ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems
Vol. 4, No. 2, 2026, pages 139 - 155
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

Adaptive Attention Fusion Network for Automated Liver and Tumor Segmentation in CT Images

Rabia Bibi, Arfan Jaffar, Muhammad Imran Ali, Zeeshan Naeem, Usman Mohyud Din Chaudhary
[1]Faculty of Computer Science and Information Technology, Superior University Lahore
[2]Faculty of Computer Science and Information Technology, Superior University Lahore
[3]Faculty of Computer Science and Information Technology, University of Southern Punjab, Multan, Pakistan
[4]Faculty of Computer Science and Information Technology, University of Southern Punjab, Multan, Pakistan
[5]Faculty of Computing & Emerging Technologies, Emerson University, Multan

ARTICLE INFO

Article History:

Received: May 30, 2026
Revised: June 20, 2026
Accepted: June 25, 2026
Available Online: June 30, 2026

Keywords: Liver Segmentation, Tumor Segmentation, CT Imaging, Deep Learning, Attention Mechanism, AAF-Net, Medical Image Analysis.

Classification Codes:

Funding: N/A

ABSTRACT

Accurate segmentation of liver parenchyma and hepatic tumors from computed tomography (CT) images is important for clinical applications such as surgical planning and radiation therapy. However, automated liver and tumor segmentation remains challenging because of severe class imbalance, small tumor regions, unclear boundaries, and variations in tumor size, shape, and morphology. Conventional encoder-decoder networks commonly combine shallow encoder features and deep decoder features through fixed skip connections, which may introduce irrelevant background information and reduce boundary representation. To address this issue, this study proposes an Adaptive Attention Fusion Network (AAF-Net) incorporating an Adaptive Attention Fusion Module (AAFM). The proposed module dynamically evaluates the contextual relevance of encoder and decoder features at each spatial location and adaptively combines them through a learnable attention mechanism. The empirical evaluation was conducted using the 3D-IRCADb-01 dataset, comprising 2,823 axial slices from 20 patient volumes. Patient-level data splitting was used to reduce data leakage. The methodological pipeline included liver-specific Hounsfield Unit (HU) windowing, intensity normalization, spatial resampling, stochastic geometric augmentation, and a composite loss consisting of weighted categorical cross-entropy and multi-class Dice loss. AAF-Net was evaluated against U-Net, Attention U-Net, and DeepLabV3+ using Dice, Intersection over Union (IoU), Precision, and Recall on a fixed test set and through 3-fold patient-level cross-validation. On the fixed test set, AAF-Net achieved a liver Dice score of 0.8126, IoU of 0.7589, Precision of 0.7894, and Recall of 0.9597. For hepatic tumor segmentation, it achieved a Dice score of 0.7048, IoU of 0.6841, Precision of 0.7621, and Recall of 0.9052. Although U-Net and Attention U-Net obtained higher tumor Dice scores on the fixed test split, AAF-Net achieved a higher mean tumor Dice of 0.7639 compared with 0.7519 for U-Net in 3-fold cross-validation. The ablation study showed that adaptive AAFM improved tumor Dice from 0.7279 to 0.7992. Overall, the findings demonstrate that adaptive feature fusion provides an alternative to fixed skip-connection fusion for liver and hepatic tumor segmentation under the evaluated experimental conditions.

Corresponding Author's Email: imranali.mltn@gmail.com

1. Introduction

Medical image analysis has evolved to be an essential part of today's clinical radiology, allowing non invasive visualization of the human body's internal structures for diagnosis, treatment scheming and monitoring therapy [1]. In the field of medical imaging methods, computed tomography (CT) is highly used for abdomen investigation because of its high spatial resolution, fast acquisition time and strong ability to distinguish the anatomical structure by using the radiodensity of anatomical tissues. Within the clinical practice of oncology, computed tomography (CT) imaging is the most commonly used technique to diagnose, stage and plan the management of liver cancers, such as primary liver cancers and metastatic liver cancers [3]. High-accuracy delineation of liver and intrahepatic tumor volume from computed tomography (CT) data sets is essential for many clinical applications, such as estimating pre-major hepatectomy liver volume, planning surgical resection margin, radiofrequency ablation (RFA), transarterial chemoembolization (TACE), and radiation therapy targeting. Traditionally, delineation of organ contours and pathological lesions were drawn manually by expert radiologists, slice-by-slice. But manual segmentation is very time consuming, labour intensive and has a huge amount of intra- and inter- observer variability. A standard 3D (axial) computed tomography (CT) scan of the abdomen has dozens to hundreds of slices and manually annotating the fine anatomical boundaries across numerous patients will add significant clinical burden to operational latency. Multiple arterial phases are needed to properly evaluate the liver. To evaluate the liver, multiple arterial phases are required, and creating fine anatomical boundaries manually across dozens to hundreds of patients will create a significant amount of operational delay and burden [4]. Studies have been carried out to allow automated and semi-automated computer-assisted diagnosis (CAD) systems to be produced to address some of the limitations of manual delineation. The first automated segmentation methods were based on conventional computer vision and image processing algorithms, including intensity thresholding, region growing, active contour model and statistical shape models. While these classical techniques laid the groundwork for medical image analysis, such techniques are less effective when applied to more complicated clinical scans. Most conventional algorithms fail to perform well in low contrast conditions between neighbouring soft tissues, due to non-uniform intensity 3 fluctuations from image noise or beam hardening, and from the extreme diversity of liver lesions' morphology. Medical image segmentation has undergone great changes in recent years with the introduction of deep learning, especially of Fully Convolutional Networks (FCNs), and Convolutional Neural Network (CNN) architectures [51]. In deep learning, features can be hierarchically extracted automatically from raw image data without any special feature engineering tasks. The U-Net architecture has set the benchmark for the volumetric medical image segmentation task, with a symmetric decoder-encoder paradigm. U-Net and its variations have leveraged the ability to connect contracting encoder paths to form a high-level gradient for multi-scale feature learning, then expanding decoder paths for recovering spatial resolution to be used to achieve a high level of segmentation accuracy in a wide range of anatomical structures [6, 7]. Unfortunately, computer vision problems of liver/iliherpatic tumor accurate segmentation in CT volumes are still a hard problem in spite of all these advancements. Anatomical and pathological variation are the major obstacles in abdominal scans. The intensity level is relatively homogeneous and the geometry is coherent in the healthy liver parenchyma, while both pathological livers are characterized by changed tissue density and irregular outer liver boundary, especially in the case of cirrhosis, and liver deformation due to mass effect of large tumors. Moreover, hepatic tumors vary greatly in size, shape, location, number and contrast enhancement, from a small hypodense focal tumor to massive, hypervascular hepatocellular carcinoma [1]. Because standard deep learning models often provide only a moderate level of accuracy for both tumors and normal liver tissue, this approach is often unable to achieve high segmentation accuracy in both cases. Traditional encoder-decoder methods use shallow feature maps with high spatial resolutions and deep feature maps with high semantic levels by simply means of concatenation or addition operations in the skip connections. This immature feature fusion method assumes that all spatial regions and channel dimensions have equal contribution, which frequently leads to semantic discrepancy, spatial noise and irrelevant background noise at the decoding path [13,]. This results in a poor prognosis of a conventional model, because of boundary blurriness and false positive prediction of neighboring organs with similar values of Hounsfield Unit (HU), and false negative prediction of tumor boundaries located at low contrast [1]. In order to deal with these challenges, more recent studies examined dynamic attention mechanisms and a paradigm of adaptation to the feature fusion in Deep Neural Networks. By adding attention mechanisms, networks have the ability to adaptively adjust their feature representation, focusing more on the 4 anatomically interesting parts of the brain while neglecting irrelevant background white matter signals. Based on these concepts, the Adaptive Attention Fusion Network (AAF-Net) is proposed in this thesis [14], designed to adaptively augment feature representation via dynamically adaptive multi-class skip connections and channel-aware attention at the point of each pixel in abdominal computed tomography (CT) scans, resulting in more detailed liver and tumor segmentation.

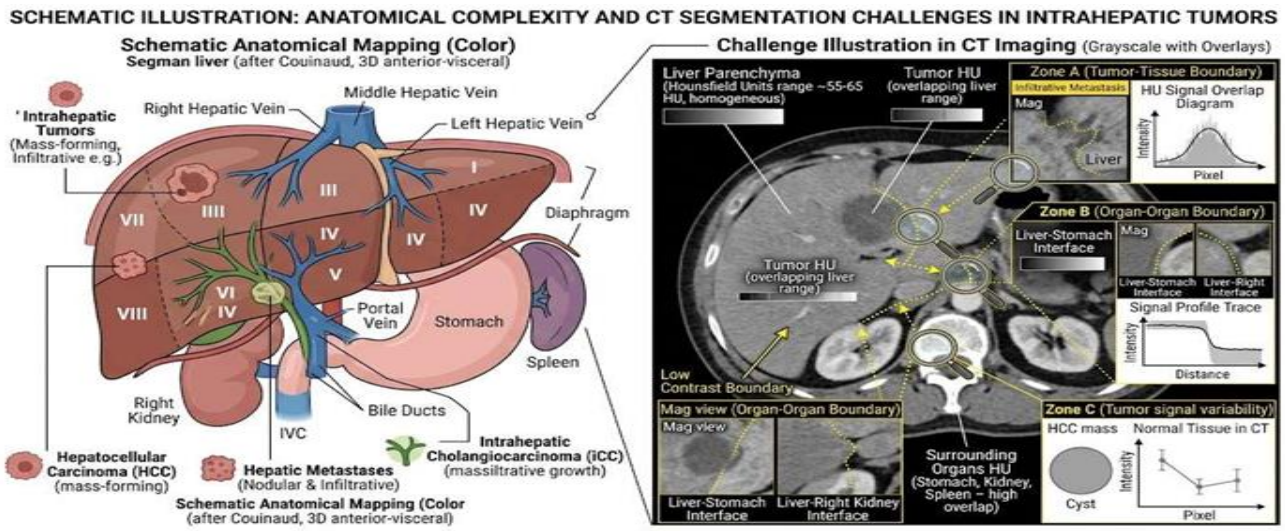

Figure 1: Schematic illustration of liver anatomy and intrahepatic tumors in abdominal CT imaging, highlighting low-contrast boundaries.

Medical image segmentation is the division of a digital medical scan, like a CT scan, Magnetic Resonance Imaging (MRI) scan, or Ultrasound scan, into non-overlapping, anatomically meaningful regions. In the strict sense, medical image segmentation is the task of partitioning medical images, given an input image domain $\Omega \subset \mathbb{R}d$ where d = 2 for planar slices and d = 3 for volumetric images, into a specific number of anatomical or pathological classes with $c \in \{0,1,2,\ldots,C-1\}$ assigned to each spatial coordinate $\mathbf{x} \in \Omega$, with C being the number of target classes including background. 5 Medical image segmentation is performed under severe imaging constraints, as opposed to natural image segmentation, where objects are expected to have well-defined borders, structured lighting and distinct colors. Medical scans have inherent acquisition noise, partial volume artifacts, motion artifacts and a small dynamic range of intensity contrast between adjacent soft tissues. Efficient modeling of the combination of localized spatial information and high-level contextual information is necessary for good medical image segmentation, thus tackling ambiguous image boundaries [6].

2. Related Work

Accurate and automatic segmentation of the liver and hepatic cancer within computed tomography (CT) images is essential in contemporary oncologic diagnostic and treatment planning and supervision of the disease. In the last decade, medical image analysis has experienced a transformation from semi-automated to deep learning-based, very complex, fully automated methods [1]. The recent decades have witnessed the rapid progress of deep learning algorithms for liver tumor segmentation, which has been closely driven by the demands for solving complex clinical problems as reported in comprehensive surveys [1, 2] or as expressed in more recent papers [4]. Some of these challenges are extreme variability among tumors in both size and shape, vague tumor boundaries with healthy tissue, and the extreme class imbalance in medical scans, in which tumors make up an extremely small fraction of the entire volume [5]. The literature review done here can be seen to have a clear history, spanning from old computer vision algorithms to sophisticated attention-driven neural algorithms. As shown in the recent bibliographic analysis, which includes a systematic review in the Journal of Medical Systems published in 2024 [5] that specifically focuses on the application of CNN, especially the encoder-decoder systems such as U-Net [11], have revolutionized the field [6, 8]. Moreover, the fusion of 'attention mechanisms' and transformerbased models has recently reached new records in the field of capturing complex contexts [30, 31]. This study discusses critically the evolution of liver and tumor segmentation techniques. It discusses traditional methods, the basics of deep learning models, deep attention models, and feature fusion methods in a systematic way. Furthermore, it outlines public datasets and evaluation metrics that underlie comparative research, reaching a conclusion that identifies the issues in research that motivated the proposed framework, Adaptive Attention Fusion Network [38]. Liver and tumor segmentation as a whole has drawn a path of decreased human involvement and higher complexity with computation. The evolution can be divided into techniques and systems based on traditional algorithms, Machine Learning (ML), and Deep Learning.

Traditional image processing methods have been used to extract hepatic structures from computed tomography images as a starting point before the new development of high-performance computing and big data. Early works mostly made use of gray-level features

and functional priors, as recorded in a comprehensive review in Artificial Intelligence Review [4] in 2025. Intensity thresholding was used to isolate the liver tissue from other organs in the abdomen by using the Hounsfield Unit values. Frequently, we encountered intense over-segmentation due to the similar intensity profile of neighboring organs such as the spleen and stomach [4]. Many region growing algorithms were used to further determine these boundaries. These were performed by hand by a clinician on a point from either the liver or the tumor and then the algorithm iteratively added neighboring pixels meeting a predefined homogeneity criterion [4]. Although it worked better than simple thresholding, region growing was quite dependent on the initial placement of the seed and tended to fall short at boundaries around tumors because of the lack of contrast at these regions, a problem known as the partial volume effect. A different approach was tried by using mathematical algorithms known as Active contour models, or snakes, which minimized an energy function to pull the curve towards the edges of the objects [4]. These methods, while mathematically elegant, were both time-consuming and often required manual initiation and were lack of strength to manage the diverse characteristics of hepatic tumors [3]. In order to alleviate the shortcomings of purely heuristic algorithms, researchers used embedded handstructured features extracted from traditional machine learning algorithm techniques [4]. In these techniques, specific feature extractors will be created by a human expert to represent aspects of liver tissue and liver lesions in terms of either local binary pattern for texture or histogram of oriented gradients for shape or statistical variance measures.

Classical classifiers were then used to classify these handcrafted features. In high-dimensional feature spaces, SVMs were often used to build the optimal separating hyperplanes between the tumor pixels and healthy tissue [4]. Random Forests based on ensembles of decision trees were shown to be more robust to the noise of medical imaging [3]. These machine learning methods minimized human effort in the segmentation process, but they had one big challenge—the feature engineering process. Extracted features were to some extent limited in terms of what was known about the human body and did not always take advantage of the fact that the spatial changes of interest (the hepatocellular carcinoma) are high dimensional, highly complex, non-linear, and hierarchical in nature and present at lower contrast levels. As described in the Data Mining and Information Security review of 2025 by Dutta et al. [3], all of these models failed due to their limited ability to generalise over various protocols and patient groups, leading to a shift towards data-driven features.

The major drawback of handcrafted features has finally been addressed by the deployment of deep learning technology, namely the convolutional neural network [10]. In another study entitled Neurocomputing for data-efficient deep learning liver segmentation [2] in 2026, it is explained how neural networks learn the hierarchical representations directly from the raw data of the pixels. In this paradigm, low-level information, like the outline, edges and textures, is captured by the initial levels of the network, while its deeper layers create high-level correspondences that represent highly complex anatomical and functional structures [6, 10]. In the medical image segmentation field, deep learning methods have completely replaced classical machine learning methods [6]. These models can compute the entire volumetric scan fully and implicitly learn the spatial context of the liver in relation to the ribs, spine and other abdominal organs [12]. Moreover, state-of-the-art deep learning networks are capable of performing multiclass segmentation of the background and healthy liver, while also segmenting various regions of the tumor, into a single forward pass [42, 43]. Specific convolutional architectures are developed to optimally fit the limitations of medical imaging and have been shown to play an important role in the success of this transition [11, 15]. Very recent advances in automated medical image analysis are enabled by the use of Convolutional Neural Networks (CNN) [6, 10]. They are very well suited to processing data with a grid structure like computed tomography scans [11] since they share weights and use a receptive field that is local. The Fully Convolutional Network is a breakthrough in dense prediction problems [51]. Before its introduction, it was mainly used at the end with fully connected layers with a single classification to the end of the image [51]. Fully Convolutional Networks eliminated these dense layers and substituted them with convolution layers to get a spatial heat map of classification scores [51]. These networks facilitated pixel-wise classification by transposing the feature maps to have the same resolution as the input image [51]. Early developments, however, would have coarse segmentation output due to the subsequent pooling stage needed to extract the features, in which the high-frequency edge information would be lost [51]. Directly propelling more complex encoder-decoder designs, particularly for biomedical signals, was precisely this lack of fine-grained spatial resolution, a feature of preserving the quality of the original signal [11]. However, few architectures have such an outstanding impact in medical image segmentation as the U-Net of Ronneberger et al., 2015 [11]. It was observed that the U-Net architecture revolutionized the method for biomedical image analysis [6] as per the analysis study published in the said paper in the International Journal of Network Dynamics and Intelligence in 2024 [8, 9]. The model is symmetric to capture the context in the contracting path (encoder), and precise localization is facilitated by the expanding path (decoder) [11]. To obtain a good grasp of the genius of the U-Net, one has to begin by understanding the skip connections [11]. These connections skip the deepest and most compact layers of the network and directly concatenate a pair of high-resolution feature maps in the encoder and upsampled feature maps in the decoder corresponding to the same layers [11]. This fusion enables the combination of the deep and coarse semantic information and shallow and fine spatial

information, thereby providing the final segmentation boundaries with sharp edges and anatomically precise information [11, 13]. After the success of the 2-D U-Net, quite a few 3-D U-Nets came along. The 3D U-Net was introduced by Cicek et al. in 2016, in which the 3-dimensional convolution layers, max pooling operation, and upconvolution layers replaced the traditional 1-dimensional operations [12]. This enabled the network to capture the volumetric representation of densities from sparsely annotated slices while taking advantage of the inter-slice spatial information that is fundamental in the case of a three-dimensional organ such as the liver [12]. In 2018, Zhou et al. further expanded the paradigm using UNet++ [13] with a nested structure that contains dense skip pathways between the encoder and the decoder, thereby sacrificing a lot of computational complexity [13]. A typical skip connection in the U-Net makes the propagation of spatial information pass through the early encoder layers, but retains all irrelevant background noise. Considering this, in 2018, Oktay et al. added Attention U-Net [14]. In this architecture, the basic structure of skip connections is modified with the addition of attention gates [14]. The attention gates have the coarse-grained high-level features from the decoder used as a gating signal to evaluate whether the fine-grained high-level features coming from the encoder are relevant [14]. The network automatically downweights the activity in irrelevant brain areas (e.g., regions around the abdomen) and focuses on relevant features related to the liver/tumor by element-wise multiplying a learned attention coefficient map [14]. This process enables the network to pay attention to the targeted area without the necessity of any explicit regional cropping, which leads to a significant improvement of sensitivity and precision of the model, especially with respect to small, hidden hepatic lesions in particular [14, 27]. One of the most important problems with deep U-Net model training is the degradation problem in which, after a few layers, the gradients of backpropagation vanish or grow very big. Researchers adopted the residual learning block proposed by He et al. in 2016 [16] and added it to the U-Net topology [18] to overcome this. This leads to an overall ResUNet structure, which consists of replacing common convolutional blocks with residual blocks [16]. A residual block is used in a shortcut connection which can have one or more convolutional layers in the way between the block that skips or bypasses and performs an identity mapping, which is added to the output of the convolutional operations [16]. This implies that only a residual mapping has to be learnt instead of the sub-function itself [17]. In this context, ResUNet allows the development of much deeper networks while enabling gradient flow to any other level without any impediment, enabling the detection of very intricate, powerful features that better indicate the presence of heterogeneous liver tumors, and converging much faster during the optimization phases [18, 20]. A comparative study between multi-phase contrast-enhanced computed tomography [19] demonstrated that multi-phase models based on ResUNet are always more robust even in the presence of image noise. Although convolutional operations are successful, they have a built-in disadvantage, since their receptive fields are local and cannot model long-range dependencies in the data [30, 31]. To achieve global context, Chen et al. proposed using a combination of the localized feature extraction capability of Convolutional Neural Networks and the global relationship modeling capability of Transformer networks in 2021 (TransUNet, [34]). TransUNet builds on the concept of the Vision Transformer developed by Dosovitskiy et al. [31] and uses a convnet to obtain a high-resolution feature map [34]. This map is then flattened into a series of patches and passed to a regular Transformer encoder [34]. Within an attention mechanism, the self-attention process in the Transformer is able to capture the relationship of all the patches at once, thereby providing an explicit global understanding of the entire abdominal cavity [30, 34]. The output sequence is then reshaped into a spatial grid and fed into a regular CNN decoder [34] with skip connections from the first few convolutional layers. In the medical imaging domain, it has proven to be state-of-the-art, combining a transformer for modeling global context and convolutions for precise localization [33,34]. Similarly, some models solely based on transformers have been developed such as Swin-Unet [33] (Cao et al., 2022) which has now shown that transformers are highly capable encoders for medical segmentation [32, 33, 40].

To this end, they noted that although architectural changes may sometimes be useful, they are less effective than appropriate hyperparameter optimization, a claim supported by U-Net architecture shoreline segmentation benchmarking that showed that nnU-Net produced the best performance with state-of-theart accuracy [15] in 2021. The "no-new-Net" (non-extensible networks) framework states that, when the overall pipeline is optimized to the max, the basic U-Net is enough to solve the problem [15]. The first approach, called nnU-Net (self-configuring neural networks), learns all its pre-processing steps (resampling, normalization), network structure (patch size, batch size and depth) and post-processing (scoring) automatically from the statistical properties of its target dataset [15]. Across many biomedical segmentation tasks, nnU-Net has been able to simplify the system by casting away the manual use of heuristics and is able to out-compete very complex and novel systems [15, 43]. The success of this project highlights the need for domain-specific handling of the data; for example, the use of the correct windowing of Hounsfield Units, as well as anisotropic spacing resolution, and rich data engineering is essential to a successful architecture rather than simply a benefit of it [15]. Medical segmentation networks deepened, and the need to explicitly pay attention to important anatomical structures rose to the fore [21]. These are akin to human visual attention, and are used to dynamically weight features according to their relevance to the task at hand [22]. Various attention mechanism modules have been developed in various domains as discussed in

a survey of attention mechanisms for pattern recognition in 2025 [21], which can be broadly classified under three categories: channel-attention, spatial-attention, and self-attention [22, 23, 24]. In deep neural networks, every channel in a feature map is a specialized detector for a certain pattern, in this case, for horizontal edges, circular textures or certain intensity gradients [21, 23]. But not all the found patterns are the same for tumor segmentation. The goal of the channel attention mechanism is to explicitly capture the interdependencies among various channels [23]. It works generally by going through a global average pooling to reduce the spatial size of each channel to create a global feature descriptor for each channel [23]. These descriptors are used by a small multi-layer perceptron which produces an attention weight vector [23]. This vector is then multiplied pointwise to the original feature map, "recalibrating" the network [23]. It enhances the passage for data with the most discriminative tumor features and diminishes the passage for redundant or background noise data [22]. This allows the network to consume its computation resources in the most useful semantic representation [21, 27]. Channel attention indicates what information is relevant, while spatial attention indicates the location of the relevant information in the image [22]. If the liver and the tumors to be observed have particular coordinates in the computed tomography (CT) scan, it is very useful to direct the reflexes of the network to them [23]. A usual method for obtaining spatial maps with a spatial attention mechanism is to perform a pooling operation with respect to the channel axis; this yields two spatial maps: maximum activations and average activations [23]. All these maps are concatenated and then come through a standard convolutional layer for obtaining a spatial attention mask, normalized through a sigmoid function [23, 24]. As it is shown with the usage of a Masked Feature Map in frameworks such as MANet for liver tumor segmentation [27] (Scientific Reports, 2023), multiplying this spatial mask with the feature map will give rise to a dynamic spatial filter [26, 27]. It enhances the signals in the hepatic regions and suppresses the signals from other organs such as lungs or intestine drastically, so that false prediction on organ boundaries is reduced [25, 27]. Originally developed by the seminal works of Vaswani et al. in NLP [30] in 2017 and later extended to computer vision [31], self-attention is now a key method in sequence modeling. Self-attentions is a mechanism that treats the spatial position of the response as a variable and applies a weight to the features at each spatial position of the image unlike localized convolutional filters [30, 31]. It consists of three steps of converting the input features into three different representations: Query matrix, Key matrix and Value matrix [30]. The algorithm calculates the Dot product for a particular Query and for all other Pixels, and creates an attention map quantifying the similarity or relationship between all pairs of pixels [30]. The map determines the amount of contribution of each pixel to the constructed representation of the desired pixel [30]. The self-attention guarantees that the network can recognize longrange anatomical dependency, where the classification of a single pixel does not only depend on the information from its local structure, but also the global context of the entire abdominal scan as applied in advanced models such as Semi-Wavelet Attention Module U-Net [28] (The Biomedical Signal Processing And Control, 2025), and various Transformer-based architectures [32, 34, 35, 36, 37]. The high-level abstract features stored in the encoder do not well align to the low-level spatial information requested by the decoder, which causes the need for a complex feature fusion algorithm to be used for accurate segmentation mask recovery [11].

Amongst the most basic fusion techniques, introduced by U-Net [11] is the direct concatenation of encoder feature map and decoder feature map [11] which is quite popular. In this method, important information in high resolution captured from the contracting path is cropped and then joined to the upsampled information from the expansive path along the channel axis [11]. These combined channels are then passed to a second convolution layer to combine them [11]. Direct concatenation is an easy but static procedure [11] that is very effective at recovering lost spatial information during pooling. This makes the network process the raw, unrefined background noises generated by the encoder along with the semantically rich decoder features – often resulting in artifacts along the edges of liver tumors [13, 14].

Hepatic tumors vary greatly in physical size, from the small nodules which are only a few cells wide, to the large ones that may take up an entire liver lobe, and therefore require simultaneous processing of 30 information at multiple scales within network processes [50]. Multi-scale fusion techniques overcome this by taking features from multiple layers of the network [50]. Such architectures can be found in recent papers and include similar modules such as Atrous Spatial Pyramid Pooling [50] (Knowledge Based Systems, 2025). The network samples the input with multiple different receptive fields by using parallel convolutional layers with different dilation rates [50]. These multi-scale representations are then combined to give a network bordering the boundaries of small lesions and still detecting the bulk structure of large, sprawling tumors despite not needing to increase network depth by as much. The multi-scale representations are then fused together; the network remains equally robust at capturing the fine boundaries of small lesions, without a correspondingly large increase in network depth to capture the bulk structure of large, sprawling tumors.

Adaptive fusion [38] is the most advanced approach toward closing the decoder-encoder gap. Instead of combining the features in a static fashion or simply by adding together what comes out of each source, adaptive fusion makes use of learned parameters that learn to combine the contribution of different feature sources in the current contents of the input image [38]. In an adaptive

mechanism, like the proposed Adaptive Attention Fusion Module [38] the network computes a pixel-wise gating coefficient based on the features received from the encoder and decoder. This coefficient controls the actual spatiotemporal information to accept from the encoder and/or the semantic information to accept inside the decoder at every single pixel [38]. On boundaries of complex tumors, the gate may be very sensitive to the high-resolution encoder features because a sharp boundary representation is crucial. When a pixel is contained on a complex tumor boundary, the gate may be sensitive to the high-resolution encoder features to ensure a sharp plot of the boundary [38]. It is also possible that the region of the liver that serves as the decoder region will choose to exhibit the decoder characteristics in areas of liver that are mostly uniform, deep within [38, 39]. In contrast to the static skip connections, this dynamic blending can be regarded as an active suppression of semantic ambiguity [21, 38] and takes place with context awareness. The widespread and ready access to large and well-annotated, public benchmark datasets [42, 43, 44] has naturally accompanied the significant progress in deep learning architectures. These datasets are used to measure each methodology in a standardized way that can be objectively compared [43].

The multi-modal imaging extension is the combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) challenge dataset [43] which extends segmentation activity to multi-modal imaging. At the outset, however, CHAOS is limited to healthy organs, the aim being to obtain a baseline anatomical segmentation; actually, it provides computed tomography and magnetic resonance imaging scans [43]. When testing models on the CHAOS benchmark or on subsets of CHAOS generated from the Medical Segmentation Decathlon [43] (Antonelli et al., 2022), a network's generalizability and its capability of transferring learned anatomical representations to completely different imaging physics and imaging contrast profiles can be assessed, which is an important factor in its versatile clinical application [43, 47, 48 , 49].

One of the most important and widely discussed benchmarks is the '3D Image Reconstruction for comparison of Algorithm Database' (3D-IRCADb) which was developed by Soler et al. in 2010 [41]. The 3D-IRCADb-01 dataset is also a set of computed tomography scans of twenty different patients [41] processed specifically for the purposes. Most importantly, it offers highly detailed, expert-approved liver and main tumor manual segmentation, as well as venous systems and internal hepatic areas [41]. The collection intentionally involves very large tumors, with significant disruption of normal liver anatomy, as well as very small, poorly-contrasted lesions which are near the absolute limit of spatial resolution [41]. 3D-IRCADb is a good test-bed for learning more sophisticated variants of attention [45; 46] because of its carefully curated annotations and extreme pathological variance. Specific statistical metrics are then applied to measure the performance of medical image segmentation networks with respect to human expert annotations in order to measure and quantify their performance in a rigorous and objective way [40, 42]. Measuring these metrics will gauge the extent of the overlap, precision and diagnostic safety of the predictions generated [40].

Table 1: Comparison of Existing Liver and Tumor Segmentation Research

Author/Year	Dataset	Method	Dice (Liver/Tumor)	Advantage	Limitation
Ronneberger et al. (2015) [11]	ISBI	U-Net	N/A (Baseline)	Established the standard encoder-decoder structure; excellent spatial recovery	Propagates noisy background features through static skip connections
Milletari et al. (2016) [40]	PROMISE12	V-Net	N/A (Baseline)	Introduced 3D convolutions and Dice Loss	High computational overhead and memory consumption for 3D volumes

Oktay et al. (2018) [14]	CT Pancreas	Attention U-Net	0.831 (Pancreas)	Attention gates suppress irrelevant background regions	Attention maps can struggle with highly heterogeneous tumor boundaries
Chen et al. (2018) [50]	PASCAL VOC	DeepLabV3+	N/A (Baseline)	Atrous Spatial Pyramid Pooling provides multi-scale contextual awareness	Lacks fine-grained localized feature recovery compared with U-Net dense skip connections
Chen et al. (2021) [34]	Synapse	TransUNet	0.853 (Average)	Models long-range global dependencies using Transformer encoders	Requires massive datasets for pre-training; struggles with localized boundary precision without convolutions
Isensee et al. (2021) [15]	MSD Decathlon	nnU-Net	0.952 / 0.738	Self-configuring framework optimizes hyperparameters for medical data	Relies on preprocessing rather than novel feature extraction
MANet (2023) [27]	LiTS	Multi-Attention Network	0.961 / 0.752	Fuses spatial and channel attention to enhance feature discrimination	Computational complexity increases with dual-attention branches
PGC-Net (2024) [46]	LiTS	Graph Convolutional Network	0.965 / 0.748	Models structural relationships using graph theory for tumor segmentation	Difficult to construct adjacency matrices for highly amorphous, non-structured tumors

MedMamba (2025) [25]	Multiple	State Space Models	0.968 / 0.760	Linear scaling for long-range dependencies	Requires specialized hardware optimization for training
SWATT U-Net (2025) [28]	BraTS / Liver	Wavelet Attention	0.970 / 0.771	Enhances frequency-domain features alongside spatial features for boundary sharpness	Complex integration pipeline; may over-amplify high-frequency noise in low-dose CT scans

3. METHODOLOGY

The overall framework for the research is built as a series of successive data processing and deep learning stages. The main goal of this methodology is to process raw, clinically obtained medical imaging data and output, after going through a number of computational layers, extremely precise segmentation masks of each medical component of the image: background, healthy liver parenchyma, and the region of the tumor.

Next, the data ingestion process begins, where DICOM (Digital Imaging and Communications in Medicine) raw volumes are extracted. Medical CT's image a property of physical tissues, so there is a special task of AVTAR for providing the data in a normalized tensor shape which is straightforward to converge to the neural networks. During training, the data is then passed through a stochastic augmentation algorithm which adds additional variation to the data set and ensures that the algorithms do not overfit specific patient anatomies.

The basic engine of the framework is AAF-Net. The network uses a contracting path during the forward pass to learn hierarchical features of the augmented image tensors. These features are then dynamically routed and blended into the large pathway, culminating in a probabilistic spatial map, provided by a final classification layer – the proposed AAFM. In the backward pass, the network computes the error between the network's predictions and the expert-annotated ground truth with a special composite loss function and adjusts 15.89 million network parameters by gradient descent optimization. Last but not least, the framework has a strict grading system. The quantitative evaluation of the predictions performed on an unseen test cohort is carried out with the help of spatial overlap and pixel-wise classification metrics. A patient-level K-Fold cross-validation approach is directly integrated into the train loop to ensure robustness of the framework and its generalizability across patients.

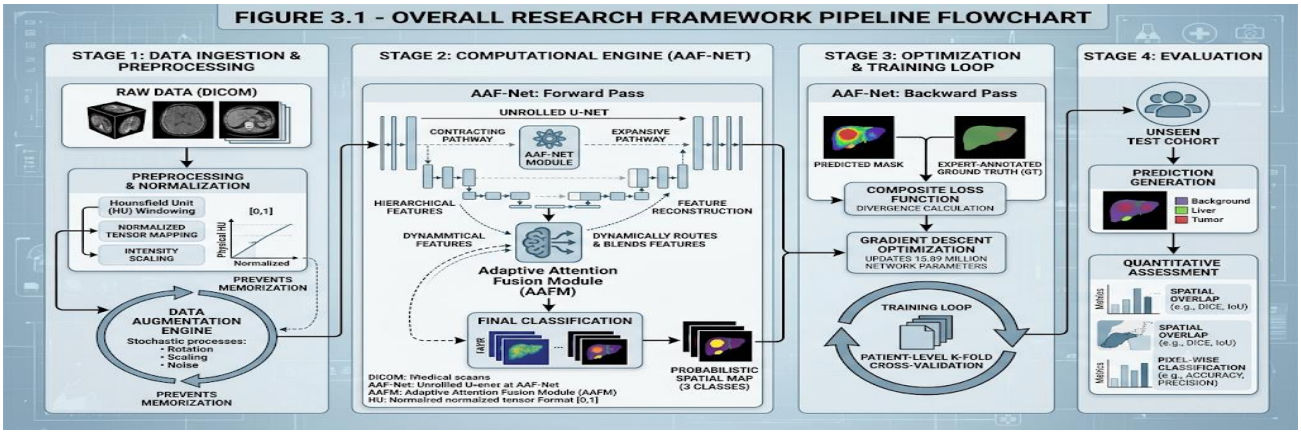

Figure 2 – Overall Research Framework Pipeline Flowchart

With the large effort needed for the annotations required to get a "ground truth" dataset, selecting an appropriate dataset is crucial to the validity of any deep learning study, especially in medical image analysis. This research utilizes an extensively examined and publicly accessible benchmark for reliability and repeatability.

The 3D-IRCADb-01 dataset are 3D abdominal computed tomography images of 20 different patients. The cohort is well balanced, comprising both 10 female and 10 male patients, presenting different anatomical and body habitus.

There are 2,823 slices in the dataset (in an axial view). The diversity of data acquisition duration among the data present in this dataset is an important element of clinical CT data. Slices are not necessarily created with the same spatial resolution and most importantly, the inter-slice distance (slice thickness) is not constant but varies patient by patient with different scanning protocols ranging from 1.25 mm to 4.0 mm. The variance means that strong spatial normalization is necessary in preprocessing the data.

These annotations were also presented here as binary mask volumes for each anatomical structure, but the two masks representing the whole liver volume and the two masks representing all the different hepatic tumors were extracted in the scope of this paper. These independent masks were then integrated to create one ground-truth multi-class mask, with an integer label for each pixel depending on the class: 0 – background, 1 – healthy liver, 2 – tumor class. The dataset contains very complicated pathological cases, both massive outgrowths and deforming contour of the liver's exterior and small, low-contrast interior nodules.

The training, validation and test sets for the dataset were split strictly at the patient level, such that every set comprised the patients recorded during the different trials. Randomized shuffling of individual 2D slices would cause great data leakage as adjacent slices within a single patient would have very similar anatomical information – thus inflating performance metrics.

Even though patient level splitting and 3-fold cross-validation was applied to minimize data leakage and model robustness, there are only 20 independent patient volumes. Thus, the results obtained may not be generalizable due to the small number of patients. The generalizability of the proposed model needs to be assessed further by validating on independent datasets.

Experience matrix 3-Fold Cross-Validation strategy was used to ensure a highly validated architecture proposal. These 20 patients were split into three folds. In every step, the final patients in a subset were assigned exclusively for final testing purpose; a small portion of it was for validation to check convergence and to identify early stopping and the rest consisted of comparison data for model training. Patient distribution by the three folds is specified in detail in Table 3.1.

Table 2. Fold Patient-Level Cross-Validation Split

Fold	Training Set	Validation Set	Testing Set (Patient IDs)	Total Test Slices
Fold 0	11 Patients	2 Patients	7 Patients (1, 10, 14, 17, 2, 5, 7)	~393 Slices
Fold 1	11 Patients	2 Patients	7 Patients (11, 12, 18, 3, 6, 8, 9)	Variable
Fold 2	12 Patients	2 Patients	6 Patients (13, 15, 16, 19, 20, 4)	Variable

Such stringent partitioning means that during every fold iteration, no patient data overlaps between training and testing and the reported quantitative measures are accurate.

The major architectural contribution of this research is the Adaptive Attention Fusion Module. It becomes a very reflective and learnable gatekeeper positioned at each of the skip connections. It aims to resolve this problem on a pixel-by-pixel basis, dynamically telling the network, whether the flowing spatial information of the encoder or the semantic information of the decoder should be trusted.

The process involves certain transformations according to a given path. Let E be the feature map received in high-resolution from the first stage of the encoder and let D be the feature map received in high-resolution from the previous stage of the decoder. To ensure mathematical compatibility with both E and D, they are first separately processed through 1×1 projection convolutions and then normalized in size. Both are first processed separately through 1×1 projection convolutions to standardise their channel dimensions and then normalised in size, to be mathematically compatible.

Subsequently, the optimal fusion ratio needs to be measured by the module. The projected features, maps E and D, are concatenated channel-wise. A combined tensor represents the spatial and the semantic information from the local region. This fusion tensor is fed into a dedicated 1x1 convolutional layer, reducing the channels to exactly one channel and with a sigmoid activation.

This is an operation that produces a spatial gating map (α). Each value in α is at the boundaries of 0 and 1 and the space of α is the same as that of the feature maps. Then, α is used to perform a mathematically precise layer-wise convex combination of the encoder and decoder features within the module. The module adaptively weights the tensors by multiplying the features E with α and the features D with $(1 - \alpha)$. The network is tasked with learning to push back α towards 1, making the strong assertion that it's learning to include the high-frequency encoder details aggressively. When the pixel falls in a noisy background region, the

network attempts to learn a value closer to 0; thereby, hiding the encoder data for these noisy pixels in the process and using only the deep semantic certainty of the decoder.

This feature is restored in the final stage of the pathway that is as large as the original input ($224 \times 224\ pixels$). But there are still many abstract feature channels in the tensor. The_final_output_layer conducts a 1×1 convolutional layer to reduce the number of channels of these abstract features to exactly C number of channels, which are that of the target classes. In this study, $C = 3$ (Background, Liver, Tumor).

This final convolution does not pass the raw linear output (logits) through an internal Softmax loss function across the network because the network is based on a combined Cross-Entropy and Dice loss function framework. Instead, raw logits are returned directly and the loss function API is numerically optimized to use stable log softmax to calculate the backpropagated loss, ensuring numerical stability when training the model.

[Placeholder: Figure 3.2 - Detailed Architectural Diagram of AAF-Net and the AAFM Module]

Convolution: The two-dimensional discrete convolution is a basic feature extraction mechanism in the network. The convoution of an input feature map X and larnable filter K $M \times N$ is a two-dimensional output feature map Y (of size MxN) whose element value at spatial coordinate (i, j) is computed by taking the dot product of X with K at spatial coordinate $(i - j, i - k)$.

$$Y(i,j) = (X * K)(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X\,(i+m, j+n) \cdot K(m,n)$$

Batch Normalization: To speed up convergence and avoid vanishing gradients with the deep architecture of 15.89 million parameters, Batch Normalization is used after every convolution. The algorithm then estimates the empirical mean μ_B and the empirical variance σ_B^2, for a mini-batch B of activations:

$$\mu_B = \frac{1}{m} \sum_{i=1}^{m} x_i$$

$$\sigma_B^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2$$

The activations x_i are then centered (subtract from the mean) and scaled to unit variance (divided by the standard deviation to maintain variance) and called $\hat{x}_i$ because a small constant ϵ was added for numerical stability.

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

Finally, the network has learnable parameters, γ and β, that recover the representational power of the network:

$$y_i = \gamma \hat{x}_i + \beta$$

The Rectified Linear Unit (ReLU) is the primary non-linear activation function used throughout the whole AAF-Net. It uses thresholds at zero to have the non-linearity that is necessary to model complex biological boundaries and avoids the vanishing gradient problem that saturating functions such as the sigmoid suffer from.

$$\mathrm{f(x)} = \max(0, \mathrm{x})$$

The Adaptive Attention Fusion Module's core mathematical logic learns the dynamics of feature blending. Suppose that E_{proj} and D_{proj} are the channel-aligned encoder and decoder feature maps, respectively, where $proj, D_{proj} \in \mathrm{R}^{\mathrm{H} \times \mathrm{W} \times \mathrm{C}}$ with H, W, and C denoting the spatial height, width, and aligned channel dimension. These tensors are concatenated along the channel dimension and passed through a 1×1 convolution W_{gate} followed by a sigmoid activation:

$$\alpha = \sigma\big(W_{gate} \cdot [E_{proj}, D_{proj}]\big)$$

The final fused output tensor F is obtained by using a pixel-wise Hadamard product (denoted as $\odot$) over the two sources, which takes a continuous convex combination of the two:

$$F = \alpha \odot E_{proj} + (1 - \alpha) \odot D_{proj}$$

The resulting fused feature map Fhas dimensions $\mathrm{R}^{\mathrm{H} \times \mathrm{W} \times \mathrm{C}}$ and $\odot$ dot denotes element-wise multiplication.

A weighted Categorical Cross-Entropy (CE) loss is used to capture the point-wise classification loss. We introduce the class weights w_c to tackle a major challenge, the class imbalance. The weighted CE loss is defined as follows for C classes (true binary indicator $y_c \in \{0,1\}$, and predicted probability $p_c \in [0,1]$:

$$L_{CE} = -\sum_{c=1}^{C} w_c \, y_c \log(p_c)$$

Within this model, the specific weights are chosen as $w_{background} = 0.2$, $w_{liver} = 0.8$, and $w_{tumor} = 3.0$ which strongly emphasize penalization of the missed tumor predictions.

structural geometry of your prediction. The multi-class Dice Loss is defined for optimizing the spatial overlap. The Dice Loss directly compares the flattened vectors of true and predicted masks ("flattened" assumes they have the same shape, or a length of N; comparing each element with one another; the length N can be configured; may become completely empty during the first epochs of training; to avoid division by zero at those epochs, a smoothing factor ϵ is added):

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} Y_i \, P_i + \epsilon}{\sum_{i=1}^{N} Y_i + \sum_{i=1}^{N} P_i + \epsilon}$$

4. Results

The proposed Adaptive Attention Fusion Network (AAF-Net) was evaluated for multi-class segmentation of liver parenchyma and hepatic tumors using the fixed test split and a 3-fold cross-validation framework. The evaluation was performed using Dice Similarity Coefficient (Dice), Intersection over Union (IoU), Precision, and Recall.

On the fixed test set, AAF-Net achieved a **Dice score of 0.8126**, an **IoU of 0.7589**, a **Precision of 0.7894**, and a **Recall of 0.9597** for liver parenchyma. For hepatic tumor segmentation, the network achieved a **Dice score of 0.7048**, an **IoU of 0.6841**, a **Precision of 0.7621**, and a **Recall of 0.9052**.

Table 3 -. Final Test Set Performance of AAF-Net

Target	Dice	IoU	Precision	Recall
Liver Parenchyma	0.8126	0.7589	0.7894	0.9597
Hepatic Tumor	0.7048	0.6841	0.7621	0.9052

The results demonstrate that the proposed network achieved a higher recall for liver parenchyma than for hepatic tumors, while both target classes obtained Dice scores above 0.70 on the fixed test set.

Comparative Performance with Existing Architectures

AAF-Net was compared with standard **U-Net, Attention U-Net, and DeepLabV3+** under the same experimental evaluation. The comparison was performed for liver and tumor segmentation using Dice and IoU. For liver segmentation, AAF-Net obtained a Dice score of 0.8126 and an IoU of 0.7589. Attention U-Net obtained 0.8104 Dice and 0.7586 IoU, while U-Net achieved 0.8091 Dice and 0.7490 IoU. DeepLabV3+ achieved 0.7465 Dice and 0.6901 IoU. For hepatic tumor segmentation, U-Net achieved the highest Dice score on the single test split (0.7791), followed by Attention U-Net (0.7750), AAF-Net (0.7048), and DeepLabV3+ (0.6775). The corresponding tumor IoU scores were 0.7598, 0.7553, 0.6841, and **0.6671**, respectively.

Table 4. Comparative Test Cohort Performance

Model	Liver Dice	Liver IoU	Tumor Dice	Tumor IoU	Parameters
DeepLabV3+	0.7465	0.6901	0.6775	0.6671	3.91M
U-Net	0.8091	0.7490	0.7791	0.7598	7.77M
Attention U-Net	0.8104	0.7586	0.7750	0.7553	7.85M
AAF-Net	0.8126	0.7589	0.7048	0.6841	15.89M

For the fixed test split, AAF-Net had a lower Dice score (0.7048) than U-Net (0.7791) and Attention U-Net (0.7750). When using the 3-fold patient-level cross validation, however, AAF-Net outperformed U-Net with a mean tumor Dice score of 0.7639, whereas U-Net had a mean score of 0.7519. This difference suggests that AAF-Net's performance was different for the patient partitions evaluated. Thus, the fixed-test result and the cross-validation result should not be interpreted as evidence of better performance on the fixed test set, but rather as evidence of better performance on the cross-validation set.

Table 5. Fold Cross-Validation Dice Performance

Model	Mean Liver Dice	Mean Tumor Dice
U-Net	—	0.7519
AAF-Net	**0.8094**	**0.7639**

An ablation study was conducted to determine the contribution of the **Adaptive Attention Fusion Module (AAFM)** by comparing fixed feature fusion with adaptive attention fusion under the same experimental conditions.

For liver segmentation, the fixed-fusion configuration achieved a Dice score of **0.8073**, IoU of **0.7520**, Precision of **0.7890**, and Recall of **0.9401**. The adaptive AAFM configuration achieved a Dice score of **0.7912**, IoU of **0.7319**, Precision of **0.8084**, and Recall of **0.9173**.

For tumor segmentation, the fixed-fusion configuration achieved a Dice score of **0.7279**, IoU of **0.7074**, Precision of **0.7789**, and Recall of **0.9140**. With adaptive AAFM, the tumor Dice increased to **0.7992**, IoU to **0.7794**, and Precision to **0.8629**, while Recall was **0.9053**.

Table 6. Ablation Results for Fixed Fusion and Adaptive AAFM

Fusion Method	Target	Dice	IoU	Precision	Recall	Parameters
Fixed Fusion	Liver	0.8073	0.7520	0.7890	0.9401	14.23M
Fixed Fusion	Tumor	0.7279	0.7074	0.7789	0.9140	14.23M
Adaptive AAFM	Liver	0.7912	0.7319	0.8084	0.9173	15.89M
Adaptive AAFM	Tumor	**0.7992**	**0.7794**	**0.8629**	0.9053	15.89M

The ablation results show that the adaptive AAFM configuration produced higher tumor Dice, IoU, and Precision than the fixed-fusion configuration, whereas the fixed-fusion configuration produced higher liver Dice, IoU, and Recall and higher tumor Recall.

The ablation results indicate a trade-off between liver and tumor segmentation performance. The adaptive AAFM achieved better tumor Dice, IoU, and Precision while causing a decrease in liver Dice, IoU, and Recall, and a slight decrease in tumor Recall. This indicates that the adaptive fusion mechanism was more useful for the discrimination of tumor features under the considered setting, but not equally effective for the two target classes.

5. Conclusion

his study proposed the **Adaptive Attention Fusion Network (AAF-Net)** with an **Adaptive Attention Fusion Module (AAFM)** for automated liver and hepatic tumor segmentation from abdominal computed tomography (CT) images. The proposed approach was developed to address the limitations of fixed skip connections by dynamically evaluating the contextual relevance of encoder and decoder feature maps at each spatial coordinate. The AAFM uses a continuous learnable gating mechanism to adaptively integrate spatial and semantic information, while emphasizing clinically relevant boundaries and reducing distracting background information.

The proposed framework was evaluated using the **3D-IRCADb-01 dataset**, comprising **2,823 axial slices from 20 patient volumes**, with patient-level data partitioning to avoid data leakage. The methodology incorporated liver-specific Hounsfield Unit (HU) windowing, intensity normalization, spatial resampling, stochastic geometric augmentation, and a composite loss based on weighted categorical cross-entropy and multi-class Dice loss. AAF-Net was evaluated against **U-Net, Attention U-Net, and DeepLabV3+** using Dice, Intersection over Union (IoU), Precision, and Recall on a fixed test set and through 3-fold cross-validation.

On the fixed test set, AAF-Net achieved a liver **Dice of 0.8126**, **IoU of 0.7589**, and **Recall of 0.9597**. For hepatic tumor segmentation, the network achieved a **Dice of 0.7048**, **IoU of 0.6841**, **Precision of 0.7621**, and **Recall of 0.9052**. Although U-Net

achieved a higher tumor Dice score on the single test split, AAF-Net achieved a higher mean tumor Dice score of **0.7639** compared with **0.7519** for U-Net in the 3-fold cross-validation. The mean liver Dice score of AAF-Net across cross-validation was **0.8094**.

The ablation study further evaluated the contribution of adaptive feature fusion by comparing fixed fusion with the AAFM configuration. For tumor segmentation, the adaptive configuration achieved a Dice of **0.7992**, IoU of **0.7794**, and Precision of **0.8629**, compared with **0.7279**, **0.7074**, and **0.7789**, respectively, for fixed fusion. These results demonstrate the contribution of the adaptive fusion mechanism to tumor segmentation performance under the evaluated experimental conditions.

Overall, the findings support the use of AAF-Net as an adaptive encoder-decoder segmentation framework for liver and hepatic tumor segmentation in CT images. The study demonstrates that pixel-wise adaptive feature fusion can provide an alternative to fixed skip-connection fusion and can improve tumor segmentation performance under the cross-validation evaluation. However, the thesis also reports limitations, including the vulnerability of the higher-capacity architecture to minority-class overfitting, increased computational requirements, elevated false-positive rates and reduced precision, and the limited variability associated with the 20-patient dataset. These limitations indicate the need for further evaluation on broader and more variable datasets before wider application.

6. References

[1] "From CNNs to SAM: A survey of deep learning techniques for liver tumor segmentation in CT images," *IEEE Access*, vol. 13, pp. 199074–199104, 2025.

[2] "Data efficient deep learning for liver and liver tumor segmentation: A comprehensive survey," *Neurocomputing*, vol. 663, Art. no. 131937, 2026.

[3] S. Dutta, A. Bhattacharya, V. E. Balas, and M. K. Hasan, Eds., "Innovations in liver tumor analysis: A review of segmentation and classification approaches," in *Data Mining and Information Security* (*Lecture Notes in Networks and Systems*), vol. 1388, Cham, Switzerland: Springer, 2025, pp. 207–227.

[4] "Advances in liver, liver lesion, hepatic vasculature, and biliary segmentation: A comprehensive review of traditional and deep learning approaches," *Artificial Intelligence Review*, 2025.

[5] "Semantic segmentation of CT liver structures: A systematic review of recent trends and bibliometric analysis," *Journal of Medical Systems*, vol. 48, Art. no. 97, 2024.

[6] "Medical image segmentation review: The success of U-Net," *IEEE Access*, pp. 10076–10095, 2024.

[7] "An in-depth analysis of U-Net variations for advancing precision in medical image segmentation," in *Proc. Int. Conf. Recent Adv. Biomed. Eng.*, 2024.

[8] "UNet and variants for medical image segmentation," *International Journal of Network Dynamics and Intelligence*, vol. 3, no. 2, Art. no. 100009, pp. 1–23, 2024.

[9] "Comparative analysis of modifications of U-Net neuronal network architectures in medical image segmentation," *Pattern Recognition and Image Analysis*, vol. 34, pp. 112–128, 2024.

[10] "Convolutional neural networks for tumor segmentation by cancer type and imaging modality: A systematic review," *Medical Image Analysis*, vol. 95, Art. no. 103245, 2025.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham, Switzerland: Springer, 2015, pp. 234–241.

[12] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham, Switzerland: Springer, 2016, pp. 424–432.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis (DLMIA)*, Cham, Switzerland: Springer, 2018, pp. 3–11.

[14] O. Oktay, J. Schlemper, L. L. Folgoc, et al., "Attention U-Net: Learning where to look for the pancreas," in *Proc. Med. Image Comput. Deep Learn. (MIDL)*, 2018.

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

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[17] S. M. Ibnul Hasan et al., "Dense U-Net: Improving feature extraction and propagation by establishing dense connections between encoder and decoder circuits," *Biomedical Signal Processing and Control*, vol. 92, Art. no. 106082, 2024.

[18] "Shape-intensity-guided U-net for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1789–1801, 2024.

[19] "Comparative study of UNet-based architectures for liver tumor segmentation in multi-phase contrast-enhanced CT," arXiv:2510.25522, 2025.

[20] "Enhanced U-Net models with hybrid attention elements for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 3, pp. 1123–1135, 2025.

[21] "A recurrent-attention mechanism for medical image segmentation," *Pattern Recognition*, vol. 158, Art. no. 111024, 2025.

[22] "Harnessing transformer-based attention mechanisms for multi-scale feature fusion in medical image segmentation," *Applied Intelligence*, vol. 55, Art. no. 1120, 2025.

[23] "Medical image segmentation via multi-scale attention guided network," *IEEE Access*, vol. 13, pp. 45231–45245, 2025.

[24] "Frequency-attentive multi-hierarchical network for medical image segmentation," *Medical Image Analysis*, vol. 99, Art. no. 103456, 2025.

[25] "MedMamba: Multi-scale deformable attention via state space models for robust medical image segmentation," *Pattern Recognition*, vol. 159, Art. no. 111124, 2025.

[26] "OPFSS: A few-shot medical image segmentation algorithm based on optimized pseudo-annotations and self-attention," *IEEE Transactions on Medical Imaging*, 2025.

[27] "MANet: A multi-attention network for automatic liver tumor segmentation in CT imaging," *Scientific Reports*, vol. 13, Art. no. 12678, 2023.

[28] "SWATT U-Net: Semi-wavelet attention module for enhancing brain tumor image segmentation," *Biomedical Signal Processing and Control*, vol. 95, Art. no. 106456, 2025.

[29] "Multi-scale attention guided network for medical image segmentation," *Knowledge-Based Systems*, vol. 298, Art. no. 112345, 2025.

[30] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[32] A. Hatamizadeh, Y. Tang, V. Nath, et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 574–584.

[33] H. Cao, Y. Wang, J. Chen, et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Cham, Switzerland: Springer, 2022, pp. 205–218.

[34] J. Chen, Y. Lu, Q. Yu, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021.

[35] E. Xie, W. Wang, Z. Yu, et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 12077–12090.

[36] "SMAFormer: Synergistic multi-attention transformer for medical image segmentation," arXiv:2409.00346, 2025.

[37] "HBFormer: A hybrid-bridge transformer for microtumor and miniature organ segmentation," arXiv:2512.03597, 2025.

[38] "A reverse Mamba attention network for pathological liver segmentation," arXiv, 2025.

[39] "HANS-Net: Hyperbolic convolution and adaptive temporal attention for accurate and generalizable liver and tumor segmentation in CT imaging," arXiv, 2025.

[40] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Fourth Int. Conf. 3D Vis. (3DV)*, 2016, pp. 565–571.

[41] L. Soler, A. Hostettler, V. Agnus, et al., "3D image reconstruction for comparison of algorithm database: A patient-specific anatomical and medical image database," IRCAD, Strasbourg, France, 2010.

[42] P. Bilic, P. Christ, H. B. Li, et al., "The Liver Tumor Segmentation Benchmark (LiTS)," *Medical Image Analysis*, vol. 84, Art. no. 102680, 2023.

[43] M. Antonelli, A. Reinke, S. Bakas, et al., "The Medical Segmentation Decathlon," *Nature Communications*, vol. 13, Art. no. 4128, 2022.

[44] "Identification of optimal semantic segmentation architecture for the segmentation of hepatic structures from CT images," *Multimedia Tools and Applications*, vol. 83, pp. 89345–89372, 2024.

[45] "DeepLabV3+ with Xception encoder for liver segmentation on 3D-ircadb-01," *Multimedia Tools and Applications*, vol. 83, pp. 45678–45695, 2024.

[46] "Parallel Graph Convolutional Network (PGC-Net) for liver tumor segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 8, pp. 2891–2903, 2024.

[47] "ResU-Net model for liver and tumor segmentation using Medical Segmentation Decathlon (MSD) liver dataset," *Biomedical Signal Processing and Control*, vol. 98, Art. no. 106789, 2025.

[48] "Automated 3D visualization and volume estimation of hepatic structures for treatment planning of liver tumours," *Computers in Biology and Medicine*, vol. 175, Art. no. 108456, 2024.

[49] "Multi-scale liver tumor segmentation algorithm on LiTS17," *IEEE Access*, vol. 13, pp. 56789–56804, 2025.

[50] L. C. Chen, Y. Zhu, G. Papandreou, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2018, pp. 801–818.

[51] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.