 (MRR 1.58, 95% confidence interval [CI] 1.19–2.11), stage 2 (MRR 2.52, 95% CI 1.85–3.42), stage 3 (MRR 3.48, 95% CI 2.51–4.83), and stage 4 (MRR 6.40, 95% CI 4.11–9.95). The pattern was even more pronounced for liver-related mortality, which increased exponentially rather than linearly across stages, with stage 4 fibrosis associated with an MRR of 42.30 (95% CI 3.51–510.34) compared to stage 0. Notably, while

elevated all-cause mortality risk was detectable even at stage 1, a statistically significant increase in liver-related mortality was only observed from stage 2 onward. Given that NAFLD affects nearly 100 million individuals in the United States alone, these findings carry major public health implications: fibrosis stage becomes a critical prognostic and management decision point, yet liver biopsy, the traditional reference standard for fibrosis assessment, remains invasive, costly, and subject to sampling variability.

Even where fibrosis has progressed to cirrhosis and established liver cancer, the tumor does not evolve in isolation. The hepatic tumor microenvironment (TME) plays an active and determining role in HCC progression, immune evasion, and therapeutic resistance. A comprehensive review of tumor-associated macrophages (TAMs) in liver cancer delineated how the hepatic immune landscape (dominated by resident Kupffer cells and infiltrating monocyte-derived macrophages) undergoes functional reprogramming toward an immunosuppressive phenotype in the presence of tumor-derived signals [7]. These TAMs, once polarized toward the alternatively activated M2 phenotype, secrete anti-inflammatory cytokines including interleukin-10 (IL-10) and TGF-β, promote angiogenesis via vascular endothelial growth factor (VEGF) and Tie2 receptor expression, enhance cancer stem cell (CSC) maintenance, and directly contribute to resistance against sorafenib and other systemic therapies [7]. Therapeutic strategies targeting TAM recruitment via CCL2/CCR2 blockade, colony-stimulating factor 1 receptor (CSF1R) blockade, or repolarization toward the tumoricidal M1 phenotype using PI3Kγ inhibitors and toll-like receptor agonists represent an emerging front in HCC treatment research.

The cumulative consequence of this biological cascade, from chronic injury through fibrosis and immune evasion, is a disease that is deeply challenging to treat. There are several barriers to effective clinical management of liver cancer, including the high prevalence of underlying cirrhosis which limits surgical candidacy, given that only 5% to 15% of patients are eligible for surgical removal, and that diminished hepatic regenerative capacity further restricts resection options [8]. Molecular heterogeneity across HCC subtypes undermines uniform therapeutic targeting, while multi-drug resistance, mediated in part by cancer stem cells (CSCs) and long non-coding RNA-driven protective autophagy, contributes to treatment failure. First-generation systemic therapies such as sorafenib, currently the standard first-line oral agent for advanced HCC, confer relatively modest survival benefits, extending median overall survival by only 3 to 5 months compared to placebo [8], with resistance typically emerging within six months of initiating the regimen. While locoregional therapies including transarterial chemoembolization (TACE) offer some disease control in intermediate stages, producing approximately a 23% improvement in 2-year survival over conservative therapy [8], their efficacy remains limited in advanced disease. Taken together, these clinical realities underscore the need for more precise tumor characterization and assessment, functions that fall within the domain of medical imaging and computational analysis.

1.1 Background

The accurate determination of liver fibrosis stage is a prerequisite for risk stratification, treatment planning, and surveillance scheduling in patients with chronic liver disease. Liver biopsy has long served as the reference standard for fibrosis staging, enabling

histological assessment using validated scoring systems such as METAVIR and Ishak, which grade fibrosis from F0 (no fibrosis) through F4 (cirrhosis) [9]. However, biopsy carries well-documented limitations: its invasive nature introduces procedure-related risks including pain, which occurs in approximately 20% of patients, and bleeding, which occurs in roughly 0.5% of cases. Furthermore, because a biopsy samples only approximately 1/50,000 of the total liver volume, it is inherently prone to sampling error, and histological interpretation is subject to both intraobserver and interobserver variability [9]. These limitations have driven intensive research into non-invasive fibrosis assessment modalities, including imaging-based approaches such as transient elastography and MR elastography, as well as laboratory-based scoring systems and combination biomarker panels, which can together eliminate the need for liver biopsy in a substantial proportion of patients [9].

The non-invasive characterization of liver cirrhosis has advanced substantially through both qualitative and quantitative imaging biomarkers. Qualitative imaging features, including liver surface nodularity, parenchymal heterogeneity, caudate lobe hypertrophy, splenomegaly, and portal hypertension signs such as varices and ascites, are generally highly specific but only moderately sensitive for cirrhosis diagnosis [10]. Quantitative approaches include the caudate-to-right lobe ratio (CRL) and liver segmental volume ratio (LSVR) derived from CT or MRI, as well as elastographic techniques such as transient elastography (TE), point shear wave elastography (pSWE), two-dimensional shear wave elastography (2D-SWE), and magnetic resonance elastography (MRE), all of which measure tissue stiffness as a surrogate for fibrosis burden. A detailed appraisal of these qualitative and quantitative imaging biomarkers reviewed their diagnostic performance and clinical applicability across ultrasound, CT, and MRI platforms. Among elastographic methods, MRE is considered the most accurate non-invasive modality, achieving area under the curve (AUC) values of 0.90 to 0.99 for cirrhosis diagnosis, while TE, the most widely available technique, achieves AUC values ranging from 0.89 to 0.94. Liver surface nodularity (LSN), measured as the mean distance between the detected liver margin and a simulated smooth liver margin, has emerged as the single most accurate and reproducible quantitative morphological biomarker, with AUC values of 0.90 to 0.96. The integration of these non-invasive assessments into clinical workflows, potentially within multistep diagnostic algorithms combining serum biomarkers and imaging parameters, may reduce reliance on liver biopsy while maintaining high diagnostic accuracy.

Once cirrhosis is established, the focus of clinical imaging shifts toward HCC surveillance and diagnosis. Ultrasound serves as the primary surveillance tool in cirrhotic patients, though its sensitivity for lesions smaller than 2 cm can be as low as 65%, with additional limitations stemming from operator dependence and altered liver echogenicity in conditions such as non-alcoholic steatohepatitis [11]. The six-monthly surveillance interval is recommended by all major guidelines, including American, European, and Asian-Pacific societies. Contrast-enhanced ultrasound, multiphase contrast-enhanced CT, and gadolinium ethoxybenzyl-diethylenetriamine pentaacetate (Gd-EOB-DTPA) enhanced MRI with diffusion-weighted imaging provide superior lesion characterization by exploiting the hallmark vascular profile of HCC, which consists of arterial phase hyperenhancement followed by washout appearance in the portal venous or delayed phases. MRI has demonstrated a pooled sensitivity of approximately 70% and specificity of 94% for HCC diagnosis regardless of tumor size, though sensitivity approaches 100% for lesions greater than 2 cm and drops to 58–65% for lesions smaller than 2 cm. The use of hepatospecific contrast agents further increases sensitivity by 5–10%. Standardized reporting frameworks, including the Liver Imaging Reporting and Data System (LI-RADS), have been developed

to reduce inter-reader variability and support algorithm-based management decisions. The diagnostic landscape has been further expanded by CT and MRI perfusion techniques, which provide quantitative functional parameters beyond morphology, as well as positron emission tomography/CT and emerging Artificial intelligence (AI) approaches that show promise for lesion characterization, grading, and treatment response assessment.

The translation of these imaging capabilities into clinical practice is guided and standardized by national and international guidelines. China's National Health Commission (NHC) guidelines for the diagnosis and treatment of primary liver cancer, updated in their 2022 edition, represent one of the most comprehensive integrated frameworks for liver cancer management and incorporate evidence from both Asian and global clinical trial data [12]. The guidelines specify that surveillance for at-risk populations should combine abdominal ultrasound with serum alpha-fetoprotein (AFP) testing at six-monthly intervals and recommend multiphasic MRI (mpMRI) as the preferred diagnostic modality given its superior soft-tissue contrast and greater accuracy in detecting and staging HCC [12]. Staging follows the China Liver Cancer (CNLC) classification, stratifying patients from stage Ia through IV based on tumor burden, vascular invasion, liver function as assessed by Child-Pugh class, and Eastern Cooperative Oncology Group (ECOG) performance status. For patients with AFP-negative disease, the GALAD model, which integrates sex, age, AFP-L3, AFP, and des-gamma-carboxyprothrombin (PIVKA-II), achieves a reported diagnostic sensitivity of 85.6% and specificity of 93.3% for early-stage HCC [12]. First-line systemic therapy for advanced HCC includes the atezolizumab plus bevacizumab combination, which demonstrated a 34% reduction in the risk of death and a 35% reduction in the risk of disease progression compared with sorafenib in the pivotal IMbrave150 trial, and is among the recommended regimens in the guidelines for patients with unresectable disease who have not received prior systemic therapy [12].

Against the clinical backdrop of this complex diagnostic and staging landscape, Deep learning (DL) has emerged as a transformative approach for automated medical image analysis. A review by Cai et al. surveyed the methodological progression from early Convolutional neural network (CNN)s (CNNs) through to architectures specifically designed for dense pixel-level prediction tasks [13]. The encoder-decoder architecture of U-Net, with its skip connections that preserve spatial information across scales, became particularly influential for organ and lesion segmentation in MRI and other imaging modalities [13]. Classification networks including AlexNet, VGGNet, and ResNet demonstrated strong performance across a range of medical imaging tasks, from retinal fundus grading to pulmonary nodule detection [13]. For segmentation specifically, the review highlighted that while FCN served as the foundational framework, U-Net's concatenation-based skip connections offered more refined feature recovery during upsampling compared to the direct addition strategy used in FCN [13]. Across application domains, DL models were shown to achieve competitive performance in tasks such as hippocampus segmentation for Alzheimer's disease diagnosis and left ventricular segmentation from cardiac MRI, with the authors' own GNNI U-Net model reporting DSC values of 0.937 and 0.957 on the Sunnybrook and LVSC datasets, respectively [13].

The architectural repertoire for AI-based medical image analysis has expanded substantially beyond convolutional models. A comprehensive survey of AI applications in medical imaging documents the rise of vision transformers (ViT) and hierarchical Shifted Window (SWIN) transformers, which employ multi-head self-attention (MHSA) mechanisms and shifted window partitioning to capture long-range spatial dependencies that CNNs

inherently struggle to model [14]. In the ViT architecture, an input image is divided into a sequence of fixed-size patches, each associated with a positional encoding, and these patches are passed through transformer encoding layers before a multilayer perceptron (MLP) head performs classification. The SWIN transformer extends this by adopting a hierarchical structure in which the image is first divided into smaller patches that are progressively merged as the network deepens, with self-attention computed within local shifted windows rather than across the entire image, reducing computational complexity while preserving both local and global feature representations. Hybrid approaches such as TransUNet combine convolution-based feature extraction with transformer-based global context modeling, and have been reported to outperform pure CNN baselines on organ segmentation tasks. Generative architectures including Generative adversarial network (GAN)s (GANs), implemented in medical variants such as MedGAN for cross-modal image translation across tasks including PET-CT translation and MRI motion artifact correction, variational autoencoders (VAEs), and diffusion models, which have been shown to outperform GANs on certain image synthesis benchmarks, have extended the reach of AI to synthetic data generation, cross-modal image translation, and image super-resolution [14]. Applications in liver disease include CNN ensemble methods integrating architectures such as ResNet152, ResNeXt101, DenseNet201, and InceptionV3 for liver and lesion segmentation on the LiTS benchmark, where the combined system outperformed each individual network. Explainability tools such as gradient-weighted class activation mapping (Grad-CAM) have been applied to make model predictions interpretable to clinicians, for example through the NDG-CAM method proposed for nuclei detection in histopathology images, and this interpretability is a desired characteristic for regulatory acceptance.

The promise of AI in medical imaging, however, must be balanced against ethical imperatives and responsible deployment practices. A contemporary review examining the intersection of AI, ethics, and clinical impact in medical imaging identified algorithmic bias as a pervasive and underappreciated risk, particularly when training data lack demographic diversity or fail to represent the full phenotypic spectrum of disease encountered in clinical practice [15]. The review further addressed challenges of model generalizability across imaging hardware, acquisition protocols, and institutional workflows, and emphasized that patient safety, data governance, explainability requirements, and robust clinical validation must be embedded into AI development pipelines from inception rather than retrofitted post-deployment [15]. These considerations underscore that methodological rigor and ethical accountability are not peripheral to AI-based medical imaging research but constitute foundational requirements.

A domain of particular clinical relevance in which AI and medical imaging directly converge is liver segmentation. The delineation of the liver from surrounding abdominal structures in CT or MRI is a prerequisite for a wide range of surgical and interventional procedures, including pre-operative volume estimation for donors and recipients in living-donor liver transplantation, surgical resection planning, and thermal ablation targeting [16]. Accurate segmentation is especially important in the context of HCC, which the World Health Organization has identified as the leading cause of cancer deaths worldwide, accounting for approximately 830,000 deaths in 2020 alone [16]. A systematic appraisal of the practical clinical utility of liver segmentation methods documented that automated segmentation algorithms reduce planning time and improve volumetric accuracy compared to manual delineation, which is time-consuming and subject to inter-observer variability [16]. For instance, automated hepatic volumetry has been shown to complete liver volume estimation in approximately 4.4 minutes compared to 32.8 minutes for manual tracing, while achieving

comparable volumetric agreement [16]. Segmentation also enables risk assessment through future liver remnant (FLR) calculations, which are critical for procedures such as portal vein embolization (PVE). In the context of oncology, precise segmentation facilitates not only treatment planning but also longitudinal tumor burden monitoring in response to systemic or locoregional therapy, including TACE and stereotactic body radiation therapy (SBRT) [16].

The growing sophistication of AI-driven liver analysis extends to a wide array of diagnostic applications beyond segmentation alone. A recent review surveyed AI tools deployed across the clinical spectrum, from fibrosis staging and steatosis quantification on imaging to histological classification of liver biopsies and treatment response prediction in HCC and cholangiocarcinoma [17]. Approaches integrating CT, MRI, ultrasound, and serum biomarkers within DL or ensemble frameworks have demonstrated strong diagnostic performance across multiple tasks, with AUC values commonly ranging from 0.84 to 0.98 depending on the modality and target condition [17]. Nevertheless, the review identified a persistent translational gap between experimental validation in retrospective datasets and prospective clinical deployment, citing insufficient multi-center validation, the reliance on single-institution cohorts, and the challenge of ensuring that AI outputs are generalizable across patient populations and imaging protocols [17]. The authors further noted that while regulatory and legislative progress has been made, fully integrating AI into routine clinical workflows remains an open challenge.

1.2 Motivation

The clinical management of HCC fundamentally depends on accurate identification and delineation of the liver and its tumors. As detailed in the preceding sections, both multiphase CT and contrast-enhanced MRI serve as essential modalities for HCC characterization and surveillance. However, the clinical reality presents a significant challenge: different imaging modalities are used in different regions and healthcare settings based on equipment availability, institutional preferences, and patient factors. In some centers, CT is the primary imaging tool; in others, MRI is preferred or mandated by local guidelines. A robust segmentation system deployed in clinical practice must therefore perform reliably across both imaging technologies to be universally applicable.

Currently, most automated segmentation methods are developed and validated on single-modality datasets. A DL model trained exclusively on CT data often shows degraded performance when applied to MRI, and vice versa. This single-modality limitation creates a significant translational barrier: algorithms that perform excellently in controlled research settings may fail when deployed in clinical environments where imaging protocols are not standardized. The presence of image artifacts, variations in acquisition parameters, differences in tissue contrast between institutions, and the distinct visual appearance of the same anatomy across modalities all contribute to this performance gap. As DL becomes increasingly central to medical imaging workflows, the need for cross-modal robustness has emerged as a prerequisite for clinical adoption.

This thesis addresses this gap by developing a cross-modal liver and tumor segmentation pipeline capable of robust performance on both CT and MRI data. The motivation

is fundamentally practical: to create segmentation tools that generalize reliably across multiple imaging modalities, thereby supporting deployment across diverse clinical settings and enabling broader clinical impact.

1.3 Objective

The fundamental challenge addressed in this thesis is understanding how domain shift between medical imaging modalities constrains DL segmentation when architectural complexity is deliberately held constant. This investigation uses a fixed-architecture experimental framework where a frozen ResNet18 encoder combined with a trainable U-Net decoder enables systematic examination of which obstacles can be overcome through improved training design and preprocessing versus which reflect irreducible misalignment between CT and MRI imaging.

This thesis pursues three complementary research objectives:

- The first objective is to establish quantitative baseline performance across both CT and MRI modalities using cross-modal evaluation configurations.
- The second objective is to systematically evaluate whether preprocessing and training design choices can reduce cross-modal performance degradation.
- The third objective is to analyze failure patterns to identify whether specific constraints require architectural modification or can be addressed through tractable improvements.

Together, these objectives provide a complete empirical characterization of what simple baseline methods achieve and where they encounter fundamental limitations.

Chapter-2: Literature Review

2.1 Prior Studies

2.1.1 Liver Segmentation

Accurate segmentation of the liver from medical images is a fundamental requirement for computer-aided diagnosis, surgical planning, and treatment monitoring. Research in this area has progressed through four distinct methodological paradigms. These are manual and semi-automatic approaches, traditional image processing techniques, classical Machine learning (ML) methods, and DL frameworks. The following subsections survey representative works from each era, tracing the evolution in accuracy, automation, and clinical applicability.

2.1.1.1 Manual & Semi-Automatic Methods

Early clinical liver segmentation relied entirely on expert radiologists manually tracing organ boundaries on each CT or MRI slice, a process that could consume up to two hours per patient. Whether semi-automatic algorithms could reduce this burden while maintaining acceptable accuracy for living donor liver transplantation was investigated through a segmentation algorithm based on geometric deformable models within a level-set framework [18]. The algorithm was augmented with an accumulative speed function to prevent contour leakage at weak boundaries, requiring users to place only a small number of initialization circles per slice before automatic propagation. This semi-automatic approach reduced mean user interaction time from 25 minutes to just 5 minutes and improved graft volume estimation accuracy in 15 of 18 cases, with interobserver repeatability improving for all major liver segments except the caudate lobe, which still required fully manual delineation [18]. The broader clinical context of liver segmentation across all major imaging modalities and use cases was systematically reviewed shortly thereafter, cataloguing manual, semi-automatic, and automatic approaches from interactive region growing and deformable contour models to early atlas-based and ML methods [19]. A critical finding from this survey was that while automatic methods performed within the inter-observer variability range for healthy livers, no single purely automatic approach had achieved consistent clinical-grade accuracy across all pathological presentations at that time, motivating continued development of robust semi-automatic pipelines [19]. Building on this need for reliable semi-automatic tools, the practical performance of two widely used semi-automatic ITK-based algorithms for preoperative 3D liver visualization was evaluated on 24 clinical

CTA datasets [20]. The pipeline integrated an ITK processing and segmentation stage with a VTK 3D rendering stage, benchmarking the Connected Threshold region-growing algorithm against the Fast Marching level-set approach. The Fast Marching method proved significantly more robust, achieving a maximum volume deviation of 45.18% compared to 87.52% for the Connected Threshold method, and consistently produced closed, smooth 3D surfaces suitable for surgical visualization [20]. However, this method systematically excluded contrast-filled intrahepatic vessels from the parenchyma, leading to volume underestimations and highlighting the need for post-processing vessel-inclusion steps [20].

The challenge of segmenting individual liver tumors, considerably harder than whole-liver delineation due to low contrast, vague boundaries, and wide variability in tumor intensity, shape, and size, led to a semi-automatic method requiring only a single user-supplied seed point per tumor [21]. From this seed, an adaptive region growing algorithm iteratively expanded the segmentation region with a dynamically updating merging criterion, enforcing a Kullback-Leibler divergence stopping condition to prevent over-segmentation while enhancing tumor contrast through Gaussian-fitting-based nonlinear intensity mapping. Graph cuts optimization then refined the boundary using a combined energy function. Evaluated on the 3Dircadb public dataset, the method achieved a DSC of 0.85 ± 0.05, outperforming several fully automatic DL networks including U-Net and Deeplabv3+ particularly for small, low-contrast tumors [21]. Meanwhile, a parallel five-stage approach grounded in probabilistic tissue modeling addressed whole-liver segmentation from a single user-defined seed point by building a subject-specific multivariable normal distribution model using local mean and standard deviation of pixel neighborhoods [22]. This model was subsequently refined by relaxation labelling to incorporate spatial context, followed by a graph-cut algorithm minimizing combined region-and-boundary energy and a novel bottleneck detection step with adjacent cross-slice contour constraints that intelligently removed false positives without discarding genuine anatomical features such as the left lobe. Validated on the MICCAI SLiver07 dataset, the system achieved a mean overall score of 72.3 for asymptomatic livers and completed full 3D segmentation in approximately 1.3 minutes per scan, demonstrating that careful probabilistic formulation can approach the accuracy of far more complex automated models with minimal user interaction [22].

2.1.1.2 Traditional Image Processing

As imaging modalities and computational resources improved, researchers pursued fully automatic segmentation methods based on classical image processing operations such as thresholding, morphological filtering, active contours, and region growing. These approaches eliminated user interaction while relying on hand-engineered rules and features. One of the early automatic CT liver segmentation pipelines for clinical volume measurement began with multilevel HU thresholding based on prior anatomical knowledge to restrict the liver search region [23]. Multiscale morphological filtering with flat structuring elements removed scattered non-liver noise, and a modified k-means algorithm with three classes refined the candidate region. A deformable contour traced the liver boundary on a morphological gradient-label map by evaluating eight-directional candidate pixels with a local cost function weighted across gradient direction, initial boundary position, and intensity pattern. The method achieved a mean segmentation correctness of approximately 96% against manual tracings with volume measurement errors within $\pm$ 3.7% [23]. However, its reliance on hardcoded thresholds and empirical weights limited robustness across patients

with atypical anatomy or severe pathology.

The region-growing paradigm was extended to volumetric texture-based features through the observation that pure intensity thresholding fails to distinguish the liver from adjacent organs with similar HU values [24]. An algorithm automatically selected a seed voxel by minimizing an objective function balancing Euclidean distance to the region centroid and pixel gradient magnitude, describing each voxel with a 91-dimensional Haralick feature vector computed from Gray Level Co-occurrence Matrices across a $7 \times 7 \times 7$ sub-volume and 13 three-dimensional directions. The growing threshold was automatically determined by fitting a Gaussian mixture model via expectation-maximization on 20,000 sampled points. Tested on five healthy volunteer scans, the method correctly identified the liver and excluded hepatic vessels with distinct local textures [24]. However, it suffered from boundary under-segmentation because fixed-size windows mixed liver and non-liver features near the organ's edge.

The challenge of ultrasound liver tumor segmentation, where speckle noise and echogenicity variations make boundary detection unreliable, was addressed through a pre-processing pipeline that transformed images into the neutrosophic domain and applied Non-subsampled Shearlet Transform despeckling [25]. Multi-scale multi-orientation texture features were extracted using Gabor filters, and a novel adaptive Otsu thresholding method constrained by independently computed intensity, Gabor texture, Fourier-reconstructed phase, and phase gradient thresholds initialized an active contour mask. A localized region-based active contour then refined the boundary using local mean energies, outperforming global Chan-Vese segmentation by approximately 10% in accuracy with a 34.02% DSC improvement over standard Otsu-based initialization alone [25]. Meanwhile, a Gradient Vector Flow snake model was applied to automatic CT liver segmentation to address conventional snakes' known failure modes of sensitivity to initial contour placement and leakage at anatomical concavities [26]. GVF forces were computed by diffusing edge gradient information across the image plane through partial differential equations controlled by a regularization parameter, allowing the contour to converge from distant initializations. A liver template bounding ellipse automatically initialized the snake, and a curvature-analysis concavity removal step prevented contour penetration into the gallbladder fossa, with refined contours automatically propagated between slices exploiting inter-slice continuity. Evaluating 20 CT volumes with 551 axial slices, the method achieved a median segmentation difference ratio of 5.3%, though performance degraded at the hepatic dome and inferior tip where slice-to-slice shape varied rapidly [26].

A method exploiting complementary contrast profiles of multi-phase CT acquisitions was introduced to simultaneously improve seed-finding robustness and suppress over-segmentation [27]. A joint histogram across unregistered arterial and venous phase images automatically located the liver's contrast uptake peak, and erosion and largest-component extraction produced reliable seed regions from which a Neighborhood-Connected Region-Growing algorithm expanded with a statistically derived intensity range. Sub-algorithms handled liver-heart separation using lung-segmentation-derived 3D coronal surfaces and portal-vein infilling via surface normal analysis. Multiple phase-specific segmentations were then affine-registered and intersected to suppress phase-specific over-segmentation, with resulting segmentations rated usable by radiologists in 94% of multi-phase cases [27]. Building on this multi-phase foundation, a fully automatic multi-phase liver segmentation algorithm addressed noise and intensity ambiguity through geometric diffusion filtering and reference-based adaptive thresholding [28]. The key novelty was an anatomy-guided

reference scheme that automatically identified a reference axial slice by tracking lung hole sizes from top to bottom, deriving intensity bounds from a 64×64 ROI at a reference point positioned midway between abdomen center and right lateral body surface. A 2.5D region-overlapping filter enforced longitudinal consistency by propagating the largest segmented region across adjacent slices, retaining only candidates that spatially overlapped with the reference region. Evaluated on 45 subjects, the method achieved a volumetric overlapping ratio of 87.7% and a volume correlation coefficient of 98.1% [28]. However, its geometric heuristics proved susceptible to errors in patients with severe left-lobe hypertrophy or major anatomical anomalies.

Progressive refinement of region-based approaches led to a two-stage pipeline combining coarse region growing with a refinement step using a novel signed pressure force level-set model [29]. Region growing from an automatically identified liver seed produced an initial binary mask initializing the zero-level-set. The SPF function applied a sigmoid transformation to local image intensity, bounding its magnitude and reversing its sign near the true boundary to cause the contour to expand inside the liver and contract outside in a self-regulating manner. This eliminated re-initialization steps required in traditional level-set methods and improved boundary precision at concave anatomical regions and near adjacent high-intensity vessels while reducing iterations required for convergence compared to both standalone region growing and conventional level-set approaches [29]. A graph-cuts-based approach then emerged that propagated shape and intensity constraints dynamically along the slice direction, eliminating the need for offline-trained probabilistic atlases [30]. After manual segmentation of a single high-diameter start slice, the algorithm iteratively derived hard foreground and background seeds through erosion and dilation, constructing an intensity histogram model from the previous liver region. A narrow-band bounding box restricted the search area around the previous mask, reducing computation. Tested on 10 highly pathological CT images, the method achieved a MICCAI-2007 score of 81.7 and required substantially less user interaction than statistical model approaches [30]. However, sequential constraint propagation remained sensitive to early mis-segmentation cascading.

2.1.1.3 Classical ML Methods

Classical ML approaches introduced learned classifiers and feature representations into liver segmentation, reducing dependence on hand-tuned rules and opening the door to data-driven generalizability. The sensitivity of Fuzzy C-Means clustering for liver CT segmentation to the number of clusters was examined through dynamic determination using silhouette values computed from fuzzy membership distances [31]. For each candidate cluster count, silhouette values compared how tightly clustered pixels were against how well-separated they were from other clusters. Interestingly, the numerically optimal silhouette score did not consistently correspond to the best anatomical liver boundary [31]. This highlighted that intensity-only clustering metrics cannot adequately capture anatomical separability when organs share similar density ranges, motivating researchers to integrate spatial constraints or texture features into clustering methods.

A combination of watershed pre-segmentation with SVM classification was employed to achieve an interactive liver tumor segmentation system that was orders of magnitude faster than voxel-wise classifiers [32]. After extracting the liver via a statistical shape

model and enhancing its histogram, a watershed transform partitioned the volume into homogeneous catchment basins. Feature vectors of four intensity statistics were extracted per basin rather than per voxel, and the user interactively labeled tumor and healthy tissue using a 3D brush, training an SVM with a linear kernel which was then applied to all unlabeled basins followed by morphological post-processing. Evaluated on the MICCAI 2008 dataset, the method achieved a Volumetric Overlap Error of 31.14% and reduced total runtime to approximately 30 seconds compared to 7 minutes for voxel-based SVMs, by classifying at the much coarser basin level [32]. A fully automatic pipeline built on wavelet-transform features and RBF-kernel SVM classification followed this same coarser-level approach [33]. Each pixel's feature vector was formed by concatenating its intensity with multi-scale, multi-orientation wavelet coefficients providing orientation-sensitive spatial frequency information. A coarse-to-fine grid search over the penalty parameter and kernel width identified the optimal hyperplane, and because pixel-wise SVM produced noisy scattered outputs, morphological erosion and dilation cleaned the SVM mask which then seeded a region-growing algorithm that snapped the final boundary to the true anatomical edge. The integration of wavelet features with SVM classification and region-growing refinement demonstrated effective liver delineation consistent with expert segmentations, though pixel-wise wavelet feature extraction across full 3D CT volumes incurred substantial computational cost [33].

A cascade classification pipeline combining Simple Linear Iterative Clustering super-pixels with two sequential AdaBoost classifiers trained on raw grey-level image patches was introduced [34]. SLIC partitioned CT images into compact, homogeneous super-pixel regions, and the first AdaBoost classifier trained from 100 decision-tree weak learners detected and removed the spinal cord via active contour refinement and image subtraction. The second AdaBoost classifier identified liver super-pixels, guided by a heuristic anatomical location mask and finalized by a region-growing algorithm. Evaluated on 16 CT images, the method achieved a DSC of 92.13%, a Jaccard Index of 85.8%, and a classification accuracy of 97.91%, demonstrating that a staged spinal cord removal step followed by focused liver classification can achieve strong segmentation results with simple patch-level features [34]. An approach converting raw CT intensity images into liver-likelihood probability images using an AdaBoost classifier trained on 12-dimensional GLCM texture features was proposed [35]. The AdaBoost output was sigmoid-transformed into a per-pixel liver probability map, and the random walks algorithm computed edge weights from both original CT intensity differences and probability differences. Crucially, foreground and background seeds were generated automatically from the probability map using adaptive thresholding and morphological erosion, removing the standard requirement for manual user input. Validated on the MICCAI 2007 Grand Challenge, the method attained a Volumetric Overlap Error of 8.76% and a MICCAI score of 72, outperforming several semi-automatic competitors despite operating fully automatically [35].

The random walk framework was substantially enriched with a multi-descriptor feature representation and a hybrid SVM-AdaBoost classifier [36]. For each pixel, four complementary feature families were computed: Local Binary Patterns, GLCM statistics, Haar-like filter responses, and Histograms of Oriented Gradients. Principal Component Analysis reduced the concatenated vector dimensionality, and AdaBoost with RBF-SVM weak learners generated the liver probability map with each SVM's scale parameter automatically decreasing each boosting round. The random walk edge weights were redesigned to jointly incorporate CT intensity differences and probability map differences with automatic foreground seeds derived from the histogram peak within the 125 to 155 HU range. Evaluated on the

MICCAI 2007 test set, the method achieved an accuracy of 95.18% and the smallest Mean Surface Distance among compared methods, notably generalizing to clinical cirrhosis volumes outside the training distribution and demonstrating robust cross-domain performance [36]. Liver segmentation in abdominal MRI presented unique low-contrast challenges where classical methods became unreliable, leading to an approach that overcame watershed over-segmentation through a multi-layer perceptron neural network feedback loop that iteratively tuned the watershed's scaling parameter [37]. Six MLP networks were trained to predict six geometric liver shape features (center of mass, perimeter, area, and axis ratios) from row-averaged intensity profiles. Starting from a fully over-segmented watershed output, the algorithm reduced its scaling parameter while computing a Multiplicative Squared Feature Error measuring the difference between current segmentation shape and MLP-predicted shape, stopping when error was minimized. Evaluated on 115 abdominal MRI images, the method achieved a Jaccard similarity coefficient of 0.94, surpassing a standard active contour at 0.92 and demonstrating that neural network-guided parameter adaptation substantially improves classical watershed robustness [37].

The feasibility of general-purpose statistical shape modeling methods including Point Distribution Models with PCA, Active Shape Models, and Constrained Local Models for achieving competitive liver segmentation without requiring large liver-specific training corpora was investigated [38]. The shape model was fitted to CT data by iteratively searching along profile normals for boundary candidates and projecting the resulting shape back onto the PCA shape subspace to enforce plausible deformations, initialized at expected liver location using anatomical priors. Benchmarked on the MICCAI 2007 Liver Segmentation Challenge, the general-purpose models achieved competitive DSC values and surface distances against liver-specific counterparts, demonstrating that domain-specific shape training is not strictly required for reasonable accuracy [38]. However, liver-specific fine-tuned models retained an advantage for complex pathological cases, suggesting a natural transition point where template-based approaches would give way to learned representations.

2.1.1.4 DL Methods

The advent of DL brought a fundamental shift in medical image segmentation as end-to-end learned representations replaced handcrafted feature pipelines, capturing hierarchical spatial context through deep convolutional architectures. A lightweight Gaussian-initialized CNN achieving competitive liver segmentation accuracy with significantly reduced computational complexity processed axial CT slices as overlapping 32×32 patches through three convolutional layers with 7×7, 5×5, and 3×3 filters, each followed by Local Response Normalisation, ReLU activation, and max-pooling [39]. Two fully connected layers and a softmax classifier labeled each patch as liver or background. A key contribution was the use of random Gaussian weight initialization, which preserved angle ratios among data classes and promoted faster convergence on low-dimensional patch data. Evaluated on the MICCAI SLiver07, 3Dircadb01, and LiTS17 datasets, the model achieved DSC values of 95.0%, 92.9%, and 97.31% respectively with an overall mean of 95.07%, substantially reducing training time and memory requirements compared to deeper models such as VGG16 [39]. The first controlled empirical benchmark comparing multiple contemporary U-Net variants under identical training conditions provided direct evidence about which architectural modifications matter most [40]. Standard U-Net, Attention U-Net, U-Net++, Residual

U-Net, and DoubleU-Net were evaluated for liver CT segmentation, directly countering confounding variables that prevent fair comparison across independently published studies. The experimental results confirmed that attention-based and nested variants consistently outperformed standard U-Net for boundary localization, with Attention U-Net's gates suppressing irrelevant feature responses at skip connections and U-Net++'s densely nested decoder sub-networks providing multi-scale feature aggregation. These improvements came at the cost of increased GPU memory and training time, with the standard U-Net remaining a competitive baseline for healthy, well-defined liver cases [40].

Both liver volume and hepatic vessel segmentation were addressed within a single two-stage pipeline enabling automatic Couinaud segment zoning for surgical planning [41]. In the first stage, a Dense U-Net operating on full $512 \times 512 \times 120$ voxel volumes achieved an average liver DSC of 0.903 under leave-one-out cross-validation on the IRCAD dataset. In the second stage, three 3D architectures were evaluated: standard 3D U-Net, 3D Dense U-Net, and 3D MultiRes U-Net in full-volume, slab-based, and box-based input configurations for intrahepatic vessel segmentation. The 3D MultiRes U-Net in a slab-based configuration achieved the best overall vessel DSC exceeding 70% with the most continuous vessel boundaries and least noise, confirming that volumetric context is more important than increased patch sampling quantity [41]. A framework integrating Fully Convolutional Network predictions into an Active Contour Model addressed the FCN's poor boundary localization and the ACM's sensitivity to initialization [42]. The FCN was trained to produce both a pixel-label map and layered boundary proximity information from which a novel external constraint force was derived with magnitude increasing near the true liver boundary and sign determined by the contour's position relative to the liver interior. The ACM could converge from any initialization, and cross-dataset generalization was validated by training the FCN on 73 clinical CT scans and testing without fine-tuning on SLiver07 with mean DSC of 96.2% and LiTS with mean DSC of 94.3%, demonstrating that deformable ACM components effectively compensate for inter-dataset intensity distribution shifts [42].

A GAN-enhanced Mask R-CNN framework addressing poor boundary completeness replaced manually preset Region Proposal Network anchors with k-means-derived anchors by clustering height-to-width aspect ratios of liver CT images to automatically determine anchor dimensions matching real liver morphology [43]. The GAN discriminator trained adversarially to distinguish expert annotation masks from Mask R-CNN-generated outputs, enforcing the output mask distribution to closely match real annotations. Evaluated on 378 training images from the Codalab liver CT competition, the GAN-augmented model achieved a DSC of 95.3%, outperforming base Mask R-CNN at 92.4% and the k-means-enhanced variant at 93.3% [43]. A semi-supervised 3D liver segmentation framework overcame training data scarcity by embedding an improved 3D U-Net as the discriminator within a GAN, enhanced with Squeeze-and-Excitation modules for channel-wise attention and a Pyramid Pooling Module at the bottleneck for multi-scale receptive field aggregation [44]. A novel generator based on a feature restoration strategy was designed where unlabeled images were encoded, a portion of the feature map was randomly masked, and the generator reconstructed the full fake 3D volume from incomplete features, forcing the network to learn realistic anatomical distributions without labeled supervision. The discriminator simultaneously classified volumes as background, liver, or generated fake data across a combined loss. Evaluated on LiTS-2017, the semi-supervised GAN achieved a DSC of 0.942, demonstrating improved performance at the challenging head and tail regions of the liver [44]. A hybrid Transformer-GAN framework exploited the Vision Transformer's

global self-attention mechanism to capture long-range spatial dependencies without the very large receptive fields required by standard CNNs [45]. The Transformer generator processed radiology images as a sequence of patch tokens and computed multi-head self-attention to aggregate contextual information globally before producing a segmentation mask, while a convolutional discriminator trained adversarially enforced adherence to global topological and shape properties beyond pixel-wise loss objectives. Evaluated on the ICPAI 2021 benchmark on liver CT and MRI data, the model achieved a DSC of 0.9433, recall of 0.9515, and precision of 0.9376, outperforming other Transformer-based approaches and confirming that GAN discriminator training provides measurable improvements by enforcing global shape plausibility [45].

2.1.2 Tumor Segmentation

While the preceding subsections traced the evolution of whole-liver delineation from manual tracings to GAN-enhanced deep architectures, precise segmentation of individual tumors within the hepatic parenchyma imposes substantially greater challenges. Liver tumors exhibit profoundly heterogeneous HU profiles where hypovascular lesions appear iso- or hypodense to surrounding parenchyma in the portal venous phase, hypervascular hepatocellular carcinoma shows intense arterial enhancement with rapid washout, and necrotic cores match the attenuation of perilesional fat. These heterogeneous characteristics render intensity-based criteria routinely unreliable, and boundary ambiguity is further compounded by the anatomical proximity of tumors to hepatic vessels, bile ducts, and the gallbladder, all of which share density ranges overlapping with lesion interiors. Additional complications include tumor multiplicity, wide intra-patient variability in lesion size from sub-centimeter microlesions to masses spanning multiple Couinaud segments, irregular non-convex morphology, and post-treatment fragmentation following chemoembolization or ablation. The following subsections review landmark contributions addressing these challenges across the same four methodological eras surveyed above for liver segmentation.

2.1.2.1 Classical Methods

Early liver tumor segmentation methods relied on hand-designed rules, morphological operations, and low-level image features to isolate lesions. These pipelines were interpretable and computationally efficient but exposed fundamental limitations of intensity-only approaches when confronted with heterogeneous and boundary-ambiguous targets. The MICCAI 2008 3D Liver Tumor Segmentation Grand Challenge established the standardized evaluation framework using Volumetric Overlap Error, Relative Volume Difference, and symmetric surface distances against which these early methods were benchmarked. The challenge of low-contrast lesion delineation was addressed by embedding minimal user interaction in the form of a two-point bounding box around each tumor slice into a 2D region-growing framework augmented with knowledge-based stopping constraints [46]. After 3×3 median filtering, a seed was automatically placed at the ROI center and the growing criterion accepted 4-connected neighbors whose intensity difference to the current region average fell below an adaptive threshold. A knowledge-based rule addressed premature termination: if the segmented area fell below half the bounding-box area indicating an unrepresentative bright seed, the algorithm automatically enlarged the intensity-averaging neighborhood and fused the two growing results via logical OR,

substantially reducing under-segmentation on patchy heterogeneous lesions. Morphological closing with a disk of radius 15 pixels smoothed the binary masks before 2D results were stacked into a 3D volume. On 10 MICCAI 2008 challenge test tumors, the method achieved an average total score of 64 with a VOE of 39.40%, reflecting the difficulty of homogeneous seed initialization for lesions of non-uniform appearance and motivating the shift toward adaptive multi-stage pipelines [46].

Responding to the clinical need for rapid volumetric monitoring of hepatic metastases, a semi-automatic pipeline was proposed that prioritizes processing speed and robustness while handling the wide diversity of metastasis appearances including hypo-dense, hyperdense, necrotic, calcified, and rim-enhancing lesions [47]. A single user-drawn stroke across the tumor region defines an ROI whose intensity histogram is analyzed to identify the lesion type and derive adaptive thresholds isolating it from surrounding parenchyma. Three-dimensional region growing then segments the lesion, but a critical challenge arises when tumors abut hepatic vessels of similar density. In these cases the growing algorithm leaks into the vascular tree, which was addressed by an adaptive morphological opening. A distance-map analysis quantifies the minimum diameter of connecting bridges and a matched structuring element applies erosion followed by dilation-based reconstruction to sever vessel contacts without eroding tumor tissue. The recovered boundary is then intersected with the initial region-growing mask to prevent false tissue recovery. On 10 MICCAI 2008 challenge tumors, the method achieved an average score of 72, VOE of 30.55%, and mean absolute surface distance of 1.55 mm at a processing time of approximately 2 seconds per lesion, demonstrating that a carefully engineered adaptive morphological pipeline can substantially outperform simpler region-growing approaches when vessel proximity is the primary failure mode [47].

A decade later, fully automated liver tumor segmentation was revisited with emphasis on the surgical planning use case where accurate three-dimensional visualization of hypovascular portal-venous-phase lesions is required to assess resection feasibility [48]. The pipeline began with Edge-Enhancing Diffusion filtering to sharpen tumor-parenchyma interfaces while suppressing intratumoural noise without blurring boundaries, followed by a localized mean shape model registered via mutual information confining the search to the hepatic volume. Otsu multi-level adaptive thresholding on the filtered image extracted an initial tumor mask, and a localized region-based level-set active contour refined this mask by deforming according to local intensity statistics at the evolving boundary rather than global image gradients. A novelty central to the pipeline was a circularity-based false-positive discriminator where a minimum bounding circle was fit to each segmented candidate and the ratio of its volume to the bounding-circle volume was computed. The system exploited the approximately spherical morphology of hypovascular lesions to reject elongated vessel artifacts and gallbladder inclusions. On the 3D IRCAD database comprising 111 tumors, the fully automated pipeline achieved a DSC of 74.96%, an Absolute Relative Volume Difference of 11.43%, and a Jaccard Index of 60.16%, establishing a competitive fully automatic baseline while highlighting the under-segmentation bias inherent in sphericity-constrained rejection for irregularly shaped malignant lesions [48].

The representational limitation of classical intensity-based methods was addressed through transformation of CT images into the Neutrosophic Set domain, which explicitly encoded the uncertainty inherent in overlapping tissue densities through three separable membership functions [49]. Each pixel was mapped to a Truth image representing the probability of belonging to an object, an Indeterminacy image reflecting boundary ambiguity, and a

Falsity image characterizing the background, all computed from local mean intensity and spatial gradient homogeneity. After 3×3 median filtering and histogram equalization, NS-domain adaptive thresholding and morphological operations on the Truth channel isolated preliminary structures while the Indeterminacy and Falsity channels served as internal markers and background seeds for a marker-controlled watershed algorithm. Over-segmentation was resolved by retaining the largest connected component as the liver mask. Fast Fuzzy C-Means clustering was then applied within the extracted liver ROI, grouping voxels into parenchyma, tumor, and vessel classes by minimizing a fuzziness-weighted sum of squared distances with accelerated centroid updates, circumventing the per-iteration full reassignment cost of standard FCM. On a dataset exceeding 105 patients, the NS-WS-FFCM pipeline achieved an accuracy of 94.98%, a DSC of 92.88%, a Jaccard Index of 86.84%, and a correlation coefficient of 91.66%, substantially outperforming standalone adaptive thresholding and region-growing baselines and demonstrating that explicitly modelling pixel-level uncertainty through neutrosophic domain transformation markedly improves segmentation consistency when hepatic and adjacent organ intensities overlap severely [49].

2.1.2.2 ML Methods

The adoption of ML classifiers introduced data-driven feature representations and adaptive decision boundaries into liver tumor segmentation pipelines, enabling greater tolerance to tumor heterogeneity and inter-patient variability than hand-tuned morphological rules while preserving the interpretability of engineered feature spaces. Among the first approaches to frame hepatic lesion extraction as an ensemble classification problem, AdaBoost was adapted to produce a weighted majority vote over weak segmentation hypotheses on three-dimensional CT volumes [50]. The pipeline extracted 54 voxel-level features per candidate structure from three complementary families: CT-value statistics computed over $3 \times 3 \times 3$ to $7 \times 7 \times 7$ neighborhoods, 3D convergence index filter responses estimating radially symmetric intensity fall-off as a proxy for lesion proximity, and standard edge filter responses. An atlas-based method restricted the search to liver-probability-positive voxels, and Expectation-Maximization normalization of CT intensities to per-patient liver mean and standard deviation reduced inter-patient variance before feature extraction. Two separate AdaBoost models were created, one for lesions under 30,000 voxels and one for large masses, each selecting feature-threshold weak hypotheses minimizing weighted training error at each boosting round. Leave-one-out validation on 16 volumes showed the Jaccard Index rising from 58.3% at a single boosting stage to 78.8% at 100 stages, with the MICCAI 2008 challenge submission achieving an average total score of 65 and a notably low false-positive rate of 0.7%, demonstrating that multi-round ensemble feature voting can substantially suppress false detections while approaching semi-automatic performance on heterogeneous hepatic lesions [50].

Building on the same MICCAI 2008 challenge setting, the prohibitive computational cost of voxel-wise SVM classification in three-dimensional CT was overcome by shifting classification to the level of watershed catchment basins [32]. After liver isolation via a statistical shape model and contrast enhancement through histogram stretching, a 3D watershed transform partitioned the hepatic volume into compact homogeneous basins, reducing the sample count requiring SVM evaluation by a factor of approximately 18. The user interactively labeled tumor and healthy tissue using a 3D brush tool from which

four per-basin features trained a linear SVM. Predictions were refined by morphological post-processing seeded from tumor annotations. On 10 MICCAI 2008 training tumors, the system achieved a VOE of 31.14% and a mean symmetric surface distance of 1.56 millimeters with segmentation runtime reducing from approximately 7 minutes for voxel-based classification to 30 seconds, confirming that region-level SVM classification offers a practical trade-off between accuracy and interactivity speed [32]. Clinical adoption of automated tumor segmentation further demanded reliable performance in longitudinal monitoring, where the same lesions must be re-segmented consistently across serial CT examinations, which was addressed by leveraging the baseline segmentation mask from the patient's prior scan as a spatial and intensity prior to constrain automated lesion search in the current acquisition [51]. A two-stage deformable registration pipeline combined 2D affine registration with 2D cubic B-spline free-form deformation minimizing normalized cross-correlation, aligning the baseline slice to the follow-up volume, and template matching localized the lesion's updated position. Three complementary constraint masks progressively refined the adaptive region-growing search space by excluding tissues adjacent to the baseline lesion boundary, removing follow-up pixels whose intensities deviated substantially from baseline lesion mean, and isolating ribs and muscles via mean-shift clustering. Seeds were derived from n-fold erosion of the registered baseline contour with the growth threshold optimized to maximize boundary compactness under intensity and standard deviation agreement constraints, and 2D results were propagated through adjacent slices by overlap-ranked region competition. Evaluated on 105 lesion cases across 79 patients, the system achieved a 90% lesion-matching rate, an average DSC of 0.83 ± 0.08, and 88% agreement with RECIST 1.1 clinical measurements, demonstrating that registration-constrained baseline-informed segmentation can largely automate the radiological follow-up workflow [51].

Beyond longitudinal monitoring, ML methods were also applied to differential diagnosis through a Computer-Aided Diagnosis system for benign-to-malignant tumor classification from multi-phase CT [52]. It combined three complementary feature families: noise suppression via median and curvature anisotropic diffusion filters preceded confidence-connected region growing from a user-supplied seed, 3D GLCM features captured texture heterogeneity, and a novel 3D elliptic shape model derived geometrical descriptors including axis ratios, volume covering ratios, angularity indices, and compactness measures quantifying morphological irregularity. Kinetic curve features characterized each tumor's multi-phase contrast enhancement trajectory through Fuzzy C-Means clustering of voxel-wise phase intensity sequences, extracting peak enhancement, time to peak, uptake rate, and washout rate. Backward elimination selected the optimal feature subset for binary logistic regression with leave-one-out cross-validation over 71 tumors, with the combined texture-shape-kinetic model achieving an accuracy of 81.69%, sensitivity of 81.82%, and AUC of 0.8713, whereas texture and shape descriptors alone yielded 71.82% and 69.01% respectively, confirming that multi-phase contrast dynamics contribute essential discriminative information that neither morphological nor textural representation alone can provide [52].

Refocusing on quantitative treatment assessment, a semi-supervised multi-phase framework for estimating the tumor necrosis rate in hepatocellular carcinoma following locoregional therapy combined supervoxel over-segmentation with a hierarchical multi-scale Random Forest [53]. Multi-phase DCE-CT acquisitions were registered to the early venous phase via a symmetric diffeomorphic method, and a 3D SLIC algorithm generated a compact supervoxel partition; for the hierarchical variant, this partition was iteratively subdivided into a multi-resolution tree. Each fine-scale supervoxel's feature vector concatenated

its own 20 dynamic features including spatial intensity statistics and temporal contrast dynamics with corresponding features from all hierarchical parent supervoxels, providing self-adaptive multi-scale context without explicit scale-selection logic. A Random Forest trained on approximately 15 user-labelled supervoxels per class then classified all remaining supervoxels as parenchyma, active tumour, or necrosis. On 8 DCE-CT exams, the hierarchical multi-phase approach achieved DSC values of 80.3% for active tumour and 86.0% for necrosis with a TN-rate estimation error of only 4.08%, substantially outperforming single-scale and single-phase baselines and confirming that scale-adaptive supervoxel hierarchies with rich multi-phase temporal features are highly effective for quantifying intratumoural composition [53]. As ML techniques matured, their application extended to large-scale radiomics analysis for tumor characterization and radiotherapy planning through a rigorous comparative evaluation of six mainstream ML classifiers applied to CT radiomics features for distinguishing the Gross Tumor Volume from normal liver parenchyma [54]. PyRadiomics extracted 1,395 features spanning First Order, GLCM, GLDM, GLRLM, GLSZM, and NGTDM descriptors from 104 HCC patients, and a two-stage dimensionality reduction using LASSO regularized logistic regression followed by Variance Inflation Factor filtering reduced the feature space to 7 robust predictors. Models trained with 200 repetitions of 5-fold cross-validation were evaluated on both discrimination and calibration, with XGBoost achieving the highest discrimination with an AUC of 0.9975 and MCC of 0.9369, while SVM delivered superior calibration with a Brier score of 0.0370 despite a slightly lower AUC of 0.9846, illustrating the clinically important distinction between statistical discriminability and probabilistic reliability frequently obscured by purely accuracy-centric tumor classification benchmarks [54]. Complementing this radiomics study, a systematic empirical comparison was undertaken to isolate the relative strengths of ML classifiers versus DL architectures, both constrained to operate on pre-extracted handcrafted feature vectors [55]. From 1,200 CT images, five non-overlapping ROIs per scan were preprocessed by Gabor filtering and histogram equalization before Mazda software extracted six feature families per ROI, producing 70-dimensional feature vectors. Both classical ML classifiers and standard DL architectures were trained on identical flat feature vectors, while pre-trained DL models were evaluated on raw CT images. The results were clear: Boost reached 99.7% and Random Forest 99.6% accuracy on handcrafted feature vectors, while CNN and Bi-LSTM performed near chance at 54.0% on the same inputs. This demonstrated that gradient-based deep networks cannot leverage their spatial inductive biases when applied to pre-abstracted tabular inputs, with DL accuracy recovering to 97.0% when operating on raw CT images, confirming that DL's comparative advantage is intrinsic to end-to-end spatial feature learning rather than to its classifier architecture per se [55].

2.1.2.3 DL Methods

The application of deep convolutional and attention-based architectures to liver tumor segmentation delivered transformative gains in detection recall and boundary precision, gains that came primarily from enabling end-to-end representation learning from volumetric CT data and accommodating the extreme class imbalance and multi-scale structural variability characteristic of hepatic lesions. The cascaded Fully Convolutional Network paradigm for joint liver and hepatic lesion segmentation established that for the first time a two-stage end-to-end deep pipeline could achieve clinically relevant accuracy across both CT and MRI modalities without architecture modification [56]. The first-stage U-Net

segmented the liver and produced an ROI crop focusing the second-stage U-Net exclusively on the hepatic parenchyma, dramatically reducing the proportion of background pixels and simplifying the learning task for the lesion detector. An adaptive weighted cross-entropy loss term compensated for the extreme class imbalance between rare lesion pixels at less than 1% of the input and background, preventing convergence to a trivial all-background prediction. 2D slices were processed independently with soft probability maps stacked into 3D volumes followed by dense 3D Conditional Random Field refinement penalizing spatially incoherent label assignments and intensity-discordant transitions. On the 3D-IRCAD CT dataset, the pipeline achieved a liver DSC of 94.3% and a lesion DSC of 56%, while on a multi-centric MR-DWI dataset the lesion DSC reached 69.7%, establishing the cascaded FCN as the dominant architectural paradigm for subsequent literature [56].

The distinct contrast enhancement dynamics of hepatic lesions across multi-phase CT acquisitions were exploited through architectural fusion of all three available phases within a single network [57]. The Multi-Channel Fully Convolutional Network processed arterial, portal-venous, and delayed phase slices simultaneously through three independent parallel encoder channels, with high-level semantic feature maps fused in deeper network layers before a shared softmax decoder. The base architecture adapted AlexNet as an FCN with eight convolutional layers, three subsampling layers, and three deconvolution layers, augmented with two cross-layer feature fusion connections concatenating coarse deep features with fine shallow features to recover spatial localization lost during subsampling. Pre-training on the whole-liver segmentation task before fine-tuning on tumor targets stabilized training despite limited multi-phase annotation. On the clinical JDRD multi-phase dataset, the MC-FCN achieved a VOE of 8.1%, RVD of 1.7%, and ASD of 1.5 millimeters, significantly outperforming single-phase variants with VOE of 15.6%, confirming that fusing multi-phase temporal contrast dynamics within the FCN architecture provides substantial discriminative benefit for hepatocellular carcinoma segmentation [57].

While the cascaded FCN paradigm greatly improved tumor recall, residual false positives per case remained a persistent limitation directly addressed by appending an object-level Random Forest classifier to a 2D U-Net [58]. False-positive suppression shifted from the pixel level to the level of 3D connected components. The U-Net employed both long skip connections and short residual-style connections, trained with a soft Dice loss restricted to a pre-computed 10 millimeter-dilated liver mask with class-balanced mini-batches managing lesion-background imbalance. Post-inference, each 3D connected component was characterized by 36 hand-crafted shape, statistical, and spatial features including PCA eigenvalue ratios, eccentricity, bounding-box extent, and standard deviation of component distances to the liver boundary. The Random Forest classified each object as true-positive tumor or false-positive artifact, reducing false positives by 85% from 4.6 to 0.7 per case while improving tumour DSC from 0.51 to 0.58, placing third in the MICCAI 2017 LiTS challenge and providing direct quantitative reference by benchmarking against a highly experienced Medical Technical Radiology Assistant who achieved a DSC of 0.70 [58].

The architectural tension between local boundary detail and global spatial context was tackled through a Three-Dimensional Dual-Path CNN simultaneously processing two differently-scaled 3D patches centered on the same voxel [59]. A local path operated on $43 \times 43 \times 43$ patches for fine-grained texture and boundary detail while a global path received a $129 \times 129 \times 129$ patch downsampled to $43 \times 43 \times 43$ to encode long-range anatomical context. Both paths shared an identical eight-block residual architecture with $3 \times 3 \times 3$ Conv3D layers, Batch Normalization, and PReLU activations with element-wise-

summed outputs passed through $1 \times 1 \times 1$ convolutional layers before softmax classification, avoiding the parameter explosion of traditional fully connected layers on volumetric feature maps. A Fully Connected Conditional Random Field refined the CNN probability output post-inference by minimizing a Gibbs energy function. On the LiTS benchmark, the TDP-CNN plus CRF pipeline achieved a tumor DSC of 0.689 and a Hausdorff Distance of 7.69 millimeters, dramatically lower than 65.38 millimeters without CRF post-processing, ranking first for liver and second for tumor in the MICCAI 2017 LiTS challenge and confirming that explicit multi-scale dual-path 3D processing combined with CRF spatial regularization provides a strong inductive bias for volumetric tumor delineation [59].

Simultaneous liver and tumor segmentation was further advanced through a cascaded three-stage pipeline combining a 2D coarse localization network with two 3D Residual Attention U-Nets providing sequential liver and tumor delineation [60]. The central architectural contribution was the Attention Residual Module replacing standard U-Net skip connections with a dual-branch structure where a trunk branch passed encoded feature maps forward and a soft mask branch learned a spatial attention map through convolution, max-pooling, and sigmoid gating to suppress background activations. Stage 1 employed a 2D RA-UNet for coarse liver bounding-box prediction, Stage 2 applied a 3D RA-UNet on $224 \times 224 \times 32$ liver patches for precise organ delineation, and Stage 3 used a second 3D RA-UNet on $128 \times 128 \times 32$ patches sampled within the predicted liver volume for tumor extraction, avoiding interpolation that would obscure sub-centimeter lesions. Overlapping patch voting aggregated predictions from multiple crops. On the LiTS test set, the method achieved a liver DSC of 0.961 and a tumor DSC of 0.595, while on the 3DIRCADb dataset the tumor DSC reached 0.830, demonstrating that residual attention mechanisms within a cascaded 3D architecture substantially improve performance on small, low-contrast hepatic lesions [60].

A hybrid ResUNet combining residual skip connections of ResNet with the encoder-decoder architecture of U-Net performed simultaneous liver and tumour segmentation on the 3D-IRCADb-1 dataset [61]. Residual blocks within the encoder addressed the vanishing gradient problem limiting training depth on small medical datasets while symmetric decoder skip connections recovered spatial resolution lost during pooling. Extensive preprocessing included HU windowing between -100 and 400, histogram equalisation, normalisation, and standardisation standardizing the 10 available CT volumes with reflection-and-rotation augmentation expanding the effective training set. Individual tumour masks were merged into unified binary labels and the model was trained first on liver segmentation then on tumour segmentation within the predicted liver ROI. On the 3D-IRCADb-1 test set, the method achieved reported tumour confusion-matrix accuracy of 99.6% and DSC of 99.2%, though the authors explicitly acknowledged that these figures are inflated due to severe background-to-tumour class imbalance intrinsic to CT volumes, emphasizing the need for more diverse datasets and alternative evaluation protocols to reliably estimate generalisation to clinical populations [61].

Several recent architectural advances were consolidated into a unified hybrid attention-aware U-Net designed to simultaneously address local boundary detail and long-range global context in liver and tumor delineation [62]. The Optimized Transformer Unit computed global multi-head self-attention across full feature maps and local window-restricted attention in parallel branches with outputs fused to capture spatial dependencies at both coarse and fine spatial scales. The Enhanced Feature Linkage module replaced standard skip connections with multi-level feature fusion blocks combining encoder outputs

from multiple decoder stages to prevent the semantic gap between coarse encoder and fine decoder representations limiting standard U-Net skip connections for small lesion recovery. Improved MBConv blocks employed Leaky-ReLU activations and atrous separable convolutions expanding the effective receptive field without increasing parameter count. An inter-sample slice-continuity learning strategy further exploited volumetric coherence across adjacent CT slices. Evaluated on combined LiTS2017 and 3Dircadb benchmarks with a 70% training, 30% test split, the model achieved a liver DSC of 98.53%, liver Intersection over Union (IoU) of 95.83%, tumour DSC of 95.52%, and tumour IoU of 89.30%, improving over the best-compared baselines by at least 1.53% in liver DSC and 10.4% in tumour DSC with training completing in 2.56 hours, establishing the current state of the art in joint liver-tumour segmentation on standard benchmarks as of 2025 [62].

2.1.3 Cross-Modal Liver Tumor Segmentation

The challenge of low-contrast tumor visibility in diagnostic CT imaging motivated the development of cross-modal contrast enhancement techniques leveraging complementary information from concurrent MRI acquisitions. One approach introduced a novel formulation of two-dimensional histogram specification as an optimization problem and extended it to multi-modal medical imaging for the first time [63]. The method incorporated a Structural Similarity Index Measure gradient to maintain structural fidelity during enhancement alongside an adaptive stopping criterion based on 2D entropy. The method computed Grey Level Co-occurrence Matrices from both CT and MR images and extracted cumulative distribution functions to establish cross-modal intensity transformations. The enhanced CT image was iteratively refined through SSIM gradient-based updates until entropy stabilization. Evaluated on 99 CT-MR image pairs from 10 patients acquired on multi-detector CT and T1-weighted MRI, the Optimized Guided Contrast Enhancement method improved tumor boundary visibility substantially. Subsequent Seeded Region Growing segmentation achieved mean DSC of 0.493 $\pm$ 0.061, Hausdorff Distance of 8.77 $\pm$ 2.48 millimeters, and sensitivity greater than 0.75 with specificity greater than 0.98 [63]. However, the reliance on pixel-level gradient-driven region growing without explicit boundary smoothness constraints or anatomical shape priors resulted in discontinuous segmentations with low DSC values and incomplete boundary overlap, indicating that enhancement alone cannot compensate for the lack of regularization mechanisms inherent in classical image processing pipelines.

Building on the insight that multi-phase CT acquisitions provide complementary enhancement signatures for tumor characterization, DL approaches began to integrate cross-modal feature fusion, advancing the field with a reciprocal cross-modal guidance module coupled with multi-scale feature fusion for segmenting liver lesions from arterial-phase and portal-venous-phase CT images that may have been acquired with incomplete spatial overlap or misalignment from respiratory motion [64]. The architecture employed parallel ResNet-34 encoders for each phase with features extracted at five resolution levels. Reciprocal guidance was implemented through cosine similarity between phase-specific tumor feature vectors and decoder representations, generating soft attention masks accommodating phase misalignment without requiring explicit deformable registration. Trained on 45 patients with 1,200 annotated 2D slices from arterial and portal phases acquired over 3 to 5 minute intervals, the method achieved mean DSC per case of 0.753 $\pm$ 0.062, voxel-level DSC was 0.912 $\pm$ 0.041, Asymmetric Surface Distance was 8.15 $\pm$ 2.45 millimeters, and

sensitivity was 0.751, outperforming multi-window U-Net and Pyramid Attention ResNet segmentation baselines [64]. While the introduction of learnable cross-modal attention and multi-scale feature aggregation substantially improved robustness to phase misalignment, limitations remained including constraint to single-center data, limited tumor type diversity, and the fundamental challenge of severe class imbalance with tumors occupying less than 1% of pixels, alongside the absence of explicit mechanisms to quantify or enforce segmentation confidence in uncertain boundary regions.

The successive improvement in DL architectures directed attention toward the temporal structure inherent in multi-phase dynamic contrast-enhanced imaging, with a pressing need emerging to quantify segmentation uncertainty in regions where pathological boundaries are ambiguous. A comprehensive framework was developed to perform simultaneous segmentation and geometric quantification of center location, maximum diameter, circumference, and area of hepatocellular carcinoma and hemangiomas while addressing severe class imbalance and non-linear enhancement kinetics across four dynamic MRI phases [65]. The Uncertainty-guided and Cross-modality Attention Network introduced a Cross-Modality Attention Pyramid Module computing enhancement and washout differences between pre-contrast and contrast-enhanced phases to isolate tumor-specific temporal signatures. A Fusion Transformer with non-local spatial attention via 3D patch self-attention and phase-aware temporal attention modelling inter-phase relationships without spatial decomposition was employed. An Uncertainty-Guided Auxiliary-Primary dual-segmentor architecture quantified boundary uncertainty via Kullback-Leibler divergence with cross-entropy loss adaptively reweighted to emphasize uncertain voxels. Evaluated on 265 subjects from a single Siemens 3.0T center with 185 training, 40 validation, and 40 test scans, the network achieved mean DSC of 0.8544 ± 0.0215 and Hausdorff Distance of 2.28 ± 0.68 millimeters, substantially exceeding nnU-Net, TransUNet, and baseline methods with quantification metrics achieving center localization MAE of 1.85 ± 0.92 millimeters and diameter measurement MAE of 1.90 ± 1.12 millimeters [65]. Despite these architectural advances, limitations persisted including single-center, single-manufacturer training on exclusively one 3.0T MRI system restricting cross-vendor generalization, support for only T1-weighted dynamic sequences limiting integration of complementary DWI and T2 information, and the absence of clinical outcome correlation meaning that segmentation accuracy remained decoupled from downstream prognostic utility or treatment response prediction.

The progression toward increasingly sophisticated architectures encountered a fundamental barrier when addressing the cross-modality transfer problem under feature absence, exemplified by the clinical challenge of intra-operative tumor localization on non-contrasted or minimally-contrasted CT where pathological boundaries are physically undetectable despite clear visibility on pre-operative MRI. An end-to-end registration-segmentation framework was presented in [66] to investigate whether anatomical localization could be transferred from MRI (where tumors are visible) to CT (where they are invisible) through weakly supervised learning. The method combined MSCGUNet, a multi-scale U-Net with self-constructing graph latent for deformable cross-modality registration, with a U-Net segmentation module, generating pseudo-labels for CT by warping MRI ground truth tumour masks through learned deformation fields. The framework was formally analyzed through information-theoretic lens, asserting that mutual information $I(z; Y) \approx 0$ between CT voxel intensities and tumor classes, characterizing tumor invisibility as aleatoric (irreducible, data-inherent) rather than epistemic (reducible with more data) uncertainty. Cross-validation on CHAOS dataset (40 subjects) yielded sequential training DSC 0.72 ± 0.04 and Jaccard 0.71 ± 0.00 compared to supervised baseline DSC 0.92 ± 0.02,

a performance gap of approximately 0.20 attributed to registration imprecision and cross-modality feature transfer difficulty. When evaluated on clinical data (7 usable volumes from Universitätsklinikum Magdeburg), the weakly supervised framework achieved DSC 0.16 ± 0.14, nearly identical to the supervised CT baseline DSC 0.19 ± 0.06, demonstrating that the dramatic drop from CHAOS (DSC 0.72) to clinical (DSC 0.16) was intrinsic to feature absence rather than method failure. This critical finding indicates a fundamental frontier in cross-modal medical image analysis. Even perfect registration and sophisticated architectures cannot enable tumor segmentation when pathological features are physically absent in the target modality. This observation shifts debate from algorithmic optimization toward information-theoretic impossibility.

2.2 Challenges Faced

Liver tumor segmentation presents multiple well-documented technical obstacles that make automated delineation inherently difficult. First is the problem of extreme class imbalance, where lesions occupy less than one percent of voxels in the entire hepatic volume while the majority of voxels represent healthy liver tissue or background anatomy. This severe imbalance creates a straightforward statistical problem for learning algorithms. Beyond class imbalance, the visual characteristics of tumors in medical images introduce additional complexity. Hypovascular lesions often appear iso-dense or hypodense relative to surrounding parenchyma in portal-venous phase CT, making intensity-based segmentation unreliable when the tumor tissue has similar brightness to healthy tissue. Boundary ambiguity arises because tumors frequently develop near hepatic vessels, bile ducts, and the gallbladder, all of which share overlapping intensity ranges with lesion interiors. Standard image processing approaches that rely on intensity gradients struggle when multiple tissue types have indistinguishable brightness. Tumor appearance itself varies substantially depending on tumor type and phase timing. Hypervascular hepatocellular carcinoma exhibits arterial enhancement and portal-venous washout, while hemangiomas show persistent enhancement across dynamic phases. Additionally, intra-patient size variability is pronounced, with tumors ranging from sub-centimeter microlesions to masses spanning multiple Couinaud segments within the same patient. Segmentation architectures must therefore operate effectively across this extreme range of tumor sizes.

Cross-modal liver tumor segmentation introduces a different class of problems that are rooted in the fundamental physics of different imaging modalities. CT imaging measures X-ray attenuation, generating HU values that represent tissue density. MRI, in contrast, measures tissue-specific T1 and T2 relaxation times during nuclear magnetic resonance recovery, a completely different physical phenomenon. When a neural network encoder is trained on CT imagery, it learns to classify tissues based on HU patterns and relationships. These learned representations correspond to density-based tissue discrimination. In MRI data, identical anatomical tissues produce signals governed by different physical principles. A tumor that appears hypodense in CT might appear hyperintense in T1-weighted MRI sequences depending on the relaxation times and pulse timing, not because of any physical density difference. This means feature representations learned from CT data do not directly transfer to MRI data. An encoder that has learned what a tumor looks like in HU space cannot simply apply that knowledge in T1 or T2 relaxation time space without substantial modification. This encoding mismatch represents a fundamental barrier to cross-modal

knowledge transfer that distinguishes it from the purely anatomical-level challenges present within single modalities.

2.3 Scope of Study

Liver tumor segmentation is a critical task in clinical practice, supporting surgery planning, volumetric measurements, and disease monitoring. Cross-modal liver tumor segmentation addresses a genuine clinical need that arises from the fact that different imaging modalities offer complementary strengths. MRI typically provides superior soft-tissue contrast and often reveals hepatic tumors with high visibility and clear boundaries. CT, in contrast, offers superior spatial resolution, lower cost, and faster acquisition, making it more practical for surgical planning and follow-up studies, yet CT frequently depicts tumors with low contrast against background liver tissue. In clinical practice, patients undergo both modalities, creating a natural research problem about whether representations learned in one modality can transfer effectively to another. The fundamental questions are whether segmentation models can transfer bidirectionally between modalities, whether representations learned from the modality with superior tumor visibility can improve segmentation in the modality with poor visibility, and whether encoder-decoder architectures can share learned features across these fundamentally different imaging physics. These are important clinical problems that existing cross-modal literature rarely addresses systematically.

This thesis establishes a baseline analysis of cross-modal liver tumor segmentation with the goal of understanding why complex architectural innovations are necessary. This represents an important gap in the current literature. While sophisticated cross-modal approaches exist including uncertainty quantification modules, attention mechanisms, and information-theoretic frameworks, they do not establish what performance levels can be achieved with simpler, more interpretable baseline methods. Without knowing baseline performance, it is impossible to determine whether architectural sophistication actually provides meaningful improvements or whether simpler approaches would suffice. The research approach is to constrain the architecture deliberately by freezing the encoder, allowing only the decoder and training strategy to change. This frozen encoder architecture holds feature extraction constant and learned representations from a single modality static, ensuring that any observed performance differences across modalities arise from the encoder mismatch problem itself rather than from the network's ability to fine-tune its feature extraction. By measuring performance under this constraint, the thesis systematically investigates preprocessing variations (morphological label cleaning, HU windowing, ROI extraction strategies), training strategies (learning rate scheduler, early stopping, dataset selection), and dataset-specific tuning to understand which failure mechanisms can be mitigated without encoder modification and which remain intractable due to the fundamental encoding mismatch rooted in cross-modal physics differences. Furthermore, because previous studies frequently relied on data from a single center or a single equipment manufacturer, their models often experienced a significant drop in accuracy when applied to external clinical data. To address this limitation, the this study includes expanding dataset coverage to incorporate more diverse data sources. Evaluating the models across a wider variety of data allows the analysis to isolate performance failures caused by restricted training diversity from those caused by the inherent physical differences between imaging modalities.

Chapter-3: Dataset

3.1 Dataset Description

This study uses five medical imaging datasets that collectively provide comprehensive coverage of different imaging modalities, patient populations, and clinical scenarios. The datasets include both healthy and pathological cases, with imaging data acquired from CT and MRI protocols. Each dataset brings specific advantages to the training and validation of segmentation models. Together, they represent the diversity encountered in clinical practice and enable robust evaluation of cross-modal segmentation approaches.

3.1.1 LiverHCCSeg

The LiverHCCSeg dataset is a publicly available resource containing multiphasic contrast-enhanced MRI scans from The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) collection [67]. The dataset provides manual 3D segmentations of both the liver and HCC tumors. Segmentations were created by two board-certified abdominal radiologists, with high inter-rater agreement demonstrated by an average DSC of 0.953 between the two raters, ensuring quality and reproducibility of the annotations. The dataset includes 17 cases with complete liver segmentations and 14 cases with corresponding tumor masks. This mix provides a solid foundation for training and evaluating segmentation algorithms. The MRI scans are T1-weighted sequences with intensity values ranging from 0 to 2108.091, and an average maximum intensity of approximately 800.12 across all images. This relatively moderate intensity range is characteristic of T1-weighted MRI protocols. A summary of the dataset is presented in Table 3.1, and example images are shown in Figure 1.

3.1.2 CHAOS

The Combined Healthy Abdominal Organ Segmentation (CHAOS) dataset provides both CT and MRI data [68]. The complete dataset includes scans from 40 CT and 40 MRI subjects. However, this study utilizes only the 20 MRI subjects that have publicly available ground truth annotations (originally provided as the challenge training set). The MRI data includes T1-DUAL and T2-SPIR sequences with annotations for four abdominal organs: the liver, spleen, and right and left kidneys. The 16-bit Digital Imaging and

Table 3.1: Summary of LiverHCCSeg Dataset

Attribute	Details
Modality	MRI (T1-weighted)
Number of Subjects	17
Number of Data	17
Total Slices	1378
Tumor Size Range	1.17 mm^3 to 12.3 mm^3
Original Format	NIfTI
Mean Age	61 ($\pm$ 10.77) years
Number of Males	11
Number of Females	6
Minimum Intensity	0
Maximum Intensity	2108.091
Average Minimum Intensity	0
Average Maximum Intensity	800.12

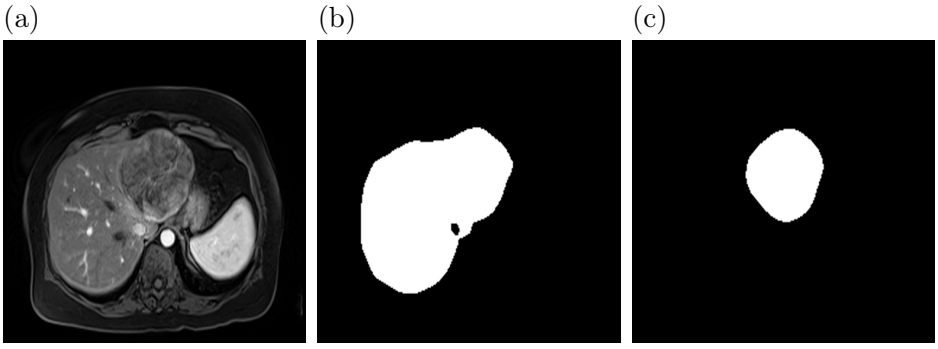


Figure 1: Example Images From the LiverHCCSeg Dataset. The Images Show (a) a T1 Weighted MRI Scan, (b) Segmented Liver Mask, and (c) Segmented Tumor Mask.

Communications in Medicine (DICOM) images present realistic clinical challenges. These include varying intensity ranges from contrast agents, significant shape differences across patients, and similar intensity values among neighboring organs. The MRI scans exhibit intensity values ranging from 0 to 2662, with an average maximum intensity of 1167.35 across the dataset. This intensity range is comparable to other MRI sequences but slightly higher than LiverHCCSeg, reflecting differences in acquisition parameters and contrast enhancement protocols. These factors make automatic organ segmentation substantially more difficult. A summary of the CHAOS dataset is provided in Table 3.2, and example images are shown in Figure 2.

3.1.3 LiTS

The Liver Tumor Segmentation Benchmark (LiTS) dataset is a large-scale resource comprising 131 contrast-enhanced abdominal CT scans for training and 70 scans for testing [69]. This dataset was created through collaboration among seven hospitals and research institutions. It provides detailed segmentation masks for the liver and various tumor types. The image data is highly diverse, including both primary tumors such as hepatocellular carcinoma and secondary metastatic lesions with varied sizes, shapes, and density levels. Some lesions appear denser than their background while others appear less dense. Because

Table 3.2: Summary of CHAOS Dataset

Attribute	Details
Modality	MRI (T1-DUAL)
Number of Subjects	40
Number of Data	20
Total Slices	647
Original Format	DICOM
Minimum Intensity	0
Maximum Intensity	2662
Average Minimum Intensity	0
Average Maximum Intensity	1167.35

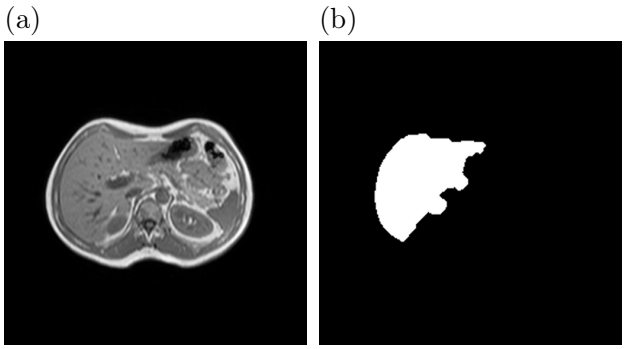


Figure 2: Example Images From the CHAOS Dataset. The Images Show (a) a T1 DUAL MRI Scan, and (b) Segmented Liver Mask.

LiTS is a CT dataset, it uses HU for intensity representation. The scans exhibit a wide intensity range from -10522 to 27572 HU, with an average minimum intensity of -1767.52 and average maximum intensity of 3285.41. This large dynamic range is substantially broader than MRI datasets and reflects the diverse tissue types and intensity variations captured across the 131 training volumes from multiple hospitals. Such variation creates both opportunities and challenges for segmentation algorithms. This heterogeneity makes LiTS an important resource for both 2D and 3D tumor segmentation as well as tumor burden estimation. A summary of the LiTS dataset is presented in Table 3.3, and example images are shown in Figure 3.

Table 3.3: Summary of LiTS Dataset

Attribute	Details
Modality	CT
Number of Subjects	201
Number of Data	131
Total Slices	58638
Tumor size Range	38 mm^3 to 1231 mm^3
Original Format	NIfTI
Minimum Intensity	-10522
Maximum Intensity	27572
Average Minimum Intensity	-1767.52
Average Maximum Intensity	3285.41

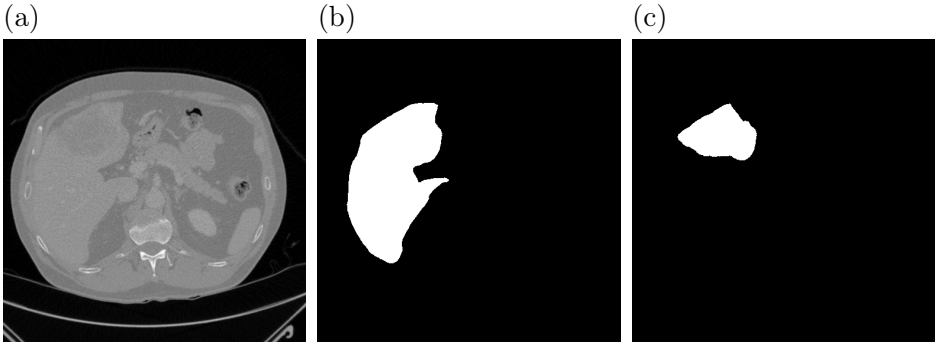


Figure 3: Example Images From the LiTS Dataset. The Images Show (a) a Contrast Enhanced CT Scan, (b) Segmented Liver Mask, and (c) Segmented Tumor Mask.

3.1.4 3D-IRCADb-01

To expand the training data and include pathological cases with tumors, the 3D-IRCADb-01 dataset was incorporated [70]. It consists of 3D CT scans from 10 female and 10 male patients, with a hepatic tumor incidence rate of 75%. Provided by the IRCAD institute, this dataset includes detailed medical images with structures segmented by clinical experts. In addition to liver and tumor masks, it provides segmentations for numerous surrounding anatomical structures including the venous system, gallbladder, lungs, and kidneys. As a contrast-enhanced CT dataset, the scans have intensity values ranging from -2048 to 3247 HU, with an average minimum intensity of -1177.55 and average maximum intensity of 1294.35. The intensity range is more moderate compared to LiTS but still demonstrates the variation expected from clinical CT acquisitions with different contrast protocols. This dataset also includes tumor location classifications using Couinaud's segmentation, which provides valuable information for preoperative planning and surgical guidance. The key characteristics of this dataset are summarized in Table 3.4, and example images are shown in Figure 4.

Table 3.4: Summary of 3D-IRCADb-01 Dataset

Attribute	Details
Modality	CT (Contrast-enhanced)
Number of Subjects	20
Number of Data	20
Total Slices	2823
Original Format	DICOM
Number of Males	10
Number of Females	10
Minimum Intensity	-2048
Maximum Intensity	3247
Average Minimum Intensity	-1177.55
Average Maximum Intensity	1294.35

3.1.5 SLiver07

The SLiver07 dataset originated from the MICCAI 2007 Grand Challenge and contains contrast-enhanced CT scans acquired during the portal venous phase [71]. Data came from

multiple clinical partners. The dataset provides 20 annotated CT volumes for training and 10 unannotated volumes for testing. As a CT dataset, it presents intensity values ranging from -1024 to 3071 HU. The average minimum and maximum intensities are both -1024 and 2097.45 respectively, indicating relatively consistent preprocessing and acquisition protocols across the 20 training volumes. This narrower intensity range compared to LiTS suggests more standardized scanner configurations, though still representing realistic clinical variation. The dataset includes segmentation masks of liver region. Characteristics of the dataset are summarized in Table 3.5, and example images are shown in Figure 5.

Table 3.5: Summary of SLiver07 Dataset

Attribute	Details
Modality	CT (Contrast-enhanced)
Number of Subjects	20
Number of Data	20
Total Slices	4159
Original Format	MHD (MetaImage)
Minimum Intensity	-1024
Maximum Intensity	3071
Average Minimum Intensity	-1024
Average Maximum Intensity	2097.45

3.2 Dataset Pre-processing

Given the diversity of imaging modalities, file formats, and acquisition protocols across the five datasets, a preprocessing pipeline was necessary to standardize all data into a consistent format suitable for training the segmentation models. The datasets arrived in three different file formats: Neuroimaging Informatics Technology Initiative (NIfTI) for LiverHCCSeg and LiTS, DICOM for CHAOS and 3D-IRCADb-01, and MHD for SLiver07. Additionally, the scans came from multiple hospitals and scanners, resulting in variation in pixel spacing, intensity distributions, and other acquisition parameters. A unified preprocessing approach made it possible to train models on combined data and to make fair comparisons across datasets.

The preprocessing pipeline began by converting all volumetric data to an intermediate NIfTI format. This standardized format consolidates three-dimensional volumetric information

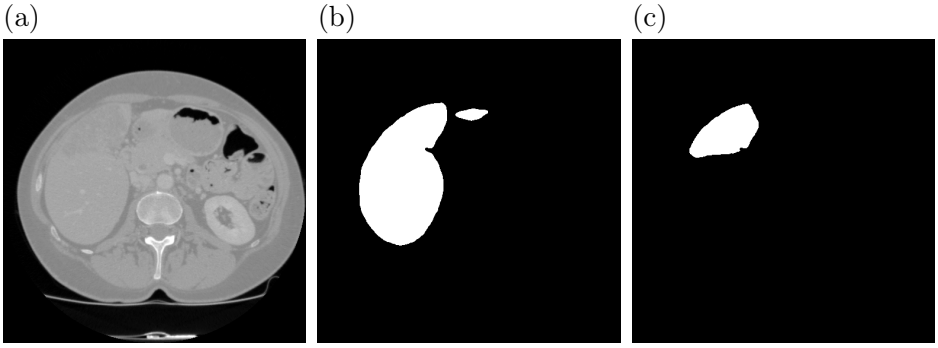


Figure 4: Example Images From the 3D-IRCADb-01 Dataset. The Images Show (a) a Contrast Enhanced CT Scan, (b) Segmented Liver Mask, and (c) Segmented Tumor Mask.

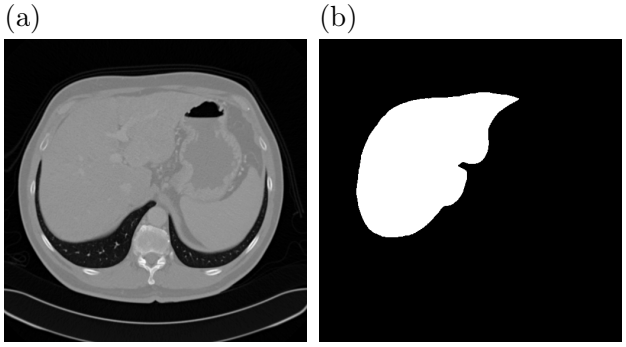


Figure 5: Example Images From the SLiver07 Dataset. The Images Show (a) a Contrast Enhanced CT Scan, and (b) Segmented Liver Mask.

and made subsequent processing steps more efficient. After conversion, each volume was decomposed into individual two-dimensional axial slices and saved as PNG images. This slice-based approach reduced memory requirements during model training and allowed the use of two-dimensional segmentation networks pre-trained on ImageNet.

A crucial step was intensity normalization. Each slice was normalized to a pixel value range of 0 to 255 to ensure that data from different scanners shared similar intensity distributions. This step was particularly important for transfer learning, since the pre-trained encoder networks expect input values within the same range as the ImageNet dataset. Without this normalization, the different intensity scales of MRI and CT images would create unnecessary challenges for the models.

Segmentation masks required careful processing tailored to each dataset. For 3D-IRCADb-01, the masks were stored in DICOM format, so the same slicing and normalization operations applied to image volumes were also applied to the corresponding masks. In contrast, CHAOS masks arrived as PNG files but contained pixel values between 55 and 70 instead of the expected binary values of 0 and 1. These masks were binarized to ensure compatibility with the evaluation metrics and loss functions used during training and validation. SLiver07 masks came in MHD format and were converted to individual PNG slices using the same normalization pipeline. This dataset-specific processing ensured that all masks ended up as binary images with 0 indicating background and 1 indicating liver.

The LiTS dataset required additional corrective steps beyond standard preprocessing. During quality inspection, it was discovered that CT scans from subjects volume-28 through volume-47 had a different spatial orientation than the remaining volumes. Similarly, masks from subjects volume-48 through volume-52 exhibited the same orientation mismatch. To correct both cases, a horizontal flip was applied to restore the proper orientation, as shown in Figure 6. These corrections ensured consistency within the LiTS dataset and prevented orientation differences from confounding the model learning process.

Despite the diversity of input formats and acquisition protocols, this preprocessing strategy produced a unified dataset suitable for model training. All images were two-dimensional PNG slices with normalized intensity values, and all masks were binary representations. This standardization enabled the training of cross-modal segmentation models that could leverage the complementary information provided by the five different imaging datasets.

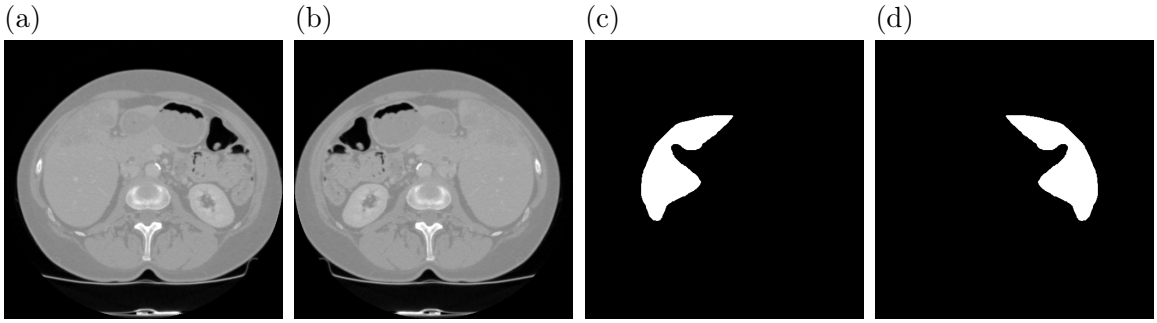


Figure 6: Visual Representation of the Spatial Orientation Correction Applied to Specific Cases Within the LiTS Dataset. The Images Show (a) a CT Scan Before Correction, (b) a CT Scan After Correction, (c) Uncorrected Liver, and (d) Corrected Liver.

3.3 Dataset Preparation

This study trained segmentation models using three of the five available datasets while reserving the remaining two for evaluation and testing. This section describes how each training dataset was partitioned into separate splits for model training, validation, and testing. The split proportions vary according to dataset size and the number of available subjects. A key principle underlying all splitting strategy was that divisions were performed at the subject level rather than the slice level. This approach ensures that all slices from a given patient belong to only one subset, preventing data leakage and maintaining the independence necessary for fair model evaluation.

3.3.1 Overview and Rationale

Proper handling of dataset splits is critical in medical imaging research. Each medical imaging volume contains hundreds or even thousands of individual slices from a single patient. When dividing data for training and evaluation, it is essential to keep all slices from a single patient within the same split. If slices from the same patient were distributed across training, validation, and test sets, the model would have opportunity to learn patient-specific features during training and then encounter the same patient again during testing. Such an arrangement would violate the independence assumption underlying model evaluation and artificially inflate reported performance metrics. Subject-level stratification reflects the clinical reality that meaningful model evaluation should measure generalization to new patients rather than simply recognizing new images from known patients. Across all datasets used for training in this study, the splitting strategy observed this principle strictly, ensuring that no patient data appeared in multiple splits.

3.3.2 LiverHCCSeg

The LiverHCCSeg dataset contained only 17 subjects with complete liver segmentations and 14 subjects with tumor masks, making it the smallest of the datasets used for training. Given this constraint, the dataset was divided using a 60% to 40% split, allocating 60% of subjects to training and 40% to testing. This resulted in approximately 10 subjects

reserved for training and 7 subjects for testing and validation. Because validation data is necessary to monitor model training progress and make decisions about early stopping, the test set was repurposed to also serve as the validation set. During model training, the same 7 subjects were used to validate the model and select the best-performing checkpoint based on validation loss. After training concluded, these same 7 subjects were then used for final evaluation. While using identical data for both validation and testing is not ideal in principle, the practical constraints of working with a small dataset made this approach necessary. The splitting was performed at the subject level, ensuring no patient data leakage between the training set and the validation or test set.

3.3.3 CHAOS

The Combined Healthy Abdominal Organ Segmentation dataset consisted of 20 subjects, providing more data than LiverHCCSeg but still a relatively modest collection for DL. The dataset was divided using a 60% to 20% to 20% split, allocating 12 subjects to training, 4 subjects to validation, and 4 subjects to testing. This balanced allocation of validation and test subjects ensured that model selection during training was performed on an appropriately sized independent subset, and that final evaluation also had sufficient data to provide reliable performance estimates. The split was performed at the subject level, with no subject appearing in more than one subset.

3.3.4 LiTS

The Liver Tumor Segmentation benchmark dataset was substantially larger than the preceding training datasets, containing 131 training subjects made available for the original challenge. The larger size allowed for a different split strategy that could allocate a higher proportion of data to training. An 80% to 10% to 10% split was applied, allocating approximately 105 subjects to training, 13 subjects to validation, and 13 subjects to testing. This allocation reflects the general practice in DL of dedicating the majority of available data to training when dataset size permits, while reserving appropriately sized held-out sets for validation and testing to ensure rigorous and reliable model evaluation. Subject-level stratification was maintained throughout, with no subject appearing in multiple splits.

Chapter-4: Methodology

4.1 Proposed Method

This work proposes a two-stage, cross-modal liver and tumor segmentation pipeline designed to operate robustly across both MRI and CT domains, as shown in Figure 7. The two-stage approach decouples two distinct tasks: the anatomically simpler liver localization and the more challenging intra-hepatic lesion segmentation. In Stage 1, two variants of morphological closing are investigated to clean ground-truth liver annotations by filling intra-organ voids and smoothing boundary irregularities. Minor fragment removal is applied to eliminate small disconnected components that introduce training noise. These label-cleaning techniques are evaluated under two settings: a Before setting that computes metrics on raw predicted masks, and an After setting that applies morphological closing as a post-processing step. The segmentation model is built using an encoder-decoder architecture combining a frozen pretrained ResNet18 encoder with a U-Net-style decoder.

Stage 2 extracts a ROI from liver masks to focus tumor segmentation on the hepatic region. Ground-truth liver masks are used during training, while Stage 1 predicted masks are used during validation. Two extraction strategies are designed and tested: a fixed center-based approach that anchors a 224×224 pixel bounding box at the mask centroid, and a morphologically dilated strategy that retains only pixels within a dilated mask boundary. Within this constrained spatial domain, tumor segmentation is performed through systematic preprocessing investigation. Four HU windowing configurations are evaluated to identify the optimal intensity range that enhances tumor-to-parenchyma contrast. Two slabbing strategies are designed and evaluated to encode volumetric context by populating input channels with progressively averaged neighborhoods of adjacent slices, enabling the model to exploit inter-slice continuity without requiring a fully 3-D architecture. To generalize windowing to new datasets without manual re-tuning, three domain adaptation approaches are tested, with ratio-based transfer normalizing boundaries to each subject's intensity distribution. The same encoder-decoder architecture is employed for tumor segmentation, trained on tumor annotations rather than liver masks.

This focused two-stage approach yields higher effective resolution for tumor segmentation, reduces false-positive rates by constraining the search space to hepatic tissue, and mitigates the extreme class imbalance characteristic of lesion detection. The systematic preprocessing investigation establishes empirical baselines that quantify where simple methods succeed and where they encounter fundamental cross-modal limitations.

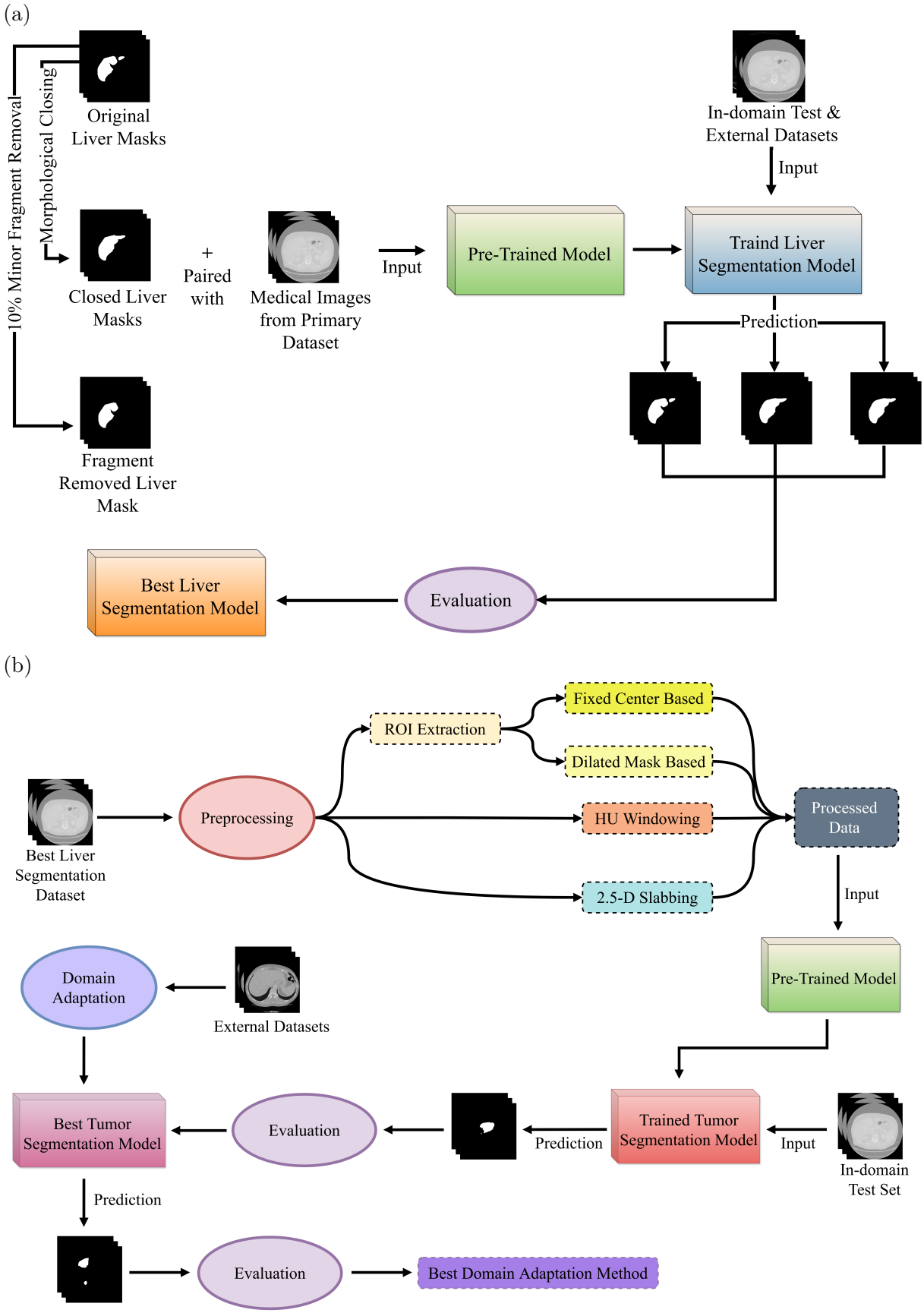


Figure 7: An overview of the Proposed Method. (a) Stage-1: Liver Segmentation, and (b) Stage-2: Tumor Segmentation.

4.1.1 Liver segmentation

The first stage of the two-stage pipeline produces a binary liver mask for each input slice. Multiple liver segmentation models are trained and evaluated on a heterogeneous set of public datasets spanning modalities (MRI, CT) and acquisition protocols. The objective is twofold: to maximize in-domain liver segmentation performance on a chosen training set, and to identify the model that generalizes best across modalities and datasets. Cross-domain performance is measured by testing each liver model on out-of-domain datasets within the same modality, while cross-modal performance is measured by testing on datasets from the other modality. The selected liver model is then used to generate masks for all cases prior to tumor modeling.

4.1.1.1 Preprocessing

Before training the liver segmentation model, ground-truth annotations are processed through a cleaning pipeline. This ensures that the model learns from high-quality supervision targets rather than reproducing manual annotation artifacts. The pipeline addresses two main issues: irregular boundaries and small disconnected fragments. Each issue is addressed with a dedicated preprocessing step described below.

Morphological Closing Ground-truth liver masks contain small intra-organ holes and irregular boundary fragments because annotators delineate each 2-D slice independently. The resulting label volumes frequently exhibit jagged inter-slice boundaries, small cavities inside the organ, and disconnected fragments at peripheral slices where the liver tapers. Training directly on such noisy labels causes the model to reproduce annotation artifacts rather than the true smooth boundary, degrading both quality and volumetric consistency.

To address this issue, morphological closing (a dilation followed by erosion with the same structuring element) is applied to training masks. This operation fills intra-organ voids and smooths boundary indentations without substantially displacing the outer surface, yielding cleaner targets that better reflect true anatomy.

Two variants of morphological closing were investigated to understand where the label-cleaning step provides the most benefit:

- The first variant applies closing only to training data. The training-split masks are smoothed while validation and test masks remain unmodified. The model learns from smoothed labels but is evaluated against original annotations. This isolates the effect of cleaner training supervision. If the model learns more coherent shapes from smoothed targets, its predictions should be more regular at inference without post-processing. Improved training-time label quality should translate into higher DSC values and reduced boundary error.
- The second variant applies closing to both training and validation data. Both splits use smoothed labels, ensuring consistent label quality throughout training and validation. This produces more stable validation scores since the reference

no longer penalizes smooth, anatomically plausible predictions for annotation-level imperfections. Consistent quality across splits also reduces the risk of selecting checkpoints that overfit annotation noise.

Throughout the Results chapter, each of the two label-cleaning variants above is reported under two evaluation settings. The Before setting computes metrics directly on the raw predicted masks. The After setting applies the same per-slice morphological closing operation as a post-processing step to the predicted masks prior to computing the metrics.

In both variants, closing is applied per slice. Each 2-D axial slice is processed independently using a 3×3 square structuring element, which confines influence to the immediate neighborhood and avoids merging adjacent structures. The number of iterations ranges from 10 to 20 depending on the dataset, reflecting differences in annotation granularity and slice thickness. As shown in Figure 8, this approach produces cleaner training targets that better reflect true anatomy.

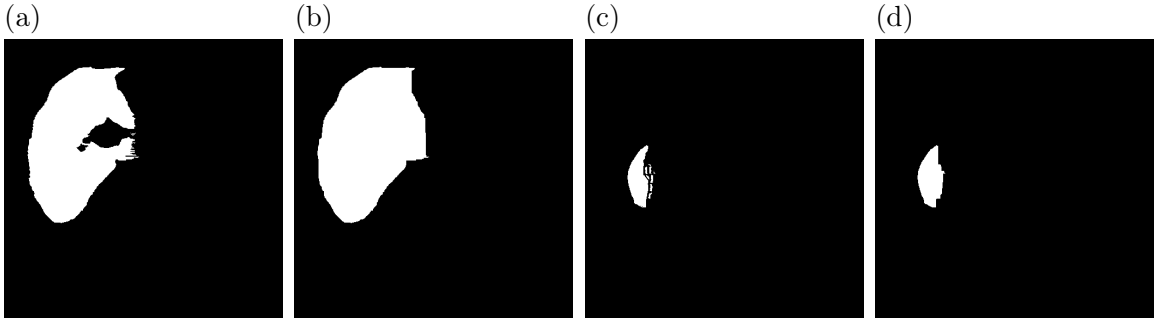


Figure 8: Examples of Morphological Closing Applied to Ground-Truth Liver Mask. The Images Show (a) Liver Mask with Intra-Organ Hole Before Applying Closing, (b) Liver Mask After Applying Closing to Fill Intra-Organ Hole, (c) Liver Mask with Minor Boundary Fragments Before Closing, and (d) Liver Mask After Applying Closing to Merge Fragments with the Main Region.

Minor Fragment Removal Ground-truth masks contain small, isolated pixel groups spatially disconnected from the main liver on each slice. These fragments arise from two sources. First, at the superior and inferior poles where the organ occupies few pixels per slice, annotators sometimes mark isolated patches not contiguous with the main body. Second, adjacent structures with similar intensity could be mislabeled. Training on masks with fragments teaches the model to reproduce disconnected predictions, increasing false positives and degrading precision.

To address this issue, fragment removal is applied per slice using connected-component analysis with 8-connectivity. Any component with area less than 10 percent of the largest connected component (LCC) is discarded. The LCC and any secondary component meeting this criterion are retained. The 10% threshold is conservative: large enough to eliminate spurious micro-fragments that cause training noise, yet small enough to preserve genuine secondary tissue detachments due to anatomical variation.

The relative (percentage-based) formulation is preferred over an absolute voxel-count threshold because it adapts automatically to variations in liver size across subjects,

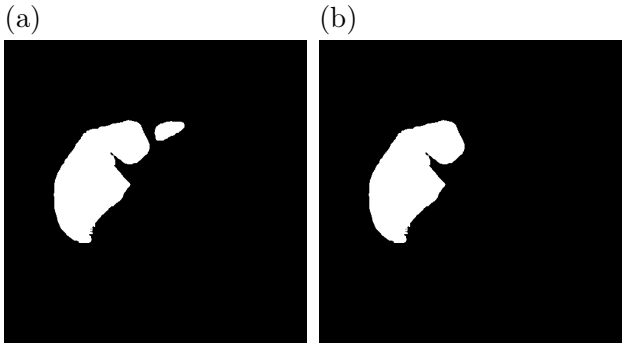


Figure 9: Example of Minor Fragment Removal from Ground-Truth Liver Mask. The Images Show (a) Liver Mask with an Isolated Minor Fragment, and (b) Liver Mask After Removing the Minor Fragment.

imaging resolutions, and modalities. An absolute threshold would require separate tuning per dataset and per voxel spacing, whereas the 10% relative rule generalizes across the heterogeneous dataset suite used in this thesis. By providing the model with cleaner, more topologically consistent training labels, fragment removal reduces false-positive predictions and improves both precision and volumetric DSC, as demonstrated in Figure 9. Together with morphological closing, it constitutes a lightweight but effective label-cleaning pipeline applied uniformly to all training-split ground-truth masks before any model is trained.

4.1.1.2 Model Development

With clean training labels in place, the segmentation model is built using an encoder-decoder architecture. The encoder extracts hierarchical feature representations from the input image while the decoder reconstructs a spatial segmentation map from those representations. This two-part design enables the model to combine the benefits of both detailed spatial information and high-level semantic understanding.

Encoder: ResNet18 ResNet18 [72], pretrained on ImageNet [73], serves as the encoder backbone. This choice addresses the scarcity of annotated medical imaging data. Training from scratch on a small medical dataset risks overfitting and slow convergence. ImageNet pretraining provides general-purpose visual representations useful for medical tasks: low-level edge and texture detectors in early layers, and progressively abstract object-part representations in deeper layers. Although natural and medical images differ substantially, these learned features transfer well to segmentation and enable efficient decoder training.

- ResNet18 is built on residual (shortcut) connections, where each block computes $\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}$. Here, $\mathcal{F}(\mathbf{x})$ is a stack of convolutional, batch normalization, and ReLU layers, and $\mathbf{x}$ is the identity shortcut. This residual formulation allows gradients to flow through many layers without vanishing. The architecture consists of an initial 7×7 convolutional layer, max pooling, and four residual stages ($\text{conv}2_x$ through $\text{conv}5_x$). Each stage progressively doubles feature channels (64, 128, 256, 512) and halves spatial resolution. With approximately 11 million parameters, ResNet18 is considerably smaller than ResNet50 or ResNet101, reducing overfitting risk when training data is limited.

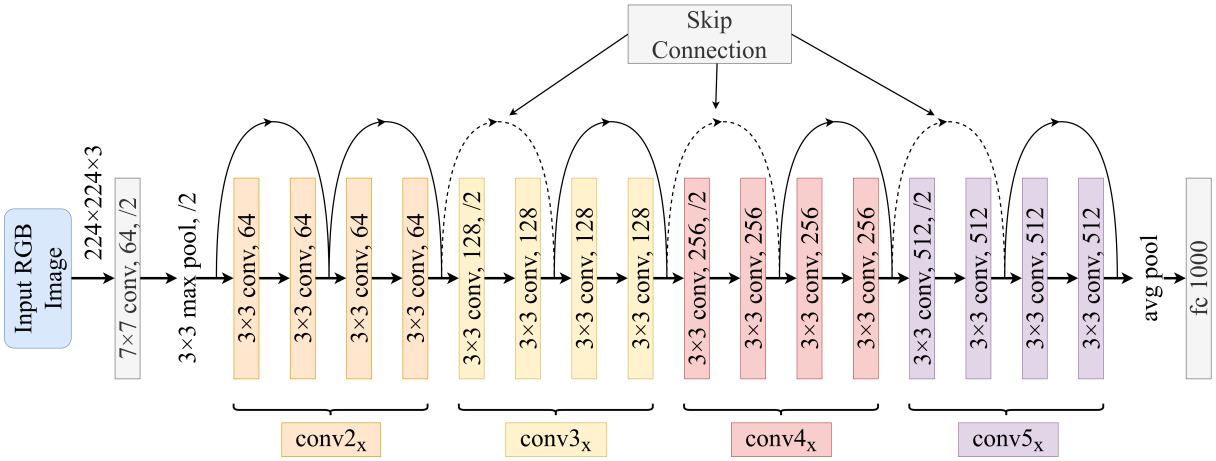


Figure 10: ResNet18 Architecture Showing the Encoder Backbone with Residual Blocks Progressively Downsampling Spatial Resolution Through conv2$_x$ to conv5$_x$ Stages. Feature Maps Extracted at Each Stage Serve as Skip Connections to The Decoder.

- For integration into a segmentation pipeline, feature maps are extracted from multiple intermediate stages simultaneously rather than using only the final stage output. The spatial feature maps produced after conv2$_x$, conv3$_x$, conv4$_x$, and conv5$_x$ are each forwarded to the decoder as skip connections. This multi-scale extraction is essential because different stages encode complementary information. Shallower stages preserve fine-grained spatial detail and boundary localization cues, while deeper stages encode semantically rich, context-aware representations required for distinguishing liver tissue from visually similar neighboring structures.

- The standard ResNet18 is designed for three-channel RGB input. Medical images are typically single-channel grayscale, so the single input channel is replicated across all three input channels of the first convolutional layer, enabling the use of pretrained weights without any architectural modification.

- Transfer learning is central to this design. The ResNet18 encoder is frozen during training; no gradients flow through it and no parameters are updated. This frozen-backbone strategy has two benefits. First, ImageNet pretraining encodes useful visual primitives including edges, textures, and object parts that transfer to organ boundary detection. Retraining from scratch on small datasets would overwrite this knowledge, risking overfitting and slower convergence. Second, freezing substantially reduces trainable parameters, which matters given limited annotated data. During training, only the decoder and encoder-decoder projection layers are updated via gradient descent.

The ResNet18 architecture shown in Figure 10 extracts features at multiple resolution levels. By combining features from different scales through skip connections, this approach preserves both the fine spatial details needed for accurate boundary localization and the semantic information necessary to distinguish the liver from neighboring tissues. The use of ImageNet-pretrained weights further enhances this multi-scale representation without requiring retraining from scratch.

Decoder: U-Net U-Net [74] is a well-established architecture for medical image segmentation. The network consists of a contracting path that extracts features while reducing

spatial resolution, and a symmetric expanding path that progressively recovers full image resolution. The key innovation is that high-resolution features from the contracting path are combined with the upsampled features through concatenation, allowing the network to make precise predictions while using contextual information. This architecture addresses the fundamental trade-off between localization accuracy and the use of image context.

- The contracting path follows the standard convolutional network pattern, consisting of repeated applications of two 3×3 convolutions (unpadded), each followed by ReLU and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels doubles. The expanding path mirrors this structure: each step consists of an upsampling of the feature map followed by a 2×2 up-convolution that halves the number of feature channels, then concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions each followed by ReLU. The cropping is necessary because the unpadded convolutions reduce the spatial extent of feature maps at each layer. The concatenation of features from the contracting path to the expanding path is essential to the U-Net architecture. Without it, the expanding path would have to reconstruct fine details from only the downsampled bottleneck representation. These skip connections carry spatial and boundary information that enables sharp localization even when tissue contrast is subtle. The symmetric design with concatenation allows the network to combine context from deeper layers with fine spatial details from shallower layers.
- In the proposed architecture, the frozen ResNet18 encoder plays the role of the contracting path, and the U-Net-style decoder serves as the expanding path. Since the encoder parameters are not updated during training, the decoder must learn to interpret the fixed encoder feature maps using only its own trainable parameters. The skip connections from encoder to decoder are particularly valuable here, providing rich spatial cues at every resolution level that would otherwise be difficult for the decoder to recover from only the bottleneck representation.
- At the final layer, a 1×1 convolution maps the decoder features to multiple channels, one per class. A softmax activation function generates probability distributions over the class labels at each pixel location. The final segmentation mask is produced by taking the argmax over these probabilities, assigning each pixel to its most likely class.

The U-Net decoder shown in Figure 11 is well suited for this encoder-decoder based architecture. By reusing encoder features at every resolution level through skip connections, the decoder can combine high-level semantic information from the frozen ResNet18 backbone with fine-grained spatial details needed for accurate organ boundaries. This architecture allows precise segmentation despite having a limited number of trainable parameters in the decoder alone.

4.1.2 Tumor segmentation within liver ROI

The second stage operates on hepatic ROIs to focus tumor segmentation on the liver. During training, ground-truth liver masks define the ROI, while during validation, Stage 1

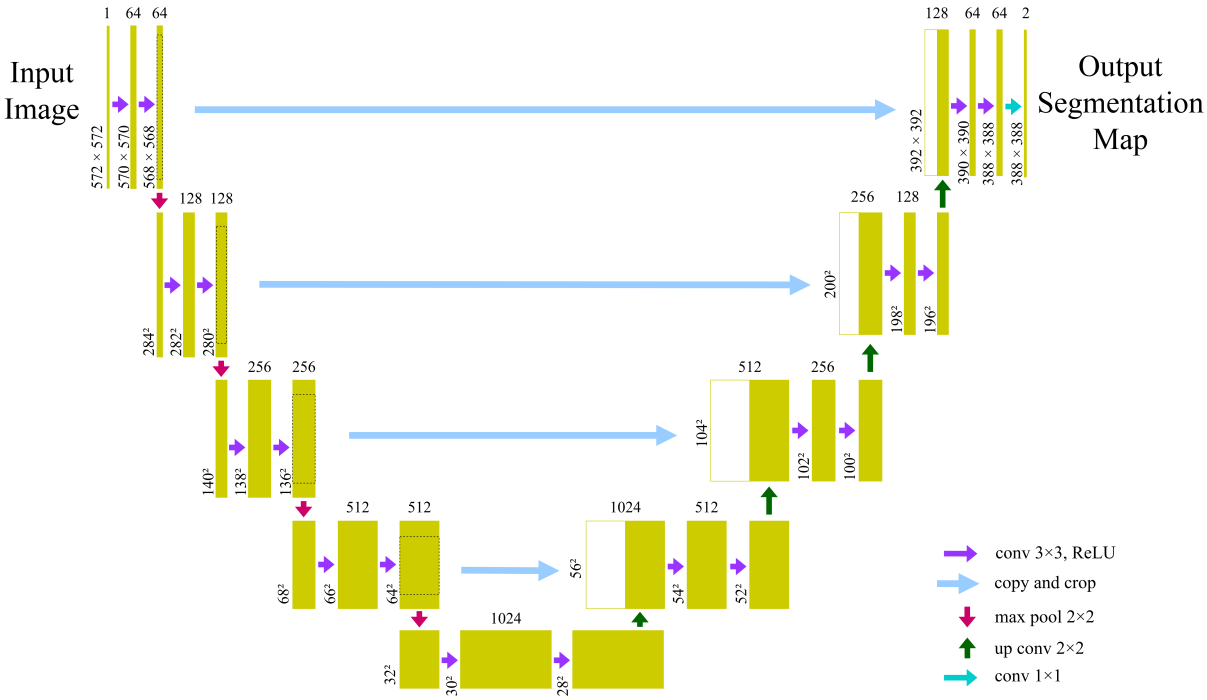


Figure 11: U-Net Decoder Architecture Showing Progressive Upsampling with Skip Connections from The Encoder. Features are Upsampled at Each Stage and Concatenated with Corresponding Encoder Outputs at Matching Resolution Levels.

predicted masks are used. This distinction reflects realistic deployment conditions where ground-truth masks are unavailable. Restricting the model to the hepatic parenchyma addresses extreme class imbalance: tumors occupy only a small fraction of total volume. Processing a hepatic crop rather than the full slice reduces computational cost and eliminates false positives outside the liver. A main limitation is error propagation from Stage 1: under-segmented livers may exclude peripheral lesions. The ROI extraction strategies below use sufficient margin around predicted boundaries to mitigate this risk.

4.1.2.1 Liver ROI Extraction

For ROI extraction, ground-truth liver masks are used during training and Stage 1 predicted masks are used during validation. This constrains the tumor search to the liver region. Since the Stage 1 model produces masks at its internal resolution, predictions are resized to match original image dimensions before extraction to ensure correct correspondence to native scan coordinates.

Two extraction strategies were designed and tested:

- The fixed center-based strategy computes the liver mask centroid per slice and places a 224×224 pixel box symmetrically around it (112 pixels in each direction). This provides consistent input dimensions, simplifying network design. However, when liver extent exceeds the box (large livers or small field of view), the fixed crop truncates tissue. In such cases, extraction falls back to a tight bounding box from the actual mask extent, plus 10 pixel padding to retain perihepatic context and

reduce artifacts. As shown in Figure 12(a), this approach ensures consistent input dimensions while preserving anatomical context.

- The morphologically dilated strategy dilates the mask by 10% using itself as the structuring element. The element is an isotropic version of the mask resized to 10 percent of its original area, ensuring dilation scales with organ size. A tight bounding box is computed from the dilated mask. Crucially, after cropping, only pixels within the dilated mask are retained as ROI; others are set to background. This differs fundamentally from the first method, which supplies the full rectangular region. By supplying only liver-interior pixels, this approach eliminates background and perihepatic tissue, reducing false positives while preserving the full liver boundary, as shown in Figure 12(b). The dilated box simply defines the crop extent; the masked output, not the full box, goes to tumor segmentation.

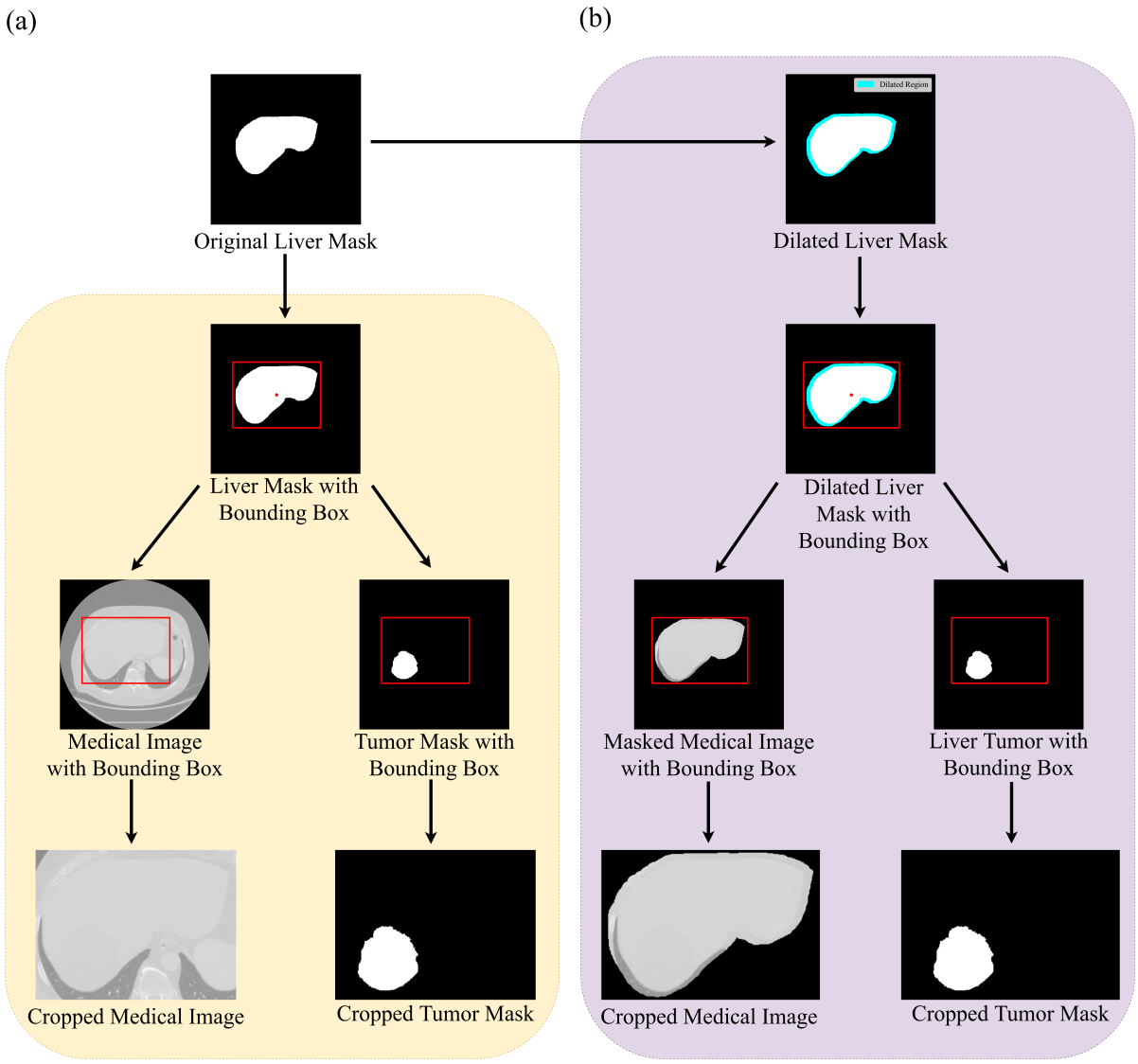


Figure 12: ROI Extraction Strategies for Tumor Segmentation. (a) Fixed Center-Based ROI Extraction Using a 224×224 Bounding Box Centered on the Liver Mask Centroid, with Fallback to Tight Bounding Box When Liver Extent Exceeds the Box. (b) Dilated Mask-Based ROI Extraction Using Morphological Dilation to Define Spatial Extent, Retaining Only Pixels within the Dilated Mask to Eliminate Background Tissue.

Both extraction strategies are evaluated with ground-truth liver masks during training and Stage 1 model predictions during validation. The respective ROI crops are used as inputs to the tumor segmentation model. The choice between strategies is treated as an experimental variable, and the quantitative effect on downstream tumor segmentation performance is reported in the results chapter.

4.1.2.2 Preprocessing

Based on the cross-domain CT-to-CT evaluation reported in the results chapter, LiTS was selected as the primary dataset for the tumor segmentation stage. The LiTS-trained liver model demonstrated the strongest cross-domain CT-to-CT generalization, making it the natural foundation for both stages of the pipeline.

- HU Windowing addresses limited contrast between hepatic lesions and liver parenchyma in CT imaging. Tissue appearance is characterized by HU: water is 0 HU, air is approximately negative 1000 HU. Liver parenchyma exhibits HU values within the soft-tissue range, while hepatic tumors exhibit varying HU depending on composition, vascularity, and acquisition phase. Hypovascular lesions like cysts appear as lower HU values, while hypervascular lesions may match or exceed parenchymal HU in arterial-phase scans. When CT volumes with their full intensity range are compressed into network inputs, tumor-to-parenchyma differences occupy only a small fraction of the input scale, effectively burying discriminative cues. Initial experiments on raw HU data confirmed this limitation: models produced imprecise boundaries and missed low-contrast lesions. This motivated systematic investigation of windowed representations.

  Windowing concentrates contrast resolution on clinically relevant tissue. Input intensity is clipped to a sub-interval $(HU_{\min}, HU_{\max})$ and rescaled to 0-255. Values below $HU_{\min}$ are clamped to $HU_{\min}$; values above $HU_{\max}$ are clamped to $HU_{\max}$. The clamped values map linearly to 0-255. Structures outside the window collapse to boundary values rather than being discarded, avoiding hard discontinuities. To identify optimal tumor-parenchyma contrast for LiTS, four window settings were evaluated, forming a 2 by 2 design space:

  - Negative 100 to 400 HU (500 HU span) encompasses the full soft-tissue spectrum: hypo-attenuating lesion cores through vascularized parenchyma. The lower bound slightly below zero retains cystic content sensitivity. This wide baseline enables comparison with narrower windows.
  - Negative 100 to 800 HU (900 HU span) extends the upper bound to include calcified foci and arterial-phase hyperenhancement. The wider span reduces per-unit soft-tissue contrast but retains highly attenuating components, trading soft-tissue discrimination for completeness.
  - Negative 20 to 400 HU (420 HU span) tightens both bounds, excluding air-containing structures and deeply hypodense content. Normal hepatic parenchyma (50-60 HU) is centered near the input range, where gradient-based learning is most sensitive. This placement maximizes discrimination for tumors with HU close to parenchyma.

– Negative 20 to 800 HU (820 HU span) combines the tight lower cutoff with the extended upper bound, balancing soft-tissue discrimination with coverage of attenuating structures.

Varying the lower bound (negative 100 versus negative 20) tests whether sensitivity to low-density cystic or necrotic content improves or harms overall segmentation. Varying the upper bound (400 versus 800) tests whether retaining high-attenuation vascular and calcified structures assists or distracts the model. All configurations apply the same clamp-and-rescale procedure, mapping the clamped HU values to (0, 255). The best-performing configuration is identified from the tumor segmentation experiments in the results chapter and carried forward to all subsequent experiments, including the 2.5-D slabbing study described below. As shown in Figure 13, the contrast variations across different window settings guide the selection of optimal windowing parameters for downstream experiments.

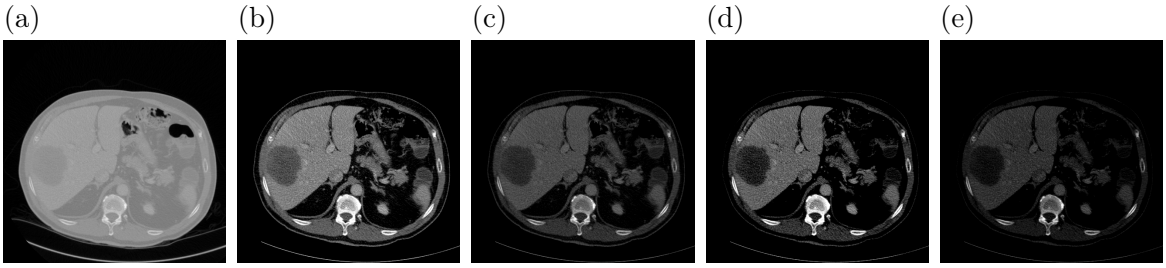


Figure 13: Comparison of HU Window Configurations for Tumor Contrast Enhancement. The Images Show (a) Original CT Scan, (b) HU Window of −100 to 400, (c) HU Window of −100 to 800, (d) HU Window of −20 to 400, and (e) HU Window of −20 to 800.

- 2.5-D Multi-Slice Input (Slabbing) enables the model to exploit inter-slice context in volumetric CT. Standard 2-D segmentation processes each axial slice independently, discarding adjacent-slice information. However, tumor appearance is strongly correlated across adjacent slices, and the three-dimensional lesion extent provides structural continuity cues that single-slice processing cannot access. Fully 3-D networks capture this context directly but require substantially greater computation. A 2.5-D approach offers a practical middle ground: the three input channels are populated with intensity information from multiple slices centered on the target, encoding local volumetric context without requiring a fully 3-D architecture.

  Two slabbing strategies were designed and evaluated using the best-performing HU window. In both strategies, each input channel is formed by averaging a symmetric neighborhood of slices centered on the target slice t, with channel assignment following the RGB convention to preserve compatibility with the pretrained ResNet18 encoder.

  – Narrow neighborhood slab (Strategy 1) uses progressively wider local averages in successive channels. The red channel preserves the target slice itself (slice t with no averaging), capturing fine spatial detail. The green channel provides a smoothed view by averaging three consecutive slices (t-1, t, t+1), which suppresses artifacts and reinforces structures consistent across adjacent slices. The blue channel extends the context by averaging five consecutive slices (t-2 to t+2), capturing gradual tumor attenuation changes and boundary continuity. The three channels together form a local multi-scale pyramid. The network exploits this structure through standard 2-D convolutions, amplifying

coherent structures while attenuating noise. Tumor ground-truth masks are processed identically. For Strategy 1, masks combine the full five-slice range $(t-2, t-1, t, t+1, t+2)$ aligned with the blue channel. This produces a composite mask encoding the lesion extent across the neighborhood.

– Wide neighborhood slab (Strategy 2) substantially expands the neighborhood window. The red channel preserves the windowed intensity of slice t (no averaging). The green channel averages eleven consecutive slices (t-5 to t+5). The blue channel averages twenty-one consecutive slices (t-10 to t+10). Wide averaging suppresses slice-specific noise more aggressively than narrow slabs. It provides a globally integrated view of the tumor volume along the axial direction, which improves boundary delineation for large lesions spanning many slices. The trade-off is significant: averaging over many slices blurs fine anatomical detail and reduces discriminability for small, punctate lesions. Therefore, Strategy 2 complements Strategy 1. The narrow-neighborhood slab is more informative for fine boundary detail in small lesions. The wide-neighborhood slab offers more stable context for larger, diffuse tumors. For masks, Strategy 2 combines the full twenty-one-slice range $(t-10, \ldots, t, \ldots, t+10)$, aligned with the blue channel. The wider neighborhood captures lesion continuity over a larger span, providing volumetric ground-truth context for large tumors.

Both strategies are applied on a per-slice basis. For target slice t, the required neighboring slices are accessed from the dataset, averaged as specified, and assembled into a three-channel tensor. At the volume boundaries (superior and inferior poles), where fewer than the required slices are available, zero-padding is applied. Missing slices are replaced with zero-valued slices to maintain uniformity. Zero-padding is applied identically to the tumor masks, with missing slices at boundary positions replaced with zero-valued mask slices. This parallel mask-slabbing pipeline ensures that supervision targets incorporate the same inter-slice information as inputs, enabling the model to learn the relationship between multi-slice representations and multi-slice ground truth. The parallel processing of CT intensities and tumor annotations is illustrated in the Figure 14.

4.1.2.3 Model Development

The tumor segmentation model employs the same frozen pretrained ResNet18 encoder and U-Net-style decoder used in the liver segmentation stage. No structural modification is introduced; details of the architectural design, rationale for the ImageNet-pretrained frozen backbone, decoder block structure, and projection mechanism at the encoder-decoder interface are as described in the Liver Segmentation section above. The 2.5-D slabbing strategies described in the Preprocessing section produce a three-channel input that is immediately compatible with the ResNet18 first-layer configuration. The sole operational distinction from Stage 1 is the supervision target: whereas Stage 1 is trained on binary liver masks, Stage 2 is trained on the LiTS tumor annotations to produce binary lesion probability maps within the hepatic ROI crop delivered by Stage 1.

4.1.2.4 Domain Adaptation

The HU windowing configurations evaluated above were derived from the LiTS CT dataset and expressed as absolute HU boundary values. Applying these thresholds directly to other datasets is problematic for two reasons. First, intensity ranges are not uniform across subjects within the same dataset, so a fixed absolute boundary may occupy markedly different positions within the tumor-to-parenchyma contrast region for different subjects. Second, intensity statistics shift substantially across datasets due to differences in scanner manufacturer, field strength, imaging protocol, and contrast phase. An absolute HU cutoff optimized on LiTS cannot be expected to isolate the clinically relevant hepatic window when applied to data from a different source.

To generalize windowing to new datasets without manual re-tuning, three approaches were tested: (1) no adaptation, (2) fixed-window applied uniformly to all targets, and (3) ratio-based transfer normalizing boundaries to each subject's intensity distribution. The third approach provides robustness to inter-subject and inter-dataset variability and is the primary contribution described here.

For each of the four boundary values (negative 20, negative 100, 400, 800 HU), a subject-level ratio is computed: boundary value added to the absolute volume minimum, then divided by total intensity range. This is expressed in Equation 4.1:

$$r_{i,b} = \frac{|S_{\min,i}| + b}{S_{\max,i} - S_{\min,i}}, \tag{4.1}$$

where b is the signed HU boundary value, $S_{\min,i}$ and $S_{\max,i}$ are the minimum and maximum intensities for subject i, and $r_{i,b}$ is in the range $(0, 1)$ representing the boundary position as a fraction of the full dynamic range. Since CT volumes have $S_{\min,i} < 0$, the expression $|S_{\min,i}|+b$ is algebraically equivalent to $b-S_{\min,i}$: the linear distance from volume minimum to the boundary value. To obtain population-level anchors robust to individual variability, the minimum ratio $r_{\min,b}$ across training subjects is computed for lower boundaries (negative 20, negative 100 HU) ensuring conservative windowing that does not exclude low-HU lesions. Similarly, the maximum ratio $r_{\max,b}$ is computed for upper boundaries (400, 800 HU) ensuring inclusive windowing that preserves the full extent of clinically relevant tissue. Population-level anchor ratios are presented in Table 4.1.

Table 4.1: Population-level anchor ratios for HU window domain adaptation.

Boundary Type	HU value	Population-level Ratio
Lower (minimum)	−20 HU	$r_{\min} = 18.61\%$
	−100 HU	$r_{\min} = 17.88\%$
Upper (maximum)	400 HU	$r_{\max} = 79.42\%$
	800 HU	$r_{\max} = 89.11\%$

When adapting to a new dataset, the volume-level minimum and maximum intensities $[S_{\min}, S_{\max}]$ of each target subject are computed. For lower-boundary values, the adapted boundary is recovered using the minimum population-level ratio $r_{\min,b}$. For upper-boundary values, the maximum ratio $r_{\max,b}$ is used. Equation 4.2 gives the adapted boundary value:

$$b_{\text{adapted}} = r_{\text{pop},b} \times (S_{\max} - S_{\min}) - S_{\min}, \tag{4.2}$$

where $r_{\text{pop},b}$ denotes the appropriate population-level ratio. Correctness can be verified by substituting Equation 4.1 back in; the $(S_{\max} - S_{\min})$ terms cancel exactly, confirming

perfect round-trip recovery. To avoid false precision and preserve conservative intent, an asymmetric rounding rule is applied. Lower window boundaries are rounded down using $\lfloor \cdot \rfloor$; upper boundaries are rounded up using $\lceil \cdot \rceil$. This ensures adapted windows remain at least as inclusive as population-level anchors, preserving sensitivity to boundary-region tissue.

4.2 Training Protocol

This section documents the training setup used for all experiments: loss function, optimizer, learning rate schedule, batch size, epoch count, and checkpointing strategy.

Loss Function Both pipeline stages are trained using the multi-class soft Dice loss. DSC is a region-overlap metric measuring the ratio of twice the intersection to the sum of predicted and ground-truth regions. Maximizing DSC is equivalent to minimizing 1 minus DSC, the training objective used here.

In the multi-class setting, the model produces separate probability maps per class via softmax over raw outputs. Softmax ensures predicted probabilities across all classes sum to one at every pixel. Dice loss is computed independently per class and averaged uniformly, as shown in Equation 4.3.

$$L_{\text{Dice}} = 1 - \frac{1}{C} \sum_{c=1}^{C} \frac{2 \sum_{i} p_{i,c}\, g_{i,c} + \varepsilon}{\sum_{i} p_{i,c} + \sum_{i} g_{i,c} + \varepsilon}, \tag{4.3}$$

In Equation 4.3, $p_{i,c}$ is the predicted probability that pixel i belongs to class c, $g_{i,c}$ is the corresponding one-hot ground-truth label, and ε is a small smoothing constant that prevents division by zero when a class is absent from a given image and improves numerical stability throughout training.

The soft formulation is essential for gradient-based learning. A hard binary DSC computed from thresholded predictions is piecewise constant and almost everywhere non-differentiable, providing no gradient signal. Soft Dice loss operates on continuous predicted probabilities, yielding dense, well-defined gradients at every pixel and training step.

Averaging Dice loss equally across all classes means each class contributes equal weight regardless of pixel frequency. This makes soft Dice loss inherently robust to class imbalance. Rare foreground classes like liver lesions, occupying small fractions of scan volume, receive the same gradient contribution as the dominant background class. This directly addresses the severe foreground-background disparity characteristic of organ and tumor segmentation tasks.

Optimizer All models use the Adam optimizer [75] with initial learning rate 0.0001. Adam maintains per-parameter estimates of first and second gradient moments. The first moment (exponentially weighted moving average of past gradients) acts as momentum,

smoothing updates and accelerating convergence. The second moment (exponentially weighted moving average of squared gradients) estimates gradient variance, enabling per-parameter adaptive scaling. This makes Adam well-suited to the frozen-encoder, trainable-decoder configuration.

Only the decoder and encoder-decoder projection layers are updated during training. These newly initialized layers exhibit heterogeneous gradients across roles. Adam's per-parameter scaling automatically adjusts step sizes without manual tuning. Momentum accelerates learning in consistently reinforced directions, reducing updates needed for decoder initialization to converge to useful segmentation.

Learning Rate Schedule The initial learning rate is annealed using a cosine schedule. The rate at epoch e decays smoothly from the initial rate to a minimum of 0.000001 at the final epoch, following one half-period of a cosine curve.

The cosine schedule is preferred over step-wise or exponential decay. Step-wise decay introduces abrupt reductions that can destabilize training at inopportune moments. Exponential decay causes the rate to fall too quickly early and become negligible before convergence. Cosine decay avoids both issues: it decays slowly early (keeping learning rate high while the decoder learns broad structure), then accelerates decay in middle epochs before plateauing near convergence. This profile aligns with typical deep segmentation loss dynamics: rapid early progress followed by slow boundary refinement. The non-zero minimum learning rate ensures the optimizer retains ability to make corrective updates in final epochs rather than halting entirely.

Batch Size A fixed batch size of 32 slices is used across all datasets and both pipeline stages.

Training Epochs The number of training epochs varies by dataset, reflecting differences in size and convergence rates. LiverHCCSeg and CHAOS use 300 epochs. These datasets are comparatively small, requiring longer schedules to allow cosine decay to reach its minimum and to prevent premature convergence before the model observes sufficient training variation. LiTS uses 50 epochs. LiTS is substantially larger, so the model encounters more absolute training samples per epoch. Preliminary experiments showed validation loss plateaued and early stopping triggered well before 300 epochs, making a shorter budget appropriate. In all cases, the cosine schedule period matches the dataset-specific epoch count so the learning rate schedule is fully utilized.

Checkpointing and Early Stopping A single best-model checkpoint is saved by monitoring validation loss. The filename encodes epoch number, validation IoU, and validation loss. Training terminates early if validation loss does not improve by at least 0.0001 for 10 consecutive epochs, preventing overfitting and reducing unnecessary computation once convergence occurs.

4.3 Evaluation Framework

Both the liver and tumor segmentation models are evaluated using four standard metrics measuring complementary accuracy aspects. These metrics are computed separately for liver and tumor classes on held-out test sets unseen during training or validation, ensuring unbiased generalization assessment.

4.3.1 DSC

DSC is a region-overlap metric widely used in medical image segmentation. DSC measures the similarity between predicted and ground-truth segmentations by computing twice the number of correctly predicted foreground pixels divided by the sum of all predicted foreground pixels and all ground-truth foreground pixels. This metric is particularly useful for imbalanced segmentation tasks because it treats foreground and background symmetrically, preventing the metric from being dominated by the larger background class. Unlike pixel-count accuracy metrics, DSC remains meaningful even when the target structure occupies a small fraction of the image.

DSC is computed as shown in Equation 4.4:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{4.4}$$

DSC ranges from 0 to 1, with 1 indicating perfect overlap. The numerator counts TP twice because it corresponds to the intersection in both the predicted and ground-truth sets, providing a symmetric formulation that balances both false positives and false negatives.

The soft Dice loss used during training (Equation 4.3) is the complement of this metric applied to probabilistic outputs. By optimizing soft Dice loss, the network maximizes DSC on validation data. This alignment between training objective and evaluation metric makes DSC a natural choice for final performance measurement.

4.3.2 IoU

IoU, also called the Jaccard Index, measures spatial overlap by computing the ratio of correctly predicted foreground pixels to the union of predicted and ground-truth foreground regions. Unlike DSC, which counts the intersection twice to create a symmetric weighting, IoU counts the intersection once in the numerator and the union in the denominator. This makes IoU stricter than DSC for the same segmentation error: both false positives and false negatives are penalized equally and without the symmetric doubling present in DSC, making IoU more sensitive to localization errors.

IoU is defined in Equation 4.5:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{4.5}$$

IoU ranges from 0 to 1, with 1 indicating perfect overlap. The denominator represents the union of predicted and ground-truth foreground regions, so any misprediction, whether an incorrect voxel included or a true voxel missed, directly reduces the metric. This makes IoU particularly useful for assessing boundary delineation precision.

IoU and DSC are mathematically related through $\text{IoU} = \frac{\text{DSC}}{2-\text{DSC}}$, so DSC is always greater than or equal to IoU for the same segmentation. Both metrics are reported because IoU is sensitive to localization accuracy and provides complementary information about boundary precision.

4.3.3 Precision

Precision measures the proportion of predicted positive pixels that are actually positive in ground truth. This metric quantifies the false-positive rate and addresses the clinical need to determine the proportion of identified lesions that are real. High precision is important in medical imaging because false positives can lead to unnecessary clinical intervention, additional imaging studies, or patient anxiety. In lesion detection, high precision means the model is conservative and confident in its predictions.

Precision is computed as shown in Equation 4.6:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4.6}$$

Precision ranges from 0 to 1, with 1 indicating all predicted positives are correct. Note that precision does not account for false negatives; a model can achieve perfect precision by predicting only a few highly certain pixels as foreground, which is clinically unacceptable.

Precision and recall represent a fundamental trade-off. A model can achieve high precision by making only confident predictions and rejecting marginal cases, but it may miss true instances and lower recall. The balance between precision and recall is adjusted through the decision threshold applied to model output probabilities.

4.3.4 Recall

Recall, also called sensitivity or the true positive rate, measures the proportion of ground-truth foreground pixels that are correctly predicted. This metric quantifies the false-negative rate and addresses the clinical need to determine the proportion of actual lesions that the model successfully detects. High recall is essential in medical imaging because false negatives can lead to missed diagnoses, delayed treatment, and adverse patient outcomes. In clinical practice, recall is often prioritized over other metrics to ensure pathological findings are not overlooked.

Recall is computed as shown in Equation 4.7:

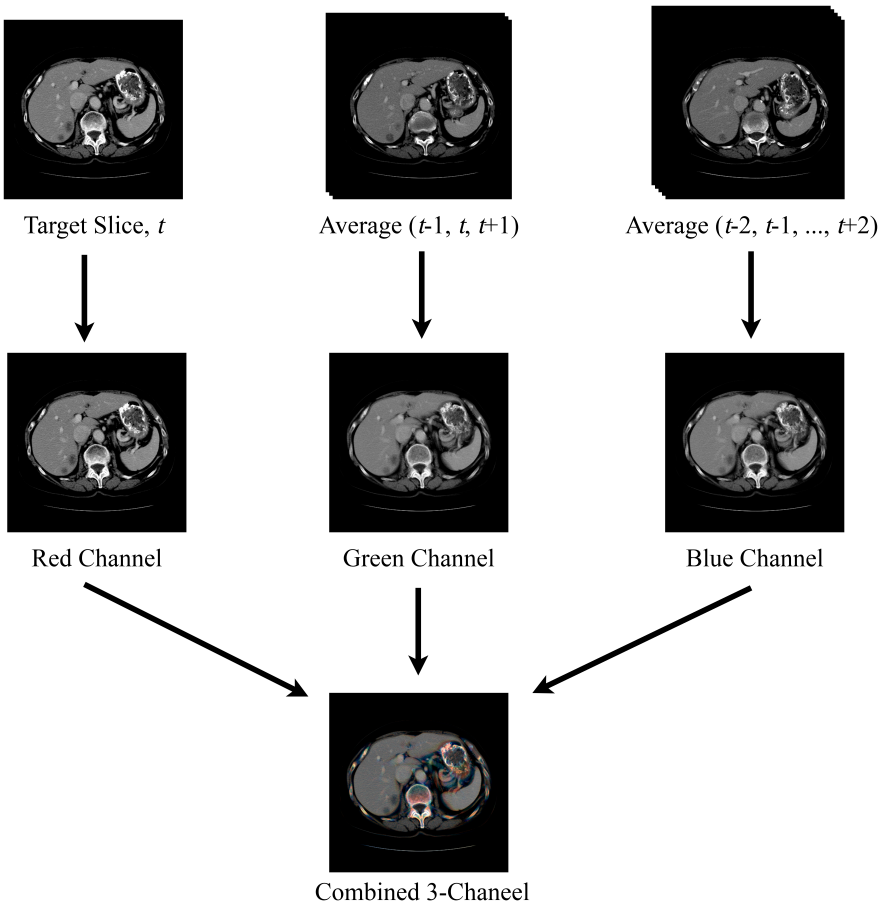
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4.7}$$

Recall ranges from 0 to 1, with 1 indicating all ground-truth foreground pixels are detected. Note that recall does not account for false positives; a model can achieve perfect recall by predicting everything as foreground, which produces an unusable high false-positive rate.

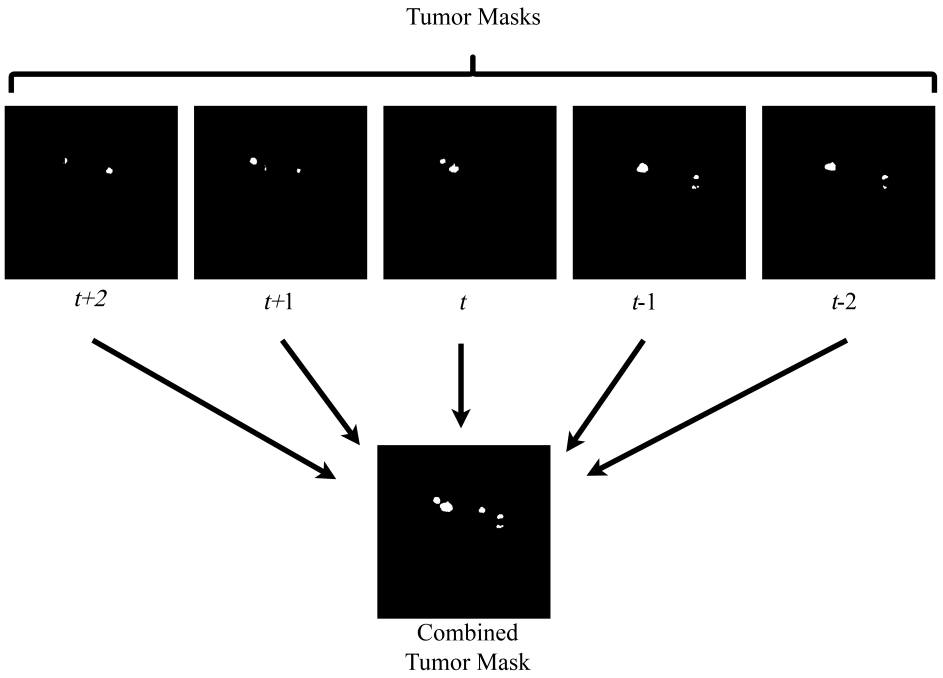
The complementary nature of precision and recall is important for understanding model behavior. A model can achieve perfect recall by predicting all pixels as foreground, but this results in poor precision. Conversely, a model can achieve perfect precision by predicting very few pixels as foreground, but this results in poor recall.

In all the aforementioned metrics, TP denotes true positives, which are correctly predicted foreground pixels. FP denotes false positives, representing pixels incorrectly predicted as foreground. Finally, FN denotes false negatives, which correspond to foreground pixels incorrectly predicted as background.

(a)



(b)



(a) Tumor mask with five-slice range ($t-2$ to $t+2$).

Figure 14: Example Images from 2.5-D Slabbing Strategies. The Images Show (a) a Narrow Neighborhood Slabbing for Medical Images, and (b) Narrow Neighborhood Slabbing for Tumor Masks.

Chapter-5: Results and Analysis

5.1 Liver Segmentation

5.1.1 LiverHCCSeg as the Primary Dataset

The liver segmentation begins with LiverHCCSeg as the primary training dataset. Table 5.1 shows that the baseline model achieves a competitive in-domain DSC of 0.925 on the held-out test split. This strong performance is expected since both training and test data come from the same MRI acquisition distribution. However, an important limitation applies: LiverHCCSeg contains only 17 patients. The small size resulted in the held-out test set also serving as the validation set. Consequently, model selection via early stopping uses the same data instances that are later reported as test performance. This differs from CHAOS and LiTS, which maintain distinct validation and test sets. Despite this limitation, external evaluation reveals clear modality bias in the model's generalization. In the cross-domain MRI-to-MRI setting, the model generalizes reasonably well to CHAOS (DSC 0.804). However, it struggles considerably on cross-modal MRI-to-CT benchmarks. Performance on 3D-IRCADb-01 (0.788) and SLiver07 (0.769) remains moderate, while LiTS (0.560) shows a more pronounced domain gap. This degradation on LiTS is consistent with the large variability characteristic of the LiTS challenge corpus.

Applying morphological closing only to the training ground truth produces small gains on two CT benchmarks. On 3D-IRCADb-01, DSC reaches 0.821 in the After variant, which is a 4.2% improvement over the baseline of 0.788. On SLiver07, DSC reaches 0.806 in the Before variant, which is a 4.8% improvement over the baseline of 0.769. Precision follows the same pattern. On 3D-IRCADb-01, precision increases from 0.696 to 0.753 in both variants. On SLiver07, precision increases from 0.662 to 0.726 in the Before variant. These gains suggest that cleaner training masks encourage tighter liver predictions with fewer false-positive fragments. In-domain performance remains stable at DSC 0.926 in the After variant. LiTS does not benefit from this strategy, where DSC drops from 0.560 at baseline to 0.512 in the After variant, which is an 8.6% relative decrease. This suggests that smoothing MRI training labels does not provide useful guidance for learning the CT boundary patterns and texture variability present in LiTS.

When closing is applied to both train and test sets, the CT results degrade sharply relative to the train-only setting. In the train-only setting, the model is trained on morphologically closed training masks, while early stopping is performed on the original, unmodified validation annotations, which for LiverHCCSeg correspond to the held-out

Table 5.1: Liver Segmentation Performance Across Preprocessing Strategies (LiverHCCSeg)

Preprocessing	Variant	Dataset	DSC	IoU	Precision	Recall
None	—	Test (40%)	**0.925**	**0.864**	**0.943**	0.913
		CHAOS (MR)	0.804	0.676	0.847	0.772
		LiTS (CT)	0.560	0.412	0.430	0.906
		3D-IRCADb-01 (CT)	0.788	0.657	0.696	0.927
		SLiver07 (CT)	0.769	0.644	0.662	**0.960**
Closing applied on train only	Before	Test (40%)	0.920	0.854	**0.953**	0.893
		CHAOS (MR)	0.785	0.652	0.853	0.737
		LiTS (CT)	0.512	0.371	0.386	0.916
		3D-IRCADb-01 (CT)	0.820	0.700	0.753	0.911
		SLiver07 (CT)	0.806	0.694	0.726	0.937
	After	Test (40%)	**0.926**	**0.865**	0.952	0.905
		CHAOS (MR)	0.796	0.667	0.855	0.755
		LiTS (CT)	0.512	0.371	0.386	0.920
		3D-IRCADb-01 (CT)	0.821	0.703	0.753	0.917
		SLiver07 (CT)	0.804	0.691	0.721	**0.941**
Closing applied on both train & test	Before	Test (40%)	0.923	0.859	**0.926**	0.921
		CHAOS (MR)	0.779	0.644	0.818	0.753
		LiTS (CT)	0.229	0.141	0.142	0.983
		3D-IRCADb-01 (CT)	0.471	0.319	0.321	0.983
		SLiver07 (CT)	0.416	0.276	0.277	0.992
	After	Test (40%)	**0.928**	**0.867**	0.925	0.933
		CHAOS (MR)	0.791	0.661	0.822	0.772
		LiTS (CT)	0.228	0.140	0.141	0.985
		3D-IRCADb-01 (CT)	0.469	0.317	0.319	0.985
		SLiver07 (CT)	0.413	0.273	0.274	**0.993**
10% minor fragment removed	—	Test (40%)	**0.918**	**0.854**	**0.946**	0.898
		CHAOS (MR)	0.793	0.661	0.810	0.789
		LiTS (CT)	0.335	0.221	0.224	0.958
		3D-IRCADb-01 (CT)	0.672	0.522	0.535	0.965
		SLiver07 (CT) 2	0.622	0.478	0.484	**0.979**

test split. This keeps checkpoint selection aligned with the annotation style used when evaluating external datasets. When these validation annotations are also closed, early stopping selects checkpoints against a morphologically smoothed target definition, while cross-dataset evaluation remains scored against unmodified ground truth. The resulting performance values are striking. On LiTS, DSC drops to 0.228 with Precision 0.141 and Recall 0.985. Similarly, 3D-IRCADb-01 drops to DSC 0.469 with Precision 0.319 and Recall 0.985, and SLiver07 drops to DSC 0.413. This pattern of near-unity recall and very low precision indicates severe over-segmentation. In contrast, cross-domain MRI performance on CHAOS changes only slightly under this strategy, from 0.779 to 0.791 across the Before and After variants. These results suggest that, for LiverHCCSeg, checkpoint selection should remain anchored to the original, unmodified validation annotations when cross-dataset generalization is a priority.

The fragment removal strategy (removing the smallest 10% of disconnected components) uniformly degrades cross-modal CT generalization on the LiverHCCSeg model. LiTS

reaches only DSC 0.335 (40.2% degradation from baseline). 3D-IRCADb-01 reaches 0.672 (14.7% degradation). SLiver07 reaches 0.622 (19.1% degradation). The strategy fails because the removal threshold incorrectly discards genuine liver structures rather than spurious artifacts. When a model encounters domain shift, it often produces fragmented but spatially correct predictions. Aggressive fragment removal in this scenario amplifies rather than corrects the domain gap.

In summary, the LiverHCCSeg-trained model demonstrates satisfactory in-domain performance but is limited by the dataset's small scale. Its cross-domain MRI-to-MRI generalization to CHAOS is reasonable, but cross-modal MRI-to-CT generalization remains limited, and no evaluated preprocessing strategy fully resolves this gap.

5.1.2 CHAOS as the Primary Dataset

Training on CHAOS yields the highest baseline in-domain DSC across all three primary datasets at 0.944, as reported in Table 5.2. This strong performance reflects the benefit of a dedicated train/validation/test split and the greater patient diversity of CHAOS compared to LiverHCCSeg. The baseline cross-modal CT generalization is mixed when evaluated against the LiverHCCSeg model. CHAOS outperforms on LiTS (DSC 0.613 versus 0.560, approximately 9.5% higher) but falls below on 3D-IRCADb-01 (0.751 versus 0.788, approximately 4.7% lower) and SLiver07 (0.730 versus 0.769, approximately 5.1% lower). This pattern suggests that the advantage of the larger and more diverse CHAOS cohort is most pronounced on LiTS, the most heterogeneous CT benchmark. Despite both CHAOS and LiverHCCSeg being MRI datasets, cross-domain MRI-to-MRI generalization from CHAOS to LiverHCCSeg is limited at baseline (DSC 0.631). This limited performance may reflect differences in MRI acquisition protocol, field strength, or anatomical variability between the two cohorts.

Applying closing to the training ground truth only yields consistent and universally positive improvements across all external evaluation sets. LiverHCCSeg reaches 0.766 (After), approximately 21.4% improvement over the baseline of 0.631. LiTS reaches 0.670, approximately 9.3% improvement over 0.613. 3D-IRCADb-01 reaches 0.806, approximately 7.3% improvement over 0.751. SLiver07 reaches 0.795, approximately 8.9% improvement over 0.730. All values exceed their corresponding baseline results, while the test performance remains competitive at 0.940 (After). The consistent benefit is observed across both cross-domain MRI-to-MRI and cross-modal MRI-to-CT external evaluation sets. It demonstrates that smoother and more anatomically coherent supervision targets improve the model's ability to generalize liver boundary representations to unseen domains. This supports the hypothesis that annotation artifacts in training labels introduce noise that confounds cross-dataset transfer.

Extending closing to the validation set (closing applied on both train and validation) leads to a regression in cross-modal performance on CT benchmarks relative to the train-only variant. LiTS drops from 0.670 to 0.609 (After), approximately 9.1% degradation. 3D-IRCADb-01 drops from 0.806 to 0.767, approximately 4.8% degradation. SLiver07 drops from 0.795 to 0.749, approximately 5.8% degradation. The in-domain test DSC shows a minor improvement at 0.943 (After), but this does not offset the cross-modal CT losses. The degradation occurs because the early stopping and hyperparameter selection

Table 5.2: Liver Segmentation Performance Across Preprocessing Strategies (CHAOS)

Preprocessing	Variant	Dataset	DSC	IoU	Precision	Recall
None	—	Test (20%)	**0.944**	**0.894**	**0.934**	**0.955**
		LiverHCCSeg (MR)	0.631	0.494	0.841	0.545
		LiTS (CT)	0.613	0.470	0.724	0.583
		3D-IRCADb-01 (CT)	0.751	0.610	0.794	0.740
		SLiver07 (CT)	0.730	0.597	0.805	0.684
Closing applied on train only	Before	Test (20%)	0.938	0.883	**0.940**	0.936
		LiverHCCSeg (MR)	0.765	0.629	0.789	0.762
		LiTS (CT)	0.666	0.519	0.623	0.751
		3D-IRCADb-01 (CT)	0.799	0.672	0.795	0.815
		SLiver07 (CT)	0.793	0.672	0.786	0.809
	After	Test (20%)	**0.940**	**0.888**	0.929	**0.952**
		LiverHCCSeg (MR)	0.766	0.630	0.786	0.766
		LiTS (CT)	0.670	0.524	0.629	0.753
		3D-IRCADb-01 (CT)	0.806	0.682	0.806	0.818
		SLiver07 (CT)	0.795	0.675	0.788	0.811
Closing applied on both train & validation	Before	Test (20%)	0.937	0.881	**0.950**	0.925
		LiverHCCSeg (MR)	0.741	0.606	0.864	0.673
		LiTS (CT)	0.605	0.459	0.590	0.656
		3D-IRCADb-01 (CT)	0.760	0.620	0.778	0.760
		SLiver07 (CT)	0.748	0.611	0.771	0.735
	After	Test (20%)	**0.943**	**0.892**	0.942	**0.944**
		LiverHCCSeg (MR)	0.743	0.608	0.862	0.677
		LiTS (CT)	0.609	0.463	0.595	0.658
		3D-IRCADb-01 (CT)	0.767	0.629	0.789	0.763
		SLiver07 (CT)	0.749	0.612	0.772	0.737
10% minor fragment removed	—	Test (20%)	**0.942**	**0.891**	**0.935**	**0.951**
		LiverHCCSeg (MR)	0.776	0.646	0.865	0.725
		LiTS (CT)	0.709	0.579	0.829	0.663
		3D-IRCADb-01 (CT)	0.818	0.695	0.852	0.794
		SLiver07 (CT)	0.788	0.674	0.867	0.740

are against a preprocessed annotation distribution rather than the original one. The best-selected checkpoint captures a model tuned to smoothed MRI validation boundaries but less representative of the true underlying distribution. This causes suboptimal generalization when tested against original unmodified ground truth on external datasets. It aligns with hyperparameter tuning constraint observed in the LiverHCCSeg results: when validation annotations are modified, model selection optimizes for a different target distribution than the baseline case, introducing a confound that degrades cross-domain generalization.

The fragment removal strategy produces the strongest cross-modal CT performance from the CHAOS model. On LiTS, it achieves DSC 0.709 (15.8% improvement over baseline). On 3D-IRCADb-01, it reaches DSC 0.818 (9.0% improvement). On SLiver07, it reaches DSC 0.788 (8.0% improvement). All three values exceed the corresponding closing-only approach. For cross-domain MRI performance on LiverHCCSeg, the strategy achieves DSC 0.776, a 23.0% improvement over the baseline of 0.631.

The underlying mechanism differs from LiverHCCSeg. For CHAOS, fragment removal helps by eliminating small, spatially disconnected false-positive predictions that arise from domain shift. For LiverHCCSeg, the same strategy proved harmful because it incorrectly discarded genuine liver structures. This indicates that fragment removal effectiveness depends on the specific model and data.

5.1.3 LiTS as the Primary Dataset

With LiTS as the primary training dataset, the baseline in-domain test DSC of 0.940 is comparable to the CHAOS baseline of 0.944, as reported in Table 5.3. Both large-scale datasets support competitive in-domain models. However, CT-to-CT cross-domain generalization differs vastly between them. The LiTS model achieves near-perfect performance on other CT benchmarks: 3D-IRCADb-01 reaches DSC 0.975 and SLiver07 reaches 0.936 under no preprocessing. These performance levels represent far superior performance compared to the LiverHCCSeg and CHAOS models (3D-IRCADb-01: 0.975 versus 0.788 and 0.751 respectively, 23.7% and 29.8% better; SLiver07: 0.936 versus 0.769 and 0.730 respectively, 21.7% and 28.2% better). This superior cross-domain CT-to-CT generalization reflects the large and diverse CT acquisition landscape represented by LiTS. The dataset encompasses a wide range of scanner characteristics, contrast protocols, and anatomical variability that transfer well to other CT benchmark datasets.

Cross-modal performance on MRI presents a contrasting picture. Generalization to CHAOS MRI is extremely poor at baseline with DSC 0.196. This performance is characterized by high Precision (0.841) but near-zero Recall (0.125). This pattern shows that the model produces very few positive liver predictions on MRI volumes. The model effectively does not activate on MRI tissue. However, the sparse predictions it does produce are spatially correct. This complete collapse reflects the CT-to-MRI modality shift. CT features calibrated to HU values bear no resemblance to MRI signal intensities. Generalization to LiverHCCSeg MRI is more moderate at 0.617. This reason is unclear, but certain MRI sequences may contain contrast patterns that partially overlap with features learned from CT liver boundaries.

Applying closing to the training ground truth only, with post-prediction closing applied (the After variant), yields the best in-domain test performance among all LiTS configurations at DSC 0.946 with Precision 0.934 and Recall 0.963. CT external test performance reaches 0.976 on 3D-IRCADb-01 and 0.951 on SLiver07, which represents near-ceiling performance. Most strikingly, cross-modal MRI generalization to LiverHCCSeg improves to DSC 0.763 compared to 0.617 at baseline, a 23.7% improvement. Training on morphologically refined CT masks induces spatially smoother prediction behavior that partially transfers to MRI domains. Although anatomical boundaries differ in texture across modalities, they maintain spatial coherence. The CHAOS cross-modal performance remains essentially unchanged at 0.229. This reflects the fundamental CT-to-MRI intensity gap that preprocessing liver mask cannot bridge. Importantly, the LiTS validation annotations remain unmodified under this strategy. This ensures that early stopping optimizes against the original annotation distribution, maintaining consistency between validation and test evaluation.

Extending closing to the validation set (closing applied on both train and validation) produces apparent gains on cross-modal MRI benchmarks. LiverHCCSeg improves to 0.791

Table 5.3: Liver Segmentation Performance Across Preprocessing Strategies (LiTS)

Preprocessing	Variant	Dataset	DSC	IoU	Precision	Recall
None	—	Test (10%)	0.940	0.892	0.932	0.955
		LiverHCCSeg (MR)	0.617	0.500	0.976	0.507
		CHAOS (MR)	0.196	0.123	0.841	0.125
		3D-IRCADb-01 (CT)	**0.975**	**0.952**	**0.977**	**0.973**
		SLiver07 (CT)	0.936	0.902	0.974	0.924
Closing applied on train only	Before	Test (10%)	0.937	0.886	0.925	0.956
		LiverHCCSeg (MR)	0.758	0.650	0.960	0.666
		CHAOS (MR)	0.228	0.143	0.803	0.147
		3D-IRCADb-01 (CT)	0.963	0.928	0.958	0.967
		SLiver07 (CT)	0.946	0.910	0.958	0.944
	After	Test (10%)	0.946	0.901	0.934	0.963
		LiverHCCSeg (MR)	0.763	0.657	0.961	0.672
		CHAOS (MR)	0.229	0.145	0.798	0.149
		3D-IRCADb-01 (CT)	**0.976**	**0.952**	**0.976**	**0.975**
		SLiver07 (CT)	0.951	0.920	0.962	0.950
Closing applied on both train & validation	Before	Test (10%)	0.934	0.880	0.921	0.953
		LiverHCCSeg (MR)	0.786	0.694	0.956	0.715
		CHAOS (MR)	0.465	0.328	0.909	0.344
		3D-IRCADb-01 (CT)	0.963	0.929	0.960	0.967
		SLiver07 (CT)	0.961	0.927	0.960	0.963
	After	Test (10%)	0.942	0.895	0.930	0.961
		LiverHCCSeg (MR)	0.791	0.702	0.957	0.723
		CHAOS (MR)	0.469	0.332	0.909	0.348
		3D-IRCADb-01 (CT)	**0.977**	**0.955**	**0.978**	**0.976**
		SLiver07 (CT)	0.967	0.938	0.966	0.969
10% minor fragment removed	—	Test (10%)	0.942	0.895	0.936	0.955
		LiverHCCSeg (MR)	0.581	0.465	0.971	0.474
		CHAOS (MR)	0.285	0.184	0.910	0.189
		3D-IRCADb-01 (CT)	**0.975**	**0.952**	**0.975**	**0.976**
		SLiver07 (CT)	0.943	0.909	0.974	0.932

(After), approximately 3.6% improvement over the train-only variant of 0.763. CHAOS improves to 0.469 (After), approximately 104.8% improvement over the train-only variant of 0.229. However, these apparent gains come at a cost. In-domain performance drops slightly from DSC 0.946 to 0.942. More importantly, an internal inconsistency emerges: the best checkpoint is selected based on performance against modified validation annotations, while final test evaluation uses unmodified ground truth. The CHAOS improvement warrants caution. Both Precision and Recall increase substantially (Precision from 0.798 to 0.909, Recall from 0.149 to 0.348). However, this behavior indicates that the checkpoint was selected for a different annotation distribution than the test set uses. The apparent cross-modal improvement may not reflect genuine generalization gains.

The 10% minor fragment removal strategy maintains strong CT cross-domain performance (3D-IRCADb-01: 0.975, SLiver07: 0.943) but substantially degrades cross-modal MRI generalization. LiverHCCSeg drops to 0.581, 5.8% below the no-preprocessing baseline and 23.9% below the selected approach. CHAOS reaches only 0.285. The problem is

clear: the fragment removal threshold, calibrated for CT predictions, incorrectly removes fragmented but genuine liver predictions on MRI. When models encounter domain shift, they often produce spatially correct but structurally fragmented predictions. Removing these fragments amplifies rather than corrects the domain gap. Consequently, this strategy is unsuitable when cross-modal coverage is required.

In summary, the LiTS-trained model with closing applied on train only, with post-prediction closing applied to predictions (the After evaluation), provides the strongest and most internally consistent combination of high in-domain performance, near-perfect CT-to-CT generalization, and meaningful cross-modal MRI improvement. It is accordingly selected as the liver segmentation model for Stage 2 ROI extraction, as discussed in Section 5.1.4.

5.1.4 Comparative Analysis and Model Selection

The three sets of liver segmentation experiments collectively characterize how training dataset choice, imaging modality, and preprocessing strategy interact to determine cross-domain and cross-modal generalization performance. This section synthesizes the key findings reported in Tables 5.1, 5.2, and 5.3, and provides the rationale for the final liver model selection used for Stage 2 ROI extraction.

5.1.4.1 Effect of Primary Training Dataset on Cross-Domain and Cross-Modal Generalization

The three primary datasets represent a clear and ordered progression in external generalization capacity across both cross-domain and cross-modal settings. The LiverHCCSeg model is limited by its small cohort of 17 patients and its MRI modality. It achieves reasonable in-domain performance but exhibits the weakest cross-modal MRI-to-CT generalization of the three models. LiTS DSC peaks at only 0.560 showing the substantial domain gap, and 3D-IRCADb-01 and SLiver07 reach at most 0.821 and 0.806 under the best strategy, as reported in Table 5.1. The CHAOS model demonstrates markedly stronger cross-modal MRI-to-CT generalization despite being MRI-trained, reflecting the benefit of a larger and more diverse training cohort. LiTS reaches 0.709 (approximately 26.6% higher than LiverHCCSeg's 0.560) and 3D-IRCADb-01 reaches 0.818 (approximately 0.3% lower than LiverHCCSeg's 0.821), as reported in Table 5.2. The LiTS model, however, achieves near-perfect cross-domain CT-to-CT transfer. 3D-IRCADb-01 approaches 0.977 (approximately 19.6% higher than CHAOS's 0.818) and SLiver07 approaches 0.967 (approximately 21.9% higher than CHAOS's 0.793), as reported in Table 5.3. These performance levels are entirely unmatched by either MRI-trained model. All three models exhibit a MRI-to-CT or CT-to-MRI generalization gap. The LiTS model produces near-zero CHAOS recall at baseline, consistent with a hard modality boundary. Yet the LiTS model still demonstrates moderate cross-modal CT-to-MRI generalization to LiverHCCSeg that is substantially recovered by training on morphologically smoothed labels.

5.1.4.2 Effect of Preprocessing Strategy Across Datasets

Across all three primary datasets, morphological closing applied exclusively to the training ground truth (closing applied on train only) often improves external performance relative to the no-preprocessing baseline, although the effect depends on the target dataset. This strategy ensures that hyperparameter tuning via early stopping occurs against the original, unmodified annotation distribution, which matches the test distribution, maintaining internal consistency between model selection and final evaluation. The benefit is observed across both cross-domain (CT-to-CT) and cross-modal (MRI-to-CT or CT-to-MRI) evaluation sets, most clearly for the CHAOS- and LiTS-trained models. However, the effect is not universal. When training on LiverHCCSeg, LiTS decreases from 0.560 at baseline to 0.512 under closing on train only, despite gains on other CT benchmarks. These results suggest that cleaner supervision targets can reduce the propagation of annotation artifacts into model predictions, but they do not fully resolve modality and cohort shifts. Extending closing to the validation set occasionally produces apparent additional gains on cross-domain and cross-modal benchmarks. However, these improvements reliably co-occur with either a reduction in in-domain performance or a fundamental shift in the validation annotation distribution used for hyperparameter tuning. When the validation distribution is altered, hyperparameter selection optimizes for a different target than the baseline case, introducing an experimental confound that undermines interpretability of performance changes.

The 10% minor fragment removal strategy is inconsistent across datasets. It provides the best cross-modal CT performance for the CHAOS model but proves harmful for the LiTS model's cross-modal MRI generalization. This demonstrates that the suitability of fragment-based post-processing is architecture and domain dependent and cannot be universally recommended.

5.1.4.3 Final Model Selection

The LiTS-trained model was selected as the final liver segmentation model for downstream liver ROI extraction. Specifically, the selected model was trained on morphologically closed ground truth masks but used unmodified validation annotations. I applied morphological closing as post-processing to the model's predictions (the After variant), because it achieved the highest in-domain liver DSC. This decision was supported by the following considerations, all grounded in quantitative evidence:

- Near-perfect CT-to-CT generalization: the selected model achieved DSC values of 0.976 on 3D-IRCADb-01 and 0.951 on SLiver07, as reported in Table 5.3. This substantially outperformed the best CHAOS-trained results of 0.818 and 0.795 (approximately 19.3% and 19.6% better) from Table 5.2 and the best LiverHCCSeg-trained results of 0.821 and 0.806 (approximately 18.9% and 17.9% better) from Table 5.1. Since CT acquisition constituted the primary application domain, this superiority was the dominant selection criterion.
- Best in-domain test set performance among all LiTS variants: the selected configuration achieved the highest in-domain test DSC among all LiTS experimental variants

at 0.946. This confirmed robust generalization to unseen LiTS test volumes under the chosen preprocessing and post-processing regime.

- Meaningful cross-modal MRI improvement: despite being a CT-trained model, the selected variant achieved DSC 0.763 on LiverHCCSeg MRI. This represented a relative improvement of approximately 23.7% over the no-preprocessing baseline of 0.617. This improvement came from training on morphologically smoothed labels, which induced spatially consistent prediction behavior that partially extended to MRI boundaries. While CHAOS cross-modal performance remained limited at 0.229 due to the fundamental CT-to-MRI domain gap, the LiverHCCSeg result demonstrated that the model could acquire meaningful MRI generalization from the selected training strategy.

The selected model (LiTS, Closing on train only) therefore represented the best-evidenced choice for generating liver masks and ROI crops for the downstream stage of the pipeline.

5.2 Tumor Segmentation

With the liver segmentation model selected, the second stage of the proposed pipeline focuses on tumor segmentation within the extracted liver ROI. This section examines the performance of U-Net based tumor segmentation models trained on the LiTS dataset. The investigation encompasses the effects of different strategies for extracting the ROI from predicted liver masks, different preprocessing approaches applied to input images, and different intensity normalization methods employed to handle datasets acquired under different protocols and scanner configurations. Evaluation encompasses performance on the in-domain LiTS test set as well as external CT and MRI datasets, providing insight into both the capabilities and limitations of cross-dataset transfer for hepatic lesion detection.

5.2.1 Performance on Primary Dataset: LiTS

The second stage of the pipeline focuses on tumor segmentation. This stage operates on liver-containing regions extracted by Stage 1. Several preprocessing methods are investigated that affect tumor segmentation accuracy on the LiTS dataset. Specifically, morphological closing as post-processing applied to the ROI mask, intensity windowing strategies for handling CT data, multi-slice approaches for volumetric context, and alternative ROI extraction method. For each strategy, which methods improve performance and which harm it is determined. Understanding these effects on in-domain data helps interpret how the models generalize to external datasets.

5.2.1.1 Effect of Morphological Closing on ROI Boundaries

The Stage 1 liver segmentation model achieved the highest performance with morphological closing applied as post-processing to predicted liver masks. Whether this same closing operation should be retained during Stage 2 tumor ROI extraction requires explicit

investigation. The results presented in Table 5.4 reveal an unexpected pattern where applying closing before ROI extraction reduces tumor segmentation performance.

Table 5.4: Liver Tumor Segmentation Performance on LiTS Across Closing on ROI Configurations

Closing applied on ROI	ROI from	DSC	IoU	Precision	Recall
Yes	Ground truth	0.251	0.163	0.420	0.220
	Prediction	0.237	0.152	0.436	0.200
No	Ground truth	**0.294**	**0.194**	**0.492**	**0.263**
	Prediction	0.260	0.168	0.449	0.254

Note. All experiments in this table use the fixed center-based bounding box (224×224 px) for liver ROI extraction.

The no-closing configuration with ground-truth ROI achieves DSC 0.294, IoU 0.194, Precision 0.492, and Recall 0.263, compared to DSC 0.251 with closing applied. This 14.6% degradation points to a mismatch between how the ROI is constructed for training crops and how it is constructed during validation-time checkpoint selection. In this experiment, morphological closing is not applied when constructing ROIs for the training split. Instead, it is applied in the validation pipeline during training, where validation performance is used for early stopping and checkpoint saving, and the same ROI construction procedure is then used at test time.

During Stage 2 training, the training ROI is derived from ground-truth masks without morphological modification. The procedure computes the mask centroid and anchors a fixed 224×224 pixel bounding box at this location. When morphological closing is applied to the liver masks used for validation ROI extraction, the spatial positioning of the extracted crop changes. Closing consists of dilation followed by erosion. The dilation expands the mask and fills holes, and the subsequent erosion can contract the boundary, particularly at organ boundaries. This combination changes the mask's center. The centroid computed from a closed mask differs from the centroid of the original unmodified mask. Since the bounding box is positioned relative to the centroid, any centroid shift moves the entire crop window.

This spatial repositioning creates a mismatch between training and the validation and inference pipelines. The model was trained to detect lesions within crops anchored to unmodified ground-truth centroids. During validation and test-time inference, the closing-induced centroid shift repositions the crop relative to liver anatomy. Lesions that were centrally positioned during training may now occupy peripheral regions. Since hepatic lesions frequently reside near organ boundaries, this shift can move tumor-bearing regions outside the crop window entirely. The model has not encountered lesions in these new spatial arrangements, resulting in false negatives and reduced performance.

A concrete example illustrates this problem in Figure 15. A sample test case shows the before-closing configuration achieving DSC 0.798 and the after-closing configuration degrading to DSC 0.637 (20.2% loss). Although both configurations use the same bounding box dimensions, the application of closing alters the spatial context during evaluation. The morphological operation reshapes mask boundaries and shifts the crop location, creating an ROI extraction distribution that does not match what the model encountered during

training. The visual comparison shows substantially more false negatives in the after-closing variant. This example validates the quantitative findings and demonstrates the critical importance of annotation consistency between training and inference.

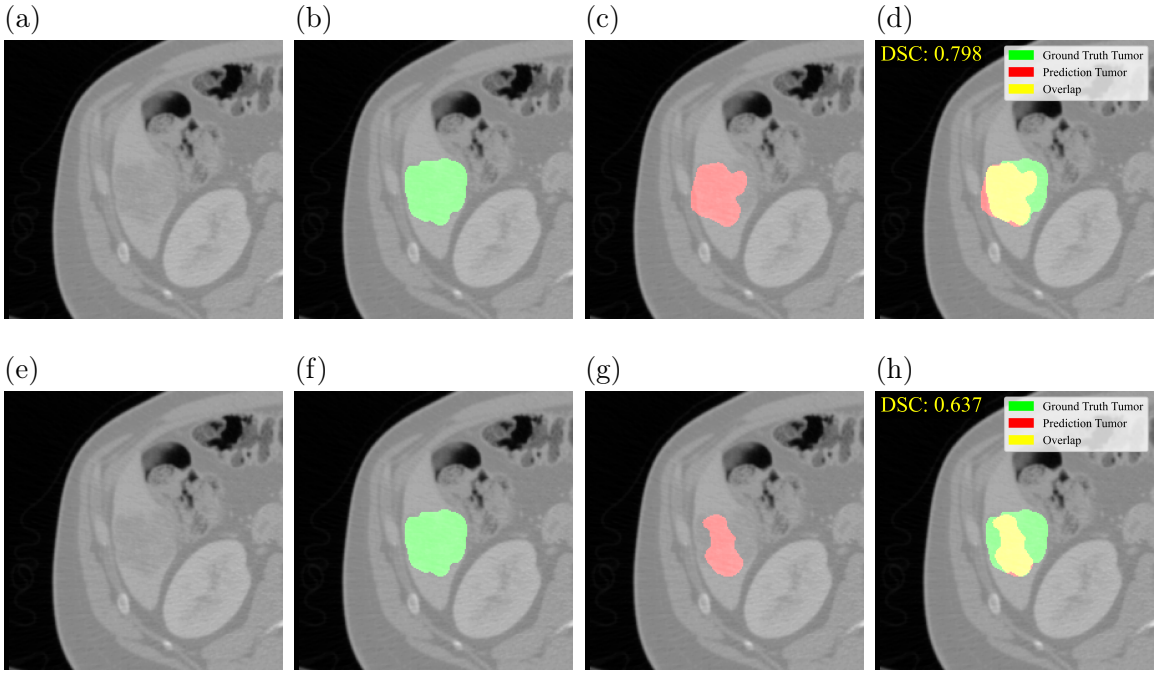


Figure 15: Effect of Morphological Closing on Tumor Segmentation. The Images Show (a) Cropped CT Using Predicted Liver ROI Before Closing, (b) Ground Truth Tumor Overlay Before Closing, (c) Predicted Tumor Overlay Before Closing, (d) Combined Ground Truth and Predicted Tumor Overlay Before Closing (DSC 0.798), (e) Cropped CT Using Predicted Liver ROI After Closing, (f) Ground Truth Tumor Overlay After Closing, (g) Predicted Tumor Overlay After Closing, and (h) Combined Ground Truth and Predicted Tumor Overlay After Closing (DSC 0.637).

Given these findings, morphological closing applied to the liver ROI mask prior to tumor ROI extraction is discarded. The non-closing approach, achieving DSC 0.294 with ground-truth ROI and DSC 0.260 with prediction ROI, serves as the baseline for subsequent experiments evaluating additional preprocessing choices and ROI extraction strategies.

5.2.1.2 Optimization of HU Windowing

Having established that morphological closing on the ROI should not be applied, the next variable under investigation is the intensity windowing applied to the CT input. The raw HU values across CT volumes span a large dynamic range, but only a narrow range is relevant for hepatic tumor detection. Table 5.5 reveals that constraining this range substantially improves segmentation performance across all four tested windowing configurations.

The best configuration applies a window of -20 to 400 HU to ground-truth ROI data, achieving DSC 0.433, IoU 0.304, Precision 0.509, and Recall 0.453. This represents a relative improvement of approximately 47.3% in DSC compared to the no-windowing baseline of 0.294 established in the previous experiment.

Table 5.5: Liver Tumor Segmentation Performance on LiTS Across HU Windowing Configurations

HU Window	ROI from	DSC	IoU	Precision	Recall
-100 to 400	Ground truth	0.398	0.280	0.473	0.423
	Prediction	0.389	0.274	0.464	0.428
-100 to 800	Ground truth	0.372	0.262	0.438	0.386
	Prediction	0.361	0.255	0.434	0.379
-20 to 400	Ground truth	**0.433**	**0.304**	**0.509**	0.453
	Prediction	0.427	0.299	0.503	**0.455**
-20 to 800	Ground truth	0.354	0.241	0.495	0.333
	Prediction	0.337	0.225	0.476	0.328

The fundamental benefit of windowing stems from the large dynamic range of unprocessed CT data. The full HU scale spans approximately 3285 HU on average across the LiTS cohort, yet the contrast between tumor tissue and surrounding parenchyma occupies only a narrow band within this range. When a CT volume presents its full intensity range compressed into the network input, the discriminative signal separating tumors from parenchyma occupies only a small fraction of the effective input scale. This signal becomes partially submerged by irrelevant intensities from structures outside the hepatic window. Windowing solves this problem by remapping the clinically relevant hepatic intensity range to the full input dynamic range. This amplifies the relative intensity contrast between parenchyma and lesion, providing the network with more discriminative input representation.

The superiority of the -20 to 400 HU window emerges when examining each boundary independently. The lower bound of -20 HU outperforms -100 HU because it more aggressively excludes deeply hypodense structures. Subcutaneous fat appears near -100 HU and fluid near 0 HU. These structures carry no discriminative information and introduce background variation. The -20 HU bound retains hypodense lesions while excluding confounding signals. The upper bound of 400 HU outperforms 800 HU because it excludes calcifications, hemorrhage, and bone while preserving hypervascular lesion variants and normal parenchyma. Extending to 800 HU reintroduces high-attenuation tissue, compressing the soft-tissue contrast range.

The impact of these windowing bounds is evident in a representative sample case shown in Figure 16. The ground truth annotation for this slice contains three distinct small tumor regions. With the optimal -20 to 400 HU window, the model successfully segments the central lesion with strong spatial overlap, achieving a DSC of 0.401, although the two smaller peripheral regions remain undetected. Modifying the lower bound to -100 HU introduces minor boundary imprecisions at the central lesion, reducing DSC to 0.383. Extending the upper bound to 800 HU produces more pronounced degradation. The -100 to 800 HU configuration yields a severely contracted prediction that overlaps only a small portion of the central lesion, resulting in DSC 0.234. When both bounds are extended to the -20 to 800 HU range, the model produces no detections on any target lesions, yielding a DSC of 0.000. This pattern illustrates a critical failure mode: extending the upper bound toward dense tissue compresses the soft-tissue contrast range to a small fraction of the

total dynamic range. This compression degrades the feature representation learned by the encoder, reducing its ability to distinguish parenchyma from lesion and preventing the decoder from recognizing valid tumor morphology.

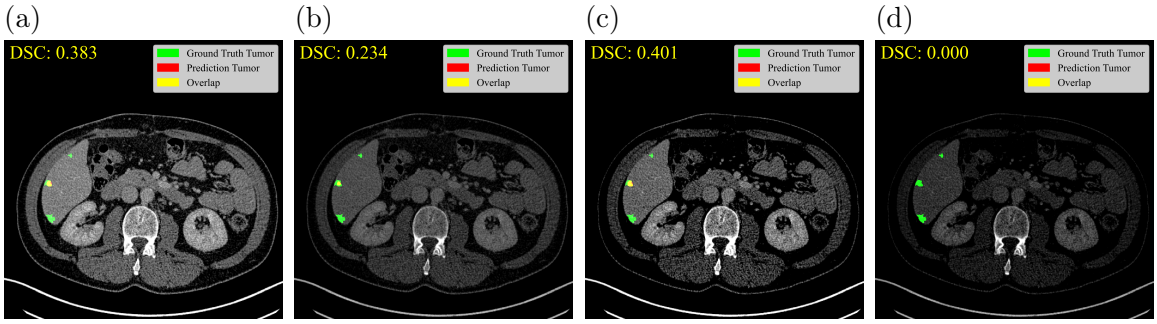


Figure 16: Effect of HU Windowing on Tumor Segmentation Using Ground Truth Liver as ROI. The Images Show Tumor Segmentation Results with Window Ranges of (a) −100 to 400 HU, (b) −100 to 800 HU, (c) −20 to 400 HU, and (d) −20 to 800 HU. Ground Truth Tumor Is Green, Predicted Tumor Is Red, and the Overlap of Ground Truth and Prediction Is Yellow.

A notable observation emerges in how ground-truth and predicted ROI results converge under optimal windowing. DSC differs by only 0.006 (1.4% relative difference), contrasting sharply with the 11.6% gap without windowing. This convergence shows that when input intensity contrast is sufficiently informative, the tumor model becomes largely robust to spatial imprecisions in the ROI crop introduced by Stage 1 errors. Based on these results, the −20 to 400 HU window is selected as the optimal configuration (DSC 0.433 with ground-truth ROI, DSC 0.427 with predicted ROI). This window is applied to all subsequent tumor segmentation experiments.

5.2.1.3 Evaluation of 2.5-D Multi-Slice Context Encoding

The tumor segmentation model operates independently on each axial slice, discarding all inter-slice structural context that could potentially assist in boundary delineation. The question of whether encoding adjacent anatomy through channel stacking can recover this volumetric information naturally arises. Table 5.6 evaluates two 2.5-D slabbing strategies that span different axial extents, both applied to the current best configuration: the non-closed fixed-crop ROI with −20 to 400 HU windowing.

Table 5.6: Liver Tumor Segmentation Performance on LiTS Across 2.5-D Slabbing Strategies

2.5-D Slabbing Strategy	ROI from	DSC	IoU	Precision	Recall
Narrow slab R: t; G: avg($t\pm1$); B: avg($t\pm2$)	Ground truth	**0.375**	**0.256**	**0.443**	0.394
	Prediction	0.373	0.255	0.435	0.400
Wide slab R: t; G: avg($t\pm5$); B: avg($t\pm10$)	Ground truth	0.300	0.198	0.290	0.432
	Prediction	0.296	0.196	0.287	**0.439**

Contrary to the hypothesis that multi-slice context improves tumor segmentation, both slabbing approaches produce degradation relative to the 2D windowed baseline. The narrow slab achieves DSC 0.375 and the wide slab achieves DSC 0.300, both lower than the baseline of 0.433. The narrow slab represents a 13.4% decrease, and the wide slab represents a 30.7% decrease.

Multiple mechanisms account for this degradation. Averaging over neighboring slices blurs high-frequency boundary information along the axial direction. The sharp tumor margins that the decoder relies on for precise localization are smoothed across multiple planes. This blurring is more severe in the wide slab configuration, where channels average slices up to ten positions away. This span covers several centimeters in typical acquisitions, encoding predominantly tumor-free parenchyma. Such anatomically inconsistent signals provide limited discriminative value.

An additional limitation comes from architectural mismatch. The frozen ResNet18 encoder was pre-trained on natural RGB images where three channels encode spectrally correlated information. Repurposing these channels to encode different axial depths violates this assumption. Pre-trained kernels produce suboptimal responses on depth-coded inputs, reducing feature quality. The wide-slab results show characteristic precision-recall imbalance: Ground-truth Precision collapses to 0.290 while Recall rises to 0.432. This pattern indicates spatially over-extended predictions where the model generates large diffuse activations that encompass tumors but substantially exceed true lesion boundaries.

Representative sample cases illustrate this behavior on clinically relevant tumors, as shown in Figure 17. A case with a single large hepatic lesion shows the differential constraints imposed by slabbing strategy. With the narrow slab, the model achieves DSC 0.858 with nearly complete true positive coverage of the lesion core while incurring predominantly false-positive errors at the lesion periphery. This aligns with the earlier analysis: neighboring slices provide sufficient information to identify the lesion core, yet the architectural constraints from repurposing RGB channels induce spurious activations in surrounding tissue. The wide slab degrades to DSC 0.760 as averaging of slices up to ten positions away smears boundary information. The ground truth in this case contains two tumor regions: one large and one small. While the model produces two predicted regions, the small region is spatially mislocalized compared to its ground truth location, resulting in a complete miss of the true small lesion positioned elsewhere. This spatial mislocation of small tumors combined with the degraded sensitivity to the large lesion core substantially reduces the overall agreement metric.

A second case with multiple small tumors reveals a more problematic failure mode, as shown in Figure 18. With the narrow slab, the model generates DSC 0.210, producing two predicted regions with only partial overlap to actual lesion anatomy and one false positive activation outside the liver boundary. The wide slab performs marginally better at DSC 0.262 with a single large predicted region showing partial overlap to one lesion while remaining totally undetected on the other two targets. These cases demonstrate that slabbing strategies amplify detection failure for small lesions, particularly those near the liver boundary where axial context becomes anatomically ambiguous. The architectural mismatch between the pre-trained encoder and the depth-coded input manifests as competing false-positive activations and degraded sensitivity to true lesion morphology.

Since neither approach improves over the 2D windowed baseline, the 2.5-D approach is not

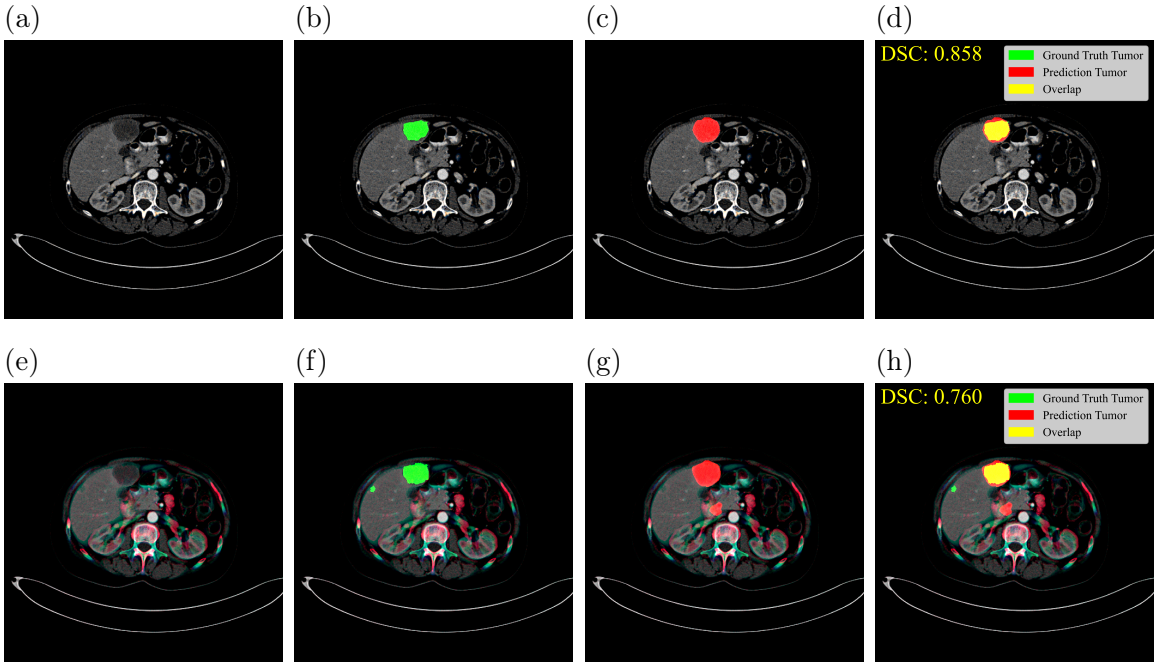


Figure 17: Effect of 2.5-D Slabbing on Tumor Segmentation with a Large Hepatic Lesion. The Images Show (a) CT Using Narrow Neighborhood Slab, (b) Ground Truth Tumor Overlay for Narrow Slab, (c) Predicted Tumor Overlay for Narrow Slab, (d) Combined Ground Truth and Predicted Tumor Overlay for Narrow Slab (DSC 0.858), (e) CT Using Wide Neighborhood Slab, (f) Ground Truth Tumor Overlay for Wide Slab, (g) Predicted Tumor Overlay for Wide Slab, and (h) Combined Ground Truth and Predicted Tumor Overlay for Wide Slab (DSC 0.760). Ground Truth Tumor Is Green, Predicted Tumor Is Red, and Overlap Is Yellow.

pursued further. The fixed-crop configuration with -20 to 400 HU windowing remains optimal. Effective multi-slice integration would require architectures that explicitly model the volumetric dimension, such as 3D convolutions or volumetric attention mechanisms, which remain beyond the current scope.

5.2.1.4 Alternative ROI Extraction Using Dilated Mask Bounding Box

The centroid-anchored fixed crop approach, while robust to Stage 1 prediction imperfections, inherently presents the tumor model with a rectangular bounding box that encompasses perihepatic anatomy beyond the organ boundary. This section investigates an alternative ROI extraction method that directly addresses this limitation by nullifying all pixels outside the liver boundary. Table 5.7 compares this masked approach against the best fixed-crop configuration established previously.

Table 5.7: Liver Tumor Segmentation Performance on LiTS Using Masked ROI

ROI from	DSC	IoU	Precision	Recall
Ground truth	**0.516**	**0.394**	**0.587**	**0.500**
Prediction	0.391	0.296	0.438	0.422

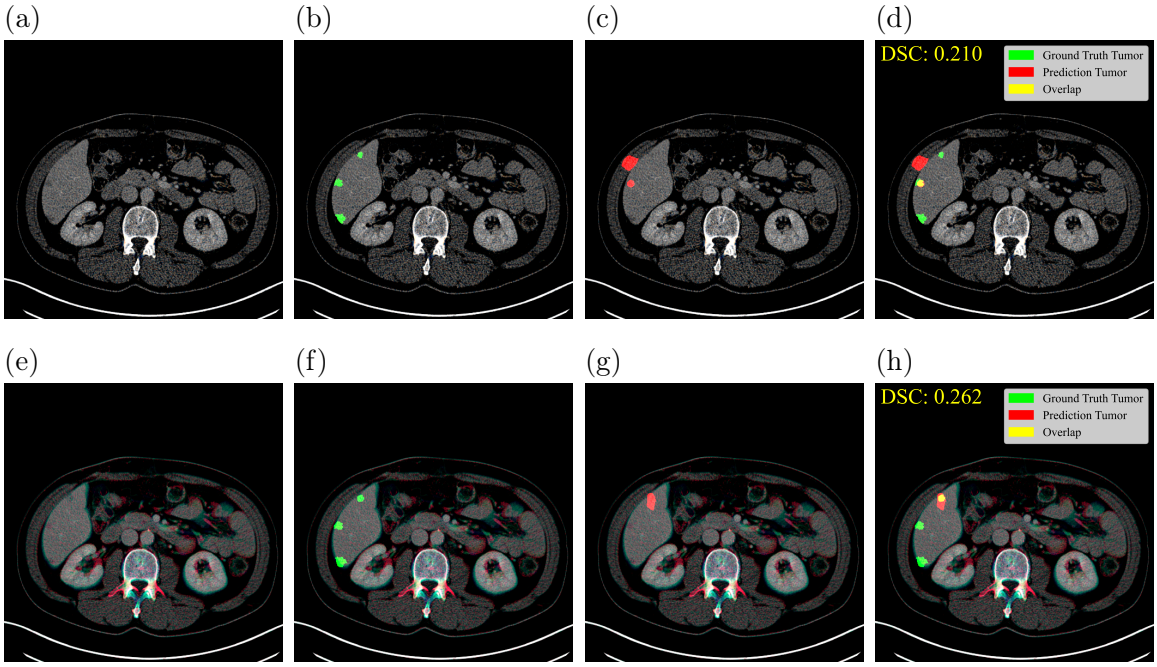


Figure 18: Effect of 2.5-D Slabbing on Tumor Segmentation with Multiple Small Hepatic Lesions. The Images Show (a) CT Using Narrow Neighborhood Slab, (b) Ground Truth Tumor Overlay for Narrow Slab, (c) Predicted Tumor Overlay for Narrow Slab, (d) Combined Ground Truth and Predicted Tumor Overlay for Narrow Slab (DSC 0.210), (e) CT Using Wide Neighborhood Slab, (f) Ground Truth Tumor Overlay for Wide Slab, (g) Predicted Tumor Overlay for Wide Slab, and (h) Combined Ground Truth and Predicted Tumor Overlay for Wide Slab (DSC 0.262). Ground Truth Tumor Is Green, Predicted Tumor Is Red, and Overlap Is Yellow.

With ground-truth ROI, the masked approach achieves DSC 0.516, IoU 0.394, Precision 0.587, and Recall 0.500, representing a relative improvement of approximately 19.2% in DSC over the fixed-crop ground-truth result of 0.433. The underlying mechanism for this improvement lies in the different input presented to the tumor model. In the fixed-crop strategy, the model receives the full rectangular bounding-box region containing adjacent organs such as the spleen, kidneys, stomach, and bowel loops. The HU values of these organs partially overlap with the hepatic window. The model must therefore implicitly suppress activations from these perihepatic structures while simultaneously detecting intra-hepatic lesions. This dual task diverts representational capacity from the primary objective and can introduce false-positive detections at organ boundaries.

The masked-ROI strategy eliminates this ambiguity by zeroing all non-liver pixels before presenting the crop to the tumor model. Every non-zero input pixel is guaranteed to lie within the liver, and the uniform background value carries no competing anatomical signal. The model can therefore allocate its full discriminative capacity to distinguishing hepatic parenchyma from intra-hepatic lesions, substantially improving both precision and recall simultaneously.

The situation differs markedly when using predicted rather than ground-truth ROI. Under these conditions, the masked approach achieves DSC 0.391, which is lower than the DSC 0.427 obtained with the fixed-crop method using predicted ROI. This represents a relative degradation of approximately 8.4%. This regression reflects a stricter dependence of

the masked approach on liver mask quality compared to the fixed-crop method. In the fixed-crop strategy, imperfect liver boundaries shift the crop centroid by a small amount without substantially altering the overall crop content. Most liver tissue and most lesions remain within the crop window despite boundary errors. In the masked approach, an imperfect predicted mask directly zeros out boundary-region pixels. Wherever Stage 1 under-segments the liver periphery, the corresponding image region becomes background before the tumor model processes the crop. This means that peripheral lesions residing in mispredicted boundary zones are completely suppressed prior to Stage 2 inference. This mechanism amplifies Stage 1 prediction error into Stage 2 input corruption and accounts for the performance drop observed when predicted rather than ground-truth masks are used under the masked strategy.

A representative sample case clearly illustrates this vulnerability when Stage 1 segmentation is imperfect. The liver in this patient has complex anatomy with two distinct lobes: one large primary lobe and one secondary lobe extending to the side. The tumor is a notable lesion located within the secondary lobe. With ground-truth liver masks, the secondary lobe is cleanly isolated and the tumor model successfully detects the tumor with DSC 0.836, as shown in Figure 19. With Stage 1 predicted masks, performance completely fails at DSC 0.000, as shown in Figure 20. The Stage 1 model under-segments the secondary lobe severely, cutting it off at the boundary. Because the masking operation zeros out all non-liver pixels, the tumor tissue that falls within this mispredicted boundary is eliminated from the input before the tumor model can see it. With nothing to detect, complete failure occurs. The contrast is striking: the only difference is the quality of the liver mask, yet one achieves strong performance while the other catastrophically fails.

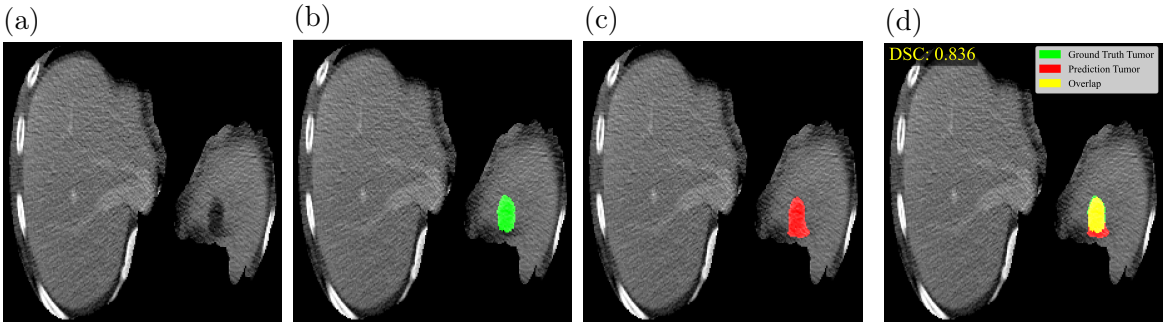


Figure 19: Effect of Masked ROI Extraction with Ground-Truth Liver Mask. The Images Show (a) Masked CT Using the Ground-Truth Liver Mask, (b) Ground-Truth Tumor Overlay on the Masked CT, (c) Predicted Tumor Overlay on the Masked CT, and (d) Combined Ground Truth and Predicted Tumor Overlay (DSC 0.836). Ground Truth Tumor Is Green, Predicted Tumor Is Red, and Overlap Is Yellow.

Consequently, the masked approach requires highly accurate liver segmentation and introduces vulnerability to Stage 1 errors that the fixed-crop method avoids. This vulnerability is particularly severe for peripheral lesions, where Stage 1 errors tend to occur at organ boundaries. When the liver is under-segmented at these boundaries, they are removed from the mask, and lesions near these regions can disappear entirely from the input.

The practical difference between the two approaches is substantial. The fixed-crop method achieves DSC 0.427 with predicted masks and maintains reasonable performance despite Stage 1 errors. The masked approach achieves only DSC 0.391 with predicted masks. This lower average conceals a deeper problem: performance varies widely across cases. Some

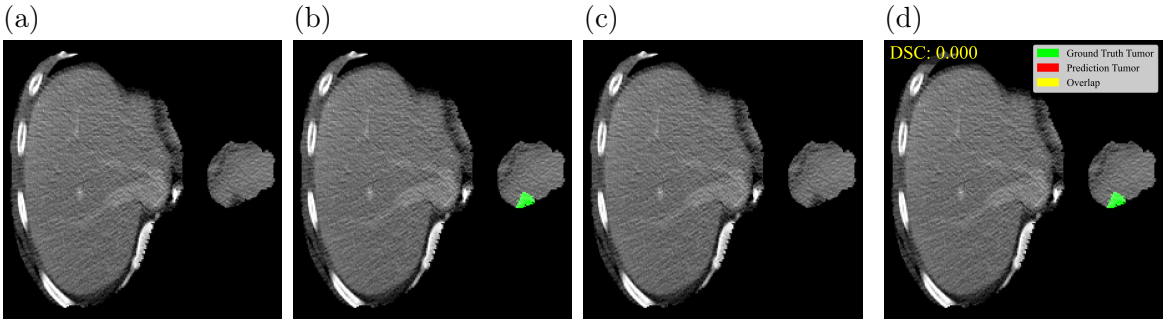


Figure 20: Effect of Masked ROI Extraction with Predicted Liver Mask from Stage 1. The Images Show (a) Masked CT Using the Predicted Liver Mask, (b) Ground-Truth Tumor Overlay on the Masked CT, (c) Predicted Tumor Overlay on the Masked CT, and (d) Combined Ground Truth and Predicted Tumor Overlay. The Example Demonstrates Catastrophic Failure (DSC 0.000) Caused by Stage 1 Under-Segmentation that Removes Tumor Tissue from the Input. Ground Truth Tumor Is Green, Predicted Tumor Is Red, and Overlap Is Yellow.

cases perform reasonably well while others fail completely, depending on whether lesions happen to fall within mispredicted boundary zones. The sample case in Figure 20 shows this catastrophic failure.

For the present pipeline, the fixed-crop configuration is the more reliable choice. Stage 1 performs best with morphological closing applied as post-processing, achieving DSC 0.946. However, the initial tumor segmentation performs better without closing, reaching DSC 0.260 with predicted masks. When closing is applied before ROI extraction, performance drops to DSC 0.237. This reveals that the preprocessing strategy optimized for Stage 1 does not transfer well to Stage 2. Under the fixed-crop approach, Stage 1 errors introduce spatial imprecision but preserve most tumor tissue in the input. The tumor model must suppress activations from adjacent organs within the crop, but it has the opportunity to detect tumors when they are present. For the masked approach, the outcome can be catastrophic: tumors in boundary regions vanish from the input entirely, leaving nothing for the model to detect. The fixed-crop configuration achieved previously with DSC 0.427 under -20 to 400 HU windowing is therefore retained as the baseline for external dataset evaluation.

5.2.2 Performance on External Datasets

While in-domain performance on LiTS provides a necessary baseline, assessment on external CT and MRI datasets is critical for evaluating the generalization capacity of the tumor segmentation model. This section evaluates the model across two external datasets under different intensity normalization strategies to characterize the domain-dependent nature of the cross-dataset transfer problem.

5.2.2.1 Cross-Domain Performance Analysis: 3D-IRCADb-01

Evaluation on the 3D-IRCADb-01 CT dataset shows how preprocessing consistency affects performance similarly to LiTS results, but with different absolute values. Three intensity normalization strategies were evaluated to assess whether the −20 to 400 HU windowing protocol proved optimal on LiTS generalizes to independent CT acquisitions from different scanners and acquisition protocols. The results presented in Table 5.8 demonstrate both the effectiveness and limitations of the windowing approach.

Table 5.8: Tumor Segmentation Performance Across Domain Adaptation Methods (3D-IRCADb-01)

Window variant	ROI from	DSC	IoU	Precision	Recall
Untouched	Ground truth	9.55×10^{-5}	4.78×10^{-5}	4.90×10^{-4}	7.20×10^{-5}
	Prediction	1.61×10^{-5}	8.03×10^{-6}	1.03×10^{-4}	1.18×10^{-5}
Clipped 18% to 80%	Ground truth	0.008	0.004	0.015	0.009
	Prediction	0.009	0.005	0.016	0.010
Clipped -20 to 400	Ground truth	0.467	0.377	0.589	0.547
	Prediction	**0.470**	**0.380**	**0.590**	**0.554**

Without explicit windowing, performance collapses dramatically. DSC reaches only 9.55×10^{-5} for ground-truth ROI, resulting in near-zero tumor detection. Despite 3D-IRCADb-01 being a native CT dataset, raw intensity values differ substantially between scanners and acquisition protocols. The encoder was pre-trained on natural images and fine-tuned on LiTS with the −20 to 400 HU window applied. Its learned filters are calibrated to this specific intensity distribution. When presented with raw, full-range CT values from a different scanner, the feature extraction pipeline produces near-zero activations, yielding minimal segmentation output.

Percentile-based clipping (18 to 80%) rescales intensities based on volume distribution, producing marginally better results (DSC 0.008 and 0.009). Some intensity normalization helps, but the specific mapping diverges from encoder training. This configuration produces extreme precision-recall imbalance (Precision 0.015, Recall 0.009), with rare, isolated positive predictions that correspond poorly to tumors.

When the −20 to 400 HU window is applied to 3D-IRCADb-01 data, performance recovers significantly: DSC 0.467 with ground-truth ROI and DSC 0.470 with predicted ROI. Compared to percentile-based clipping, this is roughly a 50-fold DSC improvement (0.467–0.470 versus 0.008–0.009), demonstrating that cross-domain CT-to-CT transfer requires explicit normalization consistency. The close agreement between ground-truth and predicted ROI performance (difference of 0.003 in DSC) shows that under appropriate normalization, the tumor model becomes robust to Stage 1 liver segmentation errors. Minor spatial imprecisions in crop shape or position do not substantially degrade tumor detection when intensity contrast is sufficiently informative. This robustness indicates that Stage 2 performance decouples from Stage 1 quality, permitting end-to-end inference with modest error propagation.

5.2.2.2 Cross-Modal Performance Analysis: LiverHCCSeg

Applying the same three preprocessing strategies to LiverHCCSeg MRI data reveals fundamentally different constraints compared to 3D-IRCADb-01 CT results as presented in Table 5.9. This demonstrates a critical distinction between cross-domain and cross-modal domain shift. The MRI cohort consists of T1-weighted acquisitions with physics fundamentally different from the CT HU scale used during training. This exposes the encoder's sensitivity to modality changes.

Table 5.9: Tumor Segmentation Performance Across Domain Adaptation Methods (LiverHCCSeg)

Window variant	ROI from	DSC	IoU	Precision	Recall
Untouched	Ground truth	**0.090**	**0.065**	0.197	0.087
	Prediction	0.090	0.064	0.165	**0.089**
Clipped 18% to 80%	Ground truth	0.021	0.011	0.306	0.023
	Prediction	0.022	0.012	**0.399**	0.024
Clipped -20 to 400	Ground truth	0.078	0.056	0.152	0.072
	Prediction	0.078	0.055	0.150	0.074

Without windowing, MRI data achieves DSC 0.090, substantially higher than the 9.55×10^{-5} for untouched CT data. This 1000-fold difference is revealing. While both lack explicit windowing, MRI data preserves some task-relevant signal. LiverHCCSeg T1-weighted MRI encodes tissue responses fundamentally different from CT X-ray attenuation. Hepatic parenchyma and tumor tissue show tissue-dependent signal differences in MRI that the model can partially exploit. In contrast, untouched CT data presents such an extreme dynamic range that these differences are suppressed.

Percentile clipping (18 to 80%) emphasizes common tissue intensities, producing marked degradation to DSC 0.021 and 0.022, approximately 76% relative decline from the untouched result of 0.090. Percentile-based normalization removes rare but important signals. Small lesions often occupy extreme intensity values relative to normal tissue distribution. Removing histogram tails disproportionately eliminates lesion signals, explaining the performance drop.

Applying the LiTS-derived -20 to 400 window as numeric clipping to MRI data improves modestly to DSC 0.078 (approximately 271% improvement from the percentile-clipped result), but remains below the untouched configuration (DSC 0.090). The window was optimized for CT HU ranges and hepatic tissue attenuation. When applied to MRI signal intensities, this clipping range is an arbitrary remapping without meaningful tissue correspondence. The identical DSC values between ground-truth and predicted ROI under this configuration (both 0.078) contrast with the 0.003 difference in 3D-IRCADb-01. At such low performance levels, both ROI quality and windowing strategy are equally limiting.

Chapter-6: Discussion

6.1 Baseline Analysis: A Foundation for Cross-Modal Progress

This work provides a systematic baseline analysis of cross-modal liver and tumor segmentation using standard architectures and training strategies. Rather than proposing novel architectural innovations, the thesis establishes what is achievable with straightforward approaches: a frozen ResNet18 encoder combined with a U-Net decoder, applied to public datasets with basic preprocessing. The core contribution is demonstrating where such baseline methods succeed and, more importantly, where they fail and why. By characterizing the performance landscape across multiple datasets and modalities, this work identifies the specific architectural and methodological bottlenecks that fundamentally limit cross-modal performance. This baseline understanding is essential because it addresses a fundamental issue that existing literature often bypasses by determining what simple approaches achieve and identifying the specific obstacles they encounter before pursuing complex solutions.

Taken together, the results show that within-modality transfer depends mainly on training scale and consistent preprocessing, while cross-modal transfer remains the central challenge. In liver segmentation, CT-to-CT transfer from LiTS is very strong on external CT benchmarks, while MRI-trained models show a larger drop when evaluated on CT. In tumor segmentation, preprocessing choices matter even more, especially intensity windowing. Without windowing, the LiTS tumor model reached DSC 0.294; with a -20 to 400 HU window, performance rose to DSC 0.433. The same window recovered cross-domain CT performance on 3D-IRCADb-01 to DSC 0.467–0.470, which shows that intensity normalization needs to match the training distribution even within CT. When crossing modalities, the pattern changes. The LiTS liver model drops to DSC 0.196 on CHAOS MRI but transfers better to LiverHCCSeg MRI and improves to DSC 0.763 under train-only closing. In Stage 2 tumor segmentation, the best result on LiverHCCSeg MRI remains low at DSC 0.090 even when ground-truth liver ROIs are used. These findings motivate the analysis below, which separates the effects of training scale, label processing, and modality mismatch.

The two-stage pipeline design decouples liver localization from lesion detection, reducing computational burden and mitigating extreme class imbalance in tumor annotation. This staging approach is not novel, but systematic evaluation across five public datasets clarifies precisely when and why staging provides benefits. The investigation of preprocessing methods including morphological closing and intensity normalization demonstrates that label quality and input standardization have measurable effects, though these effects have

clear limits when crossing modalities. By establishing these limits quantitatively, this work clarifies which problems are tractable with preprocessing and which require architectural changes.

6.2 Liver Segmentation Across Modalities: Scale, Diversity, and Preprocessing Effects

In the no-preprocessing baseline, the liver segmentation experiments show a clear hierarchy driven by dataset size and modality. LiverHCCSeg contains 17 patients and only 10 training cases under the 60/40 split. It achieved strong in-domain performance with DSC 0.925 and generalizes reasonably well to CHAOS in the cross-domain MRI-to-MRI setting at DSC 0.804, but its MRI-to-CT transfer remained limited, especially on LiTS CT at DSC 0.560. CHAOS reached the best MRI in-domain DSC of 0.944, but its cross-domain MRI-to-MRI transfer to LiverHCCSeg is lower at DSC 0.631, and its MRI-to-CT transfer remains moderate. LiTS, with 105 CT training cases, achieved in-domain DSC 0.940 and near-ceiling CT-to-CT transfer, reaching DSC 0.975 on 3D-IRCADb-01 and 0.936 on SLiver07. Overall, scale and acquisition diversity provide an important foundation for within-modality transfer, but the MRI-to-MRI results also show that protocol differences within a modality can still produce a substantial domain gap. This pattern motivates the discussion below of how label and prediction processing can help or hurt generalization.

This progression demonstrates a clear principle: within-modality transfer is driven primarily by training data scale and acquisition diversity rather than architectural sophistication. The LiverHCCSeg model failed on CT not because of inadequate architecture but because the training distribution simply did not represent CT intensity ranges and texture patterns. When the LiTS model encountered different CT scanners and acquisition protocols, it generalized nearly perfectly because its training encompassed sufficient diversity in CT acquisition characteristics. The model learned to recognize liver tissue across the natural variation inherent in CT imaging, enabling robust transfer to new CT datasets.

This finding has important implications. When applied to a narrow domain with adequate training data, a frozen backbone combined with a decoder produces excellent results. ImageNet pretraining provides essential initial features, and the decoder adapts these representations to the specific task. The frozen encoder is not a liability here because the task remains within a modality where encoded features are appropriate. The bottleneck is not feature representation but training data diversity.

Beyond dataset scale and acquisition diversity, label quality and preprocessing choices can also shift performance, but only when they are applied in a way that keeps model selection and evaluation consistent. When applied exclusively to the training set, morphological closing improved external performance. For the CHAOS model, cross-modal performance on 3D-IRCADb-01 CT improved to 0.806 from 0.751 at baseline, and for the LiTS model, cross-domain CT-to-CT performance remained near ceiling at 0.976 compared to 0.975. The mechanism is interpretable: by providing the model with cleaner supervision targets lacking annotation artifacts such as intra-organ holes and jagged boundaries, closing encourages the decoder to learn spatially coherent organ representations. These coherent representations then transfer better to new datasets because they capture true anatomy

rather than annotation noise.

Fragment removal showed more complex behavior. For the CHAOS model, removing disconnected components smaller than 10% of the largest connected component improved cross-modal MRI-to-CT performance on LiTS, reaching 0.709 compared to 0.613 at baseline. For the LiTS model in the cross-modal CT-to-MRI setting, fragment removal did not consistently help. It reduced DSC on LiverHCCSeg from 0.617 to 0.581, and while it increased CHAOS from 0.196 to 0.285, performance remained low overall.

A critical methodological finding emerged: preprocessing applied to validation or test sets introduces a confound that invalidates clean interpretation. When morphological closing was applied to both training and validation data, the model selection criterion via early stopping optimized against modified annotations rather than original masks. This mismatch between the validation distribution used for model selection and the test distribution against original masks caused substantial degradation. This observation demonstrates that when evaluating segmentation models, maintaining temporal and distributional consistency between model selection and final evaluation is non-negotiable. Preprocessing should clean training targets, but validation and test evaluation should proceed against original annotations to preserve interpretability.

6.3 Tumor Segmentation Across Modalities: The Hard Boundary

Building on the liver results, the Stage 2 tumor segmentation experiments show a sharper boundary between cross-domain and cross-modal transfer. On LiTS CT, the baseline configuration without windowing reached DSC 0.294. Applying a -20 to 400 HU window improved performance to DSC 0.433. On external cross-domain evaluation with 3D-IRCADb-01 CT, the same window recovered performance to DSC 0.467–0.470. These results show that, even within CT, intensity normalization needs to match the training distribution.

Once intensity normalization is fixed, the remaining Stage 2 results are mainly shaped by how the liver ROI is constructed. Applying morphological closing to the liver mask before ROI extraction reduced LiTS tumor DSC from 0.294 to 0.251 for ground-truth ROIs, and qualitative examples show how centroid shifts can move tumor regions toward the crop boundary. Adding limited volumetric context using 2.5-D slabbing did not help and reduced DSC compared to the 2D windowed baseline. Masking all non-liver pixels improved performance when ground-truth liver masks were used, but it reduced robustness when predicted masks were used. A representative failure case shows how Stage 1 under-segmentation can remove tumor tissue from the input. Taken together, these experiments show that ROI design affects robustness, but it does not remove the underlying modality gap.

The contrast with MRI is clear. On LiverHCCSeg MRI, the best DSC is 0.090 under the untouched setting, and both percentile clipping and CT-derived clipping reduce performance. Compared with the matched-window result on 3D-IRCADb-01 CT of DSC 0.467–0.470, this is about 5-fold lower. This gap persists even when ground-truth ROIs

are used, which indicates that the decoder features learned from CT do not transfer to MRI lesion appearance.

This gap is not mainly caused by Stage 1 errors. Under matched CT windowing, the difference between ground-truth and predicted ROIs on 3D-IRCADb-01 is about 0.003 DSC. On LiverHCCSeg MRI, ground-truth and predicted ROIs yield essentially the same DSC of 0.090. This shows that ROI quality is not the limiting factor in cross-modal tumor segmentation. Instead, the limitation sits earlier in the feature extraction pipeline, which points back to the frozen encoder.

6.4 Frozen Encoder: An Architectural Asymmetry

The frozen encoder architecture, selected for practical reasons, became the binding constraint for cross-modal work. The architecture made sense initially: ImageNet pretraining provides useful low-level features such as edges and local textures, and freezing the encoder reduces trainable parameters and overfitting risk on small datasets. Within CT, this approach succeeds admirably. The decoder learned to interpret the frozen encoder's representations for detecting soft-tissue boundaries across HU units, and this learning transferred robustly across CT scanners and acquisition protocols where the underlying physics remains identical. Across modalities, the situation is categorically different. CT and MRI are based on entirely different physical principles. CT measures X-ray attenuation. MRI measures radiofrequency signal decay. A low-attenuation region in CT might appear bright, dark, or intermediate in MRI depending on tissue properties and acquisition sequence. There is no consistent mapping between intensity and tissue type across modalities.

Encoder adaptation is the missing component. Without it, the frozen encoder contains only ImageNet pretraining weights, never adapted to CT or MRI. When the frozen encoder processes MRI signal intensities, its activations follow patterns learned from natural images, not medical imaging. A fine-tuned encoder could update its kernels in response to MRI data, learning representations aligned with MRI intensity distributions. This process requires backpropagation through the encoder during training on target-domain data. A frozen encoder prevents this adaptation, leaving the encoder with an inherent ceiling on cross-modal capability. The evidence is quantitative and compelling: identical architecture and task achieving DSC of 0.467 within CT and 0.090 across to MRI cannot be explained by decoder design or training techniques. The architecture itself contains an asymmetry. The encoder is the asymmetry.

Chapter-7: Conclusion

This thesis presents a systematic baseline analysis of cross-modal liver and tumor segmentation, establishing what is achievable with straightforward neural network architectures and standard training strategies on publicly available datasets. Rather than pursuing novel architectural innovations, the work deliberately adopts simple methods: a frozen ResNet18 encoder combined with a U-Net decoder within a two-stage pipeline. The primary contribution lies not in architectural novelty but in rigorously characterizing the performance landscape across multiple datasets and modalities, thereby identifying the specific architectural and methodological bottlenecks that limit cross-modal performance.

The results reveal a clear hierarchy of capability driven by training data properties and modality alignment. Within-modality transfer succeeds remarkably well when training datasets are sufficiently large and diverse. The model trained on large-scale CT data demonstrates strong generalization to other CT benchmarks, indicating that the approach captures the natural variation in imaging physics within a single modality. CT training data encodes the full range of scanner characteristics, acquisition protocols, and tissue attenuation patterns found in CT imaging. The decoder learns to interpret the frozen encoder representations appropriately for detecting tissue boundaries across this variation, and these learned mappings transfer well to new CT datasets of the same modality.

Cross-modal transfer remains the central limitation of this baseline pipeline. When a model is trained on CT and tested on MRI, performance drops sharply, and in some cases it collapses. For liver segmentation, the LiTS-trained model reaches DSC 0.196 on CHAOS MRI, while train-only closing improves LiTS to LiverHCCSeg to DSC 0.763. For tumor segmentation, the best result on LiverHCCSeg MRI is DSC 0.090. Stage 2 experiments show that this low performance is not explained by ROI quality, since using ground-truth ROIs does not improve the score. These outcomes are consistent with the physics of imaging. CT intensities represent X-ray attenuation in HU units, while MRI intensities represent radiofrequency signal behavior. Because there is no stable mapping between the two, a frozen ImageNet encoder that is not adapted to medical imaging cannot provide modality-invariant features.

Preprocessing improved performance within modality, but its effect depended on the stage and the target domain. For tumor segmentation on CT, HU windowing was essential. Using a -20 to 400 HU window increased LiTS tumor DSC from 0.294 to 0.433 and supported transfer to 3D-IRCADb-01 with DSC 0.467–0.470. For liver segmentation, applying morphological closing only to training masks improved external performance when validation annotations were kept unmodified, suggesting that cleaner supervision targets reduce the propagation of annotation artifacts. At the same time, some preprocessing choices did not transfer across stages. Closing the liver mask before Stage 2 ROI extraction

reduced tumor performance by shifting centroid-based crops. Finally, CT-derived clipping ranges did not provide a meaningful normalization for MRI signal intensities. Overall, preprocessing can help when it makes the input distribution match training, but it cannot resolve a cross-modal feature mismatch in a frozen encoder.

The two-stage pipeline still provides clear practical benefits. It limits tumor inference to a liver-centered region and reduces class imbalance. Under matched CT windowing, Stage 2 tumor performance is largely insensitive to small ROI errors. On 3D-IRCADb-01, the difference between ground-truth and predicted ROIs is about 0.003 DSC. On MRI, the same robustness does not translate, since the model remains near the same low DSC even with ground-truth ROIs. This confirms that staging can reduce error propagation within a modality, but it cannot overcome cross-modal limitations of the encoder.

7.1 Limitations

This work operates under several important limitations that constrain the generalizability of findings and suggest directions for future investigation.

First, the LiverHCCSeg dataset contains only 17 patients, creating a fundamental constraint on model development for the MRI modality. This small scale necessarily limits the potential in-domain performance and makes generalization particularly challenging. Moreover, the dataset scale forces the held-out test set to simultaneously serve as the validation set, creating an additional methodological complication. Any model selection via early stopping or hyperparameter tuning uses the same data instances that are subsequently reported as held-out test performance, preventing true separation between training and evaluation cohorts. This constraint does not affect the CHAOS or LiTS results, which maintain dedicated train, validation, and test splits, and it does not invalidate findings about cross-modal transfer, but it does introduce additional conservatism when interpreting LiverHCCSeg based conclusions.

Second, the frozen encoder architecture was selected for practical reasons: ImageNet pre-training provides useful low-level visual features, and freezing reduces trainable parameters to mitigate overfitting on limited datasets. However, this architectural choice becomes a binding constraint for cross-modal work. The frozen encoder prevents adaptation to either CT or MRI intensity characteristics, forcing the network to interpret medical imaging data using representations derived exclusively from natural images. While this limitation is acknowledged and forms a key finding of the thesis, exploring encoder adaptation strategies remains outside the scope of this baseline analysis. Future work must investigate fine-tuning approaches, domain-specific pre-training, or multi-modal foundation models.

Finally, the study employs only 2D slice-by-slice processing despite the inherently three-dimensional nature of volumetric medical imaging. A fully three-dimensional approach could exploit inter-slice coherence and provide richer spatial context, potentially improving both in-domain performance and cross-modal robustness. However, 2D methods remain computationally practical for clinical deployment and provide a clear baseline against which future three-dimensional approaches can be compared. The benefits of volumetric processing relative to 2D methods remain an open question deserving future investigation.

7.2 Future Work

There are several areas for future research. One clear path is to study fine-tuning methods that adapt encoder weights for CT and MRI values. This would help the network learn features specific to medical scans instead of natural images. Researchers could also look into attention methods that change how features are extracted based on the input type. This would allow a single model to handle different kinds of scans without needing to be trained again.

Another idea is to build new network architectures that naturally align features across different scan types. This would help close the gap between CT and MRI data directly, so there is less need for special preprocessing steps. Finally, testing full three-dimensional models would let the network use information from neighboring slices and gain a better spatial view. Moving to 3D models solves the basic limits of processing images one slice at a time.

Despite these limitations, the thesis succeeds in its objective of providing rigorous baseline analysis establishing where simple methods succeed and why they fail. This foundation is essential for justifying more sophisticated approaches and designing targeted solutions to identified bottlenecks.

Bibliography

[1] H. Sung *et al.*, "Global burden of primary liver cancer in 2020 and predictions to 2040," *Journal of Hepatology*, vol. 77, no. 6, pp. 1598–1606, 2022.

[2] T. Li *et al.*, "Global burden of liver cirrhosis 1990–2019 and 20 years forecast: results from the global burden of disease study 2019," *Annals of Medicine*, vol. 56, no. 1, p. 2328521, 2024.

[3] M. Cao, S. Liu, S. Zhang, and J. He, "Burden of liver cancer: From epidemiology to prevention," *Chinese Journal of Cancer Research*, vol. 34, no. 3, pp. 554–566, 2022.

[4] H. Rumgay *et al.*, "Global trends in hepatocellular carcinoma epidemiology: implications for screening, prevention and therapy," *Nature Reviews Clinical Oncology*, vol. 20, no. 12, pp. 864–884, 2023.

[5] D. Q. Huang, H. B. El-Serag, and R. Loomba, "Mechanisms of liver fibrosis and its role in liver cancer," *Experimental Biology and Medicine*, vol. 245, no. 2, pp. 96–108, 2020.

[6] R. S. Taylor *et al.*, "Increased risk of mortality by fibrosis stage in non-alcoholic fatty liver disease: Systematic review and meta-analysis," *Hepatology*, vol. 65, no. 5, pp. 1557–1565, 2017.

[7] K. Cheng, N. Cai, J. Zhu, X. Yang, H. Liang, and W. Zhang, "Tumor-associated macrophages in liver cancer: From mechanisms to therapy," *Cancer Communications*, vol. 42, no. 11, pp. 1112–1140, 2022.

[8] A. Vogel, T. Meyer, G. Sapisochin, R. Salem, and A. Saborowski, "Challenges in liver cancer and possible treatment approaches," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1873, no. 1, p. 188314, 2020.

[9] A. Baranova and Z. M. Younossi, "Liver fibrosis determination," *Gastroenterology Clinics of North America*, vol. 48, no. 2, pp. 281–289, 2019.

[10] G. Decker *et al.*, "Noninvasive diagnosis of liver cirrhosis: qualitative and quantitative imaging biomarkers," *Abdominal Radiology*, vol. 49, no. 6, pp. 2098–2115, 2024.

[11] E. Chartampilas, V. Rafailidis, P. Georgiadis, G. Kalarakis, A. Hatzidakis, and P. Prassopoulos, "Current imaging diagnosis of hepatocellular carcinoma," *Cancers*, vol. 14, no. 16, p. 3997, 2022.

[12] J. Zhou *et al.*, "Guidelines for the diagnosis and treatment of primary liver cancer (2022 Edition)," *Liver Cancer*, vol. 12, no. 5, pp. 405–444, 2023.

[13] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "A review of the application of deep learning in medical image classification and segmentation," *Annals of Translational Medicine*, vol. 8, no. 11, pp. 713–713, 2020.

[14] L. Pinto-Coelho, "How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications," *Bioengineering*, vol. 10, no. 12, p. 1435, 2023.

[15] P. Sharma *et al.*, "Artificial intelligence in medical imaging: bridging innovation, ethics, and clinical impact," *International Journal of Advances in Medicine*, vol. 12, no. 6, pp. 621–626, 2025.

[16] M. Moghbel *et al.*, "Practical utility of liver segmentation methods in clinical surgeries and interventions," *BMC Medical Imaging*, vol. 22, no. 1, p. 38, 2022.

[17] H. P. Antunes *et al.*, "Diagnostic applications of artificial intelligence in liver diseases," *Journal of Clinical Medicine*, vol. 14, no. 17, p. 6231, 2025.

[18] L. Hermoye *et al.*, "Liver segmentation in living liver transplant donors: Comparison of semiautomatic and manual methods," *Radiology*, vol. 234, no. 1, pp. 171–178, 2005.

[19] A. Gotra *et al.*, "Liver segmentation: indications, techniques and future directions," *Insights into Imaging*, vol. 8, no. 4, pp. 377–392, 2017.

[20] M. A. Selver, F. Fischer, N. Gezer, W. Hillen, and O. Dicle, "Semi-Automatic Segmentation Methods for 3-D Visualization and Analysis of the Liver," in *Medical Informatics Europe*, ser. Studies in Health Technology and Informatics. IOS Press, 2014, pp. 1133–1137.

[21] Z. Yang, Y.-q. Zhao, M. Liao, S.-h. Di, and Y.-z. Zeng, "Semi-automatic liver tumor segmentation with adaptive region growing and graph cuts," *Biomedical Signal Processing and Control*, vol. 68, p. 102670, 2021.

[22] D. C. Le, K. Chinnasarn, J. Chansangrat, N. Keeratibharat, and P. Horkaew, "Semi-automatic liver segmentation based on probabilistic models and anatomical constraints," *Scientific Reports*, vol. 11, no. 1, p. 6106, 2021.

[23] S.-J. Lim, Y.-Y. Jeong, and Y.-S. Ho, "Automatic liver segmentation for volume measurement in CT images," *Journal of Visual Communication and Image Representation*, vol. 17, no. 4, pp. 860–875, 2006.

[24] O. Gambino *et al.*, "Automatic volumetric liver segmentation using texture based region growing," in *2010 Fourth International Conference on Complex, Intelligent and Software Intensive Systems*, 2010, pp. 146–152.

[25] R. Sivanandan and J. Jayakumari, "Ultrasound liver tumour active contour segmentation with initialization using adaptive Otsu based thresholding," *Research on Biomedical Engineering*, vol. 37, no. 2, pp. 251–262, 2021.

[26] F. Liu, B. Zhao, P. K. Kijewski, L. Wang, and L. H. Schwartz, "Liver segmentation for CT images using GVF snake," *Medical Physics*, vol. 32, no. 12, pp. 3699–3706, 2005.

[27] L. Ruskó, G. Bekes, and M. Fidrich, "Automatic segmentation of the liver from multi- and single-phase contrast-enhanced CT images," *Medical Image Analysis*, vol. 13, no. 6, pp. 871–882, 2009.

[28] Y. Zheng, X. Yang, X. Ye, and X. Lin, "Fully automatic segmentation of liver from multiphase liver CT," in *Medical Imaging 2007: Image Processing*, vol. 6512, 2007, p. 65122X.

[29] L. Xu, Y. Zhu, Y. Zhang, and H. Yang, "Liver segmentation based on region growing and level set active contour model with new signed pressure force function," *Optik*, vol. 202, p. 163705, 2020.

[30] A. Afifi and T. Nakaguchi, "Liver segmentation approach using graph cuts and iteratively estimated shape and intensity constrains," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 395–403.

[31] A.-R. Ali, M. Couceiro, A. E. Hassanien, M. F. Tolba, and V. Snášel, "Fuzzy C-Means based liver CT image segmentation with optimum number of clusters," in *Innovations in Bio-inspired Computing and Applications*, ser. Advances in Intelligent Systems and Computing. Springer International Publishing, 2014, pp. 131–139.

[32] X. Zhang, J. Tian, D. Xiang, X. Li, and K. Deng, "Interactive liver tumor segmentation from CT scans using support vector classification with watershed," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 6005–6008.

[33] J. Lu, D. Wang, L. Shi, and P. A. Heng, "Automatic liver segmentation in CT images based on Support Vector Machine," in *2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, 2012, pp. 333–336.

[34] M. Barstugan, R. Ceylan, M. Sivri, and H. Erdogan, "Automatic liver segmentation in abdomen CT images using SLIC and AdaBoost algorithms," in *Proceedings of the 8th International Conference on Bioinformatics and Biomedical Science*, 2018, pp. 129–133.

[35] P. Zhang, J. Yang, D. Ai, Z. Xie, and Y. Liu, "Learning based random walks for automatic liver segmentation in CT image," in *Abdominal Imaging: Computation and Clinical Applications*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2015, pp. 251–259.

[36] Y. Zhang *et al.*, "Feature learning based random walk for liver segmentation," *PLOS ONE*, vol. 11, no. 11, p. e0164098, 2016.

[37] H. Masoumi, A. Behrad, M. A. Pourmina, and A. Roosta, "Automatic liver segmentation in MRI images using an iterative watershed algorithm and artificial neural network," *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 429–437, 2012.

[38] D. Spinczyk and A. Krason, "Automatic liver segmentation in computed tomography using general-purpose shape modeling methods," *BioMedical Engineering OnLine*, vol. 17, no. 1, p. 65, 2018.

[39] M. Ahmad *et al.*, "A lightweight convolutional neural network model for liver segmentation in medical diagnosis," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–16, 2022.

[40] A. Halder, A. Sau, S. Majumder, D. Kaplun, and R. Sarkar, "An experimental study of U-Net variants on liver segmentation from CT scans," *Journal of Intelligent Systems*, vol. 34, no. 1, p. 20240185, 2025.

[41] A. Affane *et al.*, "Segmentation of liver anatomy by combining 3D U-Net approaches," *Applied Sciences*, vol. 11, no. 11, p. 4895, 2021.

[42] X. Guo, L. H. Schwartz, and B. Zhao, "Automatic liver segmentation by integrating fully convolutional networks into active contour models," *Medical Physics*, vol. 46, no. 10, pp. 4455–4469, 2019.

[43] X. Wei, X. Chen, C. Lai, Y. Zhu, H. Yang, and Y. Du, "Automatic liver segmentation in CT images with enhanced GAN and mask region-based CNN architectures," *BioMed Research International*, vol. 2021, no. 1, p. 9956983, 2021.

[44] R. He *et al.*, "Three-dimensional liver image segmentation using generative adversarial networks based on feature restoration," *Frontiers in Medicine*, vol. 8, p. 794969, 2022.

[45] U. Demir *et al.*, "Transformer based generative adversarial network for liver segmentation," in *ICIAP Workshops*, 2022, arXiv:2205.10663.

[46] D. Wong *et al.*, "A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints," *The MIDAS Journal*, 2008.

[47] J. H. Moltz, L. Bornemann, V. Dicken, and H.-O. Peitgen, "Segmentation of liver metastases in CT scans by adaptive thresholding and morphological processing," *The MIDAS Journal*, 2008.

[48] O. I. Alirr, A. A. Abd. Rahni, and E. Golkar, "An automated liver tumour segmentation from abdominal CT scans for hepatic surgical planning," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 8, pp. 1169–1176, 2018.

[49] A. M. Anter and A. E. Hassanien, "CT liver tumor segmentation hybrid approach using neutrosophic sets, fast fuzzy c-means and adaptive watershed algorithm," *Artificial Intelligence in Medicine*, vol. 97, pp. 105–117, 2019.

[50] A. Shimizu, T. Narihira, D. Furukawa, H. Kobatake, S. Nawano, and K. Shinozaki, "Ensemble segmentation using AdaBoost with application to liver lesion extraction from a CT volume," *The MIDAS Journal*, 2008.

[51] A. Ben-Cohen, E. Klang, I. Diamant, N. Rozendorn, M. M. Amitai, and H. Greenspan, "Automated method for detection and segmentation of liver metastatic lesions in follow-up CT examinations," *Journal of Medical Imaging*, vol. 2, no. 3, p. 034502, 2015.

[52] C.-C. Chang *et al.*, "Computer-aided diagnosis of liver tumors on computed tomography images," *Computer Methods and Programs in Biomedicine*, vol. 145, pp. 45–51, 2017.

[53] P.-H. Conze *et al.*, "Scale-adaptive supervoxel-based random forests for liver tumor segmentation in dynamic contrast-enhanced CT scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 2, pp. 223–233, 2017.

[54] H.-w. Zhang, D.-l. Huang, Y.-r. Wang, H.-s. Zhong, and H.-w. Pang, "CT radiomics based on different machine learning models for classifying gross tumor volume and normal liver tissue in hepatocellular carcinoma," *Cancer Imaging*, vol. 24, no. 1, p. 20, 2024.

[55] M. H. Malik, H. Ghous, T. Rashid, B. Maryum, Z. Hao, and Q. Umer, "Feature extraction-based liver tumor classification using Machine Learning and Deep Learning methods of computed tomography images," *Cogent Engineering*, vol. 11, no. 1, p. 2338994, 2024.

[56] P. F. Christ *et al.*, "Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks," arXiv preprint arXiv:1702.05970, 2017.

[57] C. Sun *et al.*, "Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs," *Artificial Intelligence in Medicine*, vol. 83, pp. 58–66, 2017.

[58] G. Chlebus, H. Meine, J. H. Moltz, and A. Schenk, "Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing," *Scientific Reports*, vol. 8, no. 1, p. 15497, 2018.

[59] L. Meng, Y. Tian, and S. Bu, "Liver tumor segmentation based on 3D convolutional neural network with dual scale," *Journal of Applied Clinical Medical Physics*, vol. 21, no. 1, pp. 144–157, 2020.

[60] Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 605132, 2020.

[61] H. Rahman, T. F. N. Bukht, A. Imran, J. Tariq, S. Tu, and A. Alzahrani, "A deep learning approach for liver and tumor segmentation in CT images using ResUNet," *Bioengineering*, vol. 9, no. 8, p. 368, 2022.

[62] E. Goceri, "A hybrid attention-based deep learning model for segmentation of livers and liver tumors from CT scans," *Multimedia Tools and Applications*, vol. 84, no. 37, pp. 46 191–46 212, 2025.

[63] R. Naseem, Z. A. Khan, N. Satpute, A. Beghdadi, F. A. Cheikh, and J. Olivares, "Cross-modality guided contrast enhancement for improved liver tumor image segmentation," *IEEE Access*, vol. 9, pp. 118 154–118 167, 2021.

[64] W. Yu, M. Wang, Y. Zhang, and L. Zhao, "Reciprocal cross-modal guidance for liver lesion segmentation from multiple phases under incomplete overlap," *Biomedical Signal Processing and Control*, vol. 88, p. 105561, 2024.

[65] J. Zhao and S. Li, "Uncertainty-guided and cross-modality attention network for liver tumor segmentation and quantification via integrating dynamic MRI," *Knowledge-Based Systems*, vol. 310, p. 113021, 2025.

[66] B. Mukhopadhyay, C. Mandal, P. Tummala, N. Mahmoodian, A. Nürnberger, and S. Chatterjee, "Towards segmenting the invisible: An end-to-end registration and segmentation framework for weakly supervised tumour analysis," in *Artificial Intelligence for Biomedical Data*, ser. Communications in Computer and Information Science. Springer Nature Switzerland, 2026, pp. 229–242.

[67] P. Bilic *et al.*, "LiverHCCSeg: A publicly available multiphasic MRI dataset with liver and HCC tumor segmentations and inter-rater agreement analysis," *Data in Brief*, vol. 51, p. 109662, 2023.

[68] A. E. Kavur *et al.*, "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[69] P. Bilic *et al.*, "The liver tumor segmentation benchmark (LiTS)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

[70] L. Soler *et al.*, "3d image reconstruction for comparison of algorithm database: The IRCAD research platform," https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/, 2010, accessed: 2024.

[71] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

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

[73] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, vol. 9351. Springer International Publishing, 2015, pp. 234–241.

[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106