
COMPARATIVE STUDY OF UNET-BASED ARCHITECTURES FOR LIVER TUMOR SEGMENTATION IN MULTI-PHASE CONTRAST-ENHANCED COMPUTED TOMOGRAPHY

Doan-Van-Anh Ly

School of Computer Science and Engineering
The Saigon International University
Ho Chi Minh City 700000, Vietnam
anhldv1203@gmail.com

Thi-Thu-Hien Pham

School of Biomedical Engineering
International University
Ho Chi Minh City 700000, Vietnam
Vietnam National University HCMC
Ho Chi Minh City 700000, Vietnam
ptthien@hcmiu.edu.vn

Thanh-Hai Le

School of Computer Science and Engineering
The Saigon International University
Ho Chi Minh City 700000, Vietnam
lethanhhai@siu.edu.vn

November 25, 2025

ABSTRACT

Segmentation of liver structures in multi-phase contrast-enhanced computed tomography (CECT) plays a crucial role in computer-aided diagnosis and treatment planning for liver diseases, including tumor detection. In this study, we investigate the performance of UNet-based architectures for liver tumor segmentation, starting from the original UNet and extending to UNet3+ with various backbone networks. We evaluate ResNet, Transformer-based, and State-space (Mamba) backbones, all initialized with pretrained weights. Surprisingly, despite the advances in modern architecture, ResNet-based models consistently outperform Transformer- and Mamba-based alternatives across multiple evaluation metrics. To further improve segmentation quality, we introduce attention mechanisms into the backbone and observe that incorporating the Convolutional Block Attention Module (CBAM) yields the best performance. ResNetUNet3+ with CBAM module not only produced the best overlap metrics with a Dice score of 0.755 and IoU of 0.662, but also achieved the most precise boundary delineation, evidenced by the lowest HD95 distance of 77.911. The model's superiority was further cemented by its leading overall accuracy of 0.925 and specificity of 0.926, showcasing its robust capability in accurately identifying both lesion and healthy tissue. To further enhance interpretability, Grad-CAM visualizations were employed to highlight the region's most influential predictions, providing insights into its decision-making process. These findings demonstrate that classical ResNet architecture, when combined with modern attention modules, remain highly competitive for medical image segmentation tasks, offering a promising direction for liver tumor detection in clinical practice.

Keywords CBAM · Grad-CAM · Liver CECT · Mamba · ResNet · Transformer · Tumor Segmentation · UNet3+

1 Introduction

Liver cancer is among the leading causes of cancer-related mortality worldwide, and accurate delineation of the liver and its lesions in medical images is a prerequisite for computer-aided diagnosis, surgical planning, and treatment monitoring [1, 2]. Contrast-enhanced computed tomography (CECT), particularly in multi-phase acquisition, provides

detailed anatomical and functional information that makes it a preferred modality for liver imaging. However, manual annotation of liver structures is time-consuming, subject to inter-observer variability, and impractical for large-scale clinical deployment. This has motivated the development of several automated liver segmentation methods.

The application of deep learning has fundamentally transformed medical image analysis, with semantic segmentation being a critical prerequisite for efficient disease diagnosis and treatment by identifying organ or lesion pixels from images such as CT or MRI. Historically, this field has been dominated by Convolutional Neural Networks (CNNs). The foundational architecture in biomedical image segmentation is the UNet, known for its symmetric encoder-decoder structure and skip connections that enable the capture of both local and global contextual information for precise characterization of structures of interest [3]. While UNet architectures have shown promising results, CNNs inherently struggle to model long-range dependencies (LRDs) due to the local nature of convolutional operations. To enhance performance, researchers have focused on modifying the UNet backbone by integrating elements that improve feature extraction and information flow. One significant development involves leveraging Residual Networks (ResNet), which utilize skip connections to prevent issues like vanishing gradients and performance degradation in deep networks, facilitating successful training and effective extraction of high-level features. Hybrid models like UNet-ResNet combine UNet's structure with ResNet's deep residual learning capabilities to capture complex features and contextual details, demonstrating superior segmentation accuracy compared to conventional UNet-based approaches in tasks such as liver tumor segmentation [4]. This hybrid model was validated on a dataset consisting of 130 CT scans of liver cancers. Experimental results demonstrated impressive performance metrics for liver cancer segmentation, achieving an accuracy of 0.98 and a minimal loss of 0.10. Similarly, the UIGO model [5] designed for automated medical image segmentation for liver tumor detection, merges the UNet architecture with Inception networks (specifically InceptionV3 blocks and an Inception-ResNet backbone) to enhance multi-scale feature extraction capabilities, addressing the challenge posed by the diversity in tumor shape, size, and texture. The model was tested using the LiTS, CHAOS, and 3D-IRCADb1 datasets for liver tumor detection. UIGO demonstrated exceptional results, achieving a segmentation accuracy of 99.93%, a Dice Coefficient of 0.997, and an IoU of 0.998.

Further refinement has come through the integration of Attention Mechanisms. Models such as Attention UNet and AHCNet (Attention Hybrid Convolutional Network) integrate attention with skip connections to improve segmentation quality, enhancing the model's capacity to identify fine features and focus on informative channels [6]. AHCNet was trained using 110 cases from the LiTS dataset (after removing the 3DIRCADb subset) and subsequently evaluated 20 cases in the 3DIRCADb dataset and 117 cases in a Clinical dataset. The proposed model achieved high performance in tumor segmentation accuracy, demonstrating an 11.6% improvement in the dice global score compared to the baseline method on the 3DIRCADb dataset. The FSS-ULivR model designed for few-shot liver segmentation, employs improved attention gates, residual refinement, and multiscale skip connections in its decoder to restore spatial detail and generate accurate boundaries, selectively enhancing relevant features while suppressing irrelevant ones carried by conventional skip connections [7]. This proposed model, trained on the LiTS dataset, was robustly validated on cross-datasets including 3DIRCADb01, CRLM, CT-ORG, and MSD-Task03-Liver. The model achieved an outstanding Dice coefficient of 98.94%, an IoU of 97.44%, and a specificity of 93.78% on the LiTS test set, surpassing existing few-shot methods.

The realization that purely convolutional architectures were insufficient for capturing large-scale dependencies motivated the integration of the Transformer architecture, which excels at modeling global relationships through its self-attention mechanism. Networks like TransUNet [8] and Swin-UNet [9] combine the hierarchical structure of UNet with Transformer modules in the encoder to improve long-range dependency modeling. However, a major drawback of Transformers in high-resolution and 3D medical image analysis is the high computational cost, as the self-attention mechanism scales quadratically with the input size. This limitation prompted research into State Space Models (SSMs), particularly the Mamba architecture, known for its ability to model long sequences with enhanced computational efficiency and linear scaling in feature size [10]. Mamba-based models leverage the Visual State Space (VSS) block, which uses a Cross-Scan Module (CSM) to convert non-causal visual images into ordered patch sequences, thereby adapting Mamba to computer vision tasks [11]. SegMamba [12] is introduced as the first method specifically utilizing Mamba for 3D medical image segmentation, featuring a tri-orientated Mamba (ToM) module to enhance sequential modeling of 3D features and a gated spatial convolution (GSC) module to retain spatial information. This model was tested on the new CRC-500 dataset, BraTS2023, and AIIB2023. On the BraTS2023 dataset, it achieved 93.61%, 92.65%, and 87.71%, and HD95s of 3.37, 3.85, and 3.48 on WT, TC, and ET, respectively. VMAXL-UNet [13] fuses VSS blocks with extended LSTMs (xLSTM) within a UNet architecture to capture long-range dependencies while maintaining linear computational complexity. The model was tested on dermatological (ISIC17, ISIC18) and polyp segmentation (Kvasir-SEG, ClinicDB) datasets. The model significantly outperformed traditional CNNs and Transformers, achieving the best results with 90.1% mIoU and 95.21% DSC on the challenging ClinicDB dataset.

Recent efforts to improve liver segmentation from multi-phase CECT images have focused on building standardized datasets and developing advanced deep learning models. Luo et al. [14] published a comprehensive 3D multi-phase

CECT dataset for primary liver cancer, covering three major types: HCC, ICC, and cHCC-CCA. The dataset includes 278 cancer cases and 83 non-cancer cases, with over 50,000 annotated slices, providing a valuable foundation for training and evaluating classification and segmentation models. Meanwhile, Hu et al. [15] proposed the TMPLITS framework for reliable multi-phase liver tumor segmentation, incorporating Dempster–Shafer theory to model uncertainty and a MEMS expert-mixing mechanism to fuse information across phases. Their method outperformed existing approaches in both accuracy and reliability, especially under noisy or incomplete data conditions. Additionally, Nayantara et al. [16] developed an automatic liver segmentation model using an enhanced SegNet architecture combined with an ASPP module and leaky ReLU activation. The model achieved Dice scores above 96% in the portal venous phase and over 93% in other phases (arterial, delayed and plain CT phases), demonstrating superior performance and strong clinical applicability.

Furthermore, the general opacity of sophisticated deep learning models (often operating as "black boxes") raises concerns for clinical practitioners regarding the rationale behind segmentation outputs, necessitating greater Explainable Artificial Intelligence (xAI). The field of xAI has largely focused on image classification, but methods are being extended to computer vision domains like image segmentation, specifically in medical image analysis [17]. Seg-Grad-CAM [18] is one such extension of the popular Grad-CAM algorithm applied to segmentation. However, a key nuance associated with its utilization is that Seg-Grad-CAM assigns a single coefficient to each activation map after global average pooling of the gradient matrix, meaning it does not incorporate spatial considerations when generating explanations for specific regions within a segmentation map. To solve this, Seg-XRes-CAM [19] was proposed, taking inspiration from HiResCAM [20], to generate location-aware and spatially localized explanations for image segmentation regions, demonstrating a higher degree of visual agreement with model-agnostic methods like RISE compared to Seg-Grad-CAM.

This study employs the Primary Liver Cancer CECT Imaging Dataset [14], a well-curated resource obtained from a single medical facility. The dataset offers comprehensive 3D CECT scans with expert annotations delineating both liver and lesion regions, serving as a solid foundation for the development and validation of advanced segmentation models. The methodological approach begins with an assessment of various UNet and UNet3+ architectures. Upon selection of the most effective baseline model, a detailed investigation is subsequently performed by incorporating four distinct module variants—SE, CBAM, ASPP, and a combined CBAM-ASPP configuration—to identify the optimal architecture for precise lesion segmentation.

2 Materials and Methods

2.1 Dataset

The Primary Liver Cancer CECT Imaging Dataset [14] contains 278 liver cancer cases and 83 non-liver cancer cases, each with full 3D multi-phase contrast-enhanced CT (CECT) scans (Plain, Arterial, Venous, and Delayed phases). Both the liver and lesion regions were annotated in NIfTI format.

Preprocessing was necessary to prepare the dataset, which consisted of 3D volumes, for training. A verification pipeline was developed to ensure consistency between CT volumes and their corresponding segmentation masks of tumor regions (ground truth), as shown in Fig. 1. Each CT–mask pair underwent validation for dimensional alignment, slice count, file integrity, and the presence of annotated lesions. The pairs identified as invalid or mismatched were excluded from further analysis. The results of this verification process were documented in a metadata CSV file, which subsequently served as an index for all experimental procedures. This preprocessing pipeline produced a reliable and balanced dataset of 2D liver tumor slices, enabling efficient experimentation with different segmentation architectures.

Table 1: Distribution of training, validation and testing sets after preprocessing 3D CECT scans.

Dataset	Train	Val	Test
Primary Liver Cancer CECT Imaging Dataset	7507	2146	1073

For model training, 2D slices were extracted from the valid CT–mask pairs. Each slice was normalized to zero mean and unit variance, then resized. To enhance model generalization, data augmentation was applied during training, consisting of random horizontal and vertical flips, as well as 90° rotations. The dataset was partitioned into training, validation, and testing subsets with ratios of 70%, 20%, and 10%, respectively. Table 1 presents the number of slices in each set following validation of the CT–mask pairs from the dataset.

Figure 2 illustrates the distribution of the dataset across three subsets. For all data splits, the distribution is heavily concentrated near zero, with the 25th percentile and median pixel areas being very low. Tumor regions were uniformly

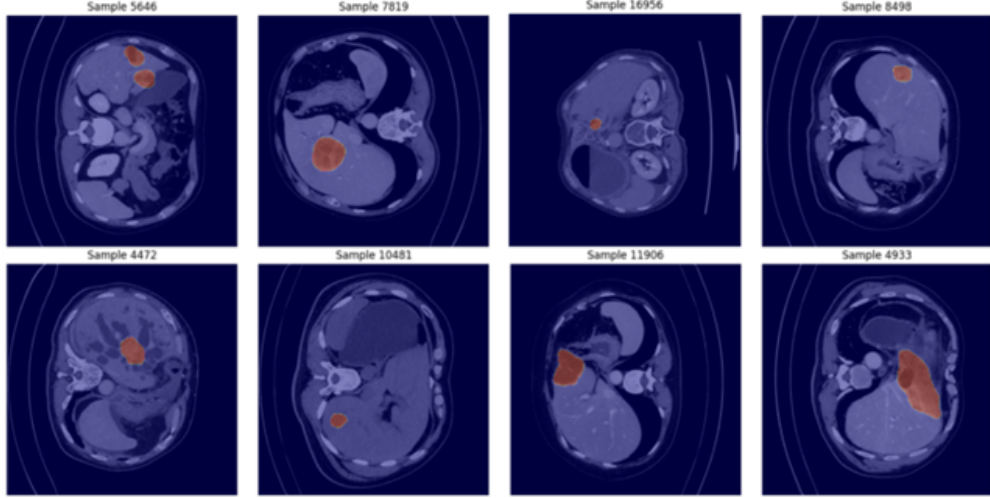


Figure 1: CT slice samples with ground truth.

measured on images with a resolution of 256×256 pixels. These findings indicate that most tumors are small, presenting inherent challenges for segmentation due to limited spatial and contextual information available for the network to leverage.

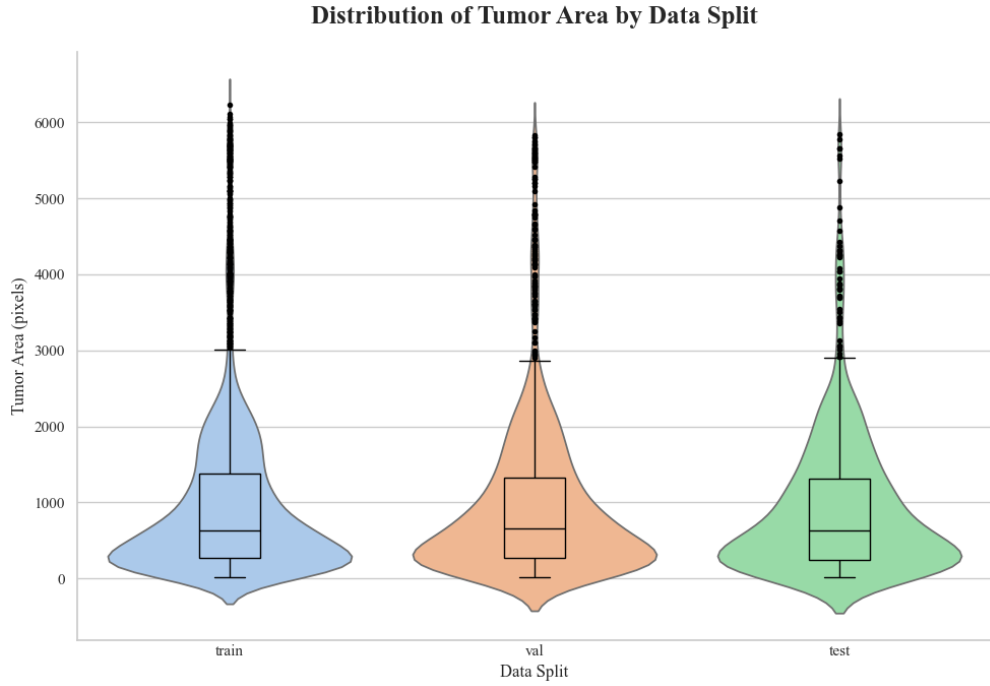


Figure 2: Distribution of training, validation and testing sets on tumor area.

2.2 Models

2.2.1 Base Model Selection

In this study, two widely used segmentation architectures were selected as the foundation for experimentation: UNet [21] and UNet3+ [22].

The UNet architecture has become the de facto standard for medical image segmentation tasks due to its encoder–decoder design with skip connections. The encoder captures hierarchical features through successive convolution and pooling operations, while the decoder progressively reconstructs spatial details. The skip connections ensure that fine-grained localization is preserved, making UNet particularly effective for segmenting small and irregular structures such as liver tumors.

Building upon UNet, UNet3+ introduces full-scale skip connections that aggregate feature maps from multiple encoder and decoder levels. This design allows for a more comprehensive fusion of semantic and spatial information, addressing the limitations of conventional UNet in balancing detail preservation with global contextual understanding. In liver tumor segmentation, this capability is especially beneficial since tumors may vary significantly in size, shape, and intensity.

To further enhance feature extraction, both architectures were integrated with different backbone networks as encoders. Backbones such as Mamba, ResNet, ... provide pre-trained feature representations that can accelerate convergence and improve generalization. Incorporating these backbones allows the models to leverage knowledge from large-scale natural image datasets, which is particularly advantageous given the limited size of medical imaging datasets.

By comparing the performance of UNet and UNet3+ under various backbone configurations, this study aims to identify a balance between segmentation accuracy, boundary precision, and computational efficiency in the context of liver tumor segmentation from CECT scans. Figure 3 presents the operational framework that will be utilized throughout this study.

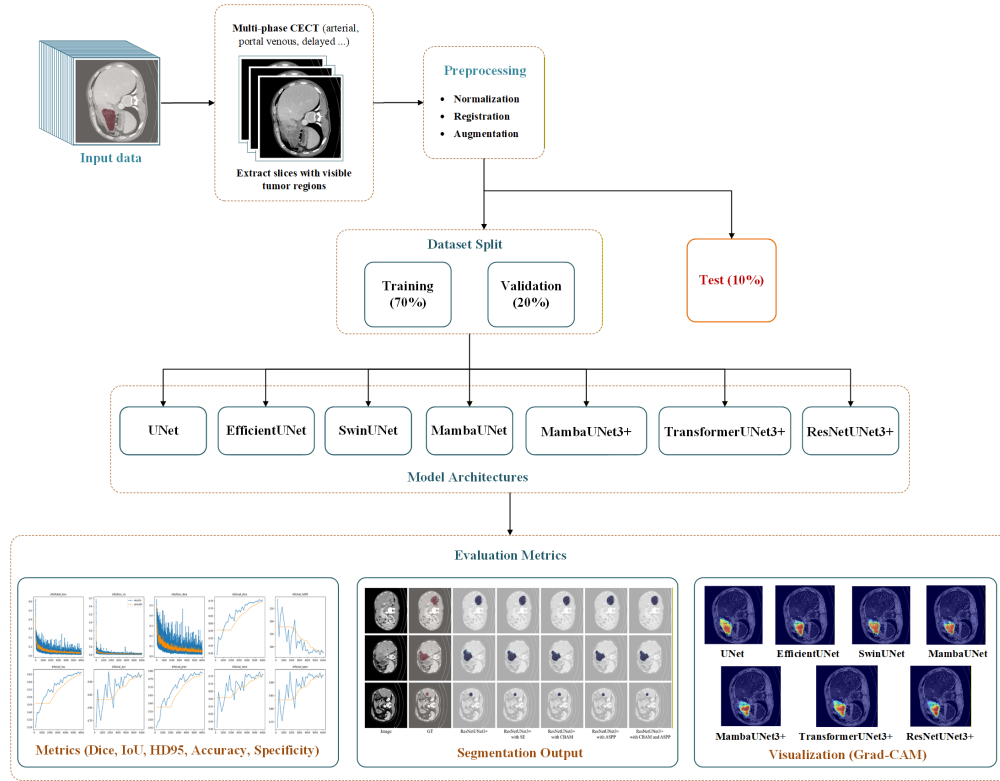


Figure 3: Research framework.

2.2.2 ResNet UNet3+ with attention mechanisms

The best-performing architecture in this study was based on the combination of a ResNet [23] encoder backbone and a UNet3+ decoder, enhanced by the Convolutional Block Attention Module (CBAM) [24, 25]. This section provides a detailed rationale for the choice of each component and their synergistic effect on liver tumor segmentation from CECT images.

The encoder network plays a critical role in extracting hierarchical features from the input images. ResNet was selected as the backbone due to its residual learning framework, which facilitates the training of very deep networks by introducing skip connections that mitigate the vanishing gradient problem. This allows the model to capture both low-level spatial details and high-level semantic features. Figure 4 presents the architecture of ResNetUNet3+ incorporating the ASPP module, illustrating one of the variants examined in this study.

In medical imaging, where tumor boundaries are often subtle and tumor appearance varies across phases (plain, arterial, venous, delayed), ResNet provides a balance of computational efficiency and robust feature extraction. Unlike lightweight networks that may underfit complex structures, or Transformer-based encoders that require larger datasets and computational resources, ResNet50 has demonstrated stable performance in transfer learning scenarios with limited annotated data. Its pre-training on ImageNet further supports generalization by offering strong initialization for low-level edge, texture, and shape features, which are essential for distinguishing lesions from healthy liver tissue.

While the original UNet has proven highly effective for biomedical segmentation tasks, it relies on single-scale skip connections that may not fully capture the wide variation in lesion size. UNet3+ extends this architecture by introducing full-scale skip connections and a redesigned decoder path. Specifically, UNet3+ aggregates feature maps from encoder and decoder layers at multiple scales, ensuring that fine-grained details (e.g., small lesion edges) are combined with high-level semantic context (e.g., overall tumor shape).

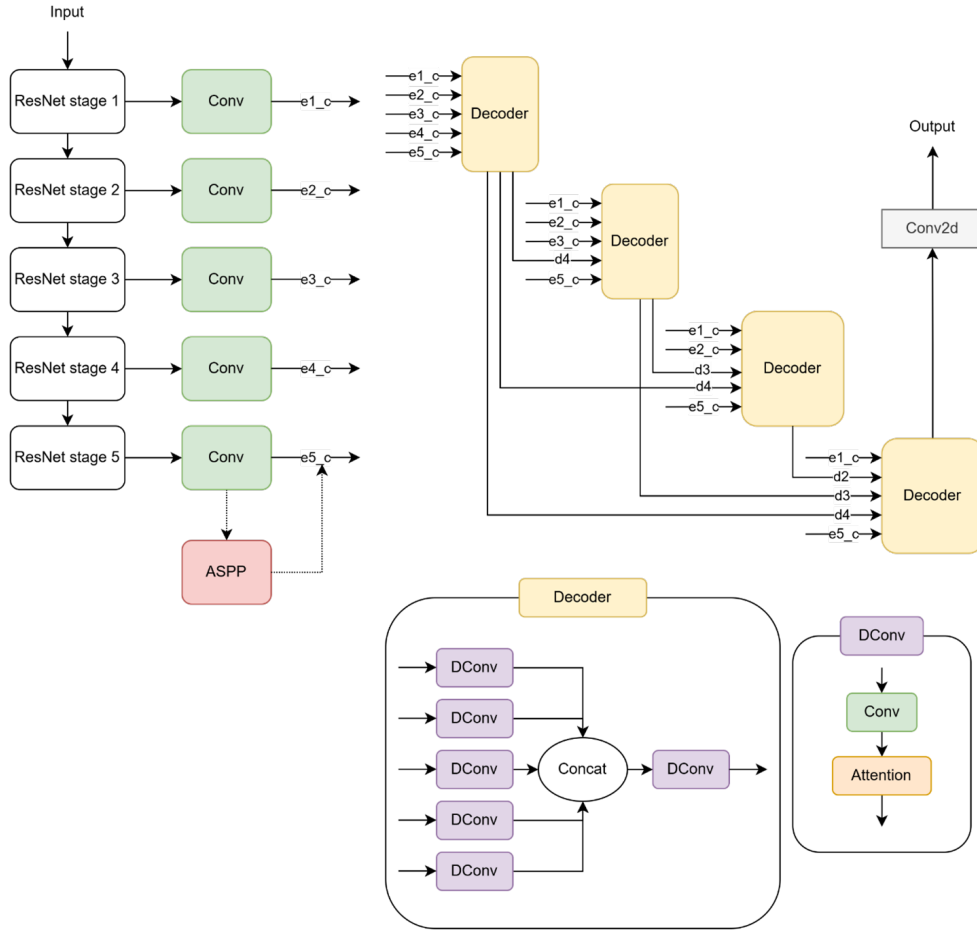


Figure 4: Illustration of ResNetUNet3+ variant.

For liver tumor segmentation, this design is particularly advantageous. Tumors can vary significantly in size and intensity across patients and CECT phases; full-scale feature fusion enables the network to remain sensitive to both microscopic nodules and large heterogeneous masses. Additionally, the redesigned decoder mitigates the semantic gap between encoder and decoder features, leading to more precise reconstruction of lesion boundaries.

Although UNet3+ with ResNet provides strong baseline performance, not all extracted features are equally relevant for tumor segmentation. To address this, attention mechanisms were integrated into the decoder to enhance feature discriminability. Two approaches were considered:

- Squeeze-and-Excitation (SE) [26]: SE modules perform global average pooling followed by channel-wise recalibration, allowing the network to assign higher weights to informative feature channels. While effective in many applications, SE does not explicitly model spatial dependencies.
- Convolutional Block Attention Module (CBAM): CBAM extends the SE concept by sequentially applying channel attention and spatial attention. Channel attention identifies which feature maps are most relevant (e.g., texture-sensitive channels for tumor edges), while spatial attention highlights discriminative regions within each feature map (e.g., the tumor area itself rather than its surroundings).

By jointly refining both channel and spatial information, CBAM directs the network’s focus toward tumor-relevant signals across phases while suppressing background noise such as blood vessels or liver parenchyma. This dual attention is critical in CECT imaging, where tumors may exhibit subtle intensity differences relative to the background.

To enhance the interpretability of the segmentation models, Gradient-weighted Class Activation Mapping (Grad-CAM) [27, 28] was applied. The last convolutional layer of each network was identified as the target for backpropagation-based saliency generation. During inference, the predicted segmentation mask was used to define the target category (foreground vs. background). Grad-CAM was then computed by propagating gradients from the target class back through the final convolutional feature maps, producing a class-specific heatmap. These heatmaps were normalized and overlaid on the original CT images, enabling qualitative assessment of whether the model focused on clinically relevant tumor regions when making predictions.

2.3 Evaluation Metrics

In this study, both region-based and boundary-based metrics were employed to evaluate segmentation performance. Each metric captures different aspects of model effectiveness, from volumetric overlap to boundary accuracy. The following notations are used:

- P : set of pixels (or voxels) predicted as lesion (positive region).
- G : set of pixels (or voxels) belonging to the ground truth lesion.
- TP (True Positives): pixels correctly predicted as lesion.
- FP (False Positives): pixels incorrectly predicted as lesion.
- FN (False Negatives): pixels belonging to lesion but missed by the model.
- TN (True Negatives): pixels correctly predicted as background.
- $d(p, g)$: Euclidean distance between a predicted boundary point p and a ground truth boundary point g .

Dice Similarity Coefficient (DSC) evaluates the overlap between prediction and ground truth. Here, $|P \cap G|$ represents the number of pixels correctly predicted as lesion (TP), while $|P|$ and $|G|$ represent the total number of predicted and true lesion pixels, respectively. DSC balances false positives and false negatives, making it one of the most reliable metrics for medical image segmentation. A higher DSC indicates a stronger agreement with the reference annotation.

$$DSC = \frac{2 |P \cap G|}{|P| + |G|} \quad (1)$$

Intersection over Union (IoU) measures the ratio of overlap between predicted and ground truth lesion areas to their combined area. The denominator $|P \cup G|$ corresponds to all pixels labeled as lesion either by the model or in the ground truth. IoU provides a stricter evaluation compared to DSC, as even small mismatches reduce the score.

$$IoU = \frac{|P \cap G|}{|P \cup G|} \quad (2)$$

Hausdorff Distance (95%) (HD95) evaluates the similarity of the predicted and ground truth boundaries. For each boundary point p in the prediction, the minimum distance to any ground truth point g is computed, and vice versa. The 95th percentile is used instead of the maximum to reduce the effect of outliers caused by noise or annotation errors.

A lower HD95 value indicates that the model’s predicted boundary closely follows the true tumor contour, which is crucial for clinical applications.

Accuracy measures the proportion of correctly classified pixels relative to all pixels. While simple to interpret, accuracy can be misleading in medical segmentation where lesion regions are small compared to the background, as models could achieve high accuracy simply by predicting background.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision indicates the proportion of predicted lesion pixels that are correct. TP represents correctly segmented lesion pixels, while FP counts pixels wrongly identified as lesions. High precision reflects the model’s ability to avoid false positives, ensuring that detected tumor regions are likely real.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Sensitivity measures the proportion of actual lesion pixels correctly identified. A high value means the model rarely misses tumor regions, though it may include some background (false positives). In clinical contexts, high sensitivity is essential to avoid missing critical lesions.

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

Specificity quantifies the ability of the model to correctly classify background pixels. High specificity means healthy tissues are not mistakenly identified as tumor. This reduces the risk of over-segmentation, which could otherwise lead to unnecessary clinical concern.

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

By combining these complementary metrics, the evaluation framework balances volumetric accuracy (DSC, IoU, Accuracy), boundary alignment (HD95), and classification behavior (Precision, Sensitivity, Specificity). This ensures both quantitative rigor and clinical interpretability of the segmentation results.

3 Results and Discussion

3.1 Results

3.1.1 Experiment setup

All segmentation experiments were performed within the Kaggle computing environment to ensure reproducibility. Detailed specifications for the hardware and training hyperparameters are provided in Tables 2 and 3, respectively. The training regimen for all models, which include selected data augmentation techniques, utilized a consistent set of hyperparameters. It is noteworthy that while most models, including the best-performing architecture, used an input image size of 256×256 pixels (as indicated in Table 2), models incorporating a Transformer-based backbone required a smaller input resolution of 224×224 pixels due to architectural constraints.

Table 2: Kaggle experimental environment.

Processor	Intel(R)Xeon(R)CPU@2.20GHz
RAM	31.4GB
Graphics card	Tesla T4 GPU
Programming language	Python 3.11.11
Deep learning framework	Pytorch 2.6.0+cu124

The learning efficacy and stability of the model are visually documented in Fig. 5, which plots key training and validation metrics across all optimization steps. The graphs confirm that the model achieved stable convergence, as evidenced by the rapid initial decline and subsequent low, stable values across all loss functions (total_loss, loss_ce,

Table 3: Kaggle experimental environment.

Hyperparameter	Value	Hyperparameter	Value
Initial learning rate	0.01	Optimizer	SGD
Batch size	4	Epoch	100
Image size	256×256	Momentum	0.9
Max iterations	6000	Weight decay	0.0001

loss_dice). Importantly, the primary validation metrics, including the Dice coefficient (val_dice) and Intersection over Union (val_iou), show a continuous, robust improvement throughout the training duration. This simultaneous behavior substantiates the model’s ability to learn the CECT segmentation task effectively while maintaining strong generalization capability on the validation dataset.

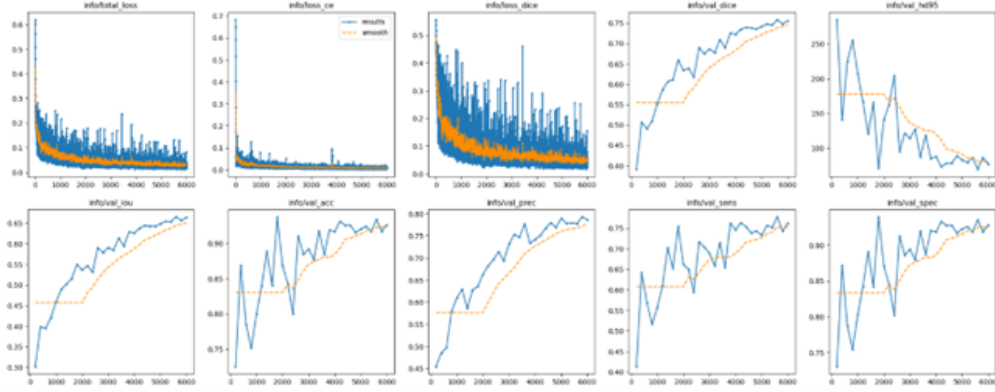


Figure 5: Training and validation process visualization of proposed model.

3.1.2 The UNet and UNet3+ with different backbones

To evaluate the effectiveness of the proposed architecture, multiple segmentation models were trained and compared on the Primary Liver Cancer CECT dataset using the evaluation metrics described earlier. Table 4 summarizes the performance of all baseline and enhanced models.

The experiments include:

- UNet based models: UNet, EfficientUNet, SwinUNet, MambaUNet.
- UNet3+ based models: MambaUNet3+, ResNetUNet3+, TransformerUNet3+.

Table 4: Performance of models on the testing set. (Note that highest and lowest scores are highlighted in blue and red fonts, respectively.)

Model	Param	Dice(DSC)	HD95	IoU	Acc	Pre	Sen	Spe
UNet	1.8M	0.506	224.710	0.491	0.786	0.563	0.532	0.789
EfficientUNet	13.2M	0.601	165.982	0.504	0.838	0.622	0.649	0.840
SwinUNet	27.6M	0.617	129.808	0.512	0.874	0.610	0.678	0.876
MambaUNet	19.1M	0.646	130.916	0.548	0.875	0.641	0.705	0.877
MambaUNet3+	36.1M	0.710	112.901	0.617	0.890	0.735	0.724	0.892
TransformerUNet3+	53.3M	0.656	159.379	0.565	0.843	0.677	0.678	0.845
ResNetUNet3+	31.1M	0.746	88.082	0.657	0.915	0.764	0.768	0.916

The original UNet achieved a Dice coefficient of 0.506, an IoU of 0.419, and a relatively high HD95 of 224.71, reflecting its limited ability to capture complex lesion boundaries in contrast-enhanced CT images. Although this baseline provided a useful point of comparison, its performance highlighted the need for more powerful architectures capable of modeling richer spatial and contextual information.

Replacing the baseline with more recent backbones demonstrated varying levels of improvement. EfficientUNet and SwinUNet achieved Dice scores of 0.601 and 0.617, respectively, showing that both convolutional and transformer-based encoders enhanced segmentation accuracy over the vanilla UNet. The MambaUNet further improved performance with a Dice of 0.646 and reduced HD95, suggesting that state-space modeling could effectively capture long-range dependencies. Notably, the MambaUNet3+ variant, which integrated the multi-scale full-scale skip connections of UNet3+, produced a substantial leap in accuracy with a Dice score of 0.710 and IoU of 0.617, demonstrating the strength of this hybrid design.

Among the tested variants, the ResNetUNet3+ consistently outperformed other backbones. With a Dice score of 0.746, IoU of 0.657, and the lowest HD95 of 88.08 at this stage, this model provided a strong balance between segmentation accuracy and boundary precision. The residual connections within ResNet, combined with the dense feature aggregation of UNet3+, appeared to enhance both low- and high-level feature representation, leading to improved lesion delineation.

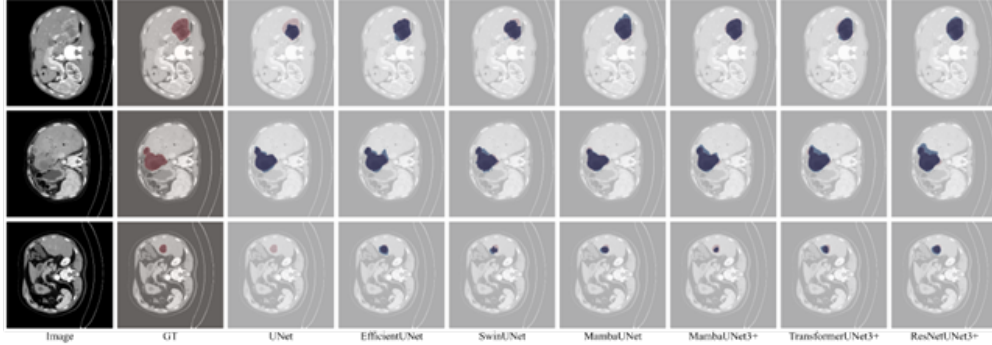


Figure 6: The visual comparison of seven segmentation methods against ground truth on the testing set. The predicted and GT are highlighted in blue and red, respectively.

Predicted masks from the best-performing model, ResNetUNet3+, show more accurate tumor boundary delineation and fewer false positives compared to other architectures. Example cases illustrate both successful segmentations and challenging scenarios, such as small lesions or lesions adjacent to vessels, where performance remained limited. Visual comparisons highlight that ResNetUNet3+ with medium size of parameters (31.1 million) is more robust in capturing tumors than baseline UNet or Mamba- and Transformer-based alternatives as shown in Fig. 6.

3.1.3 Ablation study

Based on the highest performance of ResNetUNet3+, various experiments were conducted with Attention/ASPP variants including Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), Atrous Spatial Pyramid Pooling (ASPP), and combined CBAM + ASPP. These additional modules keep the models' parameters of an acceptable value.

Table 5: Performance of ResNet Unet3+ variants on the testing set. (Note that highest scores are highlighted in blue fonts.)

Model	Param	Dice(DSC)	HD95	IoU	Acc	Pre	Sen	Spe
ResNetUNet3+	31.1M	0.746	88.082	0.657	0.915	0.764	0.768	0.916
ResNetUNet3+ with SE	31.2M	0.747	99.932	0.659	0.903	0.767	0.761	0.904
ResNetUNet3+ with CBAM	31.2M	0.755	77.911	0.662	0.925	0.768	0.777	0.926
ResNetUNet3+ with ASPP	31.3M	0.735	106.1	0.645	0.897	0.756	0.754	0.898
ResNetUNet3+ with CBAM and ASPP	31.4M	0.753	90.32	0.662	0.912	0.777	0.761	0.914

Table 5 shows introducing attention mechanisms further boosted performance. The SE-based attention improved Dice to 0.747, though boundary accuracy (HD95 = 99.93) was slightly worse than ResNetUNet3+ without attention. In contrast, the CBAM-integrated model achieved the highest overall performance, achieving a Dice of 0.755, IoU of 0.662, and the best HD95 of 77.91. These results highlight the effectiveness of jointly modeling channel- and spatial-wise dependencies, enabling the network to focus more precisely on lesion regions while suppressing irrelevant background features.

Finally, the addition of ASPP was explored both independently and in combination with CBAM. The ResNetUNet3+ with ASPP alone achieved solid performance with a Dice of 0.735 and IoU of 0.645. However, when ASPP was combined with CBAM, the results did not surpass the CBAM-only configuration, with Dice slightly decreasing to 0.753 and HD95 increasing to 90.32. This suggests that while ASPP can enhance multi-scale context aggregation, its benefits may be redundant when paired with strong attention mechanisms like CBAM, which already improve feature selectivity and spatial focus.

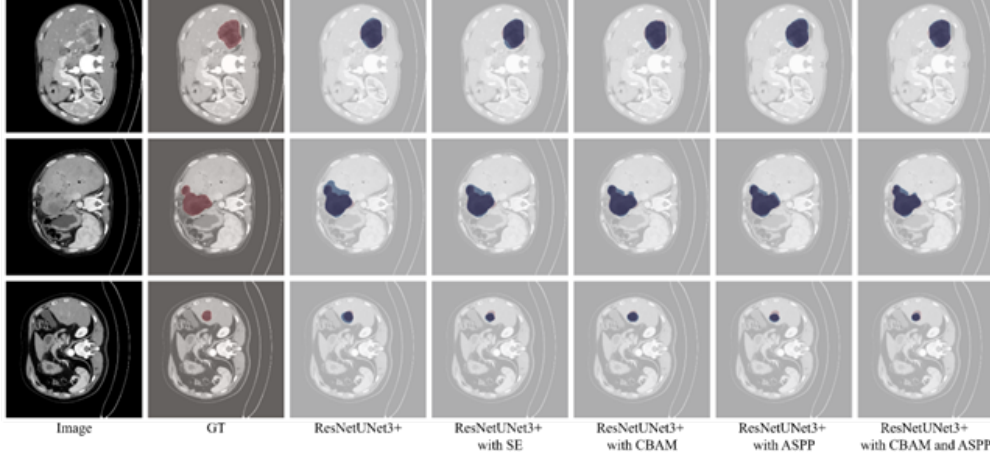


Figure 7: The visual comparison of segmentation results of ResNetUNet3+ and its variants against ground truth on the testing set. The predicted and GT are highlighted in blue and red, respectively.

Figure 7 shows a qualitative comparison of segmentation results from the baseline model and its variants, including attention modules (SE, CBAM) and a multi-scale feature extractor (ASPP), against the ground truth on three test images. The baseline model generally identifies lesions well but sometimes produces imprecise boundaries. The CBAM variant offers improved segmentation accuracy, especially in capturing irregular lesion contours, as seen in the second row. This indicates that CBAM’s spatial and channel-wise attention enhances the model’s ability to differentiate lesions from healthy tissue, resulting in more precise masks.

3.2 Discussion

The present study investigated liver tumor segmentation on multi-phase contrast-enhanced CT images using different UNet-based architectures and attention mechanisms. The comparative evaluation highlights several important insights regarding the strengths and limitations of various model configurations, as well as their potential clinical relevance.

The experimental results demonstrated clear improvements when moving from the baseline UNet to more advanced architectures. Traditional UNet achieved modest segmentation performance, reflecting its limited capacity to capture complex liver tumor boundaries in multi-phase CT data. Incorporating more powerful backbones, such as EfficientNet, Swin Transformer, and Mamba layers, consistently enhanced accuracy by enabling richer multi-scale feature extraction. However, the performance gains were most substantial when the UNet3+ framework was combined with ResNet and the Convolutional Block Attention Module (CBAM). This configuration achieved the highest Dice score and IoU, while also reducing boundary error as reflected in HD95. Importantly, the improvement was not solely due to model size, as some larger models (e.g., Swin-based UNet3+) underperformed compared to the proposed ResNet50 + CBAM design. This suggests that attention-guided feature refinement played a more crucial role than simply increasing model complexity.

To enhance interpretability, Grad-CAM visualizations were generated from the testing set (see Fig. 8). These heatmaps revealed that the proposed model consistently attended tumor regions and their surrounding contexts, confirming that the network’s predictions were guided by clinically meaningful features. In contrast, weaker models often focus attention on irrelevant background structures, explaining their higher false positive rates. The integration of Grad-CAM thus not only validates the reliability of the proposed model’s predictions but also strengthens its potential for deployment in clinical decision support. Figure 8 demonstrates that UNet-based models tend to identify two tumor regions as a single entity, whereas UNet3+-based models can accurately segment two distinct tumors. Notably, the ResNetUNet3+ model incorporating CBAM achieves boundaries most closely aligned with the ground truth.

Although quantitative metrics indicate that ResNetUNet3+ with CBAM is the leading model, qualitative evaluation of its failure cases highlights notable shortcomings, especially regarding the segmentation of small tumors as illustrated in Fig. 9. This issue is particularly significant given the dataset’s pronounced bias towards small lesions (refer to Fig. 2).

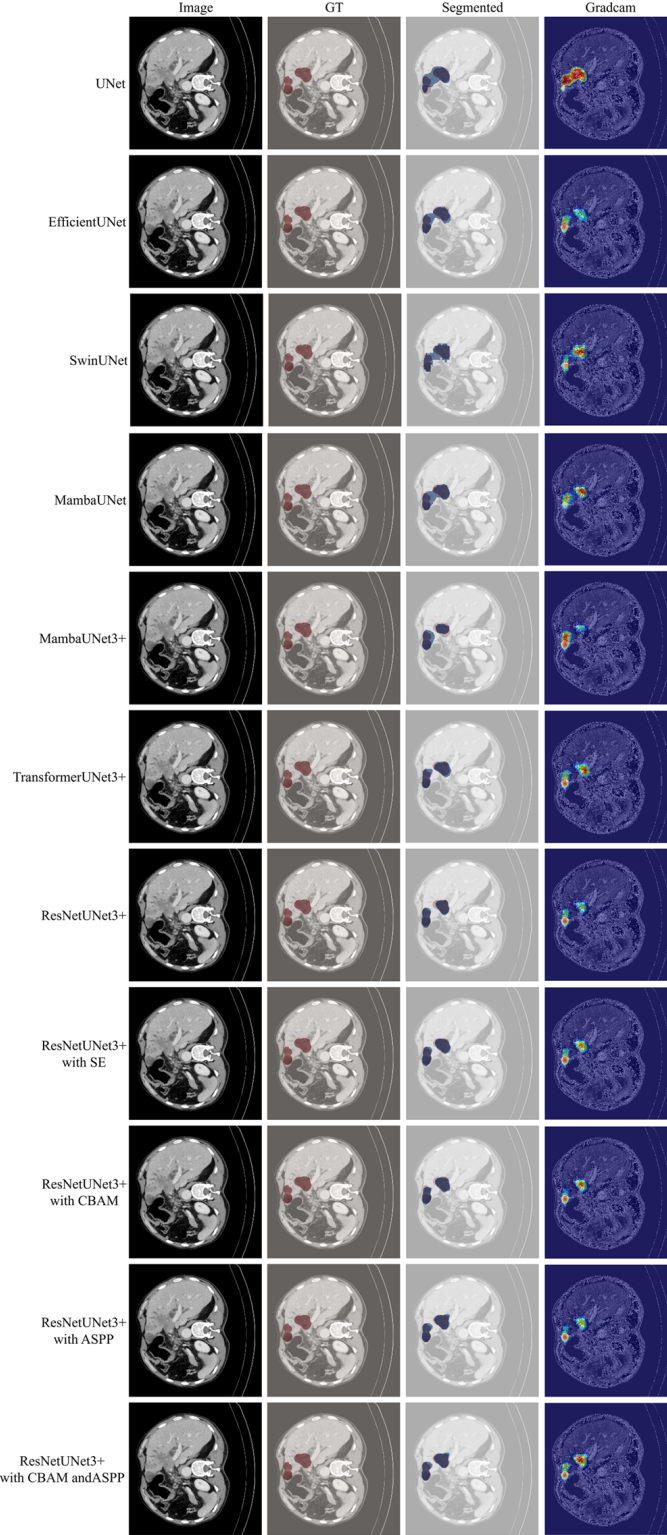


Figure 8: Sample visual explanations of all models on the testing set.

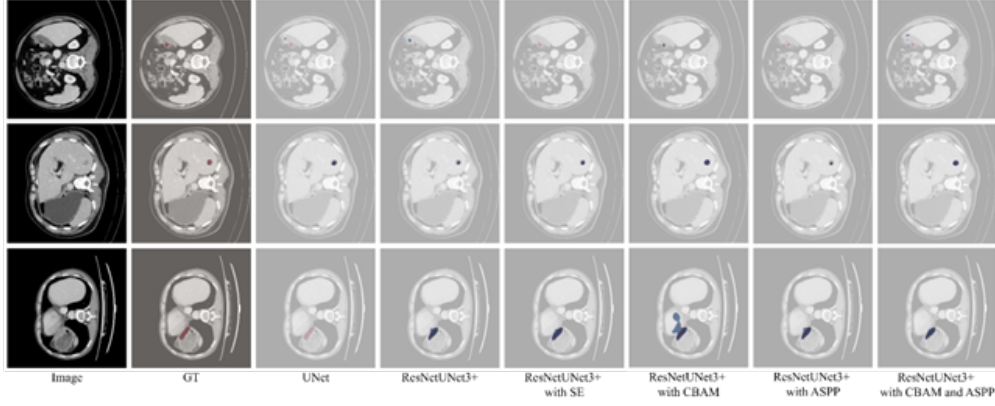


Figure 9: Some failure cases.

The error analysis highlights key obstacles encountered in segmenting small tumors (see Fig. 9):

- **False negatives (missed tumors):** The initial row of the error plot presents a scenario where a diminutive lesion exists in the ground truth. In this case, nearly all models, including the baseline, SE, and ASPP variants, fail to detect the lesion. Notably, the ResNetUNet3+ with CBAM model demonstrates partial effectiveness by detecting the approximate center of the lesion with a small mark, albeit without capturing the entire region. This outcome suggests that its attention mechanism possesses heightened sensitivity for anomaly localization, whereas other models do not recognize the lesion at all.
- **Under-Segmentation of small lesions:** The second row illustrates another challenge. All models identify the general region of the tumor; however, the baseline, SE, and ASPP variants exhibit pronounced under-segmentation, delineating only a limited portion of the lesion. By contrast, both ResNetUNet3+ with CBAM and the CBAM and ASPP variant effectively segment the lesion in its entirety. This finding highlights their superior capability to delineate the complete shape and extent of small tumors upon successful localization, thereby mitigating the under-segmentation observed with alternative approaches.
- **Arbitrary shape of tumor regions:** The third row demonstrates the segmentation error of the ResNetUNet3+ model with CBAM when addressing tumors with non-circular, irregular shapes. Compared to other methods such as ResNetUNet3+ and its variants with SE and ASPP, the proposed model tends to segment a larger region. Notably, ResNetUNet3+ with both CBAM and ASPP achieves the highest performance by accurately delineating the tumor region. This observation accounts for the highest Precision score attained by the ResNetUNet3+ with CBAM and ASPP, as reported in Table 5.

In summary, the ResNetUNet3+ with CBAM model achieves the highest overall performance, yet robust detection and segmentation of the smallest lesions remain challenging. Failures are primarily characterized either by completely missed detections (as indicated in row 1 of Fig. 9, with CBAM providing some sensitivity) or significant under-segmentation by less advanced models (as illustrated in row 2 of Fig. 9). The CBAM-equipped models consistently demonstrate considerable advantages in both lesion identification and boundary accuracy compared to the baseline. However, in certain cases involving non-circular tumor morphologies, the proposed model may exhibit an increased rate of false positives (as shown in row 3 of Fig. 9).

Despite encouraging segmentation results, several limitations warrant consideration. First, the dataset originates from a single institution, which may constrain the broader applicability to other populations or imaging settings. Second, although segmentation accuracy has improved, boundary delineation errors persist for small or low-contrast lesions, indicating the necessity for further advancements in feature representation. Lastly, while attention mechanisms have enhanced performance, integrating them with modules such as ASPP did not yield additional gains, suggesting the need for more sophisticated integration strategies.

Overall, this study establishes that utilizing multi-phase CT data in conjunction with a UNet3+ backbone, reinforced by ResNet50 and CBAM attention mechanisms, delivers superior outcomes for liver tumor segmentation. The inclusion of qualitative examples and Grad-CAM-based interpretability further supports the reliability and clinical relevance of the proposed model in tumor detection and treatment planning.

4 Conclusion

The study explored liver tumor segmentation on CT images using UNet- and UNet3+-based architectures with diverse backbone configurations. Through comprehensive experiments, it was shown that while modern Transformer- and Mamba-based backbones improved feature representation, the combination of UNet3+ with a ResNet backbone and CBAM attention provided the most consistent and superior performance across all evaluation metrics. The proposed model not only achieved higher Dice and IoU scores, but also reduced boundary errors, indicating its robustness in delineating complex liver tumor structures.

Beyond quantitative improvements, the inclusion of Grad-CAM visualizations improved the interpretability of the model, highlighting its ability to focus on clinically relevant tumor regions. These results underscore the potential of attention-augmented UNet3+ models in advancing computer-aided liver tumor detection and segmentation, which can contribute to more accurate diagnosis and treatment planning.

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was supported by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant DS2023-28-02. The funder had no role in study design, data collection and analysis, decision to publish or manuscript preparation.

Data availability statement

Data supporting the findings of this article are publicly available at <https://doi.org/10.57760/sciencedb.12207> [14].

ORCID iD

Doan-Van-Anh Ly: <https://orcid.org/0009-0006-6898-0545>

Thi-Thu-Hien Pham: <https://orcid.org/0000-0001-5808-3214>

Thanh-Hai Le: <https://orcid.org/0000-0002-3212-3940>

References

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. Accessed October 12, 2025.
- [2] International Agency for Research on Cancer. Vietnam fact sheet. <https://gco.iarc.who.int/media/globocan/factsheets/populations/704-viet-nam-fact-sheet.pdf>, 2022. Accessed October 12, 2025.
- [3] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie. From CNN to transformer: A review of medical image segmentation models. *J Digit Imaging. Inform. med.*, 37:1529–1547, 2024.
- [4] K. S. Sheela, V. Justus, R. R. Asaad, and R. L. Kumar. Enhancing liver tumor segmentation with UNet-ResNet: Leveraging ResNet’s power. *Technology and Health Care*, 33(1):1–15, 2025.
- [5] T. Banerjee, D.P. Singh, P. Kour, D. Swain, S. Mahajan, S. Kadry, and J. Kim. A novel unified Inception-U-Net hybrid gravitational optimization model (UIGO) incorporating automated medical image segmentation and feature selection for liver tumor detection. *Scientific Reports*, 15:29908, 2025.
- [6] H. Jiang, T. Shi, Z. Bai, and L. Huang. AHCNet: An application of attention mechanism and hybrid connection for liver tumor segmentation in CT volumes. *IEEE Access*, 7:24898–24909, 2019.
- [7] R.K. Debnath, M.A. Rahman, S. Azam, Y. Zhang, and M. Jonkman. FSS-ULivR: a clinically-inspired few-shot segmentation framework for liver imaging using unified representations and attention mechanisms. *J Cancer Res Clin Oncol*, 151:215, 2025.

- [8] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M.P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.
- [9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Karlinsky, L., Michaeli, T., Nishino, K. (eds) Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science*, volume 13803, 2023.
- [10] J. Ma, F. Li, and B. Wang. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprints arXiv: 2401.04722*, 2024. Accessed July 05, 2025.
- [11] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li. Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation. *arXiv preprints arXiv: 2402.05079*, 2024. Accessed July 05, 2025.
- [12] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu. SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. In *Linguraru, M.G., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. Lecture Notes in Computer Science*, volume 15008, 2024.
- [13] X. Zhong, G. Lu, and H. Li. Vision Mamba and xLSTM-UNet for medical image segmentation. *Sci Rep*, 15:8163, 2025.
- [14] J. Luo, X. Wan, J. Du, L. Liu, L. Zhao, X. Peng, M. Wu, S. Huang, and X. Nie. Comprehensive multi-phase 3D contrast-enhanced CT imaging for primary liver cancer. *Sci Data*, 12:768, 2025.
- [15] C. Hu, T. Xia, Y. Cui, Q. Zou, Y. Wang, W. Xiao, S. Ju, and X. Li. Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty. *Engineering Applications of Artificial Intelligence*, 133:108289, 2024.
- [16] P.V. Nayantara, S. Kamath, R. Kadavigere, and K.N. Manjunath. Automatic liver segmentation from multiphase CT using modified SegNet and ASPP module. *SN Comput Sci*, 5:377, 2024.
- [17] V. Lamprou, A. Kallipolitis, and I. Maglogiannis. On the evaluation of deep learning interpretability methods for medical images under the scope of faithfulness. *Computer Methods and Programs in Biomedicine*, 253:108238, 2024.
- [18] K. Vinogradova, A. Dibrov, and G. Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(10), pages 13943—13944, 2020.
- [19] S.N. Hasany, C. Petitjean, and F. Mériaudeau. Seg-XRes-CAM: Explaining spatially local regions in image segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada*, pages 3733–3738, 2023.
- [20] R.L. Draelos and L. Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprints arXiv: 2011.08891*, 2020. Accessed July 10, 2025.
- [21] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, volume 9351, 2015.
- [22] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. UNet3+: A full-scale connected Unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain*, pages 1055–1059, 2020.
- [23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pages 770–778, 2016.
- [24] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon. CBAM: Convolutional block attention module. In *Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, volume 11211, 2018.
- [25] B.T. Lieu, C.K. Nguyen, H.L. Nguyen, and T.H. Le. Enhanced small-object detection in UAV images using modified YOLOv5 model. *IET Image Processing*, 19(1):e70121, 2025.
- [26] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 7132–7141, 2018.
- [27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336—359, 2019.
- [28] J. Gildenblat and contributors. PyTorch library for CAM methods. <https://github.com/jacobgil/pytorch-grad-cam>, 2021. Accessed July 10, 2025.