

# A Lightweight Multi-Attention Deep Architecture for Liver Tumor Segmentation with Limited Samples

*Dissertation submitted in partial fulfillment of the requirements for  
the degree of*

Master of Technology  
in  
Computer Science

by

Chandradipa Nag (CS2211)

Under the Guidance of

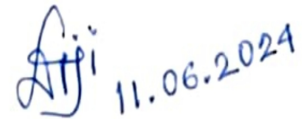
**Dr. Pradipta Maji**, Professor  
Machine Intelligence Unit  
Indian Statistical Institute



Indian Statistical Institute  
Kolkata - 700108, India  
June 2024

# Certificate

I certify that the project work titled "**A Lightweight Multi-Attention Deep Architecture for Liver Tumor Segmentation with Limited Samples**" authored by **Chandradipa Nag**, has been conducted under my supervision and guidance. The thesis represents the original work of the candidate, demonstrating thorough research and investigation. The quality of the thesis meets the expectations for the Master of Technology program in Computer Science, and I highly recommend its submission for evaluation.

A handwritten signature in blue ink, followed by the date "11.06.2024" also in blue ink.

---

Professor Pradipta Maji  
Machine Intelligence Unit  
Indian Statistical Institute  
Kolkata - 700108  
India

# Declaration

I, **Chandradipa Nag**, declare that this dissertation titled, "**A Lightweight Multi-Attention Deep Architecture for Liver Tumor Segmentation with Limited Samples**", which is submitted in fulfillment of the requirements for the Degree of Master of Technology in Computer Science, represents my own work except where due acknowledgement has been made. I further declare that it has not been previously included in a thesis, dissertation, or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed: Chandradipa Nag

Date: 12/06/2024

# Abstract

Liver tumor segmentation from CT images is of paramount importance in medical image analysis. Accurate segmentation of liver tumor is crucial for effective diagnosis and treatment planning in hepatocellular carcinoma and other liver malignancies. Manual as well as traditional segmentation approaches often struggle with the complex and heterogeneous nature of liver tumors, necessitating advanced deep learning techniques.

In this regard, the thesis introduces a supervised lightweight multi-attention deep architecture, termed as LiMAU, for liver tumor segmentation. It judiciously integrates the merits of an enhanced U-Net architecture, known as U-Net3+, traditional attention gates, and the convolutional block attention module (CBAM). The U-Net3+ represents a refined version of the traditional U-Net design, enriching it with full-scale skip connections and deep supervision, thereby enhancing its architectural sophistication. The full-scale skip connections merge low-level details with high-level semantics from feature maps at different scales, while deep supervision learns hierarchical representations from the fully aggregated feature maps. This structure is particularly beneficial for organs appearing at varying scales. The incorporation of U-Net3+ in the proposed LiMAU reduces the number of network parameters, thereby enhancing computational efficiency. The integration of traditional attention gates allows the proposed supervised model to selectively focus on relevant regions, enhancing feature learning by suppressing irrelevant background noise. On the other hand, the CBAM, which sequentially applies channel and spatial attention, further refines this focus by enhancing the model's ability to capture contextual and fine-grained details essential for precise tumor delineation. The proposed LiMAU features batch normalization layers in each double convolution block, which leads to higher segmentation accuracy.

Next, the thesis introduces a deep framework for semi-supervised learning as a promising solution for liver tumor segmentation with limited labeled samples. The proposed LiMAU serves as the cornerstone of the proposed semi-supervised approach. It integrates a novel adversarial consistency learning architecture, which effectively utilizes less labeled data while providing high segmentation accuracy. The proposed semi-supervised framework harnesses both labeled and unlabeled data to mitigate the requirement for extensive annotated data. The proposed framework judiciously integrates deep adversarial networks and the *II* model. The *II* model is based on the concept of consistency learning, which maintains the consistency of segmentation output during training across various random perturbations of both labeled and unlabeled data. The deep adversarial network consists of a segmentation network (SN) and two evaluation networks (ENs). While the SN is used for the segmentation task, the ENs are used to assess segmentation quality. The proposed LiMAU is used as the SN, while a variant of VGG16 is used for both ENs. During training, the first EN is incentivized to differentiate between annotated and unannotated image segmentation, the second one

is encouraged to distinguish between perturbed and unperturbed data, while the SN is encouraged to produce segmentations for unlabeled images similar to those for annotated ones.

The performance of the proposed supervised and semi-supervised models is evaluated on two benchmark data sets, namely, MICCAI 2017 Liver Tumor Segmentation Challenge (LiTS17) data and MICCAI-SLiver07 data, and compared with that of several state-of-the-art approaches. Experimental results demonstrate a significant improvement in segmentation accuracy over baseline models, with higher Dice similarity coefficients. This indicates that the combined use of traditional attention mechanisms and CBAM in the U-Net3+ architecture in supervised implementation as well as the semi-supervised adversarial network implementation significantly enhances the model's ability to manage the variability and complexity of liver tumor morphology. These findings suggest that the proposed models hold great potential for clinical applications, offering improved precision in liver tumor segmentation.

**Keywords:** Medical Imaging · Liver Tumor Segmentation · Deep Learning · Supervised Learning · Semi-Supervised Learning · Consistency Learning · Adversarial Learning · Attention Mechanism.

# Table of Contents

<b>1</b>	<b>Chapter 1: Introduction</b>	<b>1</b>
1.1	Problem Definition: Liver Tumor Segmentation . . . . .	1
1.2	Classical Approaches of Image Processing for Liver Tumor Segmentation . .	1
1.3	Deep Learning Based Supervised Approaches for Liver Tumor Segmentation	2
1.4	Problem of Getting Labeled Samples for Liver Tumor Segmentation . . . . .	3
1.5	Classical and Deep Learning Based Semi-Supervised Approaches for Liver Tumor Segmentation . . . . .	3
1.6	Contribution of the Proposed Work . . . . .	4
<b>2</b>	<b>Chapter 2: Literature Survey</b>	<b>6</b>
2.1	Existing Supervised Approaches . . . . .	6
2.2	Existing Semi-Supervised Approaches . . . . .	8
<b>3</b>	<b>Chapter 3: LiMAU: A New Lightweight Multi-Attention Deep Supervised Approach</b>	<b>10</b>
3.1	Batch Normalization: . . . . .	11
3.2	U-Net3+ Architecture: . . . . .	11
3.3	Deep supervision: . . . . .	12
3.4	UNet3+ with Attention Gates and CBAM Modules: . . . . .	12
3.5	Loss function . . . . .	15
<b>4</b>	<b>Chapter 4: Proposed Semi-Supervised Approach Based on LiMAU</b>	<b>16</b>
4.1	Adversarial and Consistency Learning architecture: . . . . .	17
4.2	Optimisation Function: . . . . .	18
<b>5</b>	<b>Chapter 5: Experiments</b>	<b>20</b>
5.1	Dataset . . . . .	20
5.2	Preprocessing . . . . .	20
5.3	Experimental Setup . . . . .	20
5.4	Batch-Normalization Comparison . . . . .	20
5.5	Ablation Study . . . . .	22
5.6	Comparative Performance Analysis . . . . .	26
<b>6</b>	<b>Chapter 6: Conclusions and Future Directions</b>	<b>27</b>
6.1	Future Directions: . . . . .	27
<b>7</b>	<b>References</b>	<b>29</b>

# Chapter 1: Introduction

## 1.1 Problem Definition: Liver Tumor Segmentation

Medical image segmentation plays a crucial role in various clinical applications, particularly providing analysis of tumors in different organs. Liver cancer ranks among the deadliest cancers in the world, with high mortality rates. Thus it needs fast and efficient detection and segmentation of tumors. Liver tumor segmentation involves the identification and delineation of regions of interest, namely tumors, within liver images obtained through medical imaging techniques. With advancements in imaging technologies, such as magnetic resonance imaging (MRI) and computed tomography (CT), the ability to accurately delineate tumors from surrounding tissues has become paramount for diagnosis, treatment planning, radiation therapy and monitoring of diseases [1–4]. However, manual segmentation of medical images is a labor-intensive task requiring expert knowledge, making it both time-consuming and costly and error-prone. CT technology provides visualization of tumors, yet achieving accurate segmentation presents challenges due to various factors. These include the presence of significant noise within CT images, which can blur the boundary between the tumor and the liver. Additionally, gray-scale images often exhibit low contrast between liver tumors and adjacent organs due to similar tissue density. Furthermore, the size, shape, and location of liver tumors vary among patients, further complicating the segmentation process.

Hence, to help doctors diagnose, advanced machine learning techniques involving deep learning framework are introduced to segment liver tumors without human intervention in computer-aided diagnosis and treatment research. It enables the extraction of essential organs or lesions from abnormal images.

## 1.2 Classical Approaches of Image Processing for Liver Tumor Segmentation

Classical approaches to image processing and computer vision for tumor segmentation involve various traditional techniques that predate the widespread adoption of deep learning. These methods are often based on principles of image analysis, pattern recognition, and statistical modeling. Thresholding techniques have a central role in image segmentation because of their simplicity and computational speed. Thresholding techniques involve setting a specific intensity value to separate the tumor from the surrounding tissue. The result of the segmentation is solely dependent on the intensity of the pixels in the image [5].

Active contour models (snakes) are also used for the segmentation of tumors. It uses deforming splines, also known as snakes, to fit an object in an image [6]. The snake acts on internal and external forces to minimize the composed energy dynamically. This is done by iteratively solving a partial differential equation (PDE) which deforms the curve to fit the nearest contour, which is the encircled object in the image.

**Merits:** Classical methods in image processing and computer vision offer simplicity and ease of implementation, making them accessible and easy to understand. They typically require less computational resources and are suitable for real-time applications that run efficiently on less powerful hardware. These methods are also highly interpretable, allowing users to easily understand and explain their operations and results. Additionally, many classical methods are well-established and have been extensively tested, providing a reliable baseline for various tasks. Their flexibility allows them to be adapted to a wide range of problems, often serving as useful preprocessing steps for more complex algorithms.

**Demerits:** Despite their advantages, classical methods often struggle with complex, heterogeneous, or noisy data, as they may fail to capture intricate patterns and details. Their performance heavily depends on the choice of parameters, such as threshold values and kernel sizes, which can require extensive experimentation to optimize. These methods do not adapt well to variations in data and cannot learn from new data, limiting their applicability in dynamic environments. Manual feature extraction is often necessary, which can be time-consuming and labor-intensive. Furthermore, classical methods are sensitive to noise and artifacts, leading to potential inaccuracies such as over-segmentation or under-segmentation. They may also not generalize well to new or unseen data without significant modifications.

### 1.3 Deep Learning Based Supervised Approaches for Liver Tumor Segmentation

Deep learning-based approaches have revolutionized the field of image processing and computer vision, including medical image segmentation. These methods leverage neural networks, to automatically learn complex patterns and features from data, leading to significant improvements in accuracy and robustness. Deep learning, particularly Convolutional Neural Networks (CNNs) [7], has shown remarkable promise in medical image analysis tasks. In particular, encoder-decoder networks like U-Net [8, 9], U-Net++ [10], H-DenseUNet [11], and others have shown impressive performance in this domain.

**Merits:** The U-Net based architectures has emerged as a popular choice for biomedical image segmentation due to its ability to produce high-resolution segmentations with limited training data. Deep learning models can achieve high accuracy in liver tumor segmentation by learning complex patterns and features directly from the data, outperforming classical methods significantly. These models automate the feature extraction process, eliminating the need for manual feature engineering. This makes the models more adaptable to different datasets and types of images.

**Demerits:** Despite its effectiveness, classic U-Net architectures face challenges in capturing fine-grained details and handling the complex nature of liver tumors, especially across varying scales. Traditional skip connections in U-Net aim to reduce the semantic gap between the encoder and decoder, but they often fail to fully exploit information from all scales. Low-level detailed feature maps capture rich spatial information, highlighting organ boundaries, while high-level semantic feature maps provide positional information, indicating organ locations. However, these crucial signals may become diluted during progressive down-sampling and up-sampling processes.

**Improvements:** To address these challenges, many models like UNet++, UNet++ have been introduced which focuses on the nested skip pathways, connecting each decoder layer to a subset of encoder layers. This architecture facilitates the integration of multi-scale features while maintaining computational efficiency.

However, this thesis explores U-Net3+, an improvement on UNet++, which introduces a densely connected skip pathway. This ensures that each encoder layer connects to all decoder layers, fa-



ilitating the flow of information across different scales effectively. In UNet3+ the full-scale skip connections merge low-level details with high-level semantics from feature maps at different scales, significantly improving the model’s ability to capture multi-scale features. Deep supervision further enhances the learning of hierarchical representations from fully aggregated feature maps, making the model particularly effective for organs appearing at varying scales. Additionally, this redesigned interconnection between the encoder and decoder, along with intraconnections within the decoders, captures fine-grained details and coarse-grained semantics across all scales. This design not only improves accuracy, but also reduces network parameters, enhancing computational efficiency. To further improve performance, batch normalization layers are added to the double convolution blocks of each U-Net structure. Batch normalization enhances network performance by reducing internal co-variate shift during training, leading to faster convergence and improved generalization.

In addition to architectural enhancements, we integrate attention mechanisms [13] into the segmentation pipeline to further refine the segmentation results. Traditional attention gates are utilized to enhance the propagation of semantic information through skip connections. These attention gates selectively highlight relevant features of the encoder, aiding in the segmentation task by focusing on important regions. These attention gates improve the model accuracy by focusing on relevant regions, improving the positioning accuracy by suppressing irrelevant background noises.

Moreover, to capture more valuable global information, we also incorporate the Convolutional Block Attention Module (CBAM) [14] into the decoder stage of the network, providing a more comprehensive attention mechanism. CBAM assesses both the channel-wise and spatial-wise significance of features, enabling the model to focus on identifying which features are important and where they are located within the feature maps. This comprehensive attention mechanism has the potential to lead to more refined segmentation results by effectively capturing relevant features and suppressing irrelevant ones.

#### 1.4 Problem of Getting Labeled Samples for Liver Tumor Segmentation

Despite the effectiveness of earlier methods, they rely heavily on abundant pixel-level labeled data, which is often scarce and expensive to acquire in medical imaging because of issues such as low contrast and noise interference. Obtaining labeled samples for liver tumor segmentation is challenging due to several factors. Medical data is scarce and sensitive, with privacy regulations limiting access. In addition, the annotation of medical images demands specialized knowledge, exacerbating the challenge of building large data sets with accurate labels. Also, ensuring the quality of annotations demands thorough review and verification.

In response to these hurdles, semi-supervised learning [15] emerges as a promising approach to address the inadequacy of data supervision in medical image segmentation. By leveraging a small portion of labeled data alongside a larger pool of unlabeled data for joint training, semi-supervised learning better aligns with the constraints of real-world clinical settings compared to traditional supervised learning methods. These methods aim to enhance model performance while minimizing the need for extensive manual annotation.

#### 1.5 Classical and Deep Learning Based Semi-Supervised Approaches for Liver Tumor Segmentation

Classical and Deep Learning Based semi-supervised learning methods encompass a range of techniques aimed at leveraging both labeled and unlabeled data to enhance learning performance. Within

semi-supervised medical image segmentation, there are several methodologies, broadly categorized into consistency learning [16–21], adversarial learning [22–24]. Approaches like self-training iteratively trains models on labeled data and then use them to predict labels for unlabeled data [25]. While contrastive training involves developing a model to distinguish between similar and dissimilar pairs of data samples in order to learn meaningful representations [26]. Collaborative training involves multiple models working together, often sharing information or parameters, to collectively improve learning performance on a given task [27]. This thesis specifically concentrates on consistency learning and deep adversarial learning. consistency learning employs consistency regularization with diverse perturbations to effectively train a network. On the other hand, adversarial learning, particularly leveraging Generative Adversarial Networks (GANs), involves two key components: a discriminator and a generator. The discriminator’s role is to distinguish between ground truth and generated samples, while the generator, also called the segmentation network, creates segmentation maps or predicted liver tumor masks and aims to produce outputs that are indistinguishable from ground truth, fostering mutual improvement through iterative updates. In this project, UNet3+ with Multi-Attention architecture is used as the segmentation network because of its best results in the supervision task. It helps produce outputs that closely resemble the ground truth, thus facilitating mutual improvement through iterative updates.

This study introduces an adversarial consistency training strategy utilizing double discriminators. The first discriminator learns the relationship between labeled and unlabeled data, while the second focuses on the image-level consistency of the segmentation network under varying data perturbations. Both discriminators aim to enhance the knowledge transfer capability of the segmentation network from labeled to unlabeled data, presenting a novel approach to semi-supervised medical image segmentation.

## 1.6 Contribution of the Proposed Work

The primary contribution of this thesis can be categorized into two segments: supervised and semi-supervised methodologies.

**Supervised Approach** : The thesis presents a supervised lightweight multi-attention UNet3+ based architecture, called LiMAU, for liver tumor segmentation. LiMAU consists of an encoder-decoder structure called U-Net3+, with Attention Gates and Convolutional Block Attention Module(CBAM).

1. **UNet3+:** The UNet3+ model is an advanced UNet based structure with skip connections between high-level and low-level semantics that incorporates batch normalization in each double convolution blocks. This advanced structure enhances stability and convergence, leading to improved segmentation performance.
2. **Attention Mechanisms:** The skip connections of UNet3+ integrates Attention Gates. It refines segmentation results by increasing the weight of relevant features and suppressing background noise.
3. **Convolutional Block Attention Module (CBAM):** The addition of CBAM provides a comprehensive attention mechanism, allowing the model to capture both channel-wise and spatial-wise relevance of features, resulting in more refined segmentation outcomes.

### **Semi-Supervised Approach :**

In this work a deep framework involving adversarial consistency training strategy for semi-supervised implementation is proposed which leverages both labeled and unlabeled data.

1. **Consistency Learning:** The  $P_i$  model of consistency learning produce random perturbations of both labeled and unlabeled data and then the model is trained to reduce differences between the predictions of perturbed and unperturbed data.
2. **Adversarial Learning:** It employs two discriminators and one generator. The first discriminator captures the relationship between labeled and unlabeled data, while the second ensures the segmentation network maintains consistency across different data perturbations. This approach enhances the network's ability to generalize from labeled to unlabeled data.

The proposed semi-supervised model uses a supervised U-Net3+ architecture with attention mechanisms, including traditional Attention Gates and CBAM, for supervised task, and incorporates adversarial and consistency learning for unsupervised task. Both methods are evaluated on two benchmark datasets: the MICCAI 2017 Liver Tumor Segmentation Challenge (LiTS17) dataset and the MICCAI-SLiver07 dataset. The experimental results demonstrate significant improvements in segmentation accuracy and computational efficiency compared to baseline models.

In this thesis, a detailed description of the segmentation approach, including the U-Net3+ architecture, attention mechanisms as well as the semi-supervised learning strategy is provided. A comprehensive evaluation of the model's performance is presented, highlighting its advantages in liver tumor segmentation and its potential for clinical applications.

# Chapter 2: Literature Survey

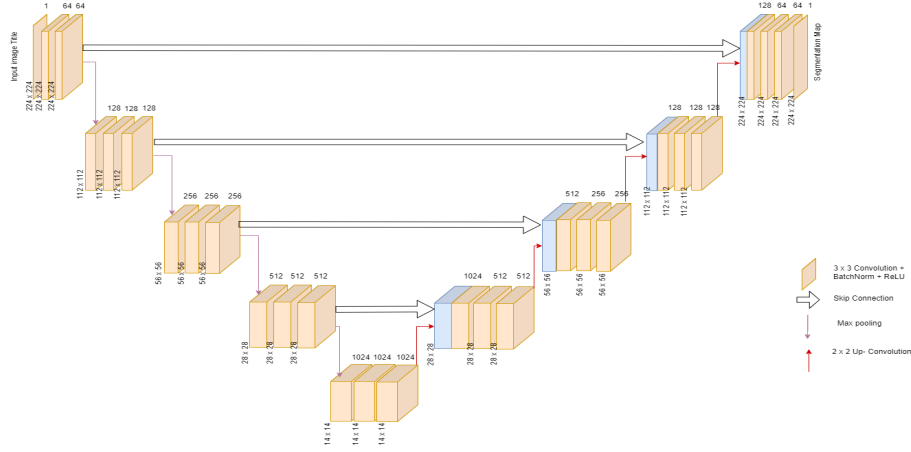
## 2.1 Existing Supervised Approaches

**Classical Approaches:** In recent years, a wide range of methodologies have been developed to address the complexities of automatic liver segmentation from 3D scans. Notable among these are graph-cut methods [28], which optimize segmentation through energy minimization techniques, deformable models [29] that adjust segment boundaries based on shape and image data, and level-set methods [30], which utilize evolving contours to define liver boundaries. There are also atlas-based approaches that employ pre-constructed anatomical models to guide segmentation. These traditional methods primarily depend on manual feature extraction, requiring significant expertise and effort to identify relevant features within the medical images. Traditional liver tumor segmentation methods involve mainly feature extraction followed by classification. Bastian et al. [31] achieved a good Dice score using intensity features, SLIC, and AdaBoost. Ali et al. [32] used first-order statistical features for liver boundary extraction and a k-Mean classifier for lesion classification. Chang et al. [33] applied binary logistic regression with texture, shape, and kinetic curve characteristics for tumor segmentation and classification. Liver analysis through medical imaging has garnered considerable interest, exemplified by competitions such as MICCAI Sliver and the 2017 MICCAI LiTS. These competitions have provided data sets to evaluate various segmentation algorithms, showcasing advances in network architectures. At the MICCAI LiTS Challenge, Bi et al. [34] highlighted the effectiveness of residual networks for liver segmentation, which employ residual connections to prevent gradient vanishing, especially in deep networks. Similarly, Huang et al. [35] proposed densely connected convolutional layers where each layer uses all preceding feature maps as inputs, and its feature maps are used as inputs for all subsequent layers. This architecture reduces gradient vanishing, enhances feature propagation, and encourages feature reuse. These methods often fail due to the complexity of tumor segmentation. Despite their effectiveness, these techniques often face limitations in handling the variability and complexity of liver anatomy in diverse populations of patients.

**Deep Learning Based Approaches:** Deep learning approaches, particularly Convolutional Neural Networks (CNNs), leverage large datasets to train sophisticated features directly from CT images and their ground-truth segmentations, leading to superior performance in liver and lesion segmentation.

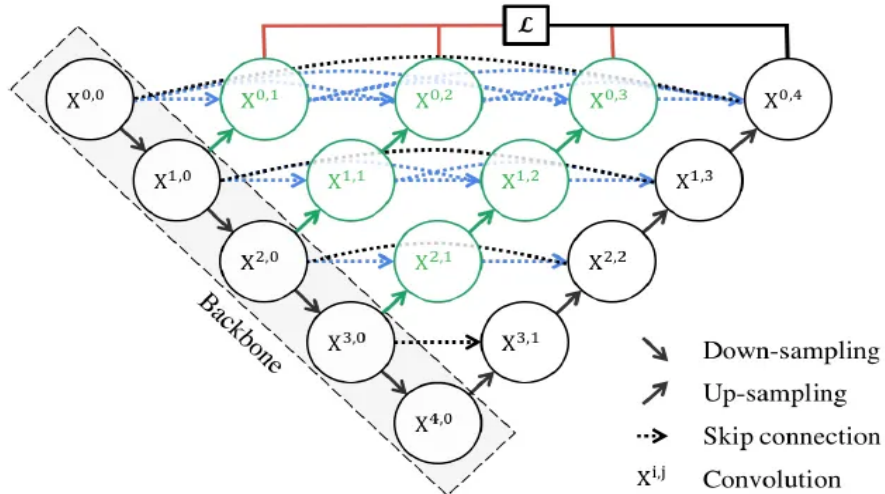
Since its introduction, the U-Net [8] architecture has become a fundamental tool in medical image segmentation, underpinning numerous studies that have modified it for various segmentation tasks. U-Net [36,37] features a dual-path structure: the contracting path and the expansive path as shown in Figure 1. The contracting path or the encoder involves a sequence of double convolutional

blocks are composed of two (3 x 3 convolution blocks, each followed by rectified linear unit (ReLU) activation) and 2x2 max pooling for down-sampling. In contrast, the expansive path or the decoder reconstructs segmentation images of the same size as the input using convolutional layers. The success of U-Net has led to many extensions, mainly aimed at improving skip connections.



**Fig. 1.** U-Net architecture

UNet++ [10, 38] marks a significant advancement over the original U-Net by incorporating redesigned dense skip connections between the encoder and decoder at multiple levels, along with nested convolutional blocks. The Figure 2 shows UNet++ structure which is built upon the traditional U-Net architecture by introducing nested and dense skip paths between the encoder. These blocks utilize multiple convolution layers to extract semantic information, with dense skip connections ensuring comprehensive feature concatenation and improved information flow. Attention



**Fig. 2.** U-Net++ architecture

mechanisms [13, 39] have further enhanced U-Net-based architectures. For instance, Attention Gates

have been introduced to selectively highlight relevant features from the encoder during segmentation, thereby improving accuracy by emphasizing important regions and suppressing irrelevant background noise. These mechanisms optimize information flow and enhance the model’s capacity to capture finer details in complex medical images.

Deep learning models, especially those with attention mechanisms, excel at focusing on target structures of varying shapes and sizes while suppressing irrelevant regions. This capability enhances the precision and accuracy of segmentation tasks, making deep learning superior to traditional methods in this domain.

## 2.2 Existing Semi-Supervised Approaches

**Classical Approaches:** Despite the promising results and state-of-the-art performances achieved by the above supervised methods with advanced architecture in medical image segmentation tasks, they still require a large amount of high-quality annotated data for training. Obtaining such large-scale, meticulously-labeled datasets is impractical, especially in medical imaging, where reliable and accurate annotations can only be provided by experts, making it a costly and difficult process. To alleviate the burden of manual labeling, significant efforts have been directed towards data-efficient deep learning methods. These methods include label generation, data augmentation, leveraging external related labeled datasets, and utilizing semi-supervised learning with unlabeled data. Among these, semi-supervised segmentation stands out as a practical approach, enabling models to make use of the more easily obtainable unlabeled data alongside a limited amount of labeled data, which is highly beneficial for real-world clinical applications.

The initial approach specifically tailored for semi-supervised segmentation employed a patch-based tree-structured method combined with the random forest algorithm [40]. Subsequently, a weighted graph-based model was introduced for the semi-supervised segmentation of 3D surfaces [41]. The random forest algorithm, known for its speed and interpretability, has been used in various semi-supervised segmentation tasks, such as abdominal magnetic resonance imaging [42]. The combination of Gaussian mixture models for probabilistic pixel modeling, random walk models for label propagation to achieve coherent segmentation, and SVMs for initial classification [43] was a notable approach used before the advent of deep learning to tackle semi-supervised segmentation problems.

**Deep Learning Based Approaches:** Since the emergence of deep-learning in recent years, it has significantly advanced semi-supervised learning methodologies, particularly in tasks such as image segmentation. Semi-supervised learning bridges the gap between supervised and unsupervised learning by leveraging a small amount of labeled data along with a large pool of unlabeled data. Pseudo-labeling involves using the model’s predictions on unlabeled data as if they were true labels. This approach iteratively refines the model by incorporating its own confident predictions into the training process. Lee (2013) [44] introduced the concept of pseudo-labeling, showing that using high-confidence predictions on unlabeled data can effectively guide the learning process.

Graph-based SSL methods leverage the natural structure of data, where data points are treated as nodes in a graph, and edges represent similarities between them. This structure helps in propagating label information from labeled to unlabeled nodes. Kipf and Welling (2017) [45] introduced the Semi-Supervised Classification with Graph Convolutional Networks (GCNs), where node features are updated based on their neighbors, facilitating effective label propagation.

Co-training and multi-view learning exploit multiple independent views of the data to improve learning outcomes. This is particularly useful when different views provide complementary information. Blum and Mitchell (1998) [46] initially introduced co-training, demonstrating its efficacy in scenarios with multiple independent feature sets.

Consistency regularization is another method that enforces the model to produce similar outputs for perturbed versions of the same input, thus enhancing the model’s robustness and performance on unlabeled data. Laine and Aila (2017) [47] presented the Temporal Ensembling method and  $\Pi$  model of consistency learning, which maintains an ensemble of predictions over different training epochs to create more stable and accurate models.

Adversarial learning is a popular approach to improve model robustness by effectively extracting potential insights from unlabeled data. For example, Zhang et al. [48] introduced a deep adversarial network (DAN) to boost the prediction quality of unlabeled data. Huang et al. [49] proposed a method where a fully convolutional discriminator differentiates predicted probability maps from ground truth segmentation distributions, considering spatial resolution. However, many semi-supervised adversarial learning methods rely on a single generator and discriminator, which can lead to low segmentation accuracy due to the dependence on the results of a single network.

Currently, semi-supervised medical image segmentation methods often use a standard encoder-decoder network as their backbone, focusing on refining learning strategies to leverage unlabeled data. However, a model with low segmentation accuracy can mislead the learning process. To address this, advanced techniques integrate both consistency learning and adversarial learning [50] to boost overall performance. This work also focuses on consistency learning and adversarial learning with Multi-Attention UNet3+ as the segmentation network to enhance performance.

# Chapter 3: LiMAU: A New Lightweight Multi-Attention Deep Supervised Approach

The thesis introduces a novel supervised architecture called LiMAU, which stands for lightweight multi-attention UNet3+. This advanced framework is specifically designed for the segmentation of liver tumors, offering a streamlined and efficient approach to medical image analysis. LiMAU architecture is a light weight model with comparatively less parameters than other advanced UNet based models. This model leverages the Batch Normalization layers within the double convolution blocks of the advanced UNet 3+ architecture, augmented with the Convolutional Block Attention Module (CBAM), traditional Attention Gates, to achieve superior segmentation performance. This method merges UNet 3+'s multi-scale feature fusion with CBAM and Attention Gates, enabling precise localization and boundary detection. It ensures focus on relevant regions while suppressing background noise for high segmentation accuracy and robustness. The inclusion of Batch Normalization further stabilizes the learning process and improves model convergence. The detailed block diagram of the proposed method is illustrated in Figure 3.

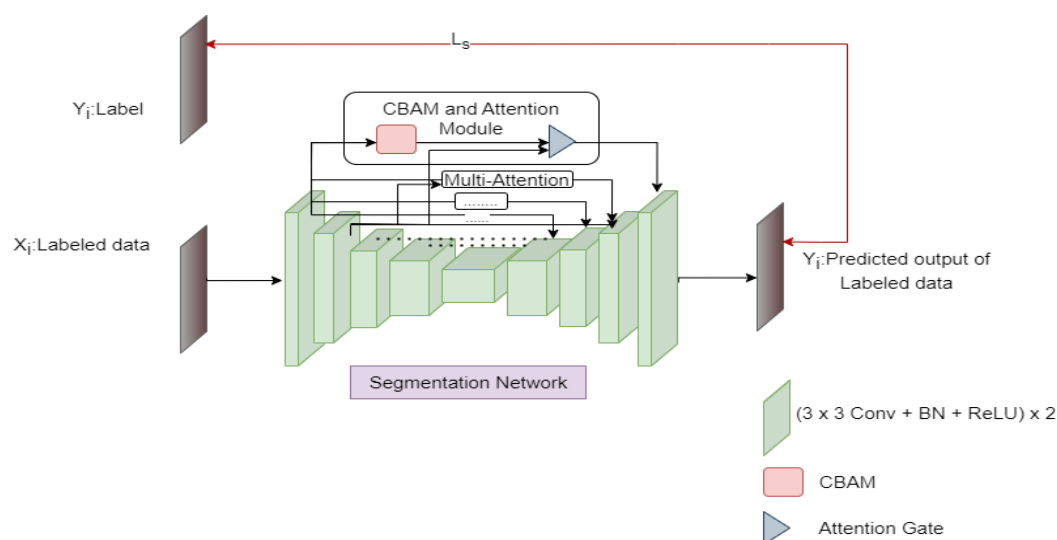


Fig. 3. U-Net3+ Block diagram



### 3.1 Batch Normalization:

A comparative study has been conducted to evaluate the impact of batch normalization on segmentation performance in different U-Net architectures: U-Net, U-Net++ and U-Net3+. The results established that models incorporating batch normalization consistently outperformed their non-batch normalization counterparts. Batch normalization significantly improved the training stability, convergence speed, and overall segmentation accuracy by reducing internal co-variate shift. This is particularly useful in medical image segmentation, where the variability in pixel intensity can be high. Additionally, BatchNorm makes the network less sensitive to the initialization of weights, increasing its robustness.

In our implementation, batch normalization layers were introduced in each double convolution block of the U-Net based structures. Specifically, the standard (Conv + ReLU) x 2 layers were modified to (Conv + BatchNorm + ReLU) x 2. This adjustment helps normalize the feature maps, ensuring that each layer receives inputs with a stable distribution, thus enhancing the model's learning capability.

### 3.2 U-Net3+ Architecture:

UNet3+ effectively captures comprehensive information across all scales of feature maps. This overcomes the limitations of both UNet and UNet++ in learning the precise target positions due to their inability to fully utilize information from multiple scales. UNet 3+ remedies this by integrating feature maps from smaller and same-scale levels of the encoder and larger-scale levels of the decoder at each decoder layer. This integration significantly enhances the precision of organ localization and boundary detection. In contrast, U-Net3+ takes these innovations one step further by implementing a more comprehensive connectivity scheme. As shown in Figure 4, in U-Net3+, each encoder layer is connected to all decoder layers, ensuring that information flows freely across the entire network. This fully connected skip pathway helps in effectively merging low-level details with high-level semantic information from feature maps at various scales, significantly enhancing the overall feature representation. By ensuring complete feature integration, U-Net3+ can better capture both fine and coarse grains, which is critical for accurate medical image segmentation.

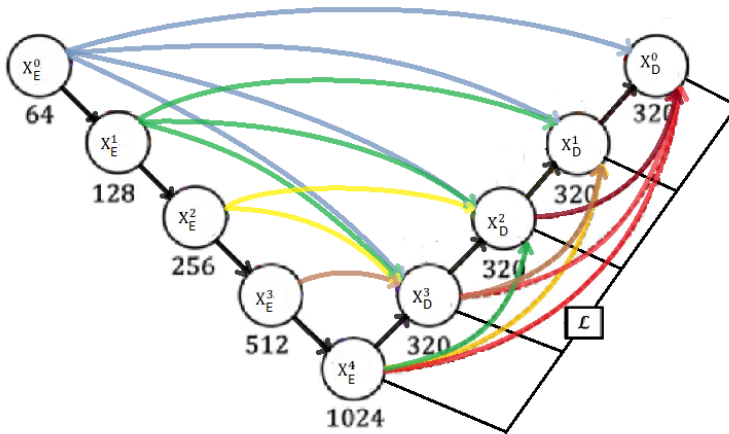


Fig. 4. U-Net3+ architecture

### 3.3 Deep supervision:

Also, U-Net3+ incorporates deep supervision, meaning that intermediate layers are directly supervised during training. In reference [12] the deep supervision of UNet3+ involves supervision of the ground truth with all the 5 decoder outputs which increases the computational complexity. However here it will be shown later if we average out the 5 outputs and then compare with the ground truth the performance is not hampered that much and also the time and complexity reduces drastically. This approach helps the network learn more robust and hierarchical representations, making it particularly effective for segmenting organs and structures that appear at varying scales. Despite increased connectivity, U-Net3+ is designed to be computationally efficient by optimizing the network architecture to reduce the number of parameters. This optimization speeds up both training and inference times without compromising accuracy, making U-Net3+ a practical and efficient choice for clinical applications.

In the UNet3+ structure, each convolutional layer is followed by a BatchNorm and then a ReLU. For example, Input comes to the first block  $X_E^0$  and passes through a block having Convolutional layer, a BatchNorm layer and a ReLU layer two times. Then it goes to  $X_D^3$  after passing through a Max-Pooling layer of stride 8 then a Convolutional layer, a BatchNorm layer and a ReLU layer. Similarly  $X_E^1$  also goes to  $X_D^3$  after passing through a Max-Pooling layer of stride 4 then again a Conv layer, a BatchNorm layer, a ReLU layer and so on. But  $X_E^3$  does not require a Max pooling layer as both  $X_E^3$  and  $X_D^3$  have same spatial dimensions, rather it goes through a layer (conv + BatchNorm + ReLU). Finally  $X_E^4$  requires a bilinear upsampling to match a dimensions with  $X_D^3$  and then a (conv + BatchNorm + ReLU) layer. Now all these 5 encode outputs are fused together. Additionally, we implement a feature aggregation mechanism on the combined feature map from five scales. This process utilizes 320 filters of size  $3 \times 3$ , followed by batch normalization and activation through a ReLU function. In this way we get all the 4 decoder outputs. The final layer of each decoder stage is passed through a straightforward  $3 \times 3$  convolutional layer, followed by bilinear up-sampling and activation with a sigmoid function. We define skip connections through the following formula: let  $i$  be the index of the output of the decoder stage and let  $N$  denote the total number of encoder layers.

$$X_D^i = \begin{cases} X_E^i & \text{if } i = N \\ H([C(\text{Down}(X_E^r))_{r=1}^{i-1}], C(X_E^i), C(\text{Up}(X_D^r))_{r=i+1}^N) & \text{if } i = 1, \dots, N - 1 \end{cases}$$

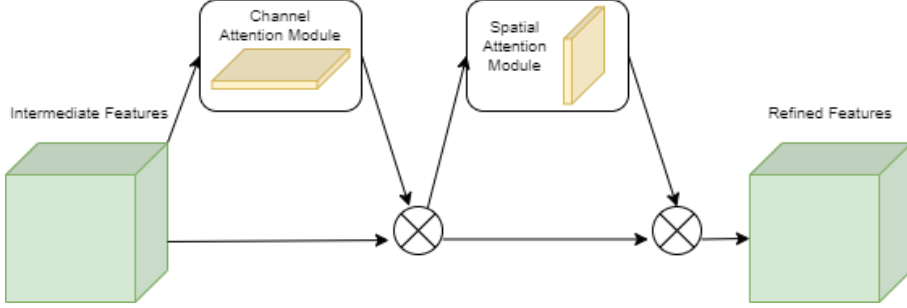
Here  $H$  is concatenation function with the feature aggregation mechanism that is made of a convolution layer followed by batch normalization and ReLU.  $C$  is a (conv + BatchNorm + ReLU) layer.  $\text{Down}$  is down-sampling and  $\text{Up}$  is up-sampling operation.

### 3.4 UNet3+ with Attention Gates and CBAM Modules:

Next we add two types of attention mechanisms to the skip connections of UNet3+. The Convolution Block Attention Module (CBAM) referred from [14], is composed of two distinct sub-modules: the Channel Attention Module and the Spatial Attention Module.

**CBAM** CBAM is designed as an efficient and versatile component that can be easily incorporated into various convolutional neural networks (CNNs). It enhances the network's performance with only a minimal increase in the number of parameters. The CBAM module begins by processing the

intermediate feature map, generating a one-dimensional channel attention map. This map is then applied to the intermediate feature map. Subsequently, a two-dimensional spatial attention map is computed and combined with the feature map from the preceding layer to refine features adaptively. Throughout this procedure, attention values are broadcasted accordingly: channel attention values extend spatially, and vice versa. CBAM is shown in Figure 5. CBAM is expressed as:



**Fig. 5.** CBAM Module

$$F' = M_c(F) \otimes F$$

$$F'' = M_s(F') \otimes F'$$

In this setup,  $F \in R^{C \times H \times W}$  represents the intermediate feature map,  $M_c \in R^{C \times 1 \times 1}$  computes the one-dimensional channel attention map, and  $M_s \in R^{1 \times H \times W}$  computes the 2D spatial attention map. The symbol  $\otimes$  signifies element-wise multiplication.  $F'$  denotes the refined feature map.

**Attention Gates** The attention mechanism initially rose to prominence in natural language processing (NLP) and quickly became a dominant approach. To effectively target and emphasize regions pertinent to the target organ, we adopt the technique introduced by reference [13] by integrating a straightforward yet powerful Attention Gate into the network architecture. The Attention Gate (AG) module, depicted in Figure 6, takes as inputs the up-sampled features from the expansion path and the corresponding features from the encoder. The former acts as a gating signal, enhancing the learning of target regions vital for the segmentation task while suppressing irrelevant areas. This mechanism improves the efficiency of semantic information propagation through skip connections. The sigmoid activation function is chosen to facilitate parameter convergence within the gate and compute the attention coefficient  $\alpha_i$  whose value lies between 0 to 1. The refined features result from multiplying the encoder features by the coefficient  $\alpha_i$ .

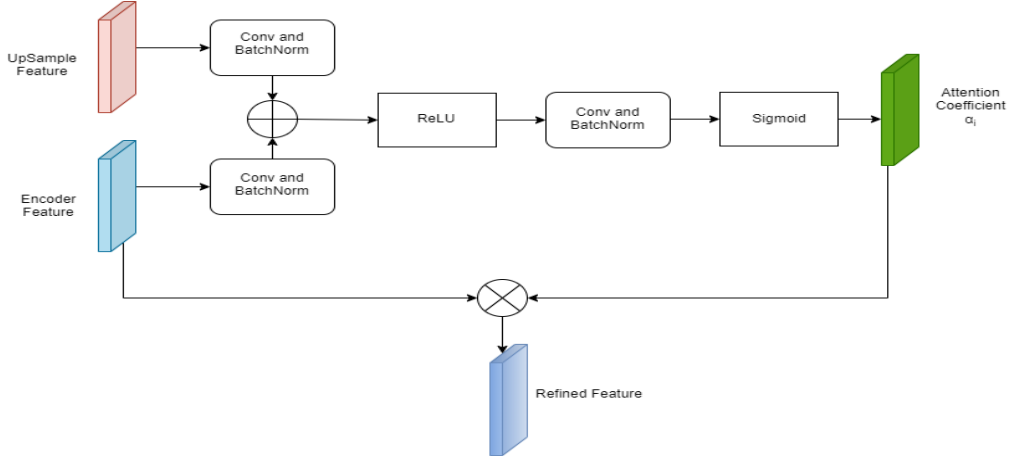


Fig. 6. Attention Module

Now we combine the two attention blocks and insert them into the skip connections as shown in Figure 7 and 8. First the output from encoder stage  $i$ ,  $X_E^i$  is given as input to CBAM Module and then the output from CBAM Module is given as input to Attention Module as the refined encoder input. Attention Gate also takes another input, the up-sampled feature from next stage encoder output,  $X_E^{i+1}$  as the gating signal. Then on the final output consecutive operations Max-pool (depending on the decoder stage), Convolution, BatchNorm, ReLU are performed to give the decoder output.

$$F''' = M_a(F'', U) \otimes F''$$

Here,  $F'' \in R^{C \times H \times W}$  represents the output from CBAM module,  $M_a \in R^{1 \times H \times W}$  computes the 2D attention map.  $U$  is up-sampled feature from next stage encoder.  $F'''$  is the attention output. The final architecture of UNet3+ with Multi-Attention is shown in Figure 7 with the detailed view of Attention blocks in Figure 8

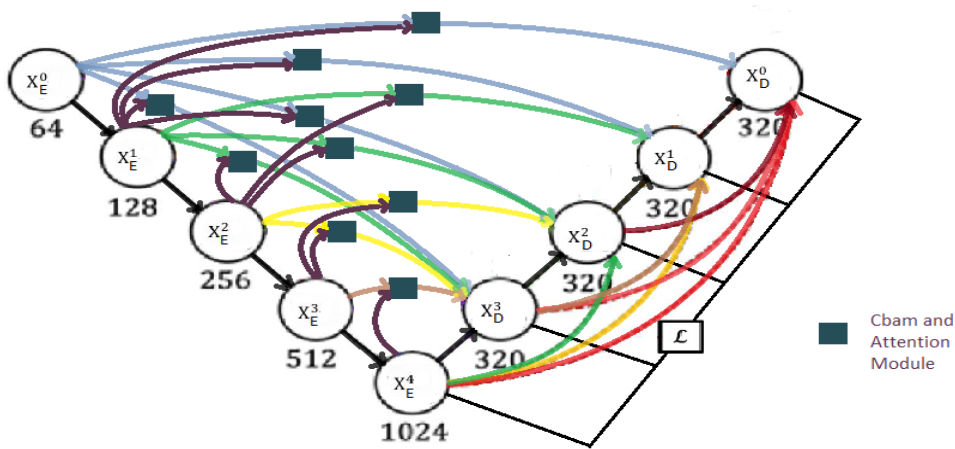
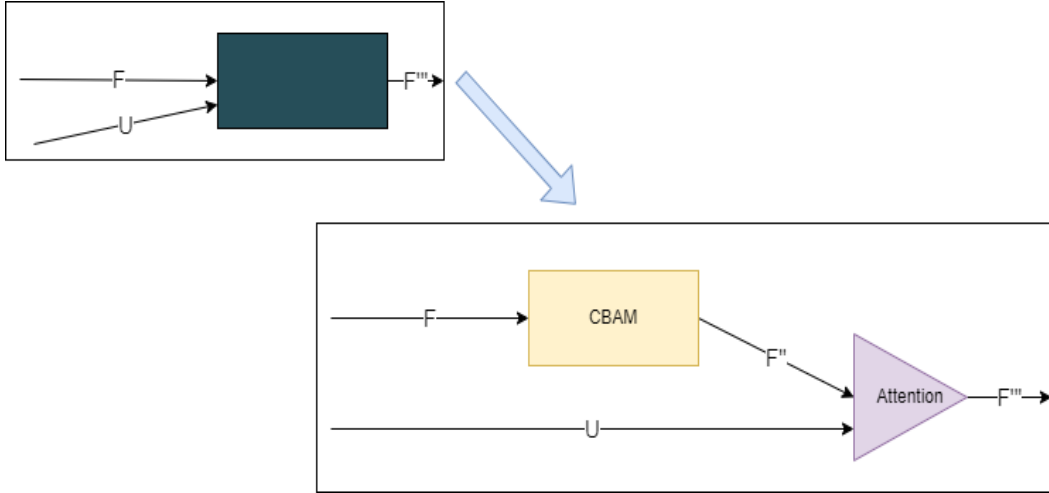


Fig. 7. Proposed UNet3+ with Multi-Attention Module



**Fig. 8.** Proposed CBAM and Attention Module

### 3.5 Loss function

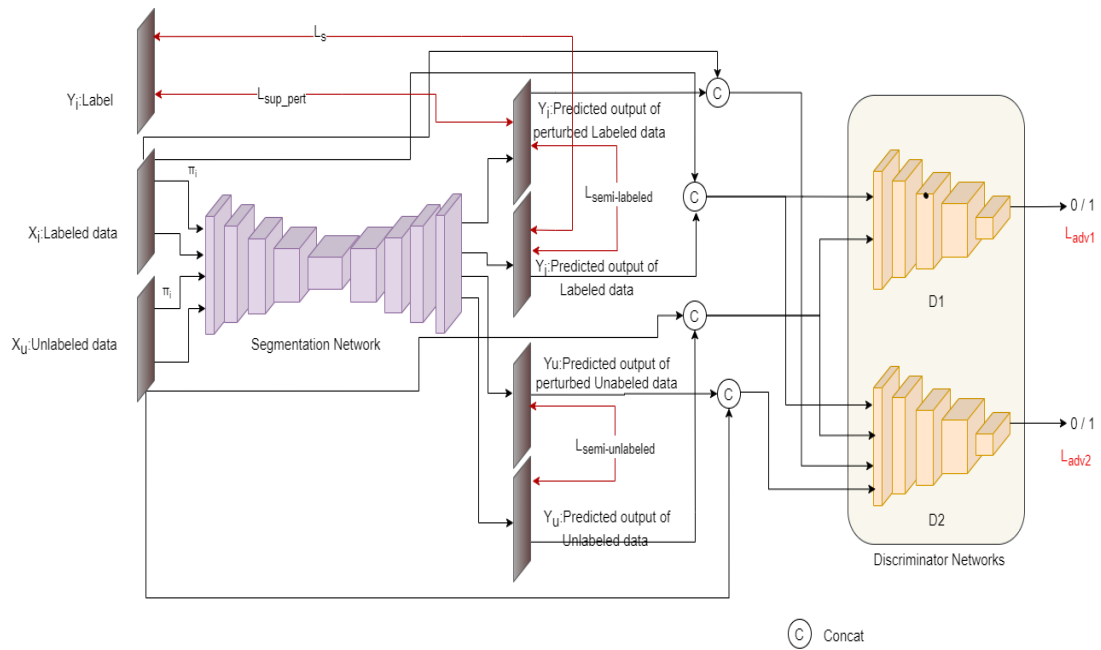
: The Loss function that have been used to perform experiments and analyze the models is binary cross entropy loss. In deep learning, the loss function measures the discrepancy between the actual value and the predicted value of the neural network output. The network chooses an appropriate loss function to enhance the accuracy of image segmentation. In this context, binary cross-entropy loss function is used, which is applied to the output at each level. The predicted map is supervised with ground truth to get the binary cross-entropy loss between them. Then back-propagation is used to update the parameter values in order to minimize the loss. The expression is given by:

$$L(Y, \hat{Y}) = -\frac{1}{N} \sum_{b=1}^N Y_b \log(\hat{Y}_b)$$

where  $\hat{Y}_b$  represents the predicted probability,  $Y_b$  represents the ground truth, and N indicates the batch size.

# Chapter 4: Proposed Semi-Supervised Approach Based on LiMAU

In this work an adversarial network is proposed that integrates *II* model of consistency learning [47] which uses perturbations and interpolations on unlabeled data as well as labeled data to enforce consistency in predictions. This approach enhances the robustness of semi-supervised learning algorithms for medical image segmentation. As illustrated in Figure 9, this architecture comprises both segmentation network(SN) functioning as the generator and two evaluation networks(EN's) working as the discriminators within the adversarial architecture. The segmentation network employs the UNet3+ with Multi-Attention model (LiMAU). Demonstrated as the top-performing model in the Experiments section, LiMAU is selected for the segmentation task due to its superior performance.



**Fig. 9.** Adversarial Network

In this work an Adversarial Consistency training strategy is proposed in reference to [50]. Although the approach in reference [50] uses two segmentation networks and two discriminator networks, only one segmentation network is used in this work to reduce the algorithm's complexity

while still achieving comparable results. Two discriminators, identical in structure, are utilized for distinct purposes. The first discriminator assesses the predicted quality consistency of the segmentation network for both labeled and unlabeled data. The second discriminator evaluates the prediction consistency of the segmentation network when given the same inputs under different perturbations.

#### 4.1 Adversarial and Consistency Learning architecture:

LiMAU, the proposed supervised model acts as the Generator(G) of this adversarial based implementation which gives the segmentation probability map. By referring [23] the VGG16 is chosen as the discriminator network which is shown in Figure 10. It is a lightweight structure due to having less parameters.

Discriminator D1 learns to distinguish between the output quality of labeled and unlabeled data, while Discriminator D2 differentiates between perturbed and unperturbed unlabeled data. The combined supervision loss  $L_s$  consistency loss,  $L_{semi}$ , and adversarial losses ( $L_{adv1}$  and  $L_{adv2}$ ) guide the network to produce high-quality segmentation results on unlabeled data. D2 and  $L_{semi}$  complement each other.  $L_{semi}$  ensures pixel-level consistency, focusing on feature map details, while D2 enforces image-level consistency, emphasizing global information. Adversarial consistency learning is achieved through alternate training meaning segmentation network and discriminators are trained in alternate manner. Medical images are input into the segmentation networks to generate prediction maps. These maps, along with the original images, are fed into the discriminators. The discriminators evaluate segmentation quality, with D1 scoring outputs from labeled data as 1 and outputs from unlabeled data outputs, and D2 scoring outputs from unperturbed data as 1 and outputs from perturbed data outputs. During the segmentation network’s training, it is encouraged to produce high-quality results for unlabeled data, aiming for outputs close to 1, while the discriminator networks are trained to effectively distinguish between different inputs.

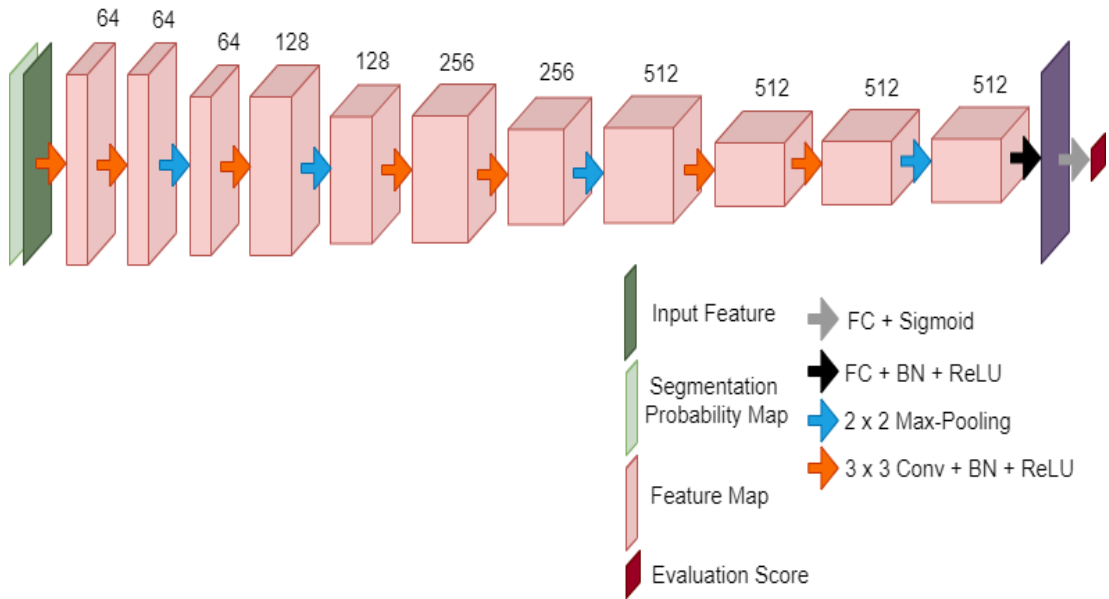


Fig. 10. Discriminator Architecture: VGG16

The input to the discriminator network consists of a unique concatenation of the segmentation result post softmax and the original image, rather than the segmentation result alone. This approach allows for a more thorough evaluation of the segmentation quality by using the original image as a benchmark to judge the alignment between the segmentation results and the ground truth. By referring [23], input for the evaluation network is prepared by combining segmentation probability maps with the input image. It allows EN to evaluate segmentation quality by analyzing their correlation. Element-wise multiplication ensures that both segmentation maps and input images influence EN’s decision-making, enhancing joint training of model parameters. To prevent EN from relying solely on the image’s appearance and to account for lower intensity structures, both the input image and its inverted version are used. For an image channel  $I$  and a probability map  $P$ , we create  $I \cdot P$  and  $(1 - I) \cdot P$ . These maps, combined from all possible pairs of  $I$  and  $P$ , form the input for EN, ensuring comprehensive evaluation.

## 4.2 Optimisation Function:

The objective function of the segmentation network  $G$  and the two discriminator networks  $D1, D2$  is defined as:

$$\min_G \max_{D1, D2} (L_G(\theta) + L_{D1}(\theta) + L_{D2}(\theta))$$

where  $\theta$  represents the parameter to be optimized. The objective function of the segmentation network,  $L_G(\theta)$ , is defined as:

$$L_G(\theta) = L_s(\hat{Y}_l, Y_l) + \lambda(\delta L_s(\hat{Y}_l^p, Y_l) + L_{semi}(\hat{Y}_u, \hat{Y}_u^p) + L_{semi}(\hat{Y}_l, \hat{Y}_l^p) + L_{adv1}(D1(X_u, \hat{Y}_u), 1) + L_{adv2}(D1(X_u^p, \hat{Y}_u^p), 1) + L_{adv2}(D1(X_l^p, \hat{Y}_l^p), 1))$$

Here,  $L_s$  is supervision loss (BCELoss),  $L_{semi}$  is consistency loss calculated using MSELoss and both  $L_{adv1}$  and  $L_{adv2}$  are BCELoss as well. Gaussian noise is introduced to (5-20)% of the total number of pixels present in the input image tensor to introduce random perturbations. Additionally, the brightness factor is varied randomly on a logarithmic scale. These perturbations are implemented to enhance the model’s robustness and generalizability. By simulating real-world conditions, which often include various types of noise and inconsistencies, the model is trained to handle diverse and imperfect data. This approach ensures that the model can effectively process and interpret data that is not always pristine, thereby improving its performance in practical applications.

In the loss function, the terms  $\hat{Y}_l, Y_l, \hat{Y}_l^p, X_l, X_l^p$  represent predicted outputs of labeled data, labeled outputs, predicted outputs of perturbed labeled data, labeled inputs and perturbed labeled inputs respectively. Similarly, the terms  $\hat{Y}_u, Y_u^p, \hat{X}_u, X_u^p$  denote predicted outputs of unlabeled data, predicted outputs of perturbed unlabeled data, unlabeled inputs and perturbed unlabeled inputs respectively.  $\lambda$  is the weighting factor and follows a Gaussian ramp-up curve,  $\lambda = \phi \exp(-5(1 - I/N)^2)$ ,  $\phi$  is a constant value and  $I$  denotes the current number of epoch and  $N$  is the total number of epochs. During the initial phase of network training, the value of  $\lambda$  is very small, so the network updates primarily based on the supervision loss. This means that in the early stages, the training heavily relies on labeled data. As training progresses, the value of  $\lambda$  gradually increases, allowing the network to achieve reliable segmentation results and generate targets for unlabeled data due to the influence of other loss functions. The discriminator networks then work to distinguish the outputs of the segmentation network. The objective functions for discriminators  $D1$  and  $D2$  are defined as follows:

$$L_{D1}(\theta) = \lambda(L_{adv1}(D1(X_u, \hat{Y}_u), 0) + L_{adv1}(D1(X_l, \hat{Y}_l), 1))$$



$$L_{D2}(\theta) = \lambda(\alpha(L_{adv2}(D1(X_u^p, \hat{Y}_u^p), 0) + L_{adv2}(D1(X_u, \hat{Y}_u), 1)) + (1 - \alpha)(L_{adv2}(D1(X_l^p, \hat{Y}_l^p), 0) + L_{adv2}(D1(X_l, \hat{Y}_l), 1)))$$

**Training:** To train the discriminators as well as the generator, both the contribution of labeled data and unlabeled data has been used in consistency loss. While training generator, labeled as well as unlabeled data is perturbed with noise and fed to the segmentation network to produce segmentation maps, which are then compared with segmentation maps for unperturbed data which gives us semi-supervised losses  $L_{semi}$ . For labeled data we get an extra loss term  $L_s(\hat{Y}_l^p, Y_l)$  by supervising segmentation maps of perturbed labeled data with ground truths. While training discriminator 1 we don't need perturbations, D1 compares outputs of unlabeled data with score 0 to train it to give low score for unlabeled data as input and compares outputs of labeled data with score 1 to train it to give high score for labeled data as input. But D2 needs both perturbed labeled data and perturbed unlabeled data and is trained to give score close to 1 to unperturbed outputs while close to 0 to perturbed outputs. The coefficients  $\delta$  and  $\alpha$  are constant values which are found experimentally to be 0.3 and 0.9 respectively.

In summary, the segmentation networks and discriminator networks engage in a strategic interplay one trying to oppose the goal of other. As the discriminators struggle to differentiate between the segmentation result and ground truth, the segmentation networks consistently achieve high-quality segmentation across labeled and unlabeled data, as well as data subjected to various perturbations. This adversarial learning framework proves effective in leveraging unlabeled data to enhance the accuracy of predicted pseudo-labels.

# Chapter 5: Experiments

## 5.1 Dataset

To assess the effectiveness of our network in segmenting liver and liver tumors, we used the MICCAI 2017 Liver Tumor Segmentation Challenge LiTS17 dataset comprising 131 training and 70 test CT scans and the MICCAI SLiver 07 dataset that has 20 training and 10 test CT scans with annotations split in a 1:1 ratio for **training** and **testing** for both datasets. The training set is then again divided into 90:10 ratio for **training set** and **validation set**. For experiments evaluating semi-supervised approach, the train set is again split where 80% is used as unlabeled data and 20% as labeled data. The inputs are in the ".nii" and ".mhd" format, respectively, with a varying number of slices per patient, and each slice is of size  $512 \times 512$ . It provides segmentation labels for liver, tumor, and background, with liver considered as the positive class.

## 5.2 Preprocessing

The ".nii" and ".mhd" files are extracted, and each of the 2D slices is taken as input images to our algorithm. The input images are grey-scale images. Preprocessing involves windowing of a DICOM (Digital Imaging and Communications in Medicine) medical image that optimizes the visualization of liver anatomy in medical images by adjusting display settings. This involves setting the specific window width (WW) at 150 and the window level (WL) at 30 for the liver to enhance contrast and brightness. Then each image is resized from  $512 \times 512$  to  $224 \times 224$  pixels. Subsequently, each image slice is normalized.

## 5.3 Experimental Setup

The experiments are conducted on Google Colab server equipped with a T4 GPUs, using 2560 CUDA cores, 16 GB GDDR6 VRAM and PyTorch 1.7. We employ the Adam optimizer to train the segmentation model as well as the discriminator model, starting with an initial learning rate of 0.001 for the segmentation model, 0.001 for discriminator 1 and 0.0001 for discriminator 2. To compare the performances of different algorithms, Dice Coefficient is used as the evaluation metrics.

## 5.4 Batch-Normalization Comparison

To assess the impact of including BatchNorm layers in the double convolution blocks, we evaluated three models: UNet, UNet++, and UNet3+. Dice Coefficients were calculated for each model both with and without BatchNorm to compare performance. As we can see in Table 1, the Dice Coefficient scores have increased significantly after adding BatchNorm layers. From the results of table we can

study that UNet++ performs better than traditional-UNet while UNet3+ performs better than both for both the datasets. Training set 2 has lesser number of training samples which explain its lower scores compared to Training Set 1. We can also see that training on both combined datasets gives higher Dice Scores than training on individual training sets for every model. UNet3+ with BatchNorm emerges as the best performing model with a Dice value of 95.26% in Test Set 1 and 93.09% in Test Set 2 after training with the LiTS17 dataset, 72.11% in Test Set 1 and 88.62% in Test Set 2 after being trained with the SLiver07 dataset and finally 95.47% in Test Set 1 and 95.09% in Test Set 2 after training is done on combined dataset.

In these results UNet3+ follows the architecture as referred in [12] where during deep supervision all the five outputs from five stages of decoder are being supervised with ground truth which increases time complexity. Hence, the output is averaged out and then compared once with ground truth. This architecture gives almost similar results as shown in Table 2 which is 95.29% on TS1 and 95.01% on TS2 after training on combined dataset which allows to follow the proposed architecture for all the upcoming experiments.

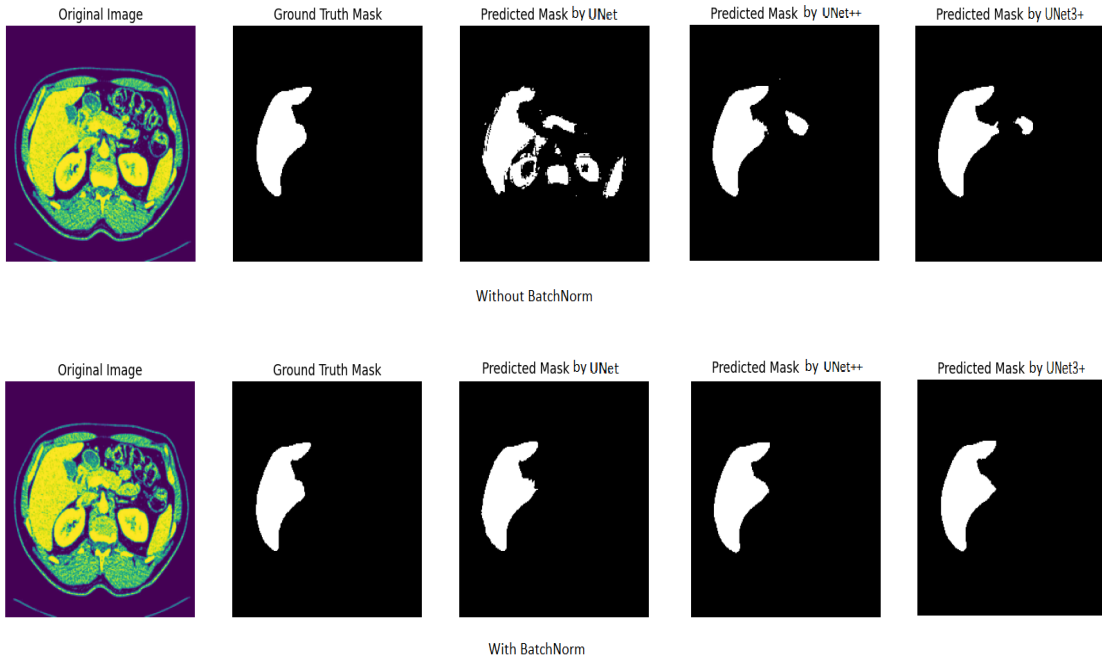
Table 1: **Batch-Normalization Comparative Study**

Experimental Setup	Training Set / Methods	Training Set1 (LiTS17)		Training Set2 (SLiver07)		Training Set1 (LiTS17) $\cup$ Training Set2 (SLiver07)	
		DI(%): Test on TS1	DI(%): Test on TS2	DI(%): Test on TS1	DI(%): Test on TS2	DI(%): Test on TS1	DI(%): Test on TS2
		<b>Without BatchNorm</b>	UNet	83.55	77.21	20.66	39.01
	UNet++	92.43	89.56	69.69	86.17	93.1	90.39
	UNet3+	94.65	92.01	71.37	87.51	94.99	93.99
<b>With BatchNorm</b>	UNet	94.17	91.49	58.77	84	94.75	92.62
	UNet++	94.29	91.93	70.49	88.27	94.88	93.78
	<b>UNet3+</b>	<b>95.26</b>	<b>93.09</b>	<b>72.11</b>	<b>88.62</b>	<b>95.47</b>	<b>95.09</b>

Table 2: **UNet3+ with Modified Deep-Supervision**

Experimental Setup	Training Set Methods	Training Set1 (LiTS17) $\cup$ Training Set2 (SLiver07)	
		DI(%): Test on TS1	DI(%): Test on TS2
	<b>UNet3+ with BatchNorm and Averaging Output</b>	<b>95.29</b>	<b>95.01</b>

Thus the results from Table 1 proves that Batch Normalization (BatchNorm) plays a crucial role in stabilizing the training process of neural networks. By normalizing the outputs of each layer, Batch-Norm ensures that the distribution of features remains consistent throughout training, mitigating issues such as internal co-variate shift, which can hinder convergence and degrade performance. Moreover, BatchNorm accelerates convergence by addressing the vanishing gradient problem, providing more stable gradients that facilitate faster and more consistent training. Overall, Batch-Norm’s ability to stabilize and accelerate training significantly contributes to the effectiveness and efficiency of model. Figure 11 shows the segmentation results of all the models on dataset 2 which follows same pattern as established in the Table 1.



**Fig. 11.** BatchNorm Study Results

## 5.5 Ablation Study

**Supervised Approach:** The ablation experiments for supervised approach are shown in Table 3. Two models UNet++ and UNet3+ are taken and three times experiments are performed for each model, once by incorporating the model with Attention Gates, next by integrating with CBAM and finally including both CBAM and Attention. Since it is already shown Batch-Norm layers are advantageous, in all the next experiments, models with Batch Norm layers are used.

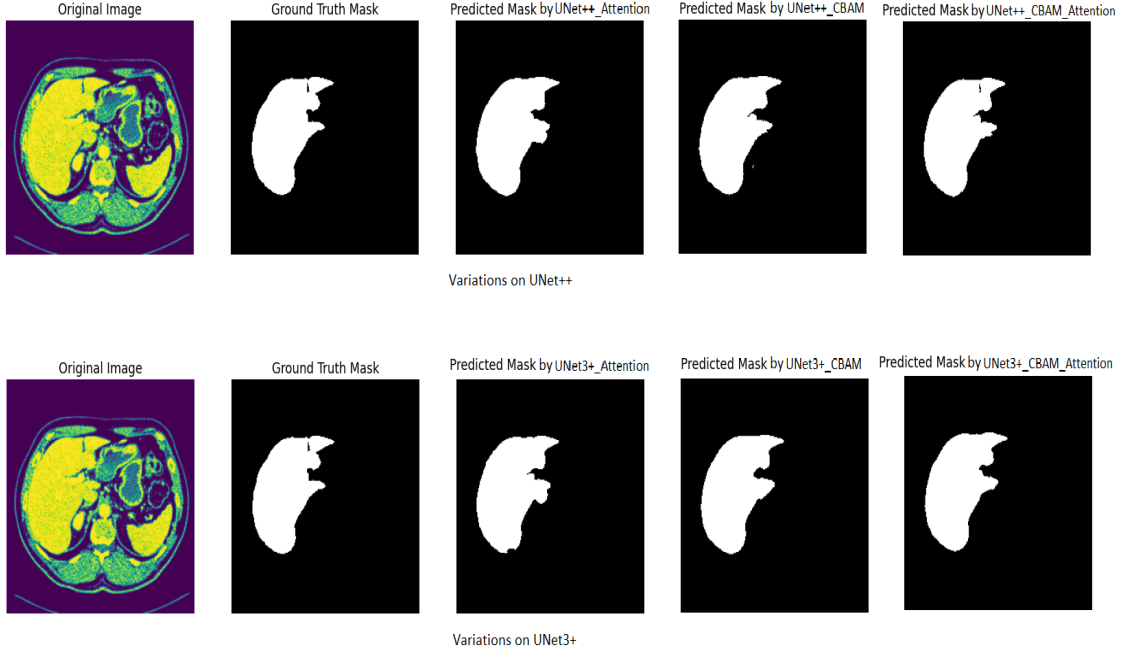
From the results it is clear that combination of both Attention and CBAM works better for both UNet++ and UNet3+ models for both the datasets as well the combination of two datasets. Similarly like the results of Batch-Normalization study, training on LiTS17 dataset proves more effective than SLiver-07 dataset but then again training on both combined datasets gives higher Dice Scores compared to individual training of datasets.

Table 3: Ablation Study for Supervised Methods

Experimental Setup	Training Set / Methods	Training Set1 (LiTS17)		Training Set2 (SLiver07)		Training Set1 (LiTS17) $\cup$ Training Set2 (SLiver07)	
		DI(%): Test on TS1	DI(%): Test on TS2	DI(%): Test on TS1	DI(%): Test on TS2	DI(%): Test on TS1	DI(%): Test on TS2
		<b>With BatchNorm</b>	UNet++ with Attention	94.97	92.83	75.46	88.27
UNet++ with CBAM	95.58		93.95	78.45	88.54	95.47	94.82
UNet++ with Attention and CBAM	95.87		94.26	79.04	90.05	95.69	95.51
UNet 3+ with Attention	95.62		93.47	79.38	89.57	95.39	94.65
UNet 3+ with CBAM	95.64		94.31	81.32	89.22	95.65	95.50
<b>UNet 3+ with Attention and CBAM</b>	<b>95.73</b>		<b>94.72</b>	<b>81.90</b>	<b>92.00</b>	<b>95.78</b>	<b>95.60</b>

Now from this ablation study and comparing different models performances with 2 datasets, it is established that UNet3+ with BatchNorm layer, Attention Gates and CBAM is the top performing model in terms of Dice Scores. It has a high Dice value of 95.78% on Test set1 and 95.60% on Test set2 after training on combined dataset which is higher than other models trained on same dataset. Similarly it gives 95.73% on Test Set1 and 94.72% on Test Set2 after it has been trained on Training Set 1 which is again higher than other models trained on same training set. Similar pattern is followed when model is trained on Training Set2 as well which gives highest scores of 81.90% on Test Set 1 and 92.00% Test Set 2. In Figure 12 the comparative results of all the models on dataset 2 has been shown and expected outcomes are reflected in the same.

Thus, it is proven experimentally that UNet3+ with Attention Gates and CBAM excels in liver tumor segmentation by enhancing multiscale feature fusion and providing superior channel and spatial Attention. This selective focus on critical regions and stabilized training through batch normalization ensures precise localization and high segmentation accuracy. Not only it gives better dice scores, it also has a lesser number of parameters, which is around 27,109,265 trainable parameters.



**Fig. 12.** Ablation Study Results for Supervised Models

It is much lesser than that of next best model which is UNet++ with CBAM and Attention having 38,740,112 parameters.

**Semi-Supervised Approach:** The ablation study results for semi-supervised approach involving four models are discussed here. Table 4 shows the results of experiment. First we follow the traditional semi-supervised approach in reference to the algorithm 1 of [51]. Initially, a segmentation model is trained using a small amount of labeled data. This model is then used to create pseudo segmentation masks for the unlabeled data. These pseudo-labeled masks are combined with the labeled data to update the model. This iterative process continues until a set number of iterations is completed. From the results it is clear it performs poorly compared to other methods. The next method is consistency learning with the  $II$  model architecture [47]. The  $II$  model generates two random augmentations of each sample, applied to both labeled and unlabeled data. During training, the model enforces consistency by minimizing the mean square loss between them, thus ensuring that the outputs of the same unlabeled sample, processed through different random perturbations, remain consistent. This method has shown considerable improvement over the previous method for training on all the two datasets and their combination individually. The third method is Adversarial Learning which is referred from [23]. In this architecture the model comprises two networks: a segmentation network (SN) for segmentation and an evaluation network (EN) for assessing segmentation quality. During training, EN learns to distinguish segmentations from annotated and unannotated images, while SN aims to produce segmentations for unannotated images indistinguishable from those of annotated ones. Through iterative adversarial training, SN learns to improve its segmentation accuracy for both types of images.

Table 4: Ablation Study for Semi-Supervised Methods

Experimental Setup	Training Set / Methods	Training Set1 (LiTS17)		Training Set2 (SLIver07)		Training Set1 (LiTS17) $\cup$ Training Set2 (SLIver07)	
		DI(%): Test on TS1	DI(%): Test on TS2	DI(%): Test on TS1	DI(%): Test on TS2	DI(%): Test on TS1	DI(%): Test on TS2
		<b>With BatchNorm</b>	Traditional Semi-Sup Approach	65.67	49.80	25.71	47.60
<i>H</i> Model of Consistency Learning	82.89		70.55	61.25	73.58	88.81	86.94
Adversarial Learning Approach	91.62		87.65	67.65	85.81	93.12	92.31
<b>Consistency with Adversarial Learning</b>	<b>93.60</b>		<b>88.50</b>	<b>75.90</b>	<b>87.59</b>	<b>94.25</b>	<b>93.22</b>

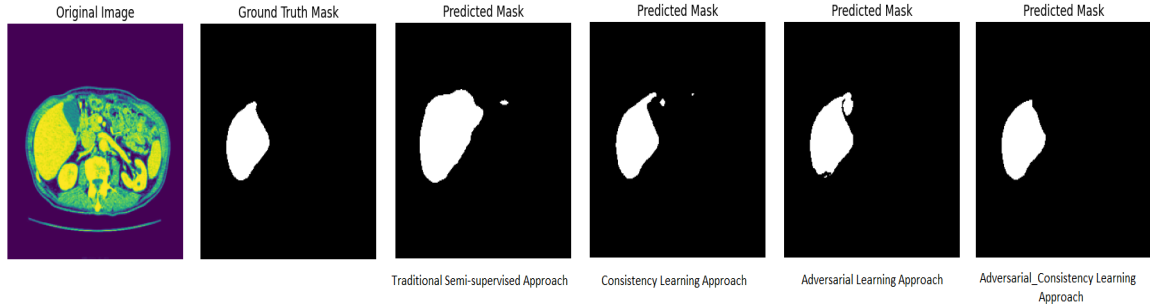


Fig. 13. Ablation Study Results for Semi-Supervised Models

In terms of Dice Metrics of Table 4 we can see performance improved largely by using adversarial networks for all the three cases. Final method is the proposed method that integrates both Adversarial and Consistency Learning. One segmentation network and two discriminators are trained alternately for a certain number of iterations. This architecture with one segmentation network and two discriminators improves segmentation accuracy using limited labeled data by leveraging adversarial training and specialized discriminators to handle both unlabeled and perturbed samples. This model gives the highest Dice Coefficient values compared to the previous state of the art methods and hence comes out as the best performing model. It gives a Dice Coefficient of 93.60% and 88.50% on TS1 and TS2 respectively after being trained on Training Set 1, 75.90% and 87.59% on TS1 and TS2 post training with Training Set 2 and 94.25% and 93.22% on TS1 and TS2 respectively after training is done on combined datasets. In Figure 13 the comparative results of all the

semi-supervised models evaluated on dataset 2 has been presented which shows that the proposed method performs better than others. It gives the output that is closest to the ground truth.

## 5.6 Comparative Performance Analysis

In this section, it is demonstrated how the proposed supervised and semi-supervised methods perform in comparison with other state-of-the-art methods in terms of the Dice Coefficient. In the reference paper [12], UNet3+ is used as the proposed approach. The LiTS Dataset is used for experiments in the paper, with 103 out of 131 volumes used for training and 28 volumes for testing, resulting in a training-to-test ratio of more than 3:1. However, in this work, the dataset is divided approximately into a 1:1 ratio, leading to a significantly smaller number of training samples being used. Despite this, the results are comparable and close to the state-of-the-art methods when the proposed model is tested on 28 samples, even with a much smaller training dataset. As shown in Table 5, the reference paper reports Dice Coefficient values of 96.01%, 96.43%, and 96.75% using three types of loss functions. The proposed approach, LiMAU, achieves a Dice Coefficient of 96.10%, demonstrating similar performance despite the reduced training dataset size.

Similarly, for the semi-supervised approach based on the proposed adversarial consistency network, it is shown that the results are similar to the current state-of-the-art methods for semi-supervised segmentation of liver tumors. The method described in the reference paper [50] is used for comparison, where experiments are conducted on the LiTS dataset with a 12:1 ratio, utilizing 121 out of 131 cases for training and 10 cases for testing. The labeled sample to unlabeled sample ratio is set at 80:20. In this thesis, the semi-supervised model is also trained with an 80:20 labeled to unlabeled sample ratio, but the overall training dataset size is significantly smaller due to an approximate 1:1 split. Nevertheless, the proposed method achieves performance that is very close to the referenced approach. As shown in Table 5, the reference paper documents a Dice Coefficient value of 95.07%. The proposed approach in this work achieves a Dice Coefficient of 94.21%, demonstrating a similar performance even with the much smaller training dataset.

Table 5: **Comparative Performance Study**

Experimental Setup	Training Set / Methods	Training Set1 (LiTS17)
		DI(%): Test on Traing Set with 28 Samples
Supervised Method	UNet3+ (focal loss) [12]	96.01
	UNet3+ (Hybrid loss) [12]	96.43
	UNet3+ (Hybrid loss + CGM) [12]	96.75
	<b>LiMAU:</b> <b>UNet 3+ (CBAM + Attention)</b>	<b>96.10</b>
Experimental Setup	Training Set / Methods	Training Set1 (LiTS17)
		DI(%): Test on Traing Set with 10 Samples
Semi-Supervised Method	ASE-Net [50]	95.07
	<b>Adversarial Consistency Learning</b>	<b>94.21</b>



# Chapter 6: Conclusions and Future Directions

In conclusion, the proposed supervised approach, integrating UNet3+ with Batch Normalization (BatchNorm), Attention Gates, and Convolutional Block Attention Module (CBAM), significantly enhances segmentation performance. This advanced architecture utilizes multi-scale feature fusion and powerful attention mechanisms to accurately localize and delineate target structures within medical images. The incorporation of BatchNorm within the double convolution blocks stabilizes feature distributions, expedites the convergence process, and mitigates internal co-variate shift, ensuring more consistent and reliable training outcomes.

Furthermore, this architecture is augmented with a semi-supervised framework that employs sophisticated adversarial and consistency learning techniques. The framework leverages a single segmentation network in conjunction with two discriminators to optimize performance. One discriminator is tasked with distinguishing between labeled and unlabeled data, effectively guiding the segmentation network to better understand the characteristics of unlabeled data. Another discriminator focuses on differentiating between perturbed and unperturbed data, reinforcing the model's ability to generalize across varying data conditions.

This iterative adversarial training process allows the network to rapidly learn and internalize the relationships between labeled and unlabeled data. By doing so, it maximizes the extraction of useful information from unlabeled data, which is often abundant but underutilized in traditional supervised learning approaches. The adversarial framework, combined with consistency learning, encourages the model to maintain robust performance even when faced with noisy or incomplete data.

By effectively harnessing both labeled and unlabeled data, this comprehensive approach significantly improves segmentation accuracy and robustness. The enhanced model performance has great potential to advance computer-assisted diagnosis and treatment planning in medical imaging. The ability to accurately segment and identify critical structures within medical images can lead to more precise diagnoses, better informed treatment decisions, and overall improvements in patient care. This approach represents a promising step forward in the application of deep learning techniques to the field of medical imaging.

## 6.1 Future Directions:

The proposed framework for liver tumor segmentation demonstrates significant advancements in accuracy and robustness by integrating UNet3+ with BatchNorm, attention gates, and CBAM,

alongside a semi-supervised learning approach using adversarial consistency learning. However, several avenues for future research and improvement remain.

One promising direction is the exploration of diffusion models. Diffusion models have recently shown great potential in generating high-quality images and could be adapted to enhance the segmentation process by generating synthetic training data that closely mimic real medical images. Thus, it can extend the limited dataset available and can introduce new variations of training samples. This approach could further reduce the reliance on extensive annotated datasets and improve the generalizability of the segmentation model.

Incorporating advanced augmentation techniques and leveraging multi-modal data (e.g., CT and MRI fusion) enhances segmentation by improving model generalization to clinical data variations. Multimodal fusion exploits complementary information from different scans, such as MRI's soft tissue contrast and CT's detailed bone structure, boosting segmentation accuracy across diverse clinical contexts and advancing diagnostic precision in medical imaging.

Another promising direction for future research lies in the refinement of adversarial and consistency learning strategies. This could involve dynamic adjustments during training, adapting to the difficulty or uncertainty of samples and predictions. By implementing mechanisms to identify and prioritize challenging cases, the model can focus its learning capacity where it is most needed, potentially improving segmentation accuracy and robustness across varying degrees of complexity in medical imaging datasets.

Finally, expanding this framework to other medical imaging tasks and datasets, such as the BraTS brain tumor dataset, will be crucial to validate its robustness and adaptability. By continuously evolving the architecture and learning strategies, the goal is to create a versatile and highly accurate tool for various medical image segmentation challenges, ultimately contributing to improved diagnostic and treatment outcomes in healthcare.

# References

1. B. Van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," *Radiology*, vol. 261, no. 3, pp. 719-732, 2011.
2. Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in the proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 297-306.
3. R. Zhang, S. Liu, Y. Yu, and G. Li, "Self-supervised correction learning for semi-supervised biomedical image segmentation," in the proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 134-144.
4. 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.
5. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. 330 Hudson Street, New York, NY 10013: Pearson, 2018, ch. 10.3, pp. 742-760.
6. M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," in the International journal of computer vision, vol. 1, no. 4, pp. 321-331, 1988.
7. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in the proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440.
8. 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, 2015, pp. 234-241.
9. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in the proceedings of International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 424-432.
10. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: A nested unet architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3-11. Springer, 2018.
11. X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663-2674, Dec. 2017.
12. Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Jian Wu1, "Unet 3+: a full-scale connected Unet for medical image segmentation," in the proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
13. Chen Li, Yusong Tan, Wei Chen, Xin Luo, Yuanming Gao, Xiaogang Jia, Zhiying Wang, "Attention Unet++: A Nested Attention-Aware U-Net for Liver CT Image Segmentation," in the proceedings of

- 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 345-349
14. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional Block Attention Module," in Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) *Computer Vision – ECCV 2018*. *ECCV 2018. Lecture Notes in Computer Science()*, vol 11211. Springer, Cham.
  15. Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.
  16. A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in the proceedings of 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 1195–1204.
  17. X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, "Transformationconsistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.
  18. K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in the proceedings of *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.
  19. X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in the proceedings of *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.
  20. L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in the proceedings of *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer*, 2019, pp. 605–613.
  21. Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in the proceedings of *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12674–12684.
  22. W. C. Hung, Y. H. Tsai, Y. T. Liou, Y. Y. Lin, and M. H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in the proceedings of *29th Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–12.
  23. Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in the proceedings of *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer*, 2017, pp. 408–416.
  24. G. Chen et al., "MTANS: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation," *NeuroImage*, vol. 244, Dec. 2021, Art. no. 118568.
  25. Y. Zhu et al., "Improving semantic segmentation via self-training," *arXiv:2004.14960*, 2020.
  26. J. Xiang, Z. Li, W. Wang, Q. Xia, and S. Zhang, "Self-ensembling contrastive learning for semi-supervised medical image segmentation," *arXiv:2105.12924*, 2021.
  27. C. Li et al., "Self-ensembling co-training framework for semi-supervised COVID-19 CT segmentation," in *IEEE J. Biomed. Health Informat.*, vol. 25, no. 11, pp. 4140–4151, Nov. 2021.
  28. Li X, Chen H, Qi X, et al. H-DenseUNet: "Hybrid densely connected UNet for liver and liver tumor segmentation from CT volumes[J]". in *arXiv preprint arXiv:1709.07330*, 2017.
  29. M. Moghbel, S. Mashohor, R. Mahmud, and M. I. B. Saripan, "Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography," in *Artificial Intelligence Review*, pp. 1–41, 2017.
  30. Linguraru, M.G., Sandberg, J.K., Li, Z., Pura, J.A., Summers, R.M.: Atlas-Based automated segmentation of spleen and liver using adaptive enhancement estimation. in the proceedings of *MICCAI 2009*, pp 1001-1008, (2009).
  31. Barstugan. R. Ceylan M, Sivri M, et al. Automatic liver segmentation in abdomen CT images using SLIC and AdaBoost algorithms. in the proceedings of *2018 8th international conference on bioscience, biochemistry and bioinformatics. ACM*; 2018, p. 129-13.

32. Ali AH, Hadi EM. Diagnosis of liver tumor from CT images using first order statistical. *Int J Eng Trends Technol (IJETT)* 2015;(20).
33. Chang CC, Chen HH, Chang YC, et al. "Computer-aided diagnosis of liver tumors on computed tomography images" in the proceedings of *Comput Methods Programs Biomed* 2017;(145):45–51.
34. Bi, Lei, Jinman Kim, Ashnil Kumar, and David Dagan Feng Feng. "Automatic Liver Lesion Detection Using Cascaded Deep Residual Networks." in *arXiv:1704.02703* (2017).
35. Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." *arXiv-prints* (2016).
36. J. G. Fernández, V. Fortunati and S. Mehrkanoon, "Exploring automatic liver tumor segmentation using deep learning," in the proceedings of 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8.
37. X. Han, "Automatic liver lesion segmentation using a deep convolutional neural network method," *arXiv preprint arXiv:1704.07239*, 2017.
38. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," in *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856-1867, June 2020
39. O.O et al., "Attention u-net: Learning where to look for the pancreas," in *arXiv:1804.03999 on Medical Imaging with Deep Learning*, 2018.
40. V. Badrinarayanan, I. Budvytis, and R. Cipolla, "Semisupervised video segmentation using tree structured graphical models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35 11, pp. 2751–64, 2013.
41. F. Bergamasco, A. Albarelli, and A. Torsello, "A graphbased technique for semi-supervised segmentation of 3d surfaces," *Pattern Recognition Letters*, vol. 33, pp. 2057–2064, 2012.
42. D. Mahapatra, P. J. Schuffler, J. A. W. Tielbeek, F. Vos, and J. M. Buhmann, "Semi-supervised and active learning for automatic segmentation of crohn's disease," the proceedings of *MICCAI*, vol. 16 Pt 2, pp. 214–21, 2013.
43. D. ping Tian, "Semi-supervised learning for refining image annotation based on random walk model," *Knowledge-Based Systems*, vol. 72, pp. 72–80, 2014.
44. Lee, Dong-Hyun, "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," in *ICML 2013 Workshop: Challenges in Representation Learning (WREPL)*
45. Thomas N. Kipf, Max Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *arXiv:1609.02907, ICLR 2017*
46. Blum, Avrim, Mitchell, Tom. (2000). Combining Labeled and Unlabeled Data with Co-Training. in the proceedings of the Annual ACM Conference on Computational Learning Theory.
47. Laine, Samuli and Timo Aila. "Temporal Ensembling for Semi-Supervised Learning." *ArXiv abs/1610.02242* (2016): n. pag.
48. Zhang, Gaowei, Pan, Yue, Zhang, Limao. (2021). Semi-supervised learning with GAN for automatic defect detection from images. *Automation in Construction*. 128. 103764. 10.1016/j.autcon.2021.103764.
49. Hung, Tsai, Liou, Lin, Yang, "Adversarial Learning for Semi-Supervised Semantic Segmentation," in *arXiv:1802.07934*, 24 Jul 2018.
50. T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan and A. K. Nandi, "Semi-Supervised Medical Image Segmentation Using Adversarial Consistency Learning and Dynamic Convolution Network," in *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1265-1277, May 2023.
51. Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, Cheng Jin, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," in *Computers in Biology and Medicine*, Volume 169, 2024