Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2025-12

A Cross-Modal Benchmark of Deep Learning Encoders for Automated Liver Segmentation

Rahman, Redwan

IUB

https://ar.iub.edu.bd/handle/11348/1077

Downloaded from IUB Academic Repository



A Cross-Modal Benchmark of Deep Learning Encoders for Automated Liver Segmentation

December 2025

Prepared by:

Redwan Rahman
ID: 2221936

Kazi Toushia Nahar
ID: 2221696

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by

Saadia Binte Alam
Associate Professor and Head
Department of Computer Science and Engineering
Independent University, Bangladesh

Attestation

We know that any form of plagiarism is not permitted and conflicts with the academic policies of our university. We attest that the work presented is original and that it was created by us. This project has some copyrighted materials and external models all of which have been cited in accordance with the accepted international citation guidelines.

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

We are most grateful to our Dr. Supervisor Saadia Binte Alam. We also would wish to acknowledge Md. Rashedur Rahman. Their suggestions, recommendations, and time played an indispensable role in the success of our project. We are immensely grateful to them due to their assistance in our academic life as it made the quality of our work significantly better.

We also wish to thank the Center of Computational & Data Sciences (CCDS) who helped and supported us. The lab environment and resources played a significant role in our research being carried out and carried out fruitfully.

Moreover, we would like to thank Mr. Zafor Iqbal and Mr. Siam Tahsin Bhuiyan who provided technical assistance in our project. We owe our gratitude to all those people who assisted us, directly or indirectly.

Finally, we used Gemini 3 Pro from Google to help polish the language and enhance the clarity of our writing. We explicitly state that the application of this tool was restricted to the completion of editorial tasks and that no part of the data analysis, result generation or graphical illustrations was produced using this technology.

Abstract

Many clinical processes, including the scheduling of liver transplants, volumetry, and management of chronic diseases, including cirrhosis and fibrosis, rely on accurate segmentation of the liver. These critical clinical situations, including incorrect tracking of disease progression, pose significant risks to patient safety when there is an incorrect segmentation. The occurrence of the domain shift phenomenon is when models that have been trained on Magnetic Resonance Imaging (MRI) fail to generalize to Computed Tomography (CT). This problem is a serious limitation to the clinical utility of such systems in healthcare facilities despite the fact that Deep Learning (DL) has demonstrated great potential in automating the task. Despite the regular introduction of the state-of-the-art methods of Unsupervised Domain Adaptation (UDA), there is not much information regarding the natural cross-modality robustness of simplistic encoder structures.

To compare adaptability, a baseline was established using four popular encoder backbones in a standard U-Net structure. These backbones include ResNet18, EfficientNetB3, DenseNet121, and MobileNetV2. The LiverHccSeg data (arterial-phase T1-weighted MRI) was used to train models to evaluate the differences in model-specific performance. They were subsequently tested on three external sets which represented varying levels of domain shift: CHAOS (MRI), 3D-IRCADb-01 and SLIVER07 (CT).

The systematic analysis indicates that in-domain accuracy does not define cross-modality stability. The best performance on the in-domain test set (Dice Similarity Coefficient [DSC] = 0.928) results in poor performance on generalization tasks, though DenseNet121 was best. Conversely, EfficientNetB3 was the most robust design, and delivered the best performance on the external MRI protocol (CHAOS, DSC 0.852) and cross-modality CT data (3D-IRCADb-01, DSC 0.825). It implies that compound scaling principles improve the achievement of cross-modality structural representations that are essential in handling various liver diseases. ResNet18, in turn, demonstrated the lowest cross-modality transfer (DSC 0.769-0.788 on CT), which indicates that it is susceptible to training modality-specific overfitting intensity patterns. Furthermore, the lightweight MobileNetV2 showed comparable generalization (DSC $\sim$0.81 on CT), demonstrating that strong performance and computational efficiency are compatible.

The use of 2D slice processing, which limits the acquisition of complete 3D volumetric context, and the comparatively small size of the training cohort ($n = 17$) are the main limitations of this study. In order to further reduce the patient risks related to domain shift, future research should concentrate on validating our results on larger, multi-center datasets and investigating 3D structures. In the end, these findings suggest that choosing the appropriate encoder is an important initial step in developing cross-modal segmentation systems that are reliable for widespread clinical application.

List of Publications

M. Z. Iqbal, R. Rahman, K. T. Nahar, S. T. Bhuiyan, A. Islam, S. B. Alam, R. Rahman and M. A. Amin, "A Cross-Modal Benchmark of Deep Learning Encoders for Automated Liver Segmentation," *2025 28th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, Dec. 19–21, 2025. (Accepted)

Letter of Transmittal

14th December, 2025

To,
Saadia Binte Alam
Associate Professor and Head
Department of Computer Science and Engineering
Independent University, Bangladesh

Subject: Submission of Final Project Report

Dear Ma'am,

With great pleasure, we present our project report, "A Cross-Modal Benchmark of Deep Learning Encoders for Automated Liver Segmentation." The completion of this report is a prerequisite for the Independent University of Bangladesh's Bachelor of Science in Computer Science and Engineering program.

We have learned a great deal from working on this project, which has enabled us to apply deep learning techniques in medical imaging. We have made every effort to guarantee that this report satisfies the necessary academic requirements, and we hope you will find it acceptable.

Sincerely,

Redwan Rahman
ID: 2221936

Kazi Toushia Nahar
ID: 2221696

Contents

List of Publications **v**

Chapter 1 Introduction **1**
1.1 Background 4
1.2 Motivation 8
1.3 Objective 9

Chapter 2 Literature Review **11**
2.1 Prior Studies 11
2.2 Segmentation Challenges of the Liver 17
2.3 Practical and Clinical Challenges 18
2.4 Scope of Study 19

Chapter 3 Datasets **21**
3.1 Primary Training Dataset: LiverHccSeg 21
3.2 External Validation Datasets 22
3.2.1 3D-IRCADb-01 22
3.2.2 CHAOS 23
3.2.3 SLIVER07 23
3.3 Data Preprocessing and Standardization 24

Chapter 4 Methodology **25**
4.1 Proposed Method 25
4.2 Segmentation Architecture: U-Net 26
4.2.1 U-Net Overview and Design Philosophy 26
4.2.2 U-Net Architecture Details 26
4.3 Pre-Trained Encoder Backbones 27
4.3.1 ResNet18: Residual Networks 28
4.3.2 EfficientNetB3: Compound Scaling 28
4.3.3 DenseNet121: Dense Connectivity 29
4.3.4 MobileNetV2: Lightweight Architecture 30
4.3.5 Transfer Learning Framework 31
4.4 Training Configuration 32
4.5 Evaluation Framework 33
4.6 Cross-Modality Evaluation Strategy 34

Chapter 5 Results **36**
5.1 Primary Training Dataset Performance 36
5.1.1 Training and Test Metrics on LiverHccSeg 36
5.2 In-Domain Generalization: CHAOS Dataset 37
5.2.1 Results and Analysis 37
5.3 Cross-Modality Generalization: CT Datasets 38

5.3.1 3D-IRCADb-01 Dataset 38
5.3.2 SLIVER07 Dataset 39
5.4 Performance Summary and Overall Trends 39

Chapter 6 Discussion **41**
6.1 Architectural Insights from Encoder Performance 41
6.2 Compound Scaling & Balanced Feature Extraction 41
6.3 Dense Connectivity & Feature Reuse 42
6.4 Residual Connections & Limited Generalization 42
6.5 Lightweight Architectures & Computational Efficiency 43
6.6 Transfer Learning & ImageNet Pre-training 43

Chapter 7 Conclusion & Future Works **44**
7.1 Summary 44
7.2 Limitation 44
7.3 Future Work 45

Bibliography **47**

List of Figures

Figure 4.1.1 Proposed Method 25
Figure 4.2.1 U-Net Architecture 27
Figure 4.3.1 ResNet18 Architecture 28
Figure 4.3.2 EfficientNetB3 Architecture 29
Figure 4.3.3 DenseNet121 Architecture 30
Figure 4.3.4 MobileNetV2 Architecture 31
Figure 4.6.1 Evaluation strategy for in-domain and cross-modality generalization. 34

List of Tables

Table 3.1.1 Summary of Primary Training Dataset (Liver Segmentation Cohort). 22

Table 3.2.1 Summary of 3D-IRCADb-01 Dataset. 22

Table 3.2.2 Summary of CHAOS Dataset (Training Set). 23

Table 3.2.3 Summary of SLiver07 Dataset. 24

Table 5.1.1 Performance Metrics for Primary Dataset. 36

Table 5.2.1 Performance Metrics for CHAOS Dataset. 37

Table 5.3.1 Performance Metrics for 3D-IRCADb-01 Dataset. 38

Table 5.3.2 Performance Metrics for SLIVER07 Dataset. 39

Table 5.4.1 DSC Performance Summary Across All Datasets. 40

Chapter-1 : Introduction

The accurate detection of liver diseases has been one of the greatest concerns in modern science [1]. Liver diseases may not be detected for years or even decades without any obvious indications or symptoms. Due to a lack of early accurate diagnosis, liver diseases often reach an advanced stage-such as liver cancer or liver cirrhosis, before they are ever identified, at which point the prognosis is low and therapy is not successful [1] and the disease has often progressed to end-stage cirrhosis. By this time, some signs appear, such as jaundice, ascites, and variceal hemorrhage. As a result, the only option left is transplantation [1]. There is an immediate need for accessible and efficient methods to detect liver disease early, when the disease may still be reversible, and long-term effects might be avoided. The asymptomatic nature of the liver has become a significant public health issue [1].

Liver disease is one of the major global health problems that lead to death even though it is silent in nature [2]. Liver disease and its complications cause more than two million deaths annually across the globe [3]. The impact of liver cirrhosis (around 1.32 million deaths in 2017 globally), and primary liver cancer (hepatocellular carcinoma, or HCC), which resulted in 830,200 and 830,000 deaths respectively in 2020 are the primary causes of these death statistics respectively [3, 4]. All these conditions collectively contribute to 4 percent of the global mortality which translates to 1 in 25 deaths worldwide [3]. Some two out of every three liver deaths are associated with men and this shows that the effects were not equally distributed among men and women [3]. The end-stage effects of chronic liver disease (CLD) are the primary cause of this considerable mortality and economic effect [5]. This impact is significant and continues to grow, as both the Global Burden of Disease study and the data provided by the World Health Organization confirmed that this issue is a reality that is growing more and more tangible over time [5]. Despite the current situation when liver disease is the eleventh most common cause of death in the world, this number could be under-reported [3]. Most of the affected groups tend to be younger and economically active. The 12th cause of disability-adjusted life-years (DALYs) between the ages of 25 and 49 is liver disease [3, 5].

There are several important etiologies that are leading to the progression of CLD (Chronic Liver Disease) [2]. Hepatitis B (HBV) and C (HCV), two types of chronic viral hepatitis, were the most common of all chronic liver diseases. Indicatively, the incidence of chronic hepatitis B has been estimated to be 0.5 -0.7 percent in Europe but in the last decade, chronic hepatitis C was found to have a rate of between 0.13 -3.26 percent prevalence in the past decade alone [6]. In Europe, it is estimated that 90 percent of the infected individuals are unaware of their condition, which is one of the primary obstacles to treating and eliminating viral hepatitis infection. However, the epidemiological shift in the world is fast. The most common causes of the rising prevalence of Chronic liver disease in most regions of the world is now alcohol-related liver disease (ALD) and non-alcoholic fatty liver disease (NAFLD), even though viral hepatitis remains a significant cause of CLD,

particularly in the Middle Eastern, Northern African, and Asian regions [5]. Specifically, NAFLD is regarded to be the hepatic expression of the metabolic syndrome and is closely associated with the global epidemics of obesity and type 2 diabetes [2].

The pathophysiology of chronic liver diseases has a significantly widespread and detrimental progression irrespective of the underlying etiology (viral, metabolic, or toxic) [7]. The process begins by long-term, persistent destruction of the hepatocytes [7]. The major defining feature of the multifaceted and protracted wound-healing reaction that is provoked by this primary and sustained harm is chronic inflammation [7]. This inflammatory condition mediated by several cytokines and chemokines offers the signals that promote the onset and subsequent progression of liver fibrosis [7]. The main pathogenic mechanism is fibrosis that is characterized by the excessive deposition of the extracellular matrix (ECM) protein and scarring that gradually changes the normal lobular structure of the liver [7]. The main type of cell which leads to this uncontrolled deposition of the matrix is the hepatic stellate cell (HSC) where vitamin A is deposited in a healthy liver by quiescent hepatic stellate cells. However, they are activated to respond to continued inflammatory stimulation transforming into a myofibroblast-like cell which is proliferative, contractile and fibrogenic. The excessive secretion of collagen type I and other ECM components by such activated HSCs leads to the eventual replacement of functional liver parenchyma by excessive scar tissue which is the primary driving force behind fibrosis [7]. It is an automatic process leading to cirrhosis in case the underlying injury is not remedied [7].

Cirrhosis or the appearance of primary liver cancer are the common late-stage manifestations of this disease process [6]. Certainly, primary liver cancer, among which the most common one is hepatocellular carcinoma (HCC), represent a worrying concern. As the statistics of 2020 show, 830,200 individuals died of primary liver cancer and 905,700 persons were diagnosed with this issue all over the world in 2020 alone [8]. HCC itself comprises of about 80 percent of these primary liver malignancies [8]. The late-stage diagnosis of HCC is highlighted by the high ratio of mortality/incidence rate [4]. As of 2020, liver cancer was among the top 5 causes of cancer related mortality in 90 countries and one of the 3 leading causes of cancer in 46 countries across the globe [4]. By the year 2040, it may increase by over 55 percent in the number of new cases and deaths due to liver cancer, or more than 1.4 million new cases and 1.3 million deaths in every year, assuming that the current incidence and mortality ratio remains unchanged [4].

Other than the malignant tumors, benign tumors of the liver are also present in the clinical scene and are increasingly being detected [9]. It is mainly believed that the number of increase in incidence is largely caused by the widespread application of sophisticated medical imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), and ultrasonography, which are commonly used for non-medical issues [9]. Hepatic hemangioma, hepatic focal nodular hyperplasia (FNH), and hepatocellular adenoma (HCA) are the most widespread ones [9]. Hepatocellular adenomas (HCAs) are a more complicated clinical challenge despite the fact that the majority of the lesions are asymptomatic and need not be treated. Inevitable bleeding or hemorrhage and in certain types of HCA, malignant progression into hepatocellular carcinoma are two major side effects of HCAs, which are interrelated with the use of oral contraceptives [9].

This wide spectrum of disease, starts with the silent phase through to the advanced end-stage cancers, presents one of the major and complicated diagnostic challenges [1, 10, 11]. At the initial stages, the asymptomatic illness cannot be detected easily [1]. The last stage

is more advanced and radiology is a key centre at the later stages. Indicatively, the present-day imaging has had tremendous improvement in the management of HCC [11]. According to a general appearance in a multiphase contrast-enhanced CT scan or MRI scan, HCC may often be non-invasively, biopsy-free, and diagnosed in high-risk patients, including cirrhosis patients [11, 12]. This characteristic appearance is a radiological signature of the wash-in/wash-out, which is hyperenhancement during the late arterial phase and hypoperfusion during the portal-venous or delayed phase of the cycle [11]. Hepatocarcinogenesis can now be diagnosed at earlier stages thanks to the introduction of hepatobiliary contrast agents for MRI and the use of ultrasonic contrast agents [11].

But even when imaging methods are strictly followed, differential diagnosis is still challenging and the diagnosis is not always straightforward [11]. Additionally, imaging is essential for staging, therapy planning, and intervention as well as for initial diagnosis [8]. Clinical imaging is a very essential supplement to physicians in order to implement appropriate interventions such as radiation, surgical resection, or transplantation [8]. This imaging data requires advanced methods of computation to be useful. Specifically, image segmentation, or the proper identification of the liver and tumor, is employed by clinicians as a preliminary measure to maximize diagnosis and treatment planning [8]. Because the accuracy of the several researches have suggested that the clinical value of such liver segmentation techniques is an important component of modern clinical surgeries and procedures after a patient develops primary liver cancer, a detailed set of guidelines is required to direct the treatment that is both fairly complicated and dependent on the performance status of a patient [8]. After the diagnosis of primary liver cancer, the treatment is somewhat complicated and must be tailored to each patient and depend on the performance status of a patient, which is why such liver segmentation techniques should receive special attention [8].

After a primary diagnosis of primary liver cancer, a comprehensive list of guidelines is required to help with the treatment, which is very complicated, and depends on the functional condition of the patient, liver cirrhosis and the staging of the tumor [12]. Every existing method is incorporated in the guidelines of the treatment and diagnosis of primary liver cancer. As an illustration, although the main method of diagnosis is imaging, patients with focal hepatic lesions who do not have the typical imaging appearance of the HCC still need the liver biopsy [12]. In case of early-stage disease, the treatment may be liver transplantation, local ablation (radiofrequency ablation), surgery with resection of the cancerous mass and local ablation (radiofrequency ablation) of the tumor mass. Treatment of the later-stage disease can include either systemic therapy or transarterial chemoembolization (TACE) [8, 12] . The available staging techniques that consider the hepatic function, vascular invasion, and tumor extent can be used to help make the best choice of treatment option available [12].

Despite its limitations, pathological diagnosis remains the standard in cases where non-invasive imaging is inconclusive [11]. Hepatocellular carcinoma (HCC) is an exception, unlike other tumors such as thyroid and pancreatic tumors, where biopsy is routinely performed [10]. It may be challenging based on morphology to provide the pathological diagnosis required to distinctly separate different types of tumors, including a well-differentiated HCC and a hepatocellular adenoma [9, 10] or a primary intrahepatic cholangiocarcinoma and a metastatic adenocarcinoma [10]. There is a diagnostic complexity which requires the use of ancillary tools. The Immunohistochemistry (IHC) that uses antibodies to recognize specific protein markers within the tissue has evolved to be an important tool. In the

present, panels of markers are commonly used by pathologists to assist them in coming up with a definite and accurate diagnosis [10].

Lastly, well-established socioeconomic and systemic disparities worsen and magnify the epidemiological, pathological, and clinical difficulties of liver disease [13]. Given that the epidemiology of liver disease is evolving, research shows that systematic disparities in racial and ethnic minorities' access to high-quality healthcare have a quantifiable, negative effect on patient outcomes and mortality rates [13]. These differences have their roots in social determinants of health rather than being random. The management, access to therapy, and final results of diseases like acute and chronic viral hepatitis continue to show significant differences. Although there is availability of this curative treatment for HCV, minority communities face obstacles to proper diagnosis. This is a clear indication and pressing need for health policy changes and community-level interventions that are particularly intended to detect, address, and take necessary steps for these essential care gaps [13].

1.1 Background

The precise and successful distinction of the liver and its interior lesions in abdominal medical images is an initial, intricate procedure in the diagnosis, treatment planning, and follow-up of a broad range of hepatic diseases, and its clinical worth is the focus of the contemporary clinical, surgical and intervention procedures, which are the accuracy of directly affected patients' results [8]. This is called segmentation where the raw pixel data is converted into clinically relevant anatomical maps, which is the key to intermediary between medical imaging and quantitative analysis. In modern clinical practice, accurate liver segmentation is vital to a number of crucial, significant tasks, which lay the foundation of the modern hepatology and oncology. To carry out the surgical resection of a hepatocellular carcinoma (HCC) and other liver tumors, segmentation allows the surgeon to carry out the needed volumetric analysis, which can be used to calculate the total liver volume, the tumor volume, which determines the tumor burden and dictates the treatment strategy and, most importantly, the future liver remnant (FLR), which is the volume of liver that will be healthy after the surgery and which will determine the post-operative liver functioning and patient survival [8, 14]. A sufficient FLR is paramount to prevent post-hepatectomy liver failure, a severe complication with high mortality. The segmentation data can be used to replicate the postoperative anatomy and calculate the volume of segments II and III for complex resections. For example, in an extended right hepatectomy, the segmentation data will constitute the FLR. For further surgical decision, this method provides a significant role. Furthermore, in the field of radiation oncology, accurate 3D outlines of the liver and intrahepatic tumors are necessary for planning advanced radiotherapies like stereotactic body radiation therapy (SBRT), ensuring that a high-conformality, therapeutic dose of radiation is delivered to the target lesion while optimally sparing surrounding healthy liver parenchyma and adjacent organs-at-risk such as the kidneys, spinal cord, and bowel [8] [15]. The construction of an accurate dose-volume histogram (DVH), a dose versus volume plot of radiation dose to a part of the part being radiated, is entirely based on good quality segmentation; a slight error may cause either under-dose to the tumor or radiation over-dose of normal tissues. This anatomical information based on imaging is used to determine the best therapeutic intervention that can either be a surgical resection, a local ablation, or the transarterial chemoembolization (TACE) based on the established staging systems [8, 12]. There are staging systems, such as the Barcelona Clinic Liver Cancer (BCLC) system, that combine tumor characteristics, liver functioning, and patient

performance level guided by measurements obtained based on segmented images.

This important delineation task has traditionally been a manual or semi-automated task of expert radiologists, but manual delineation is an important bottleneck in the clinical workflow since it is highly time-consuming, very tedious, and above all, prone to significant inter-observer and intra-observer variation. One case can take 30 minutes up to an hour to dissect diligently and various professionals might make different delimitations to the same structure, thus causing discrepancies in the calculation of volumes and the treatment regimen. Such variability may result in inconsistency and unreliability of downstream quantitative measurements and the following clinical decisions. This has led to intensive research on automated segmentation techniques where deep learning has become the most prevalent and the most promising to date and is capable of automatising the interpretation of radiological images as well as facilitating quick and consistent analysis that could supplement clinical judgements [15, 16]. These models are likely to make the process of segmentation take just a few seconds and offer a standardized and more objective result, enhancing the efficiency of the workflow and decreasing the cognitive load on clinicians.

This change in technology came with the release of the U-Net architecture, a deep learning network in which the encoder and decoder have a synergistic network, specifically in biomedical images where precise localization is necessary, usually with a limited number of annotated training images [17]. Its structure directly resolved the drawbacks of previous fully convolutional networks, specifically, by incorporating spatial detail preservation mechanisms, which are readily lost throughout the downsampling procedure. The encoder is a feature extractor, a sequence of convolutional and max-pooling layers that downsample the input image in successive layers to learn a hierarchical representation of fine contextual detail - starting with low level edges and textures at the first few layers, high-level abstract shape at the deeper layers - at the expense of fine spatial detail - at the very deepest layer, with a network of convolutional and max-pooling layers [18]. The receptive field is expanded with every following layer in the encoder so that the network can capture bigger scale contextual patterns, like the shape of the liver in general, with respect to other organs. This compacted representation is then put the images into the decoder, which consists of transposed convolutions that upsample the images successively to form a correct pixel-level segmentation map. Such upsampling is coarse in nature and without any extra information, the reconstructed map would not be fine. The U-Net innovation is, actually, the use of skip connections which feed the encoder into the decoder, allowing the network to combine the high-resolution, fine-grained spatial features with the deep semantic features, a design that has since been the foundation of many millions of models [19]. These skip connections actually provide a path through the network that plays the role of a shortcut route to high-frequency spatial information to exploit the bottleneck of the network center, allowing the accurate localization of organ and tumor boundaries.

In this context, the common and effective way to operate is to train the power of major, already trained Convolutional Neural Network (CNN) such as ResNet as the encoder, and utilize information gained on natural images to enhance performance, a process called transfer learning [18]. Such architectures, which are trained on large datasets such as ImageNet, result in a strong initiation of the feature extractor, potentially enabling faster convergence and improved generalization to the medical task than training a system out of scratch, despite the fact that the source and target domain are highly dissimilar. Nonetheless, these models have a severe limitation to their clinical relevance in the domain shift, where the performance of the model severely drops when the model is used with

data which belong to a different distribution than that the model was trained on [20]. This change may be small like variance in contrast or noise amongst scanners made by the same manufacturer, or enormous, as observed. The most extreme form of this problem occurs in the cross-modal domain shift between Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). The two modalities have absolutely different principles of operation and data acquisition, which is a core issue in the medical image field [21]. CT images are a direct map of the tissue X-ray attenuation in standardized and absolute Hounsfield Units (HU) that give the same values of certain tissues among scanners. An example is that water is 0 HU, air is -1000 HU and bone can be above +400 HU, which would form a predictable and measurable grayscale map. Conversely, MRI generates images depending on the response of protons in a magnetic field and the relative signal intensities generated have no absolute scale and differ significantly between different sequences (e.g., T1-weighted, T2-weighted), parameters (TR, TE), and scanner equipment [20]. The T2-weighted scan of a tumor can have a hyperintense (bright) pattern, whereas a T1-weighted scan of the same tumor can have a hypointense (dark) pattern, and no standard measure of the tumor intensity can be relied upon, which makes it difficult to have a model to learn stable, invariant features. The underlying data distribution in this critical disparity leads to a significant decline in performance on an MRI data trial using a CT-trained model, which is a major limitation to the development of universal, clinically practical segmentation tools [20].

To solve this, the research community has devised a number of methods under the umbrella term of Unsupervised Domain Adaptation (UDA) which involves image-to-image translation-based methods using Generative Adversarial Networks (GANs) to make one domain appearance like the other, and adversarial feature alignment which trains a domain discriminator to cause the model encoder to generate features which are modality-neutral [20]. Image-to-image translation methods like CycleGAN are trying to learn a mapping between the source domain (CT) and the target domain (MRI) with no paired data, and generate the synthetic MRIs based on the CTs. Then, these translated images are then used to train the segmentation model to achieve higher performance on actual MRI images. By contrast, feature-level alignment algorithms operate in the latent space of the model, introducing a separate auxiliary classifier of domains that attempts to predict whether the feature map was obtained by a CT scan or an MRI scan. This classifier is then trained to be fooled by the main network whereupon the main network learns to produce features that are alike in both modalities and helps the model concentrate on anatomy and not modality appearance. These endeavors have been enabled and tested using public benchmarks, including the Liver Tumor Segmentation (LiTS) Challenge, a large CT-only dataset, which has become a standard to train models and determine such metrics as tumor burden [14], and the cross-modal CHAOS Challenge, which gives paired CT and MRI data but on healthy abdominal organs [22]. LiTS has played a key role in advancing CT liver and tumor segmentation but its inability to use multi-modal data limits its use in the investigation of domain shift as compared to CHAOS multi-modal benchmarking.

Although there are these advanced methods and standards of UDA, the underlying question that has not been thoroughly studied is that, given the inherent architecture of the specific encoder (whether it be a classic CNN with heavy locality bias or a modern Transformer with a global self-attention system) what is the effect on the innermost generalization ability across modalities of a model, independent of any complex adaptation procedures? CNN has an inductive bias towards locality and translation equivariance. They perform

well for learning hierarchical features at the pixel scale but can also there is a chance of failing to capture global dependencies that are easily seen by a human reader such as the symmetrical form of the liver or the connection between a tumor and a distant blood vessel. Transformers provide self-attention mechanisms. They perform efficiently when it comes to the fact of hierarchically learning pixel-level features. On the other hand, CNNs offer the benefit of strong spatial priors manifested by such relationships. Although the available literature is full of detailed reviews on variants of U-Nets and the wide range of applications of AI to hepatology [16, 19], it is deficient in the unique, objective and organized comparison of the underlying encoder infrastructure, in a zero-shot cross-modal context, to a specific, pathological problem. In the majority of studies, a new UDA approach is presented and compared to baselines with one backbone architecture, which hides the essence of the design of the encoder itself. A better designed architecture that has a higher intrinsic ability to learn modality-independent, anatomically-grounded representations would offer a more effective basis to generalization, perhaps removing the need to find more elaborate adaptation regimes. This thesis seeks to address this major gap by offering a standardized benchmark on which the role of encoder in cross-modal generalization can be isolated and measured by performing a controlled comparison of the state-of-the-art CNN and Transformer-based encoders, trained directly on a large CT dataset, and then directly on MRI with no fine-tuning or domain adaptation, to identify which family of architecture has the better inherent robustness to this form of domain shift that is the most challenging.

The rapid development of artificial intelligence (AI), specifically deep learning, has been the driving force of a paradigm shift in contemporary medicine, providing powerful methods to process and synthesize the huge amounts of multimodal patient data that contemporary healthcare systems generate [16, 18]. This does not only cover medical imaging, but also organized data of electronic health data, genomic sequence, and electronic pathology slides, enabling the possibility of holistic patient models. In hepatology, AI methods have demonstrated significant potential across the whole clinical landscape, including anticipating disease development up to making it possible to interpret radiological and pathological images in an automated and rapid fashion [15,16]. As an example, deep learning algorithms are being trained to anticipate the probability of fibrosis progression in NAFLD patients using baseline clinical and radiographic measurements, or to detect particular genetic defects in HCC using contrast-enhanced MRI images. Semantic segmentation, which is a process of labeling each and every pixel of an image with a particular class, is one of the important tasks that can be performed in this area, and is a foundational component of computer-aided diagnosis [18, 19]. This is a more complicated set of tasks than image-level classification or object detection because it needs an accurate, dense prediction of each individual point in the input, and is crucial to the quantitative uses of the aforementioned types of application.

This task was well-prepared with the technological basis of the U-Net, which is a convolutional neural network architecture, and its design is especially suitable to biomedical images, where correct localization is key, despite the small number of annotated training images [17]. The characteristic of the U-Net is the encoder-decoder architecture, which is synergistic: a contracting path, that is, the encoder, which is a feature extractor, consisting of a sequence of convolutional and max-pooling layers that gradually downsample the input image to learn a hierarchy of more complex contextual features, starting with low-level edges and textures, and then moving to high-level abstract shapes [18]. This is similar to how a radiologist will first distinguish to the edges and textures, then the organs and

eventually, a particular pathology. Comparably, an expansive network where the decoder flips the images around the compressed expression by transposed convolutions. As a result, the decoder gradually up samples the images to reconstruct the features for exact pixel-wise segmentation map. The novelty of the U-Net is that it uses skip connections, which direct the feature maps of the encoder to the decoder so that the network can integrate the high-resolution and fine-grained spatial detail needed to accurately draw boundary lines and deep semantic content needed to recognize a classification. The U-Net is currently known as the de facto standard in medical segmentation due to its successful design. This architecture has made the foundation of many subsequent models and a rich ecology of modifications surrounding it [19]. These extensions come in residual connection (ResUNet), dense connection (DenseUNet), attention connection (Attention U-Net), and addition of transformer blocks (UNet++) all aiming at enhancing the original blueprint to a more accurate and performance-driven one.

There is a specific clinical imperative of applying such powerful technology, especially in the field of management related to liver cancer, where clinical utility of accurate segmentation methods is the essence of the modern day clinical surgeries and operations [8]. Precise 3D segmentations are paramount and stakes to several activities of the patients with hepatocellular carcinoma (HCC). They are able to give the needed volumetric analysis in the surgical planning that may be able to assess the total liver volume, the tumor volume (that dictates the tumor burden) and the future liver remnant (FLR) that is considered as one of the greatest predictors in the post-surgical survival of the patient [8, 14]. It is critical in the case of liver transplantation of living donor where the graft volume of the recipient and remnant volume of the donor are adequate. This numerical data is necessary to inform the decision about the most suitable therapeutic treatment either by performing a surgical resection, local ablation, or transarterial chemoembolization (TACE) as is presented in the established clinical guidelines [8, 12]. A typical case example would be a patient who has one, small HCC and intact liver function, which would be a candidate of curative resection or ablation, as is indicated by the segmented characteristics of tumor and liver volumes. Traditionally, this delineation has been manual in nature, which is a significant bottleneck in the clinical workflow, it is time-consuming, and it is also prone to inter-observer and intra-observer variance. This inconsistency may result in unequal treatment prescriptions and outcomes at various medical facilities. The hope behind deep learning is that it would be able to conquer this by providing a faster, reproducible, and accurate automated solution [15, 16]. The point is not to put radiologist out of the job, but rather to augment their abilities, giving them an effective means by which to cope with the monotonous and time-consuming task of measurement, so that they can be left to concentrate on the more important task of diagnostic integration and patient care.

1.2 Motivation

The four interdependent issues that influence the achievement of deep learning as a standard of care in hepatology drive this study. These issues provide a foundation of a solution which should not only be accurate, but be also efficient, standardized, and intrinsically robust.

1. Clinical Bottlenecks and Efficiency: According to the established fact, the inter-observer variability and time-consuming characteristics of manual evaluation provide an urgent necessity of the automation. But available automated methods (e.g. complex 3D U-Nets) are not only computationally expensive, but also not easily ported to resource-constrained

settings. The absence of efficient and lightweight baselines that can provide fast inference without needing to have specific supercomputing facilities is very pronounced.

2. The "Missing Baseline" for LiverHccSeg: Although LiverHccSeg dataset is a useful open source to study hepatocellular carcinoma, it does not have a detailed performance benchmark at the moment. No standardized performance metrics of this data is provided to researchers in the future who might be trying to use them. It is necessary to establish a stringent standard on this particular database that would foster gradual yet steady improvements in the field.

3. The Domain Gap and Complexity Trap: MRI models often fail on CT data because of contrast difference as mentioned earlier. The current research direction in solving this is based on more sophisticated methods of adaptation (such as GANs) that tends to obscure the inherent abilities or limitations of the underlying network per se. This study is driven by the necessity to take away this complexity and define which encoder architectures have a natural robustness to domain shift.

1.3 Objective

The primary objective of the research is to implement a thorough and rigorous research project on the rampant and severely limiting issue of domain shift in the specific application area of automated liver segmentation using medical images. This issue is a serious constraint to both clinical implementation and practical application of deep learning models. Here, this is done by constructing a comprehensive, quantitative benchmark that thoroughly evaluates and contrasts the intrinsic, inherent cross-modal generalization properties of four widely used, and architecturally different deep learning encoders ResNet18, EfficientNetB3, DenseNet121, and MobileNetV2, in the context of a conventional U-Net segmentation model.

This project design is architecturally organized towards three goals, each aimed at testing a certain aspect of the generalization problem and all of which are expected to deliver a full test of encoder robustness.

The first and the main goal is to design, execute and educate a semantic segmentation framework founded on U-Net with much care, reproducibility and standardization. Four assorted pre-trained encoder structures are added to this framework and each of them is associated with a distinct philosophical stage of CNN design:

- ResNet18: Chosen due to its residual connections.
- EfficientNetB3: Chosen due to its compound scaling strategy.
- DenseNet121: Chosen due to its novel dense connectivity.
- MobileNetV2: Chosen because of its lightweight and efficiency-optimized architecture.

The second critical goal is to develop a robust and realistic cross-modal benchmarking protocol. Each of the four models is trained on LiverHccSeg dataset (MRI) and thoroughly tested in a zero-shot fashion on three independent validation sets:

1. 3D-IRCADb-01 (CT): The main cross-modality generalization test.

2. CHAOS (MRI): To evaluate the in-domain generalization and resilience to changes in scanners.

3. SLiver07 (CT): A secondary CT benchmark to achieve consistency of trends.

The third objective is to conduct a full quantitative analysis with the help of accepted standard measurements: Dice Similarity Coefficient (DSC) to determine the volumetric overlap, Intersection over Union (IoU) to determine the spatial agreement, and Precision/Recall to detect the type of errors in particular cases (over-segmentation vs. under-segmentation). This analysis will determine which encoder backbone is more robust to the domain shift between the training MRI domain and the CT domains.

Chapter-2 : Literature Review

2.1 Prior Studies

Automated liver segmentation has achieved a lot of progress in line with the strong advancements of deep learning architectures, though the ability to achieve consistent performance across imaging modalities is still a major concern. Cross modality generalization is the primary objective of the study and this chapter examines the approach of various encoder architectures to cross modality generalization.

Deep learning has progressed rapidly in medical image analysis, which can be seen through extensive surveys [23] that reveal its effectiveness in segmentation, detection and classification. The clinical potential of automated segmentation can be seen as deep neural networks achieve human-level performance on challenging tasks such as dermatological skin cancer categorization [24]. This development has been of great benefit in organ segmentation where anatomical heterogeneity and pathological abnormalities present difficulties to the conventional computer vision algorithms. Deep learning automatically discovers hierarchical feature representations of raw pixel data, as compared to traditional methods which used intensity distributions and shape priors. Certain aspects have enabled such development such as exponential increases in computational ability over GPUs, growing access to annotated medical imaging data and conceptual progress that makes it feasible to train very deep networks with millions of parameters. Deep learning has shown to be useful in segmentation exercises, in which local texture variations are very important to help accurately define boundaries of organs, but a global anatomical context is essential. The liver, which has significant inter-patient differences in size, shape, and location, is a perfect test ground to gauge the robustness and estimation potential of segmentation models. Segmentation models are applied specifically in cases where generalization across a variety of imaging modalities, representing fundamentally different physical tissue properties, is required.

The incorporation of deep residual learning [25] has further strengthened encoder design. Residual blocks and skip connections allow deep networks to be trained by mitigating the vanishing gradient problem. Instead of learning direct mappings, residual blocks learn residual functions which enable gradient propagation across hundreds of layers. This depth allows networks to capture progressively short and discriminative features like early layers encode low-level edges and textures, middle layers learn anatomical structures and higher layers capture semantic concepts such as organ boundaries and spatial relationships.

The 3D U-Net [26] extends the base U-Net architecture to volumetric data by replacing 2D operations with 3D equivalents, enabling segmentation of organs like the liver in full CT or MRI volumes, as present in the 3D-IRCADb-01, SLiver07, and CHAOS datasets used in this study. 3D processing introduces substantial computational demands while maintaining

the ability to capture spatial relationships in three dimensions. The 3D U-Net preserves the encoder-decoder structure of the 2D version but uses 3D convolutions, pooling, and upsampling to learn volumetric feature representations. This volume method is especially applicable to the segmentation of the liver through the complex 3D geometry of the organ and how it is related to other organs. The increased computational complexity of 3D networks necessitates architectural choices such as patch-based training, memory-efficient operations, and optimized data pipelines. Beyond the basic U-Net, enhancements like UNet++ [27] reduce the semantic gap between multi-resolution feature maps by improving information flow between encoder and decoder using dense and nested skip routes. As a result, both fine and coarse semantic differences can be better captured through more efficient feature learning at every scale. By altering the U-Net topology rather than the encoder itself, UNet++ has been used to segment the liver in abdominal CT scans to get better results. By learning to concentrate on relevant target structures while suppressing unimportant background regions, attention techniques like attention gates improve segmentation more. By highlighting areas of interest the attention maps enable the network to assign weights to spatial locations based on their significance for precise segmentation. Because attention maps show which areas the network prioritizes and the attention mechanism is understandable. This study systematically compares encoder backbones, in contrast to studies that alter decoding circuits. The variety of design approaches is demonstrated by other designs such as V-Net for volumetric segmentation, SegNet for memory-efficient encoder-decoder processing, and DeepLab for multi-scale feature extraction with various convolutions. Finding the best architectures requires systematic comparisons, especially when working with suboptimal datasets [28],where annotations may be weak, noisy, or sparse. This underscores the importance of encoders with strong generalization like EfficientNetB3 that performed best in this study. This explains why encoders with good generalization such as EfficientNetB3 used in this research worked the best. The competition is very stiff in medical imaging where manual annotations are very expensive and also suffer inter-observer variation as well as are constrained by privacy and regulations. Labeling 3D volumes are also quite tedious and several experts are needed to make ground truth. Well trained radiologists can still have a dispute about the boundary delineation in low contrast or difficult anatomical areas. The annotation bottleneck also demonstrates the need of high generalization ability models that can retain their performance amidst small or substandard training data.

Privacy laws like HIPAA and GDPR that limit data sharing and aggregation that makes the lack of annotated data even worse by preventing the development of large-scale datasets that could improve model generalization. Also medical imaging datasets may have noisy annotations where mistakes in ground truth labels might confuse model training and are frequently unbalanced with a large number of background pixels and very few foreground pixels. These difficulties have led to the creation of learning techniques like active learning, weak supervision, and semi-supervised learning. This cross-modality domain shift problem arises directly from the generalization challenge and thus represents the primary focus of this study. The theoretical foundations of this issue are well-documented in surveys of unsupervised deep domain adaptation [29], which review techniques designed to bridge the distribution gap between an annotated source domain and a potentially unlabeled target domain.

Domain adaptation methods generally fall into three categories. First one is the feature-level adaptation learns domain-invariant representations by aligning feature distributions across domains, using methods such as maximum mean discrepancy (MMD) minimization,

correlation alignment or adversarial domain confusion. Second one is the instance-level adaptation reweights or selects samples from the source domain to better match the target distribution which effectively generating a more realistic training sample through importance weighting or sample selection. Last one is the adversarial adaptation employs adversarial training to learn domain-invariant features by training domain discriminators to distinguish between source and target features while the feature extractor attempts to confuse the discriminator. To address these limitations, researchers have proposed several strategies. One approach uses deep stacked transformations which improve generalization by applying data augmentation that simulates domain variations [30]. This approach demonstrates that domain generalization can be enhanced by carefully designed augmentation pipelines that expose models to a broader range of imaging conditions during training.

The deep stacked transformation method utilizes a sequence of learnable transformations that are able to reproduce domain shifts that contains variation of intensity, contrast, noise, and spatial distortion that replicates differences between imaging modalities. The greater the awareness of these variations by the model, the more it will be resistant to any kind of change in the domain since it would be able to recognize anatomical structures despite the modality-specific traits. The attraction of augmentation-based strategies lies especially in the fact that they do not need annotated target domain data, thus they can be useful even in privacy-preserving robust models like FedDG [31], which combines federated learning with domain generalization. This strategy addresses challenging scenarios where a generalized model must be trained on distributed source domains and remain effective on entirely unseen data. In addition to the simple encoder benchmark set in this study, it offers a novel solution. The federated learning architecture solves privacy issues and the requirement for a variety of training datasets by allowing several institutions working together to train a model without directly exchanging raw patient data. The episodic learning strategy employed in FedDG simulates domain shifts during training to create episodes in which the model must perform successfully on data from one domain after training on others that effectively teaches it to generalize to unseen domains. Frequency-based techniques exploit the fact that domain variations can manifest in the frequency domain where different imaging modalities or scanners exhibit distinct frequency characteristics but they also require exposure to a broad range of imaging conditions to learn meaningful models which is achievable only through multi-institutional collaboration. Nevertheless, federated learning introduces additional challenges, such as communication overhead, non-IID distributions across institutions, and the need for sophisticated aggregation strategies to ensure model convergence and performance.

This research's architectural framework is built on significant advancements in convolutional neural networks (CNNs), which serve as the benchmarked encoders. Four contemporary encoder backbones which are DenseNet121, ResNet18, MobileNetV2, and EfficientNetB3 are the focus of this review. By joining all layers in a concatenated, feed-forward manner, DenseNet [32] established the idea of feature reuse. This approach maximizes parameter utilization also improves feature propagation and successfully resolves the vanishing-gradient issue. Because of the rich feature representations of the dense connection which provides efficient gradient flow, each layer has direct access to the original input signals and loss gradients. Concatenation maintains all prior characteristics instead of summing them in comparison to ResNet's addition feature. In comparison to ResNet variations, the original DenseNet paper showed greater accuracy with fewer parameters, attaining comparable performance with just half the parameters. This makes it very innovative for

medical imaging applications with limited computational resources and complementing identity mapping-based skip connections in ResNet18 to address training degradation in very deep networks. Residual learning allows networks to model residual mappings rather than direct mappings and simplifying optimization in very deep architectures. Importantly, the residual function that representing the difference between desired output and input, is easier to learn than the full transformation. This architecture achieved 3.57% error on the ImageNet test set and won the ILSVRC 2015 classification challenge by establishing its effectiveness as a feature extractor for segmentation tasks. ResNet18 with 18 layers has provided a balanced trade-off between performance and computational complexity by making it suitable in terms of both accuracy and inference speed. These developments opened the way for lightweight and efficient models such as MobileNetV2 [33]which further improve performance in resource-constrained environments through an inverted residual structure and linear bottlenecks. All of these robust encoder backbones are embedded within the original U-Net [33], which remains the dominant framework for semantic segmentation in medical imaging.

Other architectural variations that extend or modify the U-Net include Attention U-Net [27] and and Attention U-Net [34] modifies internal connections to implement spatial attention or decrease semantic gaps. By improving network topology and attention methods without changing the encoder backbone, these systems improve segmentation performance. The diversity of architectural strategies underscores the complexity of segmentation performance which can be influenced by encoder and decoder design, skip connections, attention mechanisms or combinations. Systematic benchmarking is essential to identify which components contribute most under specific conditions such as cross-modality testing. Extensive literature exists on refining segmentation of liver and liver lesions in CT and MRI. The Modified U-Net (mU-Net) [35] shows object specific high-level features to improve liver and liver-tumor localization by achieving a Dice Similarity Coefficient (DSC) of 93.73% on the 3DIRCADb whole liver dataset. This demonstrates that incorporating prior knowledge of target structures can significantly enhance segmentation specially for large lesions ($>$20 mm) and small lesions ($<$10 mm) are consistent with meta-analyses on colorectal liver metastases detection [36, 37]. Moreover, semi-automated segmentations corrected by users reduce processing time and improve small lesion detection which suggests that automated approaches can enhance radiologist efficiency without significantly affecting measurement agreement.

Other studies employing fully convolutional neural networks with object-based postprocessing to automatically segment liver tumors in CT [38] report a mean DSC of 0.69 for detected tumors, comparable to the human annotators' score of 0.72 on the test dataset. This result highlights not only the potential of deep learning methods to achieve human-level performance but also the persistent challenges in segmenting small and irregular structures such as tumors. Tumor segmentation is inherently more difficult than whole-organ segmentation, as tumors are typically smaller, irregularly shaped, highly variable, and often share intensity characteristics with healthy tissue.

Residual networks have practical application in cascaded deep residual network architectures for liver lesion detection and segmentation [37], implementing a multi-stage process that applies residual networks in a structured manner. Another important strategy in medical image segmentation is the cascaded approach, in which a coarse initial segmentation is refined in subsequent stages, allowing the network to focus computational resources on the most challenging regions. This hierarchical model allows the model to add the global

detail of the coarse segmentation with the local detail of the refinement stages, which is usually much more accurate than single-stage approaches.

Even for more better approaches, including 3D multi-domain segmentation which is determined by unsupervised Wasserstein distance using stage able encoder architectures [39], utilise disentangled representations in a 3D scene to reduce discrepancies in the distribution of domains. The Wasserstein distance (earth movers distance) is a more informative standard of divergence of distributions than traditional standards such as Maximum Mean Discrepancy. Its benefits are that it gives meaningful gradients even in cases where distributions are not continuous, it is robust in high dimensional feature space and it reflects geometric relationships between distributions that other measures might fail to capture. The Wasserstein distance is an effective measure of differences between high-dimensional feature distributions used in the context of 3D medical image segmentation, which allows more precise domain alignment.

Studies [30] indicate that generalization can be enhanced by data augmentation methods specifically targeting to enhance the results of DSC, which rose to 79.4 instead of 70.5 in crossvendor datasets of ultrasound. This depicts that properly designed augmentation pipelines, which model domain variations, have the ability to induce generalization without undergoing complicated architecture adjustments or extra labeled evidence. The advantage of augmentation is especially in that it could be implemented during training without requiring a change in the network architecture to allow measurement of natural domain shift in these popular encoders in an unambiguous way, prior to applying computationally expensive domain adaptation. These results of the comparison give empirical evidence of the inherent strength of various encoder architectures, providing essential advice on the choice of backbones of domain adaptation strategies. The importance of the architectural effects, including the scaling of compounds and the effectiveness of feature extraction, in the context of enhancing cross-domain generalization is highlighted by the fact that EfficientNetB3 is the strongest encoder. This standard provides a credible baseline on which strategies of domain adaptation in the future can be assessed, and the systematic assessment of encoder architectures provides information about which features of the design can confer robustness, and helps in designing better architectures that can be used to effectively adapt to a change in domain.

Machine learning models have limited extrapolation capabilities to the training distribution, especially in high stakes tasks like medical imaging where domain shifts represented by differences in data acquisition, patient demographics, clinical practice and annotation protocols are significant and the classical i.i.d. assumption is not met. In reaction, the domain generalization (DG) paradigm has been introduced as an organized design procedure in creating designs that retain solid functioning on unseen target domains. In general, there are three classes of DG strategies: the data manipulation (such as augmentation, randomization, and synthetic data generation, which are comprehensively surveyed [40]) to enhance diversity; the representation learning (via adversarial training, explicit alignment, or disentanglement, or built on top of data manipulation strategies) which aims to improve the out-of-distribution generalization; and learning strategies (such as meta-learning and distributionally robust optimization), which directs the training process to enhance the out-of-distribution generalization.

However, empirical research in practice has always shown discrepancies between theory predictions and generalization in reality. In healthcare, it is common to fine-tune pre-trained

models on popular large datasets of natural images like ImageNet. Nevertheless, it is shown in [41] that this strategy is only of limited use: even more basic convolutional networks that are trained end-to-end can match or even surpass results of highly parameterized ImageNet models. It means that the features, acquired in natural pictures, might not directly be transferred to medical diagnostics, where the patterns, which are important, are usually local and texturebased and are not related to the semantically different images of objects in nature. The study also mentions that the perceived benefits of transfer learning can be realized through better initiation of weight and scaling, which are reasons why domain-specific and architecture based learning methods are important. The feasible effects of inadequate generalization can be seen in a clinical research [42] comparing deep learning models of pneumonia detection between three hospital settings. The models tended to make use of site-specific confounding data, like institution markers, device signatures or disparities in prevalence of disease, as opposed to actual pathological characteristics. Consequently, models are found to be doing well on data of the same institution but failing when used in other institutions. It is important to note that a model that was trained on data of two hospitals did well on data of these two hospitals but failed on data of a third hospital, which demonstrates that the aggregation of heterogeneous data can unintentionally encourage shortcut learning.

The field of medical image segmentation has experienced a revolution during the last decade due to the emergence of the concept of deep learning. The classical liver segmentation procedures, such as thresholding, region growing, active contours, and graph cuts, were computationally efficient but prone to differences in patient anatomy, imaging procedures, scanner models and pathological changes. Convolutional neural networks (CNNs) offer a solution to this, as they learn features which are flexible to this variability, and perform better than manually derived feature strategies. Precise liver segmentation is of clinical importance: the liver is the biggest internal organ, which plays a vital role in metabolism, detoxication and protein production. Many applications need segmentation such as surgical planning of liver resection, volumetric assessment of transplantation of living donors, radiation therapy planning, follow-up of diseases, and computer-aided diagnosis. The importance of liver volume measurement in surgery determines safe partial hepatectomy and remnant functional liver volumes (which are directly affected by patient safety and surgical results). In radiotherapy, the appropriate segmentation of the liver is necessary to guarantee that normal tissue is exposed to the minimum radiation as possible and the best options are Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), each having its advantages and disadvantages. The most popular CT modality used in imaging liver diseases is the high spatial resolution of this modality and the high accessibility of it. Its capacity to visualize the anatomical structures in high contrast especially with the administration of contrast agents is improved. Nevertheless, CT has some significant limitations, including its exposure to ionizing radiation, in which case it cannot be applied to specific patient groups, and its radiation dose should be optimized to a minimum. MRI, however, has better soft tissue contrast even in the absence of ionizing radiations and is therefore particularly useful with detailed liver characterization as well as with patients who have to be imaged regularly. The T1-weighted sequence, T2-weighted sequence, and contrastenhanced sequence are commonly used in the modern imaging of liver. A number of variables have led to the transition of old approaches to deep learning-based approaches. To begin with, bigger annotated medical imaging datasets now exist, but they are smaller than normal natural image datasets. Second, the recent progress in computational hardware computing, especially Graphics Processing Unit (GPUs) has allowed the training of deep neural networks on high-resolution 3D

medical volumes. Third, medical-image-aware architectures, including the U-Net, are good structures to use when performing dense prediction tasks, including segmentation. Lastly, transfer learning permits a beneficial extraction of features in big natural pictures of images, even previously when the medical imaging information is scanty.

2.2 Segmentation Challenges of the Liver

The development of automated liver segmentation systems is affected by numerous technical, clinical, and practical challenges, which have guided the development of the field. These challenges will be crucial in putting the previous studies into perspective and in bringing areas where improvements may be made in the future. These problems cut across various fronts, not only with basic concerns of data acquisition and annotation. Moreover, with more complicated issues of model generalization and integration into clinical processes, making it a complex terrain, and necessitating the use of advanced solutions. Medical imaging data are never constant and the variability can occur at various levels. The anatomy of patients varies significantly, imaging protocols, scanner manufacturers, administration of contrast agents and timing, positioning of patients and respiratory motions make changes. Liver appearance and shape are also distorted by pathological conditions. The liver itself has a great anatomical diversity of size, form, and location as compared to the other organs around in the abdomen. It is further complicated by the fact that it has similar intensities with other organs on both CT and MRI (heart, stomach, spleen, kidneys) that result in additional boundary delineation problems. Altogether, these lead to domain shift, where models trained using one dataset or imaging protocol may not be applicable to other sources, scanners, or institutions. The demographic and pathological factors add to anatomical variability. The accuracy of segmentation may be different at different ages, gender related variations, and ethnic differences in the morphology of organs. Such diseases as liver cirrhosis may result in extreme morphological alterations and textural distortions, tumors may result in abnormal edges and irregular intensity patterns, and fatty liver disease may have an impact on the appearance of tissues in a manner that defies intensity-based segmentation. Surgical procedures, including partial hepatectomy, add anatomical modifications which are not likely to be captured by training data, which point to the desirability of models with the ability to generalize to acquired anatomical shapes. Another severe limitation is the presence of incomplete and inconsistent annotations. Medical image annotation is very time-consuming, costly, and needs special knowledge. It is also a subjective process, and the inter-observer variation is also well-known. The definition of boundaries can be disputed even by experienced annotators especially in those areas where the liver borders other organs or in complicated pathological cases. This variability can lead to noise in training data and this may reduce the performance of supervised learning methods. The lack of annotated data in rare cases and in pediatric cases, and specific imaging guidelines, brings about the challenges of long-tail distribution, which further complicates the generalization. The complexity is pushed to three dimensions (3D) volumetric annotation. The annotation of a complete 3D volume may require hours per case, which in comparison with 2D images is much quicker. The consistency between slices is essential because the systematic errors during training spread. The quality of annotations may also depend on experience of the annotators, the complexity of the case, and tools which are used, so the datasets can be heterogeneous and have a negative influence on model performance. Although the effects of these challenges have been reduced in semi-supervised and weakly supervised learning approaches, the annotation bottleneck remains. Cross-modality domain shift is one of the key challenges to the medical image

segmentation. CT and MRI are two different imaging systems, which vary in terms of imaging physics, resulting in differences in the distribution of intensity, contrast, and noise patterns. The CT images are based on how X-rays are attenuated and the resulting images are Hounsfield Units (HU), which are not subject to much variation across different scanners but are subject to reconstruction methods and calibration. MRI relies on the magnetic resonance properties, and it is very sensitive to scanner specific issues like field strength, coil design, pulse sequences, as well as reconstruction methods. These discrepancies have significant distributional changes, which lead to drastic performance declines when the models trained in one modality are used on the other. CT tends to have a better spatial resolution whereas MRI has a better soft tissue contrast. There is also a difference in noise characteristics, in that CT enjoys quantum noise whereas MRI has thermal noise, variation of coil sensitivity and reconstruction artifacts. Therefore, the models have had to be trained on modality-invariant representations learning anatomical structures, but not modality-specific imaging properties. In practice, cross-modality generalization is necessary because there are numerous clinical situations where the models are need in both modalities.

2.3 Practical and Clinical Challenges

Medical imaging has always been a problem of class imbalance. The liver pixels ratio may change significantly depending on the field of view, patient anatomy, and imaging. In other instances the liver can be a minor percentage (below 5) of the whole volume which is an extreme imbalance that can hardly be well handled by normal loss functions. Besides, the liver pixels do not appear evenly, with a few areas being thick and others being thin which makes model learning more difficult. Correct boundary segmentation can be more clinically significant than an interior of the liver, but the percentage of pixels that are a boundary is even less, which makes it an imbalance problem of the nesting nature and necessitates the development of more sophisticated solutions. Although research prototypes may be more focused on accuracy, clinical systems need to be made to run close to real-time in order to become part of workflows. This need is biased towards light models and recommended inference strategies, which develops a trade-off between the complexity of the model and the computation necessary. The models which can be executed on regular clinical workstations without special hardware (including dedicated GPUs) are especially useful. MobileNetV2 is an architecture that fits such deployment well. The issues of efficiency do not only end with the speed of inferences but also with the deployment pipeline. Clinical systems are required to process an uncertain number of inputs, multiple patients in batches and connect with the current Picture Archiving and Communication Systems (PACS) and Radiology Information Systems (RIS). The overall computational load includes preprocessing and postprocessing tasks which include intensity normalization, resampling, and morphological operations that should be taken into account during system design. More activities, such as uncertainty quantification and clinical quality evaluation, add further to computational needs, and it should be noted that holistic efficiency is not due to model architecture only. Another challenge to developing a strong model of segmentation is privacy and data-sharing limitations. The Health Insurance Portability and Accountability is one of the regulations. In the United States (Act (HIPAA)) and Europe (General Data Protection Regulation (GDPR)) it is not permitted to combine datasets of multiple institutions to prevent the development of models that are applicable across different imaging settings. The idea of federated learning has become one of the possible solutions, which enables distributed training

of models without exchanging raw data, yet these approaches are not that simple, and thus do not address the issue of domain shift entirely. The privacy dilemma is also complicated by the need to have various training data. Single institution datasets do not tend to capture the differences in patient demographics, imaging regimens, and vendors of scanners yielding models that might not be generalizable to different environments. The privacy policies that restrict the aggregation of cross-institutions data establish a data privacy and model robustness trade-off. Sharing of data is also cumbersome and in most cases it has to be approved by institutional review board (IRB). Different privacy, secure multi-party computation, and federated learning are methods to reduce these problems, though each of them comes with trade-offs. There are also major challenges in the form of clinical validation and regulatory approval that can prevent wide adoption. Strong performance has been often reported in research studies on curated datasets, but to apply these findings to real-world clinical settings, large-scale validation of are needed on large patient populations, imaging protocols, and clinical conditions. Segmentation software should prove to be secure, effective, and useful to the clinical practice, which is a time-consuming and expensive process. The distance between research performance and clinical implementation is one of the key barriers.

2.4 Scope of Study

These reviews addresses the issue of automated liver segmentation, and in particular, deep learning-based methods of liver segmentation and how it performs in various imaging modalities. It analytically discusses architectural principles, design procedures and unsolved issues in the practice field giving it a rich historical background of the state of the art practices and research areas of the future. The review proceeds in chronological order, as it describes the development of ideas to the modern achievements, and emphasize the particular research issues that the benchmark study provides. The timeframe is between the initial advancement of convolutional neural networks in the middle of the 2010s and the modern innovations in the field of domain adaptation and generalization. Although it may seem that more traditional image processing algorithms existed prior to this, this review focuses specifically on the deep learning era, which has been taking over liver segmentation studies since around 2015. The coverage represents seminal systems like the U-Net and 3D variations, and more recent systems, like attention mechanisms or domain adaptation methods and federated learning plans. The literature review is a comprehensive coverage of the history of deep learningbased liver segmentation, including intermediate proof-of-concept results and full-fledged, clinically-focused results. On the architectural level, the survey analyses encoder-decoder models that are popular in liver segmentation and each include ResNet18, EfficientNetB3, DenseNet121, and MobileNetV2. It discusses how these encoders can be integrated into segmentation pipelines which were developed to classify natural images with reference to design principles, computational properties and their performance in the case of medical imaging. Other architectural materials, including UNet++, Attention U-Net and variations of the U-Net, are outlined in terms of their role in addressing the flaws of baseline architectures. Both 2D and 3D approaches are discussed with the consideration of the role of volumetric processing in the precise segmentation of the liver in CT and MRI images. The review also analyses the effects of architectural choices on segmentation accuracy, computational efficiency, and memory requirements and generalization capabilities which are important in research and clinical implementation. The review is methodologically inclusive of supervised and unsupervised approaches to learning and majorly takes supervised methods because they are prevalent and apply to

this benchmark study. It analyses strategies of training, loss, data augmentation methods and post-processing methodologies with an aim of improving segmentation performance. It gives emphasis to domain adaptation and generalization, as they target the problem of crossmodality. The review takes into account the idea of transfer learning, in which models trained on large natural image datasets are then fine-tuned on medical images, and the idea of domain-specific training in which models are trained with only medical data. The functions used in loss, such as Dice loss, cross-entropy loss, and boundary-aware loss, are investigated through their capability to solve the class imbalance and boundary precision. Various techniques of data augmentation, including geometric transformations and intensity changes, elastic distortions and mixup approaches that model domain changes are explained. Post-processing methods, including morphological operations, connected component analysis, and probabilistic refinement, are considered in terms of quality of the segmentation outcomes and alleviation of typical failure modes. The reviewed methods are mostly CT and MRI modalities that are most commonly used to study the liver. Although ultrasound and PET are occasionally employed, automated segmentation has been most commonly developed and tested on CT and MRI. Among the issues that present difficulties in cross-modality generalization are the modality-specific properties, contrast mechanisms, noise properties, and intensity distributions. CT imaging, involving X-ray attenuation, generates Hounsfield Units which in general are the same between scanners but which are variable to reconstruction and contrast agent variation. MRI imaging depends on properties of magnetic resonance and is very sensitive to scanner specific factors, such as field strength, coil designs, pulse sequences, and reconstruction protocols, which cause high intra-modality variation. The analysis looks at the impact that these differences have on the feature distributions, generalization of models, and cross-modality behaviors. The contrast-enhanced and non-contrast protocols are both taken into account, which are depicted in clinical practices. Although lesion-specific segmentation is also recognized, whole-organ liver segmentation is still of primary interest, which forms the basis of clinical activities including surgical planning, volumetric measurements, radiotherapy planning, and disease follow-ups. The review also contextualizes liver segmentation in the context of the multi-organ abdominal segmentation in that there are challenges that are specific to each organ though liver continues to be the priority segmentation. Discussed are the evaluation metrics, such as: Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Average Surface Distance (ASD). The review relates these metrics to clinical relevance with reference to the limitation of quantitative scores that are exclusively quantitative. Even though the studies of clinical validation are mentioned, the focus is done on the technical performance measures reported in research publications. The review considers the ways in which various metrics can be used to reflect the quality of segmentation: the overlap based measures such as DSC, reflect the overall segmentation accuracy, and the distance based measures such as HD, reflect the accuracy of the boundary which is usually important in clinical practice. The impact of metric choice on interpretation, sensitivity to error and usefulness in clinical deployment have been put into consideration. Though technical measures are focal, the review recognizes that the other factors like computational efficiency, ability to withstand imaging variations, and integrability into clinical processes too play a role in the general utility. The review is based on the works of academic centers, research hospitals, and industry organizations worldwide, considering that medical imaging research is always very collaborative and multi-institutional, but narrowing down to technical and methodological developments, as opposed to institutional or geographical considerations.

Chapter-3 : Datasets

This chapter provides a description of the datasets, which were all publicly available, used to train and evaluate the liver segmentation models. These datasets were chosen to define an evaluation framework that measures both the models' performance on the primary training domain and their ability to generalize to new data from diverse sources and imaging modalities. In-domain generalization (MRI to MRI) and cross-modality generalization (MRI to CT) were evaluated through an experimental design that used a single dataset for training and three separate external datasets for validation. The assessment of the effectiveness of deep learning models in adapting to the differences between imaging methods, acquisition conditions, and data distributions- significant factors in the construction of trusted segmentation systems requires a multi-dataset methodology.

3.1 Primary Training Dataset: LiverHccSeg

LiverHccSeg dataset [43] is a set of multiphasic contrast-enhanced Magnetic Resonance Imaging (MRI) scans that were obtained from The Cancer Genome Atlas Liver Hepatocellular Carcinoma Collection (TCGA-LIHC). All the scans have segmentation labels of all liver and tumor regions. To conduct this study, arterial-phase T1-weighted images were picked to be used for model training due to the presence of unique imaging properties and high contrast features among tissues that leads to a suitable foundation for segmentation algorithm development.

The data includes segmentation masks of the liver parenchyma and tumors. Annotations were created by two board-certified abdominal radiologists, guaranteeing the accuracy and dependability of the ground truth. The dual-rater annotation procedure is an indicator of quality assurance. In order to have consistency in training and testing, the ground truth was represented by segmentation masks of the first rater. As such, a single-reference method eliminates the possibility of inter-rater variability during the analysis, which is common to medical imaging research.

The images in the dataset were in Neuroimaging Informatics Technology Initiative (NIfTI) format. In order to make the correct assessment of generalization, the subject level was used to split the dataset but not slice. The training group had 854 slices (60%) of 10 individuals and the test group had 524 slices (40%) of seven individuals. Subject-level splitting is a fundamental requirement of analyzing the model generalization to new individuals that were not witnessed before because it ensures that no slices of the same subject will be found in both the test and training sets. Also, the NIfTI format preserves the spatial data and metadata about the image orientation and transformation thus it is appropriate in three-dimensional volumetric medical imaging. The key information about the dataset is presented in Table 3.1.1.

Table 3.1.1: Summary of Primary Training Dataset (Liver Segmentation Cohort).

Attribute	**Details**
Modality	MRI (Arterial-phase T1-weighted)
Number of Subjects	17 (HCC Patients)
Number of Data	17
Total Slices	1,378
Original Format	NIfTI
Mean Age	61 ($\pm$ 10.77) years
Number of Males	11
Number of Females	6

3.2 External Validation Datasets

Three different external datasets were used to test the capability of the trained models. These datasets were selected to make the imaging modalities, acquisition protocols, scanner parameters, and data distributions vary significantly and were not included in the training process. The multi-dataset validation technique assists in determining possible weaknesses of the model in a variety of domain settings and provides very strong evidence of the generalization strength.

3.2.1 3D-IRCADb-01

The 3D-IRCADb-01 dataset [44] served as the main benchmark for assessing cross-modality generalization. There is a particular difficulty and benefit of this assessment as CT and MRI are fundamentally different methods of imaging with unique physical principles and image properties. The volumetric CT scans provide three dimensional spatial data concerning the liver anatomy. A wide range of anatomical variations, including normal and diseased tissue are present in them.

The dataset is made up of expert-labeled liver and tumor segmentation masks which are used to measure the model performance. These annotations allow assessing the level of performance of the model both on abnormal tissue regions and on the general accuracy of liver segmentation. In preprocessing, the volumes were transformed to the NIfTI format to be compatible with the training pipeline and the volume slices were normalized into 2D PNG images including binary segmentation masks. The key information about the dataset is presented in Table 3.2.1.

Table 3.2.1: Summary of 3D-IRCADb-01 Dataset.

Attribute	**Details**
Modality	CT (Contrast-enhanced)
Number of Subjects	20
Number of Data	20
Total Slices	2,823
Original Format	DICOM
Number of Males	10
Number of Females	10

3.2.2 CHAOS

The CHAOS (Combined Healthy Abdominal Organ Segmentation) dataset [45], which examined the model's stability within the same imaging modality but in a different source, was used to demonstrate in-domain generalization. This test is important in assessing whether models have learnt generalizable anatomy features or that of a particular imaging modality. In the case when MRI data of other scanners of other manufacturers or other parameters is faced, CHAOS offers a useful domain shift that is an indication of realistic conditions. The T1-DUAL in-phase MRI sequences were only utilized in this study. The T1-DUAL sequence is significantly different compared with the T1-weighted images of the arteries obtained in the training because of the differences in the acquisition parameters and timing protocols. This difference provides a suitable challenge of model resistance to intra-modality confinements whilst staying inside the same imaging modality.

The validation step additionally includes healthy individuals whose tissue distributions are different from those of the training dataset. The dataset includes liver segmentation masks which are used for evaluation. In contrast to LiverHccSeg that targets patients that have certain clinical conditions, CHAOS presents a broader population with different anatomical characteristics, allowing testing the performance of the model on tissue characteristics of various types.

The preprocessing included the conversion of the images to NIfTI format followed by slicing them to normalized 2D PNG images. The proper manipulation of the segmentation masks was a required preprocessing stage. The original PNG masks that had non-standard pixel intensity values of 55 to 70 were transformed to binary masks (0 and 1) in order to be compatible with the model training pipeline. The standardization ensures that all validation data has the same representation. The key information about the dataset is presented in Table 3.2.2.

Table 3.2.2: Summary of CHAOS Dataset (Training Set).

Attribute	**Details**
Modality	MRI (T1-DUAL In-phase)
Number of Subjects	20 (Healthy)
Number of Data	20
Total Slices	647
Original Format	DICOM

3.2.3 SLIVER07

A second additional benchmark of cross-modality generalization of MRI to CT used SLIVER07 [46]. The use of several datasets to represent the same modality transition reduces the chances that findings are due to a single dataset and strengthens the assessment. This dataset indicates the modality of CT and can be used to assess the performance of the model on a different type of imaging, just as 3D-IRCADb-01. The liver segmentation masks in the dataset have been labeled by qualified radiologists.

SLIVER07 provides a standard framework validation, which can be compared with the findings of other methods. Preprocessing methodology used here was the same as of the

prior datasets: MHD files were turned into binary masks and normalized to 2D PNG slices to be compatible with the model training pipeline. Table 3.2.3 provides the information about the dataset.

Table 3.2.3: Summary of SLiver07 Dataset.

Attribute	**Details**
Modality	CT (Contrast-enhanced)
Number of Subjects	20 (Training Set)
Number of Data	20
Total Slices	4,159
Original Format	MHD (MetaImage)

3.3 Data Preprocessing and Standardization

A comprehensive preprocessing pipeline was created to standardize all datasets to one format that the segmentation models could use since the data was in various sources and file formats. This standardization step ensures the interoperability with the pre-trained encoder networks and allows making fair comparisons among datasets. The datasets were presented in 3 file formats: MHD (SLIVER07), DICOM (CHAOS and 3D-IRCADb-01) and NIfTI (LiverHccSeg). All volumetric data were initially transformed to an intermediate NIfTI format to specify a standard processing pipeline. This in-between step stores volumetric data and at the same time streamlines subsequent processing steps. After conversion, every NIfTI volume was divided into single 2D axial slices and stored as PNG files. Each slice was intensity normalized to ensure that all datasets and scanners sources had similar pixel value ranges. This normalization is particularly necessary in case of transfer learning in presence of pre-trained encoders since the models will be fed with input values that are within the same range as those between the ImageNet dataset.

Segmentation masks were handled based on their dataset to ensure consistency with the model training pipeline. In the case of 3D-IRCADb-01, masks were in DICOM format, and the same slicing and normalization operations were applied to them as to the corresponding image volumes. The extra process required by CHAOS masks is that the raw PNG masks had pixel values ranging between 55 and 70 as opposed to binary values (0 and 1). These mask were binarized to match with the methods of evaluation and the loss functions applied in training and validation. SLIVER07 masks in MHD format were transformed to 2D PNG slices by the same normalization techniques. Although the input formats may be diverse, this customized processing ensures that every mask is expressed as a binary image (0 background and 1 liver).

The preprocessing pipeline transforms the various input data to a standard format of 2D PNG images with binary segmentation masks. This standardisation is compatible with pre-trained encoders that require normalised input ranges and ensures that there is consistency among data sources. It also removes format-based differences in performance, enables easy batch processing when training and performing validation, and makes implementation easy with less potential error sources.

Chapter-4 : Methodology

This chapter discusses a U-Net architecture for liver segmentation that is based on deep learning and has several encoder backbones. In order to determine which encoder options offer superior generalization across various imaging modalities, the proposed method offers a methodical benchmark study of many pre-trained encoders integrated into the U-Net architecture. The major purpose is to generate baseline performance benchmarks that can guide the selection of suitable encoder backbones for accurate cross-modality liver segmentation.

4.1 Proposed Method

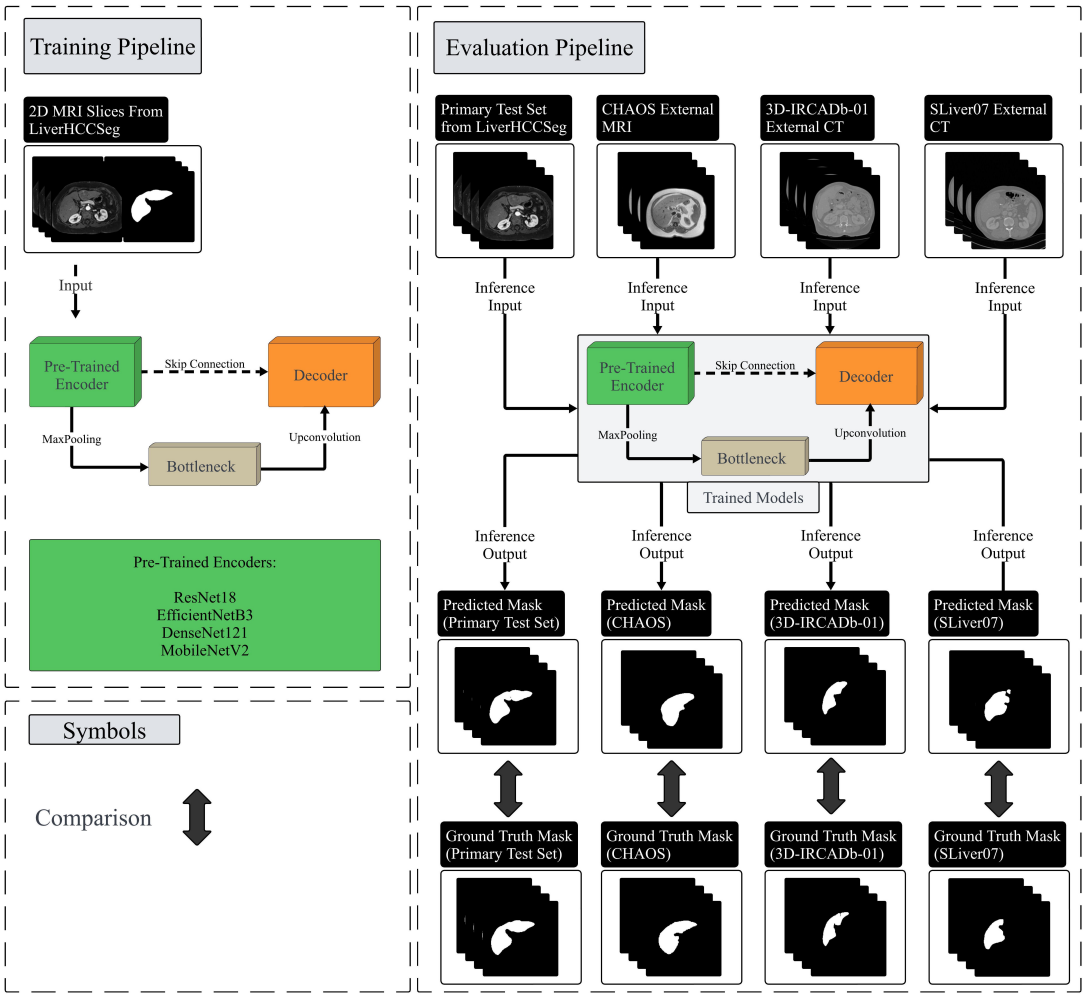


Figure 4.1.1: Proposed Method

The experimental framework has two sections as shown in Figure 4.1.1. The Training

Pipeline (left) takes the 2D MRI slices from the LiverHCCSeg dataset as input to the segmentation network. The network makes use of a U-Net architecture with the standard encoder substituted with pre-trained backbones such as ResNet18, EfficientNetB3, DenseNet121, and MobileNetV2 connected to the decoder by skip connections and a bottleneck layer. The Evaluation Pipeline (right) is used to test the generalization ability of the trained model on four different datasets, namely the primary LiverHCCSeg test set, external MRI dataset (CHAOS), and two external CT datasets (3D-IRCADb-01 and SLiver07). The inference process produces the predicted masks for every dataset and these are compared with the ground truth masks to measure the segmentation performance in intra-modality and cross-modality scenario.

4.2 Segmentation Architecture: U-Net

4.2.1 U-Net Overview and Design Philosophy

The core architecture, on which this study is based, is U-Net [17], a popular convolutional neural network, specifically designed for biomedical image segmentation. U-Net introduced the concept of encoder-decoder architecture alongside skip connections, was designed to solve the issue of medical image segmentation with small training material. Various features made the architecture a successful choice in this work; among them is its demonstrated effectiveness in a multi-medical imaging application, its high accuracy in localization owing to its multi-scale feature fusion, and its high level of data efficiency owing to its parameter-efficient design. The design principle of the U-Net is that with just a few training samples the combination of the precise spatial data in the decoder path and the contextual data in the encoder path would result in the accurate delineation of organ boundaries.

4.2.2 U-Net Architecture Details

The three main elements of the U-Net architecture consist of an encoder (contracting path), a bottleneck (lowest resolution layer), and a decoder (expanding path). The encoder path downsizes the spatial dimensions and upsizes feature depth by successively applying convolutional layers followed by max pooling operations. Two 3x3 convolutional layers with ReLU activation functions form each encoding level. Then, 2x2 max pooling is used to halve the spatial dimensions. From low-level edge and texture information to high-level semantic features, this hierarchical feature extraction captures information at multiple scales.

The bottleneck layer, which has the smallest spatial dimensions and the highest level of abstraction, is the most compressed representation of the input. For a typical U-Net handling 572x572 inputs, this layer usually processes features at 1/16th of the original input resolution. Global contextual information is encoded at the bottleneck, which serves as a feature compression stage.

Through up-sampling operations, the decoder path gradually restores spatial resolution, doubling the spatial dimensions at each step. Features are concatenated with corresponding skip connection features from the encoder path at the same resolution level after being upsampled via transposed convolutions at each decoder level. Because it merges high-level semantic features from the decoder with high-resolution spatial details from the encoder,

this concatenation step is crucial. Two 3x3 convolutional layers with ReLU activation are applied to refine the aggregated features after concatenation.

The final output layer projects the feature channels to the required number of output classes using a 1x1 convolution. This produces a single-channel output for binary liver segmentation that indicates the probability of each pixel belonging to the liver class. Due to its fully convolutional design, it can handle images of any size while maintaining spatial information throughout the network.

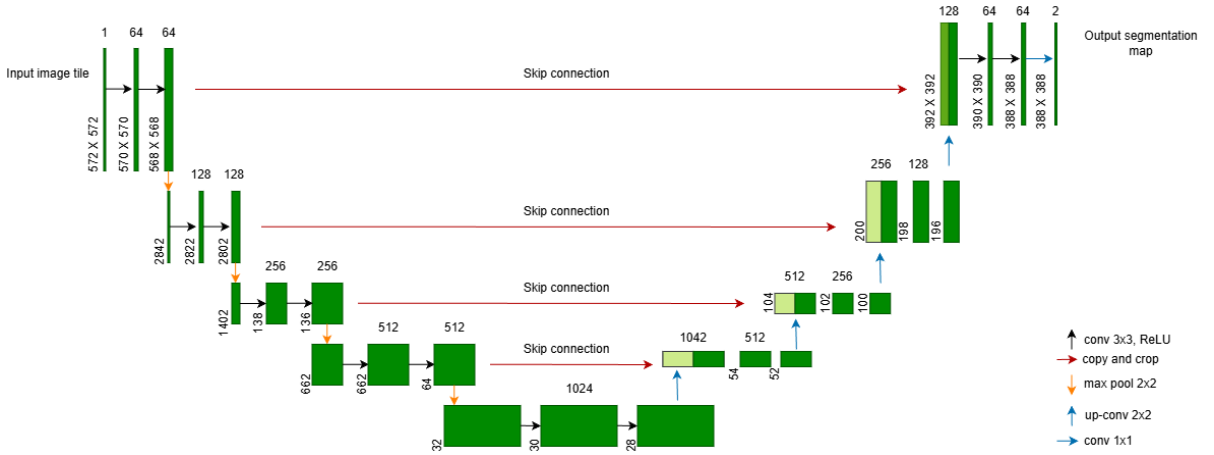


Figure 4.2.1: U-Net Architecture

As illustrated in Figure 4.2.1, the encoder path (left) progressively reduces spatial dimensions while increasing feature depth, whereas the decoder path (right) gradually restores spatial resolution. Skip connections (horizontal arrows) link encoder and decoder features at matching resolutions. The key feature that differentiates U-Net from conventional encoder-decoder designs is these skip connections. By bypassing the bottleneck, these connections provide features with direct paths from encoding levels to corresponding decoding levels. Spatial information that could disappear during the encoder's pooling operations is preserved by skip connections. Precise organ boundary segmentation requires the network's ability to combine both fine-grained spatial information and coarse-grained contextual features.

4.3 Pre-Trained Encoder Backbones

Four different pre-trained encoder architectures were used as the feature extraction backbone instead of the standard encoder in the U-Net design. By initializing encoder weights using models pre-trained on the ImageNet [47] dataset, which contains over 1.2 million natural images spanning 1000 categories, this approach uses transfer learning. It is well established in the literature that using pre-trained encoders improves both convergence speed and model generalization by allowing the network to use features learned on large-scale natural image datasets. The four encoders discussed in this study - ResNet18, EfficientNetB3, DenseNet121, and MobileNet V2 - represent various design trade-offs and architecture designs. All encoders have the traditional U-Net decoder architecture, and they ensure that the variation in performance should be attributed to the encoder architecture and not to confounding variables.

4.3.1 ResNet18: Residual Networks

The concept of residual connections, also referred to as skip connections, was proposed by ResNet [25], which greatly changed the design of deep learning architectures. Instead of learning direct mappings, residual connections enable networks to learn residual mappings, mathematically expressed as $y = F(x) + x$, where x represents the input and $F(x)$ is the learned function. The vanishing gradient issue, which occurs when training very deep networks and causes gradients to diminish exponentially during backpropagation over many layers, is mitigated by this strategy.

With eighteen layers, ResNet18 preserves the advantages of residual learning and to offer an acceptable computation efficiency / depth balance. The architecture usually has 2 layers of 3x3 convolutional filters in each stacked block of residual. Branching Each branching has skip connections every few layers. Skip connections branch off from the main path every few layers. 1x1 convolutions in the skip route project features to align with the output dimensions when spatial dimensions are reduced (via stride-2 convolutions). To evaluate whether skip connection-based architectures offer advantages for cross-modality liver segmentation compared to alternative designs, ResNet18 was employed in this study.

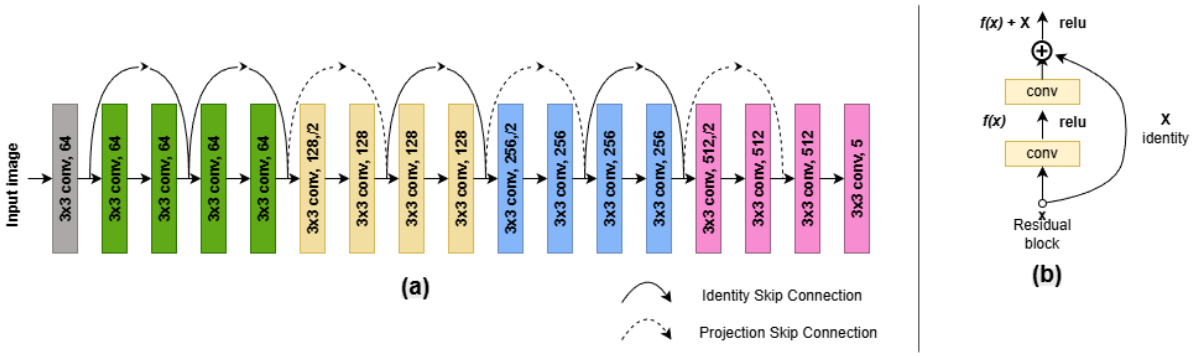


Figure 4.3.1: ResNet18 Architecture

As illustrated in Figure 4.3.1, residual blocks enable deep networks by allowing gradients to flow directly through skip connections. Each block learns residual functions $F(x)$ that are added to the input x.

4.3.2 EfficientNetB3: Compound Scaling

Leveraging the compound scaling principle, EfficientNet [48] provides an alternative strategy for network architecture. EfficientNet uses a systematic scaling technique, which scales all three dimensions in proportion to a fixed ratio, as opposed to increasing network depth, width or input resolution separately. The equation of compound scaling method is:

$$\text{depth} = \alpha^{\phi}, \quad \text{width} = \beta^{\phi}, \quad \text{resolution} = \gamma^{\phi} \tag{1}$$

where α, β, and γ are constants found during architecture search, and ϕ is a scaling coefficient that the user specifies.

The third variant in the EfficientNet series, EfficientNetB3, reflects a balanced trade-off between model capacity and computational expense. After using neural architecture search to optimize the base EfficientNet (B0), EfficientNetB3 uses the compound scaling strategy with a scaling coefficient $\phi = 3$. To reduce computational demands, the network integrates depth-wise separable convolutions with mobile inverted bottleneck blocks [33]. These

blocks consist of three layers: a depthwise convolution that performs spatial convolution on each channel individually, a pointwise convolution to project to reduced channel dimensions, and a pointwise convolution to expand the channel count.

To examine whether systematic scaling strategies provide advantages for cross-modality generalization compared to alternative architectural philosophies, particularly in balancing model capacity with computational efficiency, EfficientNetB3 was used.

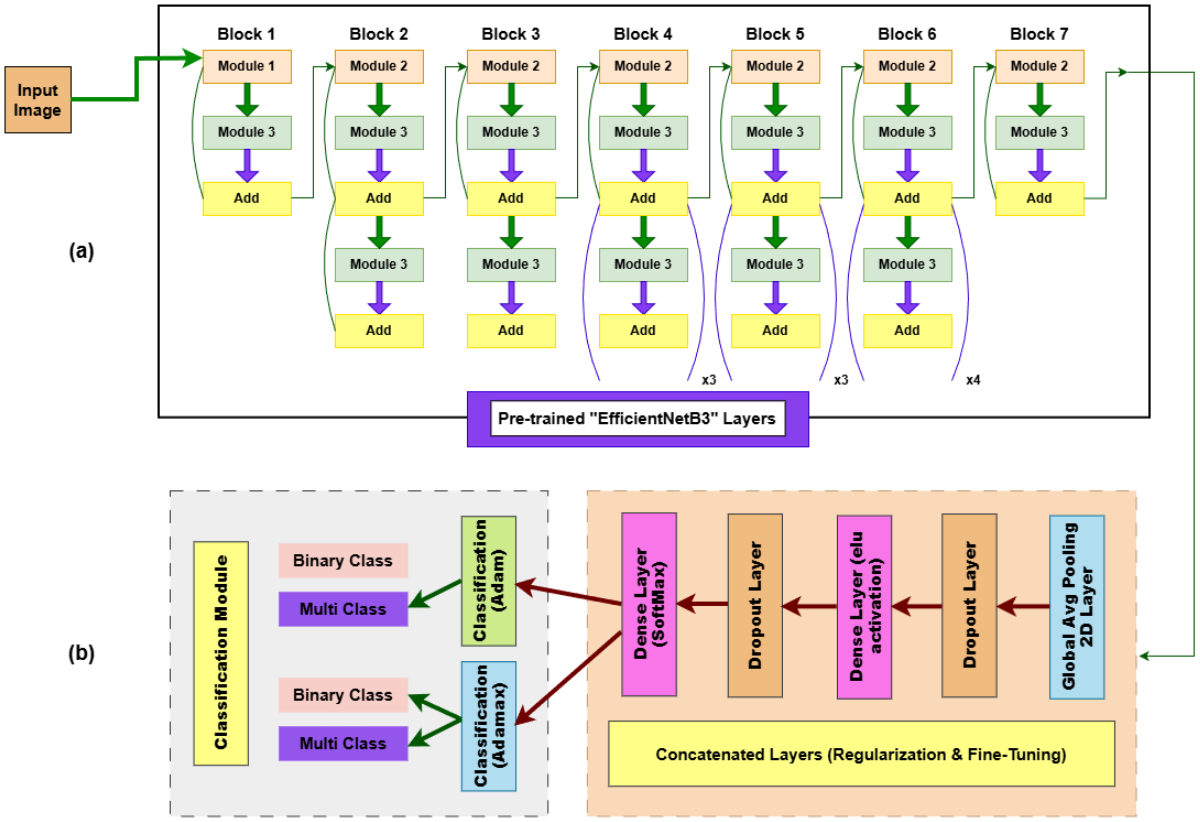


Figure 4.3.2: EfficientNetB3 Architecture

As illustrated in Figure 4.3.2, compound scaling uniformly scales network depth, width, and resolution. Mobile Inverted Bottleneck (MBConv) blocks reduce computation through depthwise separable convolutions while maintaining model capacity.

4.3.3 DenseNet121: Dense Connectivity

DenseNet [32] replaces the sequential layer stacking found in ResNet and conventional convolutional networks with dense connection patterns. Every layer in DenseNet gets inputs from every other layer in the same dense block. The expression for this connection pattern is:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{2}$$

where $[x_0, x_1, \ldots, x_{l-1}]$ represents the concatenation of feature maps from all previous layers, and H_l represents the function learned by layer l.

Numerous theoretical and practical benefits are offered by the dense connection structure. First, because gradients can spread straight from any layer to any layer before it, it keeps a

strong gradient flow across the network. Second, because all layer outputs are concatenated and handed forward, it promotes feature reuse between levels. Third, because each layer can learn more compact representations through feature reuse, it requires fewer parameters than traditional networks.

The 121 layers of DenseNet121 are arranged into dense blocks that are divided by transition layers. Transition layers use 1x1 convolutions and 2x2 average pooling to reduce dimensionality. Each layer creates 12 feature channels (the growth rate) within each dense block, which are then concatenated with the outputs of every preceding layer. In order to determine if the advantages of dense connection, including strong gradient flow and feature reuse, translate into increased robustness in cross-modality segmentation tasks, DenseNet121 was incorporated in this work.

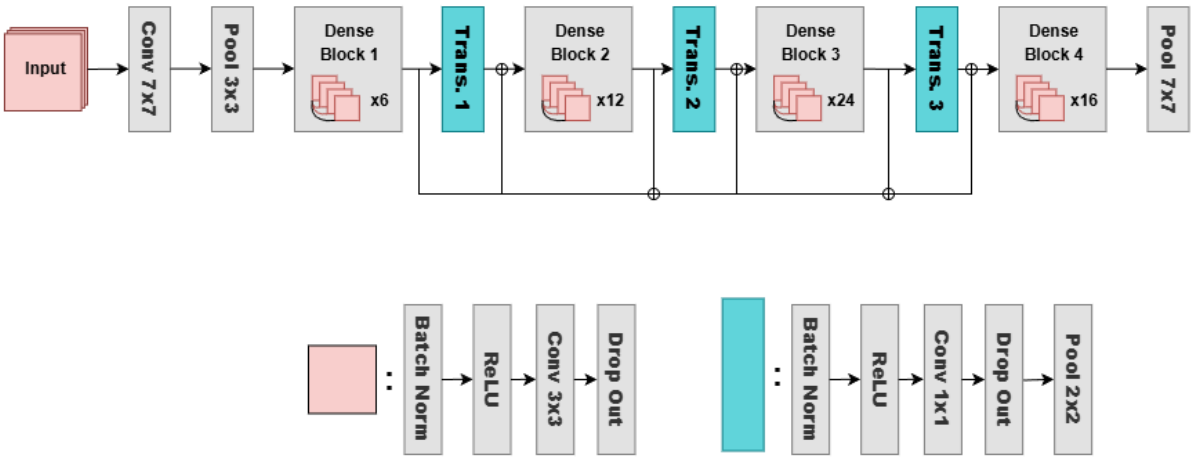


Figure 4.3.3: DenseNet121 Architecture

As shown in Figure 4.3.3, dense connectivity enables each layer to receive inputs from all previous layers within the block. This pattern maintains strong gradient flow and encourages feature reuse across network depth.

4.3.4 MobileNetV2: Lightweight Architecture

MobileNetV2 [33] was created for contexts with limited resources, like embedded systems and mobile devices. When compared to larger models, the architecture maintains comparable accuracy while resulting in substantial savings in computing cost and memory needs. Inverted residual blocks and linear bottlenecks are two primary advancements used by MobileNetV2.

The design of inverted residual blocks differs from the norm. Inverted residual blocks first expand input channels and use depthwise convolution. The representation is then contracted back to smaller number of channels. The expression for this design is:

$$\text{Expand} \rightarrow \text{Depthwise Conv} \rightarrow \text{Linear Project} \tag{3}$$

In order to provide a richer feature representation for the depthwise convolution that follows, the expansion phase employs 1x1 convolutions to raise the channel depth by a factor of six. Compared to conventional convolutions, the depthwise convolution reduces computation by applying spatial filtering independently to each channel. The linear projection projects

back to smaller channel dimensions using 1x1 convolutions without activation functions. Training stability is enhanced by the final projection's lack of non-linearity, which preserves information in the low-dimensional space.

With 19 inverted residual blocks placed in a sequential fashion, MobileNetV2 achieves roughly 300 million multiply-accumulate operations as opposed to billions for regular networks. The design maintains competitive accuracy while being about 8–9 times smaller and 6–7 times faster than conventional convolutional networks. In order to assess whether models with much lower complexity can nevertheless show generalization capabilities equal to more sophisticated alternatives, MobileNetV2 was chosen.

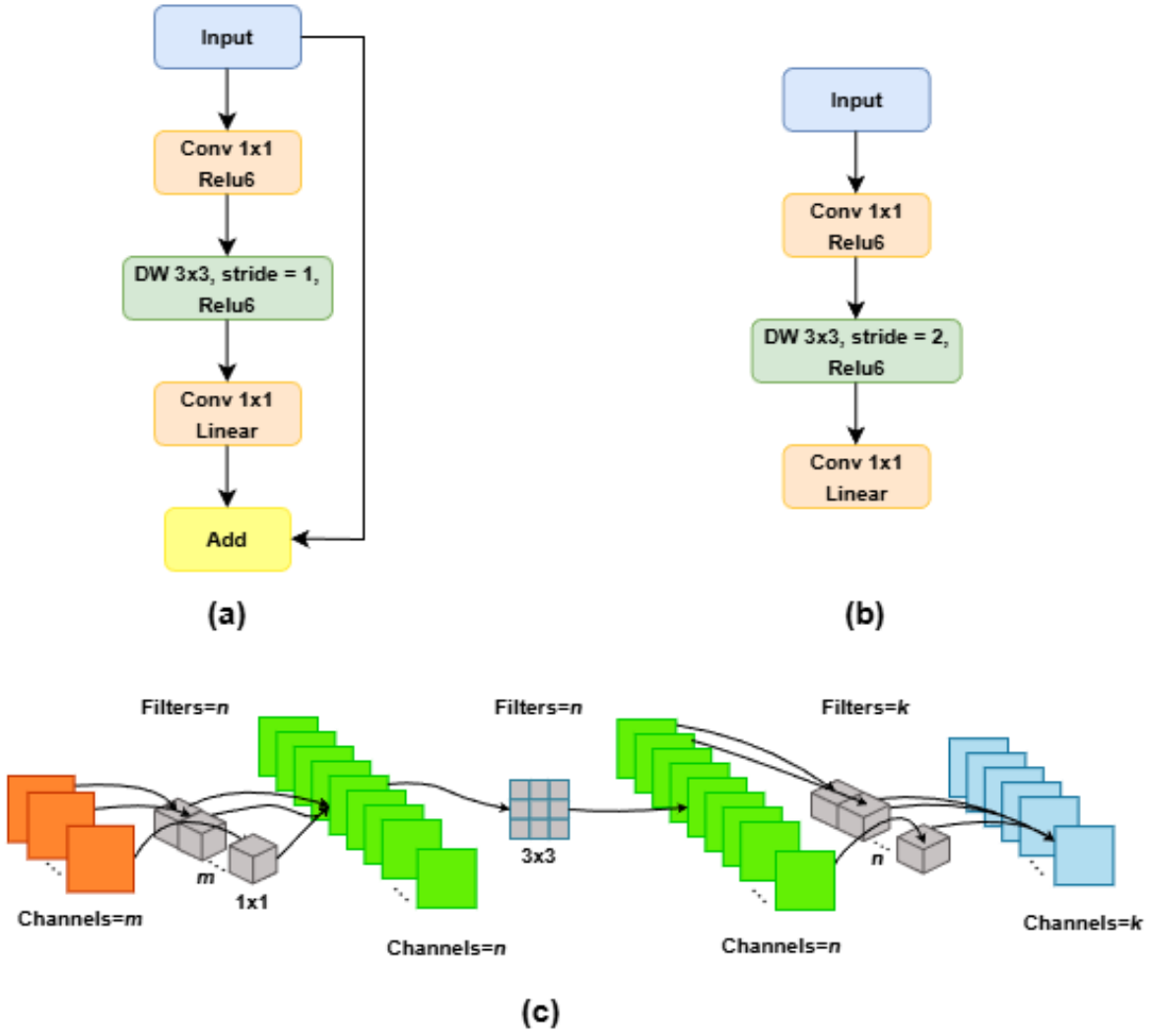


Figure 4.3.4: MobileNetV2 Architecture

As illustrated in Figure 4.3.4, inverted residual blocks expand channels before depthwise convolution, then linearly project to smaller dimensions. This design dramatically reduces computation while maintaining accuracy.

4.3.5 Transfer Learning Framework

The segmentation models benefit from features learned from 1.2 million natural images spanning 1000 object classes because all four encoders are initialized with weights pre-trained on ImageNet. In medical imaging, transfer learning provides several advantages:

first, ImageNet pre-training delivers general low-level features (edges, textures, fundamental shapes) applicable across diverse image types; second, pre-training markedly accelerates convergence on medical imaging tasks, reducing training time from hours or days to minutes; and third, on small- to medium-sized medical imaging datasets, fine-tuning pre-trained networks typically produces superior results compared to training from scratch. Through their hierarchical design, the pre-trained encoders extract multi-scale features, which the U-Net decoder then utilizes to generate segmentation masks.

4.4 Training Configuration

The LiverHccSeg dataset was used to train the model. To guarantee a consistent comparison across architectural variants, all four encoder types were trained using identical hyperparameters. To prevent overfitting via L2 regularization, the training procedure utilized the Adam optimizer with a weight decay of 1×10^{-5}. A cosine annealing scheduler was applied to control the learning rate, which started at 1×10^{-4} and decayed to a minimum of 1×10^{-6} over the course of training according to:

$$\mathrm{lr}(t) = \mathrm{lr}_{\min} + \frac{1}{2}(\mathrm{lr}_{\max} - \mathrm{lr}_{\min}) \left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \tag{4}$$

where t is the current epoch and T is the total number of epochs. The gradual decay provided by this learning rate schedule allows the model to first explore the loss landscape before fine-tuning weights as training goes on.

The training ran for 300 epochs with a batch size of 32. To meet the input size requirements of ImageNet pre-trained encoders, all input images were resized to 224×224 pixels. During training, random modifications done with predetermined probabilities were used to enhance images and masks. Among the augmentations were random rotation (between -30 and 30 degrees), random scaling (between 0.8x and 1.2x), random blurring with Gaussian filters, and contrast-limited adaptive histogram equalization (CLAHE) to improve local contrast. These procedures increase the diversity of training data and improve the model's robustness to variations in contrast and image appearance encountered in clinical practice. To guarantee that segmentation masks only included background (value 0) and foreground/liver (value 1) classes, they were binarized.

Dice Loss, a loss function that directly optimizes the Dice Similarity Coefficient—a common statistic for segmentation tasks—was used during training. The formula for dice loss is:

$$\text{Dice Loss} = 1 - \mathrm{DSC} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{5}$$

where X stands for the anticipated segmentation and Y for ground truth. Because dice loss is less sensitive to class imbalance than cross-entropy loss, it is especially well-suited for medical image segmentation when the region of interest (liver) may occupy only a small portion of the total picture area. Consistency between optimization and performance evaluation is ensured by using Dice Loss as both the training objective and the evaluation metric.

4.5 Evaluation Framework

In order to evaluate the performance of the segmentation rigorously, pixel-based classification indices were used. When binary liver segmentation is considered, the measure is based on the comparison of the two sets of pixels in the Ground Truth mask (G) and the Predicted mask (P) of the foreground.

The quantification of the comparison is done by the analysis of the overlap or the divergence of these two sets:

- **True Positive (TP):** This is the intersection of the prediction and the ground truth. These are those pixels on which the model identified the liver correctly ($TP = |G \cap P|$).
- **False Positive (FP):** Refers to the erroneous pixels in the foreground. These are background pixels which have incorrectly been labeled as liver, indicating over-segmentation ($FP = |P - G|$).
- **False Negative (FN):** These are the ground truth pixels which are foreground but not predicted by the mask. These are real liver pixels that the model did not identify, indicating under-segmentation ($FN = |G - P|$).

On these relationships, four metrics were determined:

The **Dice Similarity Coefficient (DSC)** quantifies the overlap by dividing the intersection by the mean size of the two sets. In order to convert the theoretical set notation into a calculable formula, we decompose the total cardinality of each set: the ground truth set ($|G|$) equals the sum of the detected and missed pixels ($TP + FN$), and the predicted set ($|P|$) equals the sum of correct hits and false alarms ($TP + FP$). The denominator consequently is the sum of all these elements.

$$\textbf{DSC} = \frac{2 \times \text{Area of Overlap}}{\text{Total Combined Area}} = \frac{2|G \cap P|}{|G| + |P|} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{6}$$

The **Intersection over Union (IoU)**, or the Jaccard Index, is the measure of the ratio of the intersection to the union of the two sets. The theoretical union ($G \cup P$) denotes the set of all unique foreground pixels in any of the masks. This can be mathematically computed as the sum of the sets minus the intersection ($|G| + |P| - |G \cap P|$). Replacing the terms with the confusion matrix gives $(TP + FN) + (TP + FP) - TP$, which can be simplified as $TP + FP + FN$.

$$\textbf{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|G \cap P|}{|G \cup P|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{7}$$

Precision is used to measure the purity of positive predictions by determining the ratio of the intersection to the cardinality of the set of predictions. The denominator ($|P|$) represents what the model predicted as positive: the sum of correct overlap (TP) and incorrect over-segmentation (FP).

$$\textbf{Precision} = \frac{\text{True Positives}}{\text{Total Predicted Positives}} = \frac{|G \cap P|}{|P|} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Recall evaluates the capacity of the model to cover the entire liver region and is a ratio of the intersection with the ground truth set to the ground truth cardinality. The denominator ($|G|$) represents the actual size of the liver: the total of the pixels the model discovered (TP) and the pixels the model missed (FN).

$$\textbf{Recall} = \frac{\text{True Positives}}{\text{Total Actual Positives}} = \frac{|G \cap P|}{|G|} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

4.6 Cross-Modality Evaluation Strategy

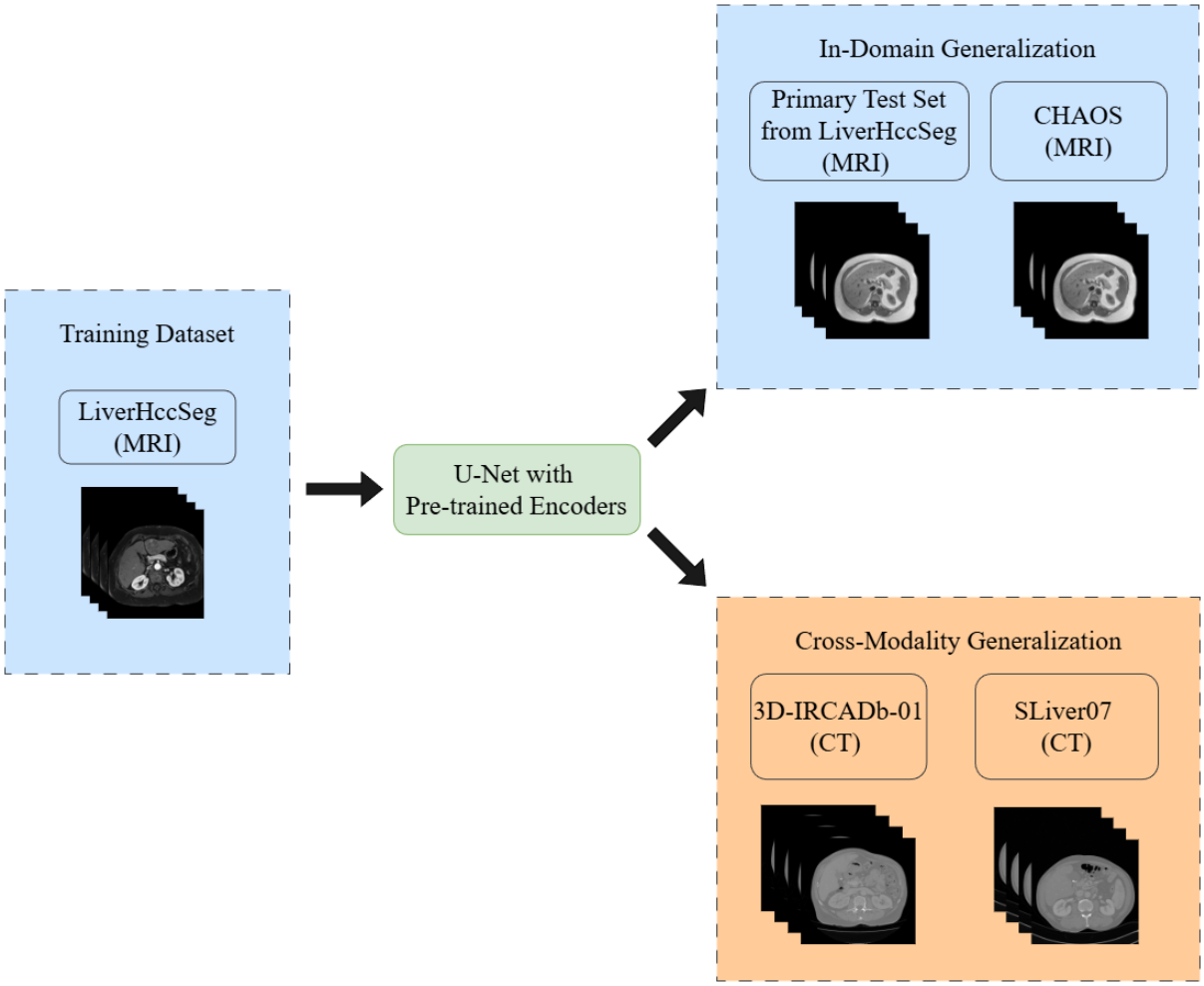


Figure 4.6.1: Evaluation strategy for in-domain and cross-modality generalization.

According to the general strategy outlined in Figure 4.1.1, this strategy is adhered to as shown in Figure 4.6.1. Training of the model is only done on the LiverHccSeg (MRI) data. The model is evaluated on three external datasets: CHAOS (MRI), 3D-IRCADb-01 and SLiver07 (both CT).

The CHAOS dataset, which contains MRI sequences from a source different from the training data but within the same imaging modality, was used to evaluate in-domain generalization. This assessment determines whether models have acquired features specific to the LiverHccSeg acquisition parameters that do not transfer to other MRI protocols, or whether models have learned imaging protocol-robust features that generalize to different MRI acquisition parameters.

Two CT datasets, 3D-IRCADb-01 and SLIVER07, were used to evaluate cross-modality generalization. By extending models trained exclusively on MRI data to markedly different imaging modalities, these datasets test the ultimate generalization capability. When models perform well on cross-modality tasks, they have developed general liver anatomical properties that are not dependent on appearance traits specific to a particular imaging modality. Conversely, poor performance on CT datasets would suggest that the acquired features heavily rely on contrast and intensity patterns unique to MRIs.

For every modality transition, multiple datasets were included to improve statistical robustness and reduce the possibility that the results are artifacts of a single dataset's features. This allows assessment of whether performance differences across models are dataset-specific or consistent across various CT acquisition protocols and patient populations.

During cross-modality evaluation, no domain adaptation methods, fine-tuning, or modality-specific adjustments were employed. Using weights trained exclusively on LiverHccSeg MRI data, models were directly applied to validation data without any modifications. This strict evaluation approach unbiasedly tests the inherent cross-modality generalization abilities of the encoder architectures. Comparative conclusions would be unreliable if adaptation strategies or fine-tuning were applied, as it would be difficult to determine whether performance improvements originated from the encoder design or the adaptation method.

Chapter-5 : Results

The quantitative evaluation outcomes of the four encoder-based U-Net models across all datasets are presented in this chapter alongside a direct examination of the results. The evaluation scenarios—primary training dataset performance, in-domain generalization across different MRI protocols, and cross-modality generalization to CT imaging—organize the results. Metrics are reported in each section, together with an immediate analysis of the findings regarding encoder performance and generalization capabilities.

5.1 Primary Training Dataset Performance

Using identical hyperparameters and procedures as described in Chapter 4, all four encoder variants were trained on the LiverHccSeg dataset. Quantitative results for both training and test splits are presented in Table 5.1.1.

5.1.1 Training and Test Metrics on LiverHccSeg

ResNet18 achieved the highest DSC of 0.980 with IoU of 0.962, Precision of 0.973, and Recall of 0.987 on the training dataset. Training DSC of 0.972, IoU of 0.945, Precision of 0.976, and Recall of 0.967 were obtained using EfficientNetB3. Training DSC of 0.974, IoU of 0.950, Precision of 0.972, and Recall of 0.977 were recorded using DenseNet121. Training DSC of 0.978, IoU of 0.957, Precision of 0.972, and Recall of 0.984 were achieved with MobileNetV2. With training DSC surpassing 0.97 for all encoders, each model demonstrated rapid convergence and effective learning.

Table 5.1.1: Performance Metrics for Primary Dataset.

Encoder	Tr. DSC	Tr. IoU	Tr. Pr.	Tr. Rc.	Te. DSC	Te. IoU	Te. Pr.	Te. Rc.
ResNet18	0.980	0.962	0.973	0.987	0.925	0.864	0.943	0.913
EfficientNetB3	0.972	0.945	0.976	0.967	0.913	0.847	0.955	0.883
DenseNet121	0.974	0.950	0.972	0.977	0.928	0.868	0.949	0.910
MobileNetV2	0.978	0.957	0.972	0.984	0.926	0.866	0.942	0.914

Abbreviations: Tr.=Training, Te.=Testing, DSC=Dice Similarity Coefficient, IoU=Intersection over Union, Pr.=Precision, Rc.=Recall.

Performance ranking on the test set differs considerably from training outcomes. DenseNet121 achieved the highest test DSC of 0.928, with IoU of 0.868, Precision of 0.949, and Recall of 0.910. MobileNetV2 reached a test DSC of 0.926 with IoU of 0.866, Precision of 0.942, and Recall of 0.914. ResNet18 scored a test DSC of 0.925 with IoU of 0.864, Precision of 0.943 and Recall of 0.913. EfficientNetB3 had the lowest test DSC of 0.913, and IoU, Precision, and Recall of 0.847, 0.955, and 0.883, respectively.

As for the test set, each of the four architectures learned generalizable features in the training domain, indicated by the difference between the best architecture (DenseNet121, 0.928) and the worst architecture (EfficientNetB3, 0.913) of 0.015. Nevertheless, the ranking of the encoders shifts between training and test: DenseNet121 ranked third and first in training and test set, respectively, and ResNet18 ranked first and third, respectively. This rearrangement shows that test performance does not always reflect training performance.

There are different train-test performance gaps in encoders. The greatest difference is in ResNet18 (0.980 to 0.925, the difference between them is 0.055), which indicates that it is more sensitive to domain change. DenseNet121 has a smaller gap (0.974 to 0.928, difference 0.046) which exhibits greater consistency in generalization within the training domain. The gap in EfficientNetB3 is also similar to that of ResNet18 at 0.059 (0.972 to 0.913). These variable generalization imply that it is likely that generalization tendencies vary.

5.2 In-Domain Generalization: CHAOS Dataset

Section 5.1 discussed the findings in the internal test set, and this section proceeds to discuss in-domain generalization. In-domain generalization, in this instance, refers to the performance of the model on data of the same imaging modality but of different origin as the training data. The model was only trained using the LiverHccSeg data, which consisted of MRI scans. The trained model was later tested on the CHAOS dataset which is also an MRI dataset.

To assess generalization to new MRI data while remaining within the same modality, the CHAOS dataset provides MRI sequences acquired using different techniques than LiverHccSeg. This analysis separates the effects of acquisition protocol variations from intrinsic modality differences.

5.2.1 Results and Analysis

The CHAOS dataset results are shown in Table 5.2.1. Encoder performance changed noticeably: EfficientNetB3 achieved the highest DSC of 0.852 with IoU of 0.744, Precision of 0.869, and Recall of 0.840. This shows that EfficientNetB3's lower performance on the primary dataset does not imply to worse generalization, as CHAOS represents its best performance (0.852) compared to its lowest on LiverHccSeg (0.913).

MobileNetV2 ranked second with a DSC of 0.836, IoU of 0.720, Precision of 0.855, and Recall of 0.825. DenseNet121 achieved a DSC of 0.823 with IoU of 0.703, Precision of 0.887, and Recall of 0.776, showing a considerable drop from its primary dataset score (0.928 to 0.823, difference 0.105). ResNet18 recorded a DSC of 0.804 with IoU of 0.676, Precision of 0.847, and Recall of 0.772.

Table 5.2.1: Performance Metrics for CHAOS Dataset.

Encoder	**DSC**	**IoU**	**Precision**	**Recall**
ResNet18	0.804	0.676	0.847	0.772
EfficientNetB3	0.852	0.744	0.869	0.840
DenseNet121	0.823	0.703	0.887	0.776
MobileNetV2	0.836	0.720	0.855	0.825

According to the CHAOS results, EfficientNetB3 shows an advantage when adapting to new MRI protocols despite its obvious shortfall on the primary dataset. The performance gap between DenseNet121 (0.823) and EfficientNetB3 (0.852) is 0.029 DSC, surpassing the internal variations observed on LiverHccSeg. This finding shows a trade-off. Models that perform best on training

data may not perform as well on other protocols. EfficientNetB3 also appears to be able to learn features which are robust to protocol differences even at the expense of slightly reduced training accuracy with the compound scaling method.

Although the model size is much smaller than that of the other encoders, the very high in-domain performance of MobileNetV2 (0.836) demonstrates that protocol robustness is not dependent on network size.

Among the encoders, ResNet18's consistent suboptimal performance on CHAOS (0.804) shows the largest proportional drop from primary dataset results (0.925 to 0.804, difference 0.121). This shows that residual connections do not intrinsically provide robustness to acquisition protocol changes, despite aiding in training of deep networks.

5.3 Cross-Modality Generalization: CT Datasets

Here, the model trained on MRI (Section 5.2) is tested on a completely different imaging modality. These MRI-trained model was tested on the 3D-IRCADb-01 and SLiver07 datasets which, unlike the in-domain testing, belong to the Computed Tomography (CT) domain. This establishes cross-modality validation in cases where training domain (MRI) and testing domains (CT) are radically different.

Cross-modality evaluation represents the ultimate generalization challenge, requiring models trained exclusively on MRI data to segment CT images. Because CT and MRI are based on fundamentally different physical principles, their intensity distributions, contrast patterns, and noise characteristics differ. Models that succeed on cross-modality tasks have learned general anatomical properties that are largely independent of modality-specific appearances.

5.3.1 3D-IRCADb-01 Dataset

The 3D-IRCADb-01 dataset comprises 2,823 CT slices for cross-modality testing. Results are summarized in Table 5.3.1.

EfficientNetB3 attained the highest DSC of 0.825, with IoU of 0.705, Precision of 0.760, and Recall of 0.911. DenseNet121 achieved a DSC of 0.824, IoU of 0.701, Precision of 0.763, and Recall of 0.910, performing almost equivalently to EfficientNetB3. Despite differences in primary dataset results, the two encoders show nearly identical cross-modality transfer (0.001 DSC difference). MobileNetV2 obtained a DSC of 0.816 with IoU of 0.695, Precision of 0.719, and Recall of 0.956. ResNet18 exhibited considerably poorer performance, with a DSC of 0.788, IoU of 0.657, Precision of 0.700, and Recall of 0.927.

Table 5.3.1: Performance Metrics for 3D-IRCADb-01 Dataset.

Encoder	**DSC**	**IoU**	**Precision**	**Recall**
ResNet18	0.788	0.657	0.700	0.927
EfficientNetB3	0.825	0.705	0.760	0.911
DenseNet121	0.824	0.701	0.763	0.910
MobileNetV2	0.816	0.695	0.719	0.956

Given their notable primary dataset performance gap (0.915 versus 0.928), the convergence of EfficientNetB3 and DenseNet121 on CT is striking. Architectural differences that drove improvements in within-domain accuracy are less impactful when facing the substantial domain shift from MRI to CT. Instead, performance depends on by architectural traits that support resilient cross-modality transfer.

Precision-recall patterns show systemic tendencies. EfficientNetB3 and DenseNet121 deliver well-balanced segmentation with similar precision (0.76–0.77) and recall (0.91). MobileNetV2 is prone to overestimate liver regions, shown by its higher recall (0.956) and lower precision (0.719). ResNet18 consistently overestimates on CT images, showing the lowest precision (0.700) and highest recall (0.927). This pattern shows that ResNet18 learned intensity features particular to MRI that cause high activation on CT images.

For medical image segmentation, the 0.037 DSC gap between the top (EfficientNetB3, 0.825) and bottom (ResNet18, 0.788) encoders is medically relevant. This gap aligns with the in-domain generalization difference (CHAOS: 0.029 DSC between best and worst), suggesting that cross-modality transfer relies on architectural components that also promote in-domain robustness.

5.3.2 SLIVER07 Dataset

SLIVER07, with 4,159 CT slices, is the largest external validation dataset. The expanded sample size strengthens the statistical validity of performance results. Results are summarized in Table 5.3.2.

DenseNet121 achieved the highest DSC of 0.806, with IoU of 0.692, Precision of 0.716, and Recall of 0.951. EfficientNetB3 obtained a DSC of 0.805, IoU of 0.690, Precision of 0.728, and Recall of 0.929 (0.001 DSC difference from DenseNet121), performing almost identically. MobileNetV2 achieved a DSC of 0.803 with IoU of 0.686, Precision of 0.699, and Recall of 0.963. ResNet18 obtained a DSC of 0.769, with IoU of 0.644, Precision of 0.662, and Recall of 0.960.

Table 5.3.2: Performance Metrics for SLIVER07 Dataset.

Encoder	**DSC**	**IoU**	**Precision**	**Recall**
ResNet18	0.769	0.644	0.662	0.960
EfficientNetB3	0.805	0.690	0.728	0.929
DenseNet121	0.806	0.692	0.716	0.951
MobileNetV2	0.803	0.686	0.699	0.963

Patterns observed on 3D-IRCADb-01 are reinforced by the SLIVER07 findings. The near-equal performance of EfficientNetB3 (0.805) and DenseNet121 (0.806) across the two CT datasets (differences of 0.001 and 0.020) provides strong evidence that their comparable cross-modality robustness is systematic rather than dataset-specific. Consistency across datasets enhances confidence in these architectural conclusions.

Across all CT datasets, ResNet18's performance gap relative to the top encoders (0.036–0.037 DSC) is stable, indicating inherent design limitations rather than dataset artifacts. Its reliably low precision (0.662) and high recall (0.960) show consistent overestimation of liver regions on CT images.

Despite model sizes differing by about 10–15×, MobileNetV2's performance (0.803) reaches about 99.6% of DenseNet121's accuracy on SLIVER07 (0.806). When architectural design highlights shared feature extraction, computational efficiency and multi-modality resilience can coexist. This compact architecture achieves multi-modalit performance comparable with much bigger models.

5.4 Performance Summary and Overall Trends

Table 5.4.1 shows DSC values across all evaluation scenarios for direct comparison and trend analysis.

Table 5.4.1: DSC Performance Summary Across All Datasets.

Encoder	**LiverHccSeg Test**	**CHAOS**	**3D-IRCADb-01**	**SLIVER07**
ResNet18	0.925	0.804	0.788	0.769
EfficientNetB3	0.913	0.852	0.825	0.805
DenseNet121	0.928	0.823	0.824	0.806
MobileNetV2	0.926	0.836	0.816	0.803

Each encoder shows clear outcomes trends across increasing domain shifts, according to the summary. As domain shift increases, ResNet18 consistently decreases, dropping from 0.925 on the core dataset to 0.804 on various MRI procedures and down to 0.769 on CT. The largest decrease among all encoders is the 0.156 DSC difference between the primary dataset and the largest domain shift (0.925 to 0.769).

In contrast, EfficientNetB3 shows a small performance reduction from CHAOS (0.852) to CT datasets (0.825, 0.805), showing lower primary dataset performance (0.913) but increased resilience to domain shifts. Compared with ResNet18's 0.156 drop, the overall decrease from the primary to the largest domain shift (0.913 to 0.805) is 0.108.

DenseNet121 exhibits a trade-off behavior, with the best results on the original (0.928) and a big decline on CHAOS (0.823) and a steadying under the CT datasets (0.824, 0.806). Compared to EfficientNetB3, multi-modality performance (CT datasets) is more comparable, but the reduction between the primary dataset and any generalization scenario is larger.

In all cases, MobileNetV2 demonstrates steadiness in its performance: high cross-modality transfer (0.816, 0.803), good in-domain generalization (0.836), and similar results in the primary dataset (0.926).

These trends indicate that multi-modality robustness is determined by principles governing architectural design to regulate features extraction and representation learning, other than the depth and size of the network.

Chapter-6 : Discussion

This chapter provides an overview of the results from Chapter 5 and discusses them in the larger context of deep learning, transfer learning, and medical image segmentation. It compares the results to existing research and synthesizes the experimental findings to develop a deeper understanding of the factors affecting multi-modality generalization.

6.1 Architectural Insights from Encoder Performance

The comparative analysis shows that generalization behavior depends on by encoder architectural traits, which has important implications for understanding deep learning in medical imaging. The study assessed four unique encoder approaches: sequential residual refinement (ResNet18), compound multi-scale scaling (EfficientNetB3), dense connectivity with feature reuse (DenseNet121), and a lightweight, efficient design (MobileNetV2).

One of the key lessons learned is the fact that multi-modality generalization and in-domain accuracy are independent of each other. DenseNet121 exhibited severe performance drop on new protocols in MRI (0.823 CHAOS), even though it attained the best accuracy on the primary dataset (0.928). Even though EfficientNetB3 had less performance in the primary dataset (0.913), the in-domain performance was higher (0.852 CHAOS). Both architectures converged on cross-modality CT tasks (DSC 0.825 and 0.806), indicating that factors contributing to in-domain performance are different to factors that facilitate cross-modality transfer.

This dissociation aligns with the concept of transfer learning, which makes a distinction between the generalizable feature learning and the specialized feature learning. Although it may incur the cost of lower specialized accuracy on the training domain, architectures with the emphasis on balanced hierarchical feature extraction as well as on widespread feature reuse appear to learn more generalizable representations.

6.2 Compound Scaling & Balanced Feature Extraction

Although the accuracy in the primary dataset is lower, the conceptual benefits of compound scaling are upheld by the fact that EfficientNetB3 has a superior in-domain generalization (0.852 CHAOS vs. 0.823 DenseNet121, a difference of 0.029). In order to prevent any dimension from acting as a bottleneck, this design process balances the network depth, width, and input resolution to create consistent topologies.

The results indicate that compound scaling is beneficial to learning multiple features on a wide range of scale and resolution that can be utilized in various acquisition protocols. The balanced architecture reduces over-specialization to specific MRI protocol features while training on LiverHccSeg data with particular acquisition settings. CHAOS performance shows this, with EfficientNetB3 surpassing DenseNet121 by 0.029 DSC.

However, on cross-modality tasks EfficientNetB3 loses this advantage. On CT datasets, EfficientNetB3 and DenseNet121 nearly converge (DSC 0.825 and 0.824 on 3D-IRCADb-01, 0.805 and 0.806 on SLIVER07). This shows that multi-modality robustness depends on by different architectural factors than resistance to MRI protocol variations. The convergence on CT shows that common features supporting cross-modality anatomical representation dominate over architectural differences when faced with significant domain shifts.

6.3 Dense Connectivity & Feature Reuse

DenseNet121's architecture, which links each layer to all subsequent layers within dense blocks, supports strong gradient flow and broad feature reuse. This connectivity strategy led to larger generalization gaps on new protocols (0.105 point reduction from 0.928 to 0.823 on CHAOS), while retaining DenseNet121's advantage on primary dataset performance (0.928).

One reason is that the architecture effectively leverages in-domain features through dense connectivity's gradient flow and feature reuse, producing strong training domain accuracy but increased sensitivity to domain shifts. Extensive feature concatenation across layers may produce representations strongly optimized for LiverHccSeg characteristics. Such highly focused representations generalize less effectively to novel MRI protocols compared with more generalizable architectures.

The decreased in-domain generalization of DenseNet121 might also stem from a different phenomenon: dense connections facilitate robust within-domain fitting but may inadvertently capture noise or artifacts unique to a particular acquisition. Features that fail to generalize to new MRI protocols can vary from those that fail across modalities, as observed in DenseNet121's improvement when moving to multi-modality tasks (0.823 CHAOS to 0.824–0.806 CT) compared to EfficientNetB3's smaller improvement (0.852 to 0.825–0.805).

6.4 Residual Connections & Limited Generalization

ResNet18's uniform underperformance across all generalization scenarios shows that residual connections do not intrinsically guarantee resilience to domain shifts, despite allowing deep network training. As domain shift increased, ResNet18's performance dropped steadily: 0.925 on the core dataset, 0.804 on various MRI protocols, and 0.769–0.788 on CT datasets. The 0.156 DSC drop from primary to greatest domain shift is significant higher than for other encoders.

ResNet18's precision-recall patterns on CT show consistent liver overestimation (high recall 0.93, low precision 0.66–0.70). It means that ResNet18 learned to pick up MRI-specific patterns of intensities, which cannot generalize to CT and lead to overactivation on CT tissue. Gradually becoming increasingly domain-specific in its representations, ResNet18 representations appear to be increasingly associated with modality-specific features, as seen in the gradual deterioration in performance.

These results challenge the widely-held belief in the medical imaging field where deeper network structures are believed to generate higher generalization. Despite the fact that the design of ResNet18 is relatively deep and that the use of residual connections makes it trainable efficiently. The mechanism of cross-modality generalization and in-domain learning appears to be different.

6.5 Lightweight Architectures & Computational Efficiency

Having far fewer parameters than the standard networks, MobileNetV2 has high performance in every scenario (0.926 primary, 0.836 CHAOS, 0.816-0.803 CT), disproving the idea that the size of a model determines its generalization capability. With roughly 1/10th the parameters, the architecture attains 97–99% of DenseNet121's cross-modality performance (DSC 0.816 vs 0.824 on 3D-IRCADb-01, 0.803 vs 0.806 on SLIVER07).

As opposed to the dense sequential processing, MobileNetV2 focuses on the efficient feature extraction through per-channel and spatial separation with depthwise separable convolutions. Such design has natural support to diverse, dispersed representations of features that are not narrow fitting. Its consistent performance during domain changes implies that computational efficiency does not require generalization robustness to decrease.

These insights have practical relevance for deploying medical imaging models in resource-limited environments (e.g., mobile devices, embedded systems, low-power edge devices). The competitive cross-modality performance demonstrates that deployment constraints do not require compromising between cross-modality robustness and computational efficiency.

6.6 Transfer Learning & ImageNet Pre-training

All encoders were set up with ImageNet pre-trained weights, which could affect measured generalization behavior. ImageNet pre-training provides general early-stage features (edges, textures, simple shapes) extracted from 1.2 million natural images across 1000 object categories. During fine-tuning on medical imaging tasks, different encoder architectures may emphasize or extract different aspects of these pre-trained features.

The findings indicate that architectures emphasizing balanced hierarchical feature extraction (EfficientNetB3, DenseNet121) may better preserve a diverse set of pre-trained ImageNet features during fine-tuning, enabling cross-modality transfer. Lightweight architectures (ResNet18) may rely more heavily on subsets of features that are susceptible to modality-specific characteristics.

Examining this hypothesis would require training encoders from scratch without ImageNet pre-training, which was not performed in this study. The influence of pre-training on measured generalization behavior continues to be an open question for future research.

Chapter-7 : Conclusion & Future Works

7.1 Summary

Though all four encoder based models had a high and competitive accuracy on the in domain MRI test set of the LiverHccSeg dataset (with DenseNet121 giving a peak DSC value of 0.928), what this research contributes most to is the ability of these models to perform well on external and unseen datasets, which are different in terms of acquisition protocols, patients populations, and most importantly, the imaging modality. The findings indicate that there are some significant difference in terms of generalization ability. The most robust and versatile encoder became EfficientNetB3 with the highest DSC scores on the external MRI dataset (CHAOS, DSC 0.852) and more challenging CT dataset (3D-IRCADb-01, DSC 0.825). It is based on its innovative scaling of compounds and efficient architecture design, which strikes a tradeoff between the capacity to extract features and computational parsimony. Conversely, DenseNet121, though a competitive in-domain and on single external CT set (SLIVER07, DSC 0.806) was extremely sensitive to modality variations, and would regularly make erratic and untrustworthy segmentations on CT data. This issue provides the fact that all popular architectures are not necessarily domain-resilient. Our research provides a deep performance foundation of cross-modality liver segmentation which proves that the selection of architecture can have a profound impact on the generalization. However, this research also has such limitations as the relatively small and homogeneous training dataset, limitations of handling 3D anatomical structures with 2D slices with independence, and some trade-offs in precision and recall with different datasets. Altogether, these results highlight the necessity to conduct further research on larger and more varied training datasets, more extensive research on more sophisticated architectures including attention mechanisms and 3D convolutions, and a more intensive study of specific architectural properties that facilitate crossmodality resilience. These initiatives are necessary to gain more reliable, generalizable and clinically deployable deep learning solutions to medical image analysis.

7.2 Limitation

The research is a good source of reference in the study of encoder performance in cross-modality liver segmentation, but its inferences are limited by a number of flaws in the study. The first limitation is the size and homogeneity of the training dataset. The training of the models was performed only using the rather small and specific LiverHccSeg dataset, which comprises of arterialphase T1-weighted MRI of a restricted number of hepatocellular carcinoma patients. Such a dataset is unable to represent the wide biological diversity of patient groups or the vast technical disparities that different scanner manufacturers, strength of magnetic fields, and acquisition regimes bring across the world. The models, therefore, were very accurate in-domain, however, the attributes learned could be specific to the intensity distributions and anatomical features available in this single dataset. This worsens the problem of domain shift and can inflate the generalization gap that is seen during the application of models to new CT or MRI data. This weakness is amplified by the fact that a 2D slice-based U-Net model is used where each axial liver image is individually processed, thus ignoring volumetric information and inter-slice spatial features that constitute the three dimensional form of the liver. The range and method of evaluation

of the study is also among the other constraints to generalization. They might not work out these ambiguities in individual slices and may limit the accuracy of segmentation, especially on tasks (such as cross-modality) where the consistency of 3D forms may be a stronger invariant than 2D textures. The test architecture purposefully leaves a traditional U-Net decoder and does not provide any of the recent improvements to it, including attention gates, transformer blocks, or other more sophisticated domain-invariance methods. As a result, the hierarchy of encoder performance that was observed may vary in a more advanced architecture. Also, although cross-modality validation is a major strength, it can only be done with a few contrast enhanced MRI and CT protocols. The robustness of models in a broader range of clinical settings, including non-contrast CT, the alternative MRI sequences, images with severe artifacts, and so on, has not been tested and remains unconfirmed. Applying aggregate measures, like the Dice coefficient, can conceal important clinical errors, such as systematic boundary errors or topological inconsistencies in segmentations, which are especially of importance when used in practice, e.g., surgical planning. Lastly, the assessment does not touch on practical deployment factors, such as computational efficiency, inference speed, and memory footprint, which are all factors that determine how well they can be implemented into clinical processes.

7.3 Future Work

Based on the contributions and shortcomings determined in this benchmarking study, it is possible to imagine a wide, multidimensional scenery of future work that would help resolve the existing shortcomings and develop clinically sound, generalizable, and ethically responsible automated liver segmentation systems. This future work may be theoretically divided into a few interrelated pillars which deal with certain challenges but introduce new opportunities to innovation.

The first pillar focuses on the use of multi-institutional, large-scale, and multi-modal datasets in order to ensure that the spectrum of clinical variability is well-represented. It involves the combination of data in healthcare centers of differing global locations to represent a broad spectrum of biological diversity, e.g., age, ethnicity, body mass index, as well as liver pathologies other than hepatocellular carcinoma such as different stages of cirrhosis, steatosis, and vascular malformations. Introducing and using these benchmarks would offer a starting point on training models to learn truly modality - and protocol-invariant anatomical representations.

The second pillar is centered on the architectural development. The research is to shift towards structures which can utilize 3D spatial context naturally rather than 2D slice-based models. This includes architectural design and optimization of fully 3D convolutional networks (i.e., 3D U-Net, V-Net) and hybrid designs of volumetric data. On volumetric medical data, future research must critically assess the next-generation encoder backbones and architectural paradigms, the models based on EfficientNet or DenseNet variants, transformer-based models, such as Swin Transformers and UNETR (U-net Tranformers) to learn the long-range inter-slice relation, and dynamically assign computational resources to the challenging regions. In addition to the core segmentation network, implementation of the following advanced modules should be provided: spatial and channel attention mechanisms to emphasize the salient organ boundaries, contrast-adaptive normalization layers to overcome the intensity variations, and the adversarial or consistency-based modules that can be trained to learn modality-invariant spaces of features.

Other areas of research would be novel loss functions that impose both topological and anatomical consistency on the output on top of voxel-wise accuracy (e.g. through persistent homology) and support robust learning paradigms. The pre-training on large, unlabeled and multi-modal datasets, which is self-supervised and unsupervised, is among the major directions; models can be trained by masked image modeling, multimodal contrastive learning, and anatomical symmetry modeling to allow them to learn intrinsic anatomical representations that facilitate data efficiency

and out-of-distribution generalization. The further research on the generalization and test-time adaptation is crucial, such as feature distribution alignment, style transfer and the meta-learning method of adapting to novel scanners or protocols without labeled data. Moreover, causal representation learning - learning to model underlying causes of imaging variability, which can be anatomy, pathology, and modality physics, can be helpful in models resistant to false correlation, leading to foundational liver segmentation models.

Bibliography

[1] I. Macpherson, K. W. Abeysekera, R. Harris, D. Mansour, S. McPherson, I. Rowe, W. Rosenberg, J. F. Dillon, and A. Yeoman, "Identification of liver disease: why and how," *Frontline gastroenterology*, vol. 13, no. 5, pp. 367–373, 2022.

[2] S. De Siervi, S. Cannito, and C. Turato, "Chronic liver disease: latest research in pathogenesis, detection and treatment," p. 10633, 2023.

[3] H. Devarbhavi, S. K. Asrani, J. P. Arab, Y. A. Nartey, E. Pose, and P. S. Kamath, "Global burden of liver disease: 2023 update," *Journal of hepatology*, vol. 79, no. 2, pp. 516–537, 2023.

[4] H. Rumgay, M. Arnold, J. Ferlay, O. Lesi, C. J. Cabasag, J. Vignat, M. Laversanne, K. A. McGlynn, and I. Soerjomataram, "Global burden of primary liver cancer in 2020 and predictions to 2040," *Journal of hepatology*, vol. 77, no. 6, pp. 1598–1606, 2022.

[5] Z. M. Younossi, G. Wong, Q. M. Anstee, and L. Henry, "The global burden of liver disease," *Clinical Gastroenterology and Hepatology*, vol. 21, no. 8, pp. 1978–1991, 2023.

[6] M. Blachier, H. Leleu, M. Peck-Radosavljevic, D.-C. Valla, and F. Roudot-Thoraval, "The burden of liver disease in europe: a review of available epidemiological data," *Journal of hepatology*, vol. 58, no. 3, pp. 593–608, 2013.

[7] S. Tanwar, F. Rhodes, A. Srivastava, P. M. Trembling, and W. M. Rosenberg, "Inflammation and fibrosis in chronic liver diseases including non-alcoholic fatty liver disease and hepatitis c," *World journal of gastroenterology*, vol. 26, no. 2, p. 109, 2020.

[8] M. Y. Ansari, A. Abdalla, M. Y. Ansari, M. I. Ansari, B. Malluhi, S. Mohanty, S. Mishra, S. S. Singh, J. Abinahed, A. Al-Ansari *et al.*, "Practical utility of liver segmentation methods in clinical surgeries and interventions," *BMC medical imaging*, vol. 22, no. 1, p. 97, 2022.

[9] K. J. Oldhafer, V. Habbel, K. Horling, G. Makridis, and K. C. Wagner, "Benign liver tumors," *Visceral medicine*, vol. 36, no. 4, pp. 292–303, 2020.

[10] Y. Takahashi, E. Dungubat, H. Kusano, D. Ganbat, Y. Tomita, S. Odgerel, and T. Fukusato, "Application of immunohistochemistry in the pathological diagnosis of liver tumors," *International journal of molecular sciences*, vol. 22, no. 11, p. 5780, 2021.

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

[12] N. H. C. of the People's Republic of China *et al.*, "Guidelines for the diagnosis and treatment of primary liver cancer (2022 edition)," *J Multidiscip Cancer Management*, vol. 8, no. 2, pp. 16–53, 2022.

[13] L. D. Nephew, G. Aitcheson, and M. Iyengar, "The impact of racial disparities on liver disease access and outcomes," *Current Treatment Options in Gastroenterology*, vol. 20, no. 3, pp. 279–294, 2022.

[14] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical image analysis*, vol. 84, p. 102680, 2023.

[15] M. D. Pomohaci, M. C. Grasu, A.-Ş. Băicoianu-Nițescu, R. M. Enache, and I. G. Lupescu, "Systematic review: Ai applications in liver imaging with a focus on segmentation and detection," *Life*, vol. 15, no. 2, p. 258, 2025.

[16] J. C. Ahn, A. Connell, D. A. Simonetto, C. Hughes, and V. H. Shah, "Application of artificial intelligence for the diagnosis and treatment of liver diseases," *Hepatology*, vol. 73, no. 6, pp. 2546–2563, 2021.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241, 2015.

[18] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine*, vol. 8, no. 11, p. 713, 2020.

[19] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net. arxiv," *arXiv preprint arXiv:2211.14830*, 2022.

[20] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[21] E. Ahishakiye, M. Bastiaan Van Gijzen, J. Tumwiine, R. Wario, and J. Obungoloch, "A survey on deep learning in medical image reconstruction," *Intelligent Medicine*, vol. 1, no. 03, pp. 118–127, 2021.

[22] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical image analysis*, vol. 69, p. 101950, 2021.

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

[24] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[25] 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*, pp. 770–778, 2016.

[26] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, pp. 424–432, 2016.

[27] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*. Springer, pp. 3–11, 2018.

[28] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical image analysis*, vol. 63, p. 101693, 2020.

[29] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[30] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[31] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1013–1023, 2021.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[34] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[35] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1316–1325, 2019.

[36] E. Vorontsov, M. Cerny, P. Régnier, L. Di Jorio, C. J. Pal, R. Lapointe, F. Vandenbroucke-Menu, S. Turcotte, S. Kadoury, and A. Tang, "Deep learning for automated segmentation of liver lesions at ct in patients with colorectal cancer liver metastases," *Radiology: Artificial Intelligence*, vol. 1, no. 2, p. 180014, 2019.

[37] L. Bi, J. Kim, A. Kumar, and D. Feng, "Automatic liver lesion detection using cascaded deep residual networks. 2017," *arXiv preprint arXiv:1704.02703*, 2017.

[38] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine, "Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based postprocessing," *Scientific reports*, vol. 8, no. 1, p. 15497, 2018.

[39] C. You, J. Yang, J. Chapiro, and J. S. Duncan, "Unsupervised wasserstein distance guided domain adaptation for 3d multi-domain liver segmentation," in *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*. Springer, pp. 155–163, 2020.

[40] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.

[41] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[42] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[43] M. Gross, S. Arora, S. Huber, A. S. Kücükkaya, and J. A. Onofrey, "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.

[44] L. Soler, A. Hostettler, V. Agnus, A. Charnoz, J.-B. Fasquel, J. Moreau, A.-B. Osswald, M. Bouhadjar, and J. Marescaux, "3d image reconstruction for comparison of algorithm database," *IRCAD Research Dataset Documentation*, vol. 13, 2010.

[45] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical image analysis*, vol. 69, p. 101950, 2021.

[46] B. van Ginneken, "Sliver07," Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2597575

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255, 2009.

[48] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, pp. 6105–6114, 2019.