

Enhanced Liver Segmentation Using Hybrid U-Net-Transformer Architecture

Shivaleela Betageri

Akkamahadevi Women's University: Karnataka State Akkamahadevi Women's University

G N Girish

`girish@iiitdwd.ac.in`

Indian Institute of Information Technology Dharwad

Rohini A. Bhusnurmath

Akkamahadevi Women's University: Karnataka State Akkamahadevi Women's University

Research Article

Keywords: Liver, Segmentation, U-Net, Transformer, Computed Tomography

Posted Date: January 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8677562/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Enhanced Liver Segmentation Using Hybrid U-Net-Transformer Architecture

Shivaleela Betageri¹, G N Girish^{2*}, Rohini A. Bhusnurmath¹

^{1*}Department of Computer Science, Karnataka State Akkamahadevi Women University, Vijayapura], ,India.

²Department of Computer Science and Engineering, Indian Institute of Information Technology Dharwad, Dharwad], ,India.

*Corresponding author(s). E-mail(s): girish@iiitdwd.ac.in;

Contributing authors: shivaleela.betageri@gmail.com;

rohinibhusnurmath@gmail.com;

Abstract

Segmentation of the liver plays a crucial role in the detection and diagnosis of liver disorders. Conventional manual segmentation methods are time-intensive, susceptible to operator-dependent inconsistencies, and ineffective for extensive clinical applications. To overcome these obstacles, this study presents a hybrid U-Net model with a Transformer bottleneck, which enhances accuracy in segmentation by capturing both local and global contextual information. The model extracts spatial characteristics by utilizing the U-Net encoder, which are then handled by the Transformer encoder to refine global representations before passing through the U-Net decoder. Through the combination of transformer architecture and convolutional neural networks (CNNs), this integration ensures reliable liver segmentation. The proposed model obtained Dice Similarity Coefficients (DSC) of 98.09% and 98.12%, respectively, when evaluated using two publicly available benchmark datasets, CHAOS and 3D-IRCADb. These findings show that the hybrid U-Net-Transformer model enhances segmentation accuracy while maintaining computing efficiency. The source code is available on (<https://github.com/shivaleelabetageri/Liver-segmentation>)

Keywords: Liver, Segmentation, U-Net, Transformer, Computed Tomography

1 Introduction

The liver is a vital organ responsible for key physiological processes, including metabolism, detoxification, and nutrient regulation. However, liver-related diseases, including hepatocellular carcinoma (HCC), cirrhosis, fatty liver disease, and hepatitis, pose significant health risks worldwide. The primary causes of these conditions include excessive alcohol intake, obesity, viral infections (such as hepatitis B and C), and genetic predispositions. Early diagnosis plays key role in managing and potentially curing liver diseases, significantly improving patient survival rates. Medical imaging techniques, particularly ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI), are extensively employed for the early diagnosis of liver diseases. Traditionally, hepatologists and radiologists manually analyze these medical images to identify and segment the liver and detect anomalies. However, manual segmentation takes lots of time and is a labor-intensive process that is significantly reliant on the clinician’s expertise, leading to inter-observer variability and potential diagnostic inconsistencies.

To overcome these limitations, deep learning driven automated liver segmentation models have been introduced to achieve precise, efficient, and consistent segmentation outcomes. Deep learning approaches, in particular CNNs, have significantly transformed medical image interpretation by achieving exceptional accuracy in segmentation tasks [1]. Automated segmentation of the liver utilizing deep learning techniques eliminates the dependency on human expertise, reducing subjectivity and improving efficiency. Within deep learning frameworks, the U-Net model has proven to be particularly instrumental in segmenting medical images due to its ability to learn hierarchical characteristics and capture spatial dependencies [2].

The U-Net model, defined by its symmetric contracting-expanding path network with lateral connections [3], has been extensively employed for liver segmentation. Its capability to effectively capture minute details while maintaining spatial context is its strongest point. U-Net effectively segments liver regions from CT scans, aiding in the precise diagnosis and planning of treatment. However, despite its effectiveness, certain constraints hinder its full potential. To improve liver segmentation performance, a number of U-Net model variants have been introduced, such as U-Net++ [4], Attention U-Net [5], and Residual U-Net [6]. These variants integrate advanced mechanisms such as attention modules, dense connections, and residual learning to improve accuracy and robustness. However, segmentation performance may be impacted by issues including the vanishing gradient, computational inefficiency, and parameter redundancy that arise as U-Net-based models are deepened.

In order to address the constraints of traditional CNN-based UNet architectures, researchers looked into the use of transformers for segmenting medical images. By leveraging the self-attention mechanism [7], transformers concentrate on the most essential features, enhancing segmentation accuracy. In medical image segmentation, transformer models provide significant advantages by capturing global dependencies, improving feature representation, and reducing information loss during down-sampling [8]. The integration of Transformers with CNN-based U-Net architectures has demonstrated promising outcomes in improving segmentation accuracy, especially in complex medical imaging tasks. Thus, in this research, a hybrid architecture is proposed

for the segmentation of the liver from CT scans. The proposed model incorporates transformer-based attention mechanisms within the U-Net framework in a bottleneck structure. This enables efficient recording of long-distance correlation and enhances delineation accuracy.

The remainder of this paper is organized into five parts. Section 2 presents the background and review of the existing work, while Section 3 details the constructed framework. Section 4 outlines the experimental setup and discusses the achieved outcomes. Finally, Section 5 includes the final observations and future research directions.

2 Background

As computational techniques have evolved, researchers have developed automated and semi-automated techniques to increase the precision and accuracy of liver segmentation. Numerous approaches, from conventional image processing methods to more sophisticated approaches built around traditional and hierarchical learning, are presented in the literature.

Prior to the advent of deep neural architectures, segmentation of the liver from CT scans was primarily performed using traditional methods. These methods often included edge detection, thresholding, region growth, and atlas-based approaches [9]. Although effective in certain cases, these techniques struggled with complex anatomical structures and inconsistent image quality. Despite their limitations, traditional approaches laid the foundation for liver segmentation.

Latest developments in neural learning based methodologies have substantially enhanced the precision and robustness of liver segmentation techniques [10]. Techniques based on deep learning, especially CNNs, leverage huge databases of medical image annotations to find patterns associated with liver anatomy. Among the most influential deep learning techniques in the segmentation of medical images is the U-Net, originally formulated by Ronneberger et al. [3], this technique outperformed conventional methods. Tao et al. [11] introduced DefED-Net, an enhanced U-Net based framework for the delineation of the liver and its tumors. Their methodology utilized deformable convolution with residual design, a novel Ladder-ASPP module, and adaptive morphological reconstruction. However, handling irregular liver shapes remains a challenge due to the limitations of fixed convolutional kernels and the spatial information degradation introduced by pooling and strided convolution operations. Lijie et al. [12] proposed the MDR-Net framework, which employs a multiscale dense residual architecture to enhance feature representation and optimize liver delineation precision. Where, the model integrates grayscale mapping, contrast enhancement, and multiscale feature extraction with channel attention and depthwise separable convolutions to increase the accuracy of segmentation. Furthermore, using CNNs and region-based features, Vanda et al. [13] created a flexible volumetric liver segmentation model for CT scans. Their approach involved pre-processing techniques such as rescaling and supervoxel segmentation, along with a modified 3D RP-U-Net with ResPath for segmentation. Despite its success, challenges persist in accurately segmenting liver tumors, particularly near the liver border, and improving evaluation metrics.

Segmenting medical images using Transformer structures has attracted a lot of interest because of their increasing success in computer vision. Transformers, which were originally created for natural language processing (NLP), are ideal for modeling long-distance correlations, which is an important problem in medical image analysis. TransU-Net [14] exemplifies an integrated design that incorporates a convolution based encoder and a transformer driven bottleneck, resulting in enhanced segmentation accuracy across multiple organs. Similarly, Swin-U-Net [15] leverages the hierarchical and locality-aware attention mechanisms of the Swin Transformer to improve the delineation of organ boundaries. MMTU-Net [16] is a U-shaped architecture that incorporates a Multi-level Fusion Transformer (MFT) in its deeper levels to increase the receptive field. It uses a Dual-Channel Attention (DCA) technique to rebalance channel weights and improve edge data extraction. In addition, the skip connections include a Double-Layer Multi-Scale Feature Fusion (DMFF) module, which merges multi-level features, minimizing feature loss and semantic gaps. Furthermore, to capture both local and global contextual information, AttTransCFusion [10] uses a dual encoder architecture made up of a CNN encoder branch and a transformer encoder branch. While a gated fusion module selectively combines local and long-range dependence information to create an accurate prediction map, an attention decoder module refines features across several network levels. Like this, various hybrid frameworks have been proposed, embedding self-attention mechanism within the U-Net framework to augment feature representation and spatial context modeling.

3 Methodology

The proposed technique is built on the traditional U-Net design [4]. Here, encoder, bottleneck, and decoder blocks are used and modified to construct transformer-based UNet, as demonstrated in Figure 1 (a).

3.1 Hybrid U-Net

(1) Encoder Block:

The encoder component of a U-Net structure improved with residual blocks. Each encoder block systematically extracts hierarchical features from the input data utilizing two convolutional residual blocks (Conv-RB) and a maximum pooling layer (see Figure 1 (b)). Each Conv-RB block processes an input tensor via two sequential 3×3 convolutional blocks, each followed by "ReLU" activation function and same padding to retain spatial resolution. To achieve dimensional consistency between the residual input and the output block, the shortcut connection undergoes a 1×1 convolution and is subsequently combined to the resultant of the second 3×3 convolution unit, followed by ReLU activation. This architectural design facilitates residual learning, thereby improving gradient propagation, accelerating network convergence, and promoting efficient feature reuse in deeper architectures.

The encoder stacks these residual blocks in three steps. Each stage comprises two residual blocks followed by a MaxPooling 2D layer that downsamples the feature representation by two. Prior to the pooling operation, a shortcut-link is established to combine the output of the current encoder block with the corresponding decoder

block, facilitating efficient feature propagation and spatial information retention. At each step, the number of filters is raised (64, 128, and 256, respectively), enabling the model to pick up more intricate and abstract properties as the network gets deeper. This hierarchical feature extraction, combined with the stability and depth provided by residual learning, makes the encoder highly effective for capturing rich contextual information necessary for accurate image segmentation in the U-Net framework.

(2) Bottleneck Block:

The bottleneck serves as a pivotal component in the proposed hybrid U-Net structure, by enabling the integration of a transformer encoder into the traditionally convolution-based framework. Its primary function is to record global contextual relationships and long range dependencies that conventional CNNs inherently lack due to their localized receptive fields. This design addresses the limitations of CNNs in modeling holistic spatial dependencies, which are vital for precise semantic segmentation tasks such as liver segmentation.

The bottleneck works on the resultant feature map generated by the encoder, which encodes rich spatial and semantic information focused primarily on local features. Before being embedded within a expanded feature space, this feature map

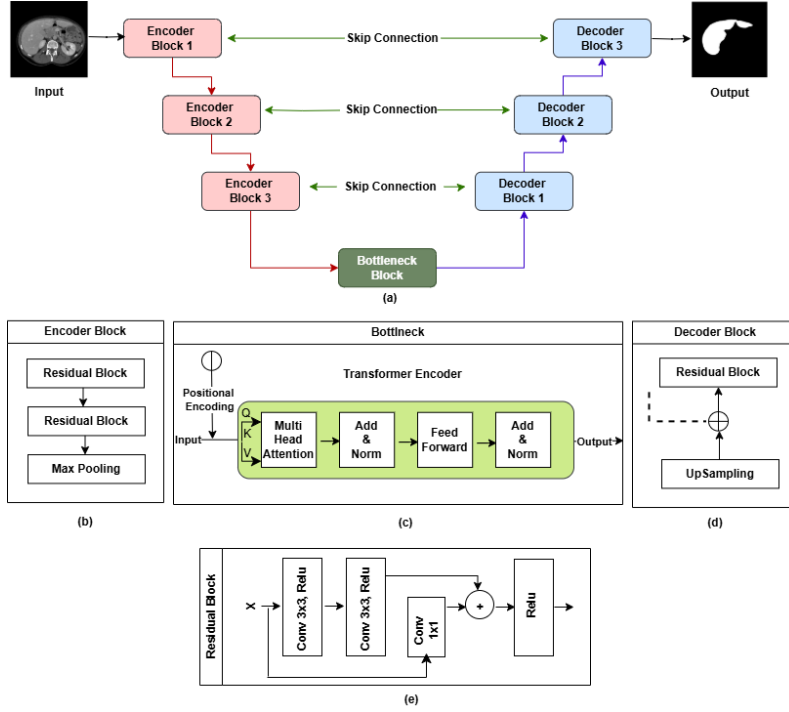


Fig. 1: (a) The proposed model's architecture (b) Encoder block (c) Bottleneck block with positional encoding and transformer encoder (d) Decoder block (e) Residual block

is partitioned into spatially discrete patches of size 16×16 . Each patch is subsequently flattened into one-dimensional vector and transformed through a linear embedding operation. This step converts the 2D spatial feature map into a sequential representation of patch embeddings, making it compatible with transformer-based processing.

To address the transformer’s lack of inherent spatial awareness, encodings of learnable positions are included in the patch embeddings. A trainable positional embedding matrix $P \in \mathbb{R}^{N \times D}$ is initialized, where the total count of patches is denoted by N and embedding resolution by D . Each positional embedding P_i is indexed by patch position $i \in [0, N - 1]$ and added element wise to the corresponding patch embedding X_i , resulting in:

$$Z = X_i + P_i \quad (i)$$

This augmentation enables the transformer to recognize spatial order and proximity within the input sequence, an essential capability for delineating anatomical structures and organ boundaries in medical imaging.

The positionally enriched sequence Z is subsequently passed via multiple transformer contractive path, where individual path containing layer normalization unit, multi-head self-attention systems and feed-forward networks. The attention mechanism dynamically computes relationships between spatially distant patches, whereas feed-forward network further improves these feature representations. Layer normalization is incorporated before and after each attention block to enhance training stability, facilitate gradient flow, and promote faster convergence. The final output of the bottleneck encapsulates a globally contextualized, spatially aware feature representation, which is forwarded to the decoder for spatial dimension recovery and segmentation reconstruction.

(3) Decoder Block:

The decoder module of the proposed hybrid U-Net framework reconstructs the segmentation output by progressively restoring the spatial resolution of the input data through three different stages. During the initial stage, the decoder takes as input the global contextual features extracted by the transformer-based bottleneck. The high-level feature map of encoder block 3 is concatenated with these features via a skip connection. The combined feature map is processed by a residual block with 256 filters, effectively blending global and local information to enhance semantic understanding and spatial coherence. Bilinear interpolation is used in the second stage to upsample the refined high-level feature representations to match the spatial resolution of the output segmentation map from Encoder Block 2. A skip connection is subsequently employed to append these upsampled feature representations with the corresponding feature maps from encoder. The merged representation is further refined by a residual block with 128 filters, guaranteeing better interaction between mid and high level feature representations. To map the spatial dimensionality of the first encoder layer, the decoder upsamples the mid level features in the third stage. These features are concatenated with the low-level encoder features, which retain fine-grained spatial characteristics such as edges and borders. This fusion is processed by a residual block with 64 filters, which incorporates comprehensive local and global contextual data

to improve segmentation accuracy. To construct liver region prediction map, a non-linear function "sigmoid" is applied to the final decoder output, succeeded by a 1×1 convolutional unit to refine segmentation outcome.

3.2 Patch Embedding

The CNN-based patch embedding module serves as a critical interface between CNNs and vision transformers (ViTs) in the proposed liver segmentation framework. Although CNNs excel at extracting localized spatial information, like edges, textures, and fine-grained anatomical details, they are naturally constrained in collecting long-range dependencies. Conversely, ViTs excel at modeling global contextual relationships but require their input in a sequential format, rather than the spatial grid structure used in CNNs. To reconcile these differences, the CNN feature map is converted into a series of patch embeddings appropriate for transformer processing by the patch embedding module. This transformation is performed using a 2D convolutional layer, where both the kernel filter dimension and step size are set to a predefined patch dimension. This process divides the input feature map to distinct parts, which are linearly reflected into a higher-dimensional embedding space. Formally, for a patch located at position (i, j) the embedding is stated as

$$P[i, j] = W \cdot F[i : i + p, j : j + p] + b \quad (\text{ii})$$

where $P[i, j]$ represents the patch embedding, F denotes the representation of the input features, p represents the patch size, W denotes the learnable weight matrix and bias with b . The convolutional projection yields an output tensor of dimensions $(B, H/p, W/p, D)$, where B denotes the batch size, H and W correspond to the spatial dimensions of the input feature map, and D is the embedding dimension. Subsequently, the tensor is reshaped into a sequential representation of size (B, N, D) , where $N = H/p \times W/p$ denotes the overall count of non-overlapping patches generated from the input feature map.

This sequence format allows the ViT to process the contained patches through self-attention mechanisms while maintaining computational efficiency due to the reduced sequence length. Importantly, the embedding preserves critical spatial information, enabling the Transformer to learn both global dependencies and spatial features. This capability is especially critical in abdominal CT imaging, where the liver is usually next to or overlapping organs like kidneys, pancreas, and stomach. These anatomical proximities lead to ambiguous boundaries and overlapping intensity profiles, making it essential for the framework to effectively acquire fine grained regional features and holistic global context. By integrating CNN patch embeddings into the segmentation pipeline, the model effectively combines the strengths of CNNs in local feature extraction and Transformers in contextual reasoning, thereby achieving more accurate and robust liver segmentation.

4 Experimental results

4.1 Dataset Description

The experimental evaluation primarily employs two publicly available benchmark datasets: the Combined Healthy Abdominal Organ Segmentation (CHAOS) [17] and the 3D Image Reconstruction for Comparison of Algorithms Database (3DIRCADb) [18]. The CHAOS data repository comprises 20 intra-abdominal CT scans acquired from different subjects, while the 3DIRCADb dataset involves 20 CT scans with corresponding ground truth annotations. Both datasets provide reference labels for training and validation purposes, and the detailed utilization of these datasets is summarized in Table 1.

Table 1: Attributes of the data source

Dataset	3D CT Scans	Total 2D Slices	Training Slices	Testing Slices	Size
CHAOS	20	7023	5618	1405	512×512
3DIRCADb	20	4680	3744	936	512×512

4.2 Experiments

4.2.1 Dataset Pre-processing

Medical images stored in the DICOM format contain pixel data that lack physical meaning unless appropriately scaled. To derive clinically meaningful information, pixel values are transformed into HU using metadata attributes such as Rescale Slope and Rescale Intercept [19]. The HU scale offers a standardized representation of tissue densities, which is essential to interpret CT scans. This transformation facilitates the differentiation of anatomical structures, as each HU value corresponds to a specific tissue type or material. For liver segmentation, the HU range of interest typically lies between $[100, 400]$, encompassing soft tissues relevant to the liver. Values above 800, such as those associated with metallic artifacts, are excluded to suppress noise and focus on meaningful regions. Mathematically, the conversion from pixel intensity to HU is expressed as:

$$\text{HU}(x, y) = \text{PixelValue}(x, y) \times \text{RescaleSlope} + \text{RescaleIntercept} \quad (\text{iii})$$

This transformation emphasizes liver-specific features while suppressing irrelevant high-intensity artifacts. After HU conversion, median filtering is applied to both images and masks using a 3×3 kernel. This operation suppresses noise without compromising edge details, which is critical for segmentation tasks where precise boundary delineation between anatomical structures is essential. To ensure uniformity across samples, all input images and corresponding masks are resampled to a standardized spatial resolution of 256×256 . This standardization is crucial for enabling batch processing and ensuring compatibility with deep learning architectures.

4.2.2 Experimental Environment

The experimental evaluation was performed on a workstation running Windows 10, configured with an AMD Ryzen 7 processor featuring integrated Radeon graphics, an NVIDIA GeForce RTX 4080 GPU, and 64 GB of system capacity. Model implementation and training were conducted using the TensorFlow deep learning framework. The model was built from the scratch, using arbitrarily initialized parameters, employing the Adam optimization algorithm with a starting step parameter of $1 \times e^{-3}$. To mitigate overfitting and enhance model generalization, an early stopping criterion was applied, where the system evaluated validation loss and stopped training after 30 consecutive epochs of no progress. Training was carried out for a maximum of 200 epochs, utilizing 2 as the batch size. These hyperparameters were empirically chosen to balance memory constraints and convergence stability throughout the training.

4.2.3 Evaluation Metric

To assess the effectiveness of the proposed liver delineation framework, quantitative metrics like dice similarity coefficient (DSC) and accuracy are utilized. The DSC, which serves as the primary evaluation metric, quantifies the shared region between the reference (A) and predicted (B) regions. The definition of it is as follows:

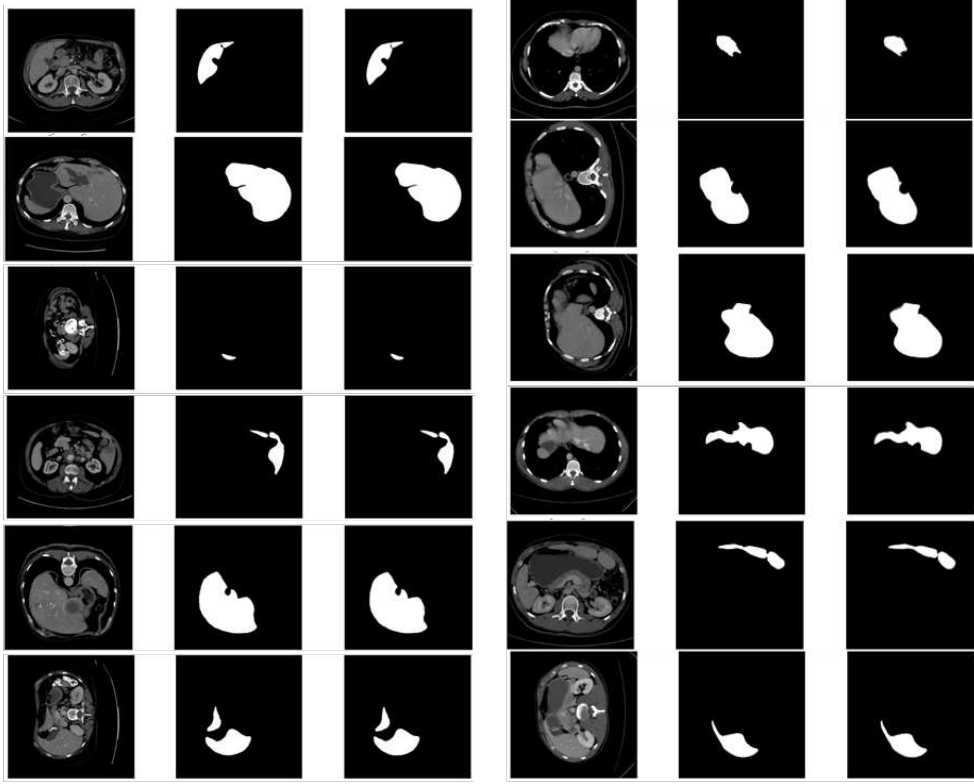
$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|} \quad (\text{iv})$$

The DSC directly evaluates the overlap quality, with higher values indicating better segmentation performance. It ensures that both over-segmentation and under-segmentation are penalized, making it highly relevant for liver segmentation tasks.

4.3 Results

Qualitative findings indicate the efficiency of the approach proposed in two datasets: 3DIRCADb and CHAOS, as visually represented in figure 2. In figure 2a, results for the 3DIRCADb dataset are shown; the first column displays input CT image slices containing the liver region. The second column presents the ground truth segmentation masks used for reference in performance evaluation, while the segmentation masks predicted by the proposed model are shown in the third.

In the figure 2a, the liver appears in the upper abdominal region (see row 1 from top), likely near the superior border. A limited region of the liver is observable, indicating that this slice is from a higher axial plane in the CT scan. The segmentation model closely follows the ground truth, with minimal deviation in boundary details. Accurate segmentation in small liver regions is crucial, as errors in this region can impact volume estimation and diagnostic assessments. Row 2 represents the central portion of the liver, where the organ reaches its maximum cross-sectional area. A large segment of the liver is visible, occupying a significant portion of the abdominal cavity. The ground truth and the anticipated mask match quite well, demonstrating well-defined borders and minimal false positives. Accurate segmentation in this region is essential for volume estimation and disease detection, as the mid-liver area



(a) Segmentation outcomes of IRCADB dataset (b) Segmentation outcomes of CHAOS dataset

Fig. 2: Segmentation results

provides key structural information for clinical analysis. In Row 3, the liver appears near the inferior border, where its size is significantly reduced. This indicates that the slice is closer to the lower margin of the liver. The model successfully identifies the liver. Proper segmentation of inferior liver slices is particularly important for detecting abnormalities in the caudate and quadrate lobes, as these regions are critical in assessing liver diseases and surgical planning.

Row 4 captures discontinuous liver segments, likely near the porta hepatis. The liver appears fragmented due to anatomical variations in this region, making segmentation more challenging. Despite this complexity, the model correctly identifies both liver fragments, demonstrating strong generalization across varying shapes. Accurate segmentation of fragmented liver regions is vital for surgical planning and analysis of liver resection, as errors in this area could lead to incorrect assessments of liver function and structure. In Row 5, the CT slice represents a lower-to-mid abdominal region where the liver exhibits complex contours. The organ appears relatively large, with an

irregular boundary. The model captures the overall shape well, though minor deviations in contour smoothness are observed. Accurate delineation of liver boundaries in such cases helps in early disease detection and improves the reliability of automated liver analysis systems.

Similarly, in Figure 2b, results for the CHAOS dataset are shown; the first column displays input CT image slices containing the liver region. The reference segmentation masks are displayed in the next column, while the predicted masks by the proposed model are shown in the third. In Row 1 (from top), the liver appears small and compact, suggesting a scan from a higher axial plane near the diaphragm. The model segments the liver accurately, preserving boundary details, which is crucial as small liver regions are prone to segmentation errors. In Row 2, the liver is surrounded by multiple organs, making segmentation challenging due to varying contrast with neighbouring tissues and the presence of blood vessels and bile ducts. Despite these complexities, the model effectively preserves the liver shape and smooth boundaries. In Row 3, the liver is partially overlapped by adjacent structures, complicating segmentation. Its irregular shape demands robust generalization across different contours. The model maintains good boundary adherence, with minor shape variations but overall reliable segmentation.

In Row 4, the liver exhibits a slightly irregular boundary, likely due to anatomical variations or adjacent structures. As liver morphology varies significantly among patients, robust generalization is essential. The model successfully identifies the liver with high segmentation quality. In Row 5, the liver is thin and elongated, surrounded by abdominal structures with similar grayscale intensities, making boundary delineation challenging. Despite these complexities, the model maintains high segmentation accuracy without fragmentation. In Row 6, the liver appears crescent-shaped with an irregular boundary. Although relatively large, its variable morphology makes smooth segmentation difficult. However, the model effectively preserves the overall liver shape, ensuring a reliable segmentation outcome. These qualitative results effectively highlight the robustness of the proposed architecture in capturing anatomical variations while preserving high segmentation accuracy across different datasets as well.

The presented framework exhibits better results relative to current state-of-the-art techniques, as evidenced by the empirical findings presented in Table 2. With a DSC of 98.09% and an accuracy of 99.61% on the CHAOS dataset, in 3D-IRCADb the proposed technique gave a DSC of 98.12% and an accuracy of 98.11% outperforming SOTA methods.

4.4 Ablation Study

The ablation study systematically analyses the influence of key factors such as the removal of the empty slices (slices without ground truth), the choice of loss functions, architectural modifications, and data augmentation on liver segmentation performance. These experiments, conducted on the 3DIRCADb dataset, aim to fine-tune the model for enhanced accuracy and robustness, as presented in Table 3. By iteratively refining the model, the study highlights how each targeted enhancement contributes to improved generalization, a higher DSC, and reduced loss. The effectiveness of these

Table 2: comparison between proposed and state-of-the-art models

Dataset	Author	DSC (%)	Accuracy (%)
CHAOS	Merna et al. [20]	81.50	-
	Tao et al. [21]	94.68	-
	Devidas T et al. [22]	95.10	-
	Lijie et al. [12]	96.89	-
	Zhen et al. [23]	96.90	-
	Firat et al. [24]	97.86	-
	Proposed model	98.09	99.61
3D-IRCADb	Zhen et al. [23]	95.20	-
	Lijie et al. [12]	96.19	-
	Devidas T et al. [22]	97.30	-
	Yuran et al. [25]	98.10	-
	Proposed model	98.12	98.11

modifications in enhancing liver segmentation performance is further demonstrated in Table 4.

Table 3: Ablation study details

Method	Modification	Reason for Improvement
Ablation 1	Removing Empty Slices	Reduced redundant data, improved focus on meaningful regions
Ablation 2	Switching from Hybrid Loss to BCE	BCE worked better for segmentation than Hybrid Loss
Ablation 3	Replacing Conv Layers with Residual Blocks	Improved feature extraction and learning
Ablation 4	Adding Augmentation	Helped generalization by introducing variations in training data

Table 4: Influence of model components on DSC

Method	DSC(%)	Method	DSC(%)
Without Empty Slices	81.35	With Empty Slices	85.91
Without BCE	87.23	With BCE	88.85
Without Residual Blocks	90.47	With Residual Blocks	91.89
Without Augmentation	92.65	With Augmentation	98.09

Comparison with Empty Slices vs. without empty slices: A substantial proportion of training samples consist of empty CT slices without liver ground truth, causing the model to overfit to background regions and underemphasize liver structures. This resulted in inefficient learning and increased computational overhead, as

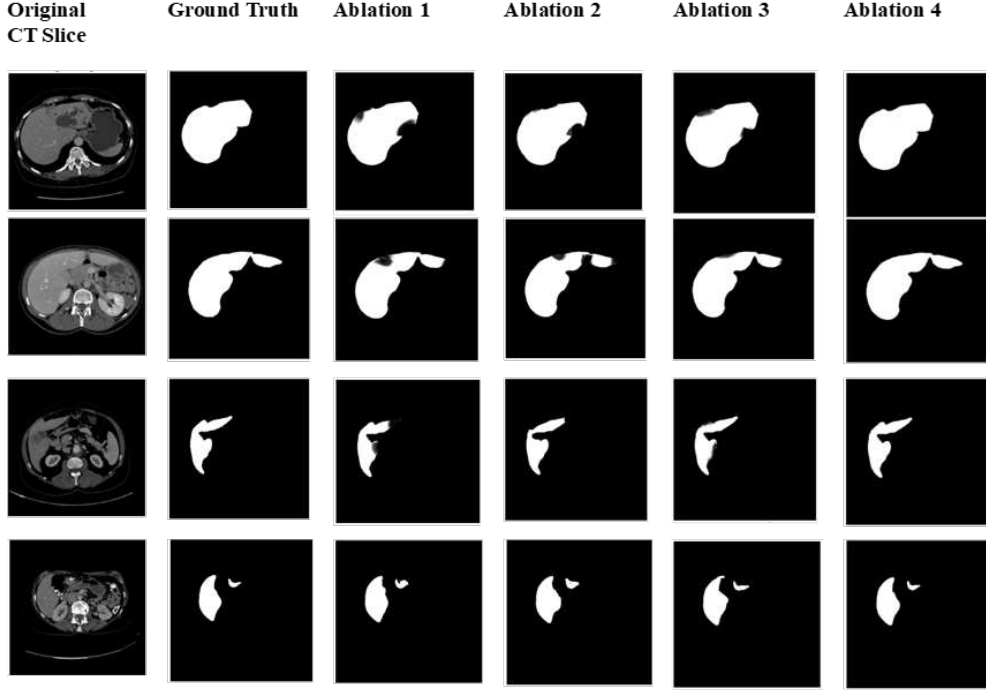


Fig. 3: Ablation study results

the model allocated resources to non-informative data. Furthermore, due to the nature of dice loss, which emphasizes object presence, training with excessive background-only slices degraded the capacity of the model to identify liver regions, yielding a DSC value of 81.35 on the 3DIRCADb dataset. To mitigate this, all empty slices were excluded, ensuring that only anatomically relevant slices were used during training. This refinement improved the foreground-to-background pixel ratio, allowing the model to concentrate on meaningful liver features. As a result, segmentation performance improved significantly, with DSC increasing to 85.91, as reported in Table 4 and qualitatively demonstrated in Figure 3 (column three).

Switching from Hybrid Loss to Binary Cross Entropy (BCE): The presented framework was initially optimized with a hybrid loss function that integrates dice loss, BCE loss, and a weighted Hungarian loss term ($0.1 \times$ Hungarian loss), designed to enhance segmentation precision by jointly minimizing region overlap errors and pixel-wise classification discrepancies. However, this complex formulation introduced over-regularization and conflicting gradients, hindering convergence and limiting the DSC to 87.23 on the 3DIRCADb dataset. Replacing the hybrid loss with BCE loss, a more stable and widely adopted metric for binary segmentation, allowed direct optimization of per-pixel classification, improving model focus on liver versus background differentiation. This adjustment led to a performance gain, with the DSC increasing to 88.85. Table 4 and column 4 of Figure 3 further demonstrate improved boundary delineation and overall segmentation quality following this change.

Residual Blocks as an Alternative to Convolutional Layers: Traditional convolutional layers, while effective in hierarchical feature extraction, suffer from vanishing gradients and loss of fine-grained spatial details in deep architectures, critical limitations for accurate liver segmentation. To overcome these challenges, residual blocks (ResBlocks) were introduced, incorporating identity skip connections that facilitate direct gradient flow and preserve spatial information. This architectural enhancement mitigates gradient degradation, promotes feature reuse, and enables deeper networks to learn complex liver structures more effectively. Empirical results demonstrate improved segmentation performance, with an increased DSC after replacing standard convolutions with ResBlocks (see Table 4 and column 5 of Figure 3). These findings confirm that residual learning enhances both model convergence and anatomical precision in liver segmentation.

Effect of Data Augmentation on Liver Segmentation: Medical image segmentation models often face overfitting due to limited and homogeneous training datasets, particularly in liver CT scans where anatomical variability is high. To address this issue, data enhancement approaches, such as horizontal and vertical flipping, and 90-degree rotations, were employed to synthetically increase training diversity while preserving structural integrity. These spatial transformations exposed the model to varied liver shapes, orientations, and intensities, improving its robustness to inter-patient variability and scanner positioning. As a result, the model demonstrated enhanced generalization, reflected in improved DSC scores (refer Table 4, Figure 3 column 6). Thus, augmentation proved essential in minimizing bias and allowing for more discriminative learning of the model and invariant liver features for accurate segmentation.

5 Conclusion

This study addressed the issue of automated liver segmentation in CT scans by proposing a hybrid U-Net-Transformer model. The development of automated approaches is required since experts driven annotations are time intensive and unpredictable. Classical U-Net framework have shown remarkable effectiveness in medical image segmentation; yet, they have drawbacks such as vanishing gradients and poor global context understanding. To overcome these challenges, the proposed model integrates a Transformer-based bottleneck to enhance feature extraction by capturing long-range dependencies, thereby improving segmentation accuracy. The efficiency of this proposed strategy is shown by the experimental findings. The model achieved DSCs of 97.91 and 98.12, respectively, when tested on two notable datasets, CHAOS and 3DIRCADb. These findings demonstrate the proposed hybrid model’s greater performance over traditional U-Net based techniques. The ablation studies conducted further confirms the impact of essential modifications such as data augmentation, residual blocks, and optimized loss functions in refining the model’s performance. Future research will concentrate on expanding this strategy to multi-organ segmentation and incorporating 3D Transformer architectures to capture richer volumetric information.

Declarations

Funding

The authors did not receive funding support from any organization for the submitted work.

Conflicts of Interests

The authors declare that they have no conflict of interest.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. The dataset used in this work is available in the literature and we have used the dataset as cited by [17] and [18].

CRedit authorship contribution statement

Shivaleela Betageri: Methodology, Software, Validation, Investigation, Visualization, Writing - original draft.

Rohini Bhusnurmath: Validation, Investigation, Supervision, Methodology, Writing, Review.

GN Girish: Conceptualization, Methodology, Formal analysis, Project administration, Resources, Writing – review and editing, Supervision.

References

- [1] Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., Xie, Z.: Deep learning and its applications in biomedicine. *Genomics, Proteomics & Bioinformatics* **16**(1), 17–32 (2018) <https://doi.org/10.1016/j.gpb.2017.07.003>
- [2] Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12), 10076–10095 (2024) <https://doi.org/10.1109/TPAMI.2024.3435571>
- [3] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, Cham (2015)
- [4] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (eds.) *Deep Learning in Medical*

- [5] Trebing, K., Stańczyk, T., Mehrkanoon, S.: Smaat-unet: Precipitation now-casting using a small attention-unet architecture. *Pattern Recognition Letters* **145**, 178–186 (2021) <https://doi.org/10.1016/j.patrec.2021.01.036>
- [6] Khan, R.A., Luo, Y., Wu, F.-X.: Rms-unet: Residual multi-scale unet for liver and lesion segmentation. *Artificial Intelligence in Medicine* **124**, 102231 (2022) <https://doi.org/10.1016/j.artmed.2021.102231>
- [7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS’17*, pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA (2017)
- [8] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in medical image analysis with vision transformers: A comprehensive review. *Medical Image Analysis* **91**, 103000 (2024) <https://doi.org/10.1016/j.media.2023.103000>
- [9] Kaur, H., Kaur, N., Neeru, N.: Evolution of multiorgan segmentation techniques from traditional to deep learning in abdominal ct images – a systematic review. *Displays* **73**, 102223 (2022) <https://doi.org/10.1016/j.displa.2022.102223>
- [10] Gul, S., Khan, M.S., Bibi, A., Khandakar, A., Ayari, M.A., Chowdhury, M.E.H.: Deep learning techniques for liver and liver tumor segmentation: A review. *Computers in Biology and Medicine* **147**, 105620 (2022) <https://doi.org/10.1016/j.combiomed.2022.105620>
- [11] Lei, T., Wang, R., Zhang, Y., Wan, Y., Liu, C., Nandi, A.K.: Defed-net: Deformable encoder-decoder network for liver and liver tumor segmentation. *IEEE Transactions on Radiation and Plasma Medical Sciences* **6**(1), 68–78 (2022) <https://doi.org/10.1109/TRPMS.2021.3059780>
- [12] Xie, L., Zhu, F., Yao, N.: Mdr-net: Multiscale dense residual networks for liver image segmentation. *IET Image Processing* **17**(8), 2309–2320 (2023) <https://doi.org/10.1049/ipr2.12793> <https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/ipr2.12793>
- [13] Czipczer, V., Manno-Kovacs, A.: Adaptable volumetric liver segmentation model for ct images using region-based features and convolutional neural network. *Neurocomputing* **505**, 388–401 (2022) <https://doi.org/10.1016/j.neucom.2022.07.024>
- [14] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A., Zhou, Y.: Transunet:

Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis* **97**, 103280 (2024) <https://doi.org/10.1016/j.media.2024.103280>

- [15] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) *Computer Vision – ECCV 2022 Workshops*, pp. 205–218. Springer, Cham (2023)
- [16] Zhou, S., Wang, M., Huang, Z., He, Y., Han, X., Zhao, Y., Zhao, Z.: Attention-guided fusion of transformers and cnns for enhanced medical image segmentation. *The Visual Computer*, 1–21 (2025)
- [17] Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: Chaos - combined (ct-mr) healthy abdominal organ segmentation challenge data (2019) <https://doi.org/10.5281/zenodo.3362844>
- [18] Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J.: 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep **1**(1) (2010)
- [19] Herrmann, M.D., Clunie, D.A., Fedorov, A., Doyle, S.W., Pieper, S., Klepeis, V., Le, L.P., Mutter, G.L., Milstone, D.S., Schultz, T.J., Kikinis, R., Kotecha, G.K., Hwang, D.H., Andriole, K.P., lafrate, A.J., Brink, J.A., Boland, G.W., Dreyer, K.J., Michalski, M., Golden, J.A., Louis, D.N., Lennerz, J.K.: Implementing the dicom standard for digital pathology. *Journal of Pathology Informatics* **9**(1), 37 (2018) https://doi.org/10.4103/jpi.jpi_42_18
- [20] Bibars, M., Salah, P.E., Eldeib, A., Elattar, M.A., Yassine, I.A.: Cross-modality deep transfer learning: Application to liver segmentation in ct and mri. In: Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T., Gordon, S. (eds.) *Medical Image Understanding and Analysis*, pp. 96–110. Springer, Cham (2024)
- [21] Lei, T., Sun, R., Du, X., Fu, H., Zhang, C., Nandi, A.K.: Sgu-net: Shape-guided ultralight network for abdominal image segmentation. *IEEE Journal of Biomedical and Health Informatics* **27**(3), 1431–1442 (2023) <https://doi.org/10.1109/JBHI.2023.3238183>
- [22] Kushnure, D.T., Tyagi, S., Talbar, S.N.: Lim-net: Lightweight multi-level multiscale network with deep residual learning for automatic liver segmentation in ct images. *Biomedical Signal Processing and Control* **80**, 104305 (2023) <https://doi.org/10.1016/j.bspc.2022.104305>
- [23] Xia, Z., Liao, M., Di, S., Zhao, Y., Liang, W., Xiong, N.N.: Automatic liver segmentation from ct volumes based on multi-view information fusion and condition random fields. *Optics & Laser Technology* **179**, 111298 (2024) <https://doi.org/10.1016/j.optlastec.2024.111298>

[//doi.org/10.1016/j.optlastec.2024.111298](https://doi.org/10.1016/j.optlastec.2024.111298)

- [24] Özcan, F., Uçan, O.N., Karaçam, S., Tunçman, D.: Fully automatic liver and tumor segmentation from ct image using an aim-unet. *Bioengineering* **10**(2) (2023) <https://doi.org/10.3390/bioengineering10020215>
- [25] Zhou, Y., Kong, Q., Zhu, Y., Su, Z.: Mcfa-unet: Multiscale cascaded feature attention u-net for liver segmentation. *IRBM* **44**(4), 100789 (2023) <https://doi.org/10.1016/j.irbm.2023.100789>

Author Biography



Mrs. Shivaleela Betageri holds an MSc in Computer Science and is currently pursuing a PhD in the same field. Her research interests include image processing, medical image analysis, deep learning, and machine learning.



Dr. Girish G. N received the Ph.D. degree in computer science and engineering from the National Institute of Technology Karnataka, Surathkal, India. He was a Postdoctoral Research Fellow at the Tearney Laboratory, Wellman Centre for Photomedicine, Massachusetts General Hospital and Harvard Medical School, Harvard University, USA. He is currently working as an Assistant Professor with the Indian Institute of Information Technology Dharwad, India. His research interests include image processing, machine learning, deep learning, optical coherence tomography, and medical image analysis.



Rohini A. Bhusnurimath, is an Assistant Professor at Akka Mahadevi Women's University, Bijapur, Karnataka. She holds a PhD degree in Computer Science and Engineering; her areas of interest are Image Processing and Pattern Recognition. She has published 45 research papers in peer-reviewed International Journals and Proceedings of International Conferences.