

Liver Segmentation Using Modified CAG-SwinUNet with Explainability

Kumar S. S.*, Vinod Kumar R. S., and Ranjith V. G.

Noorul Islam Centre for Higher Education, Kumarakoil 629180, India

e-mail: kumarss@live.com

Abstract. Accurate liver segmentation from computed tomography (CT) images is crucial for clinical applications such as tumor detection and surgical planning but remains challenging due to anatomical complexity and imaging variability. Existing deep learning models, struggle with ambiguous liver boundaries, noise sensitivity, and weak feature integration across scales, leading to segmentation errors. This study introduces Cross-Attention Gate-Shifted Window U-Net (CAG-SwinUNet), an enhanced Swin-UNet variant that incorporates a Cross-Attention Gate (CAG) in skip connections to selectively refine feature fusion. Unlike traditional concatenation, CAG dynamically enhances encoder features based on decoder context, integrating residual connections and output projection to balance local and global information. Extensive evaluation on Liver Tumor Segmentation (LiTS) and Segmentation of the Liver Competition 2007 (SLIVER07) demonstrates state-of-the-art performance, achieving 97.75% Dice Similarity Coefficient (DSC) and 2.40 mm Hausdorff Distance (HD) on LiTS, and 96.65% DSC and 3.10 mm HD on SLIVER07, respectively. To enhance explainability, gradient-weighted class activation mapping, provide visual insights into the model's decision-making process, ensuring transparency and reliability in liver segmentation.

Keywords: liver segmentation; transformer model; SwinUNet; CAG-SwinUNet; XAI.

Paper #9274 received 11 May 2025; revised manuscript received 14 Jan 2026; accepted for publication 17 Jan 2026; published online 2 Mar 2026. doi: [10.18287/JBPE26.12.010303](https://doi.org/10.18287/JBPE26.12.010303).

1 Introduction

Liver segmentation from Computed Tomography (CT) is fundamental to clinical applications such as tumor detection, surgical planning, and treatment monitoring. Accurate delineation enables precise volumetric analysis and identification of pathological regions, directly influencing diagnostic accuracy and therapeutic decisions [1]. However, segmentation remains challenging due to the liver's complex anatomy, characterized by irregular boundaries, internal heterogeneity from vascular structures and lesions, and significant inter-patient variability. These challenges are further compounded by imaging-related factors, including noise, artifacts, and variations in acquisition protocols, making consistent segmentation across diverse conditions difficult.

Traditional liver segmentation approaches have relied on manual and semi-automated techniques. Manual delineation by radiologists, is time-consuming, prone to inter-observer variability, and impractical for large

datasets [2]. Early automated methods, such as intensity-based thresholding [3], exploit the liver's distinct Hounsfield Unit (HU) range but struggle with overlapping intensities from adjacent organs and imaging noise, necessitating extensive post-processing. Region growing [4] improves upon thresholding by expanding seed points based on intensity similarity but suffers from sensitivity to initialization and boundary leakage, particularly in heterogeneous livers. Deformable models, such as active contours [5] and level sets [6], adapt to liver edges using shape priors and energy minimization, offering flexibility for irregular boundaries. However, they fail in low-contrast regions or significant pathological alterations like tumors, often requiring manual parameter tuning, which limits scalability.

The emergence of deep learning has transformed liver segmentation, with Convolutional Neural Networks (CNNs) [7] setting new benchmarks. U-Net [8], a groundbreaking encoder-decoder architecture with skip connections, enables multi-scale feature fusion and has

demonstrated strong performance in medical image segmentation. However, its limited receptive field hinders the modeling of long-range dependencies, often resulting in blurred boundaries or adjacent tissue inclusion in complex organ segmentation. Variants such as UNet++ [9] and Attention UNet [10] attempt to refine feature aggregation and attention mechanisms, yet CNN-based models remain constrained in capturing global contextual information, which is critical for accurately delineating the liver's complex morphology.

To address these limitations, transformer-based [11, 12] architectures have gained traction, employing self-attention mechanisms to model long-range dependencies more effectively. SwinUnet [13], a hybrid of U-Net and Swin Transformers [14], employs windowed multi-head self-attention (W-MSA) for improved computational efficiency and contextual awareness. By processing images as patch-based token sequences and using hierarchical feature extraction, SwinUnet captures both fine-grained details and broader anatomical context like organ shapes. However, its skip connections, which rely on direct concatenation, indiscriminately merge encoder and decoder features, introducing noise and irrelevant details mainly surrounding organ textures, thereby compromising boundary precision.

To overcome these challenges, this study presents CAG-SwinUnet, an enhanced SwinUnet variant integrating a Cross-Attention Gate (CAG) mechanism within skip connections. Unlike standard concatenation, CAG employs cross-attention to dynamically weight encoder features based on decoder context, ensuring the selective fusion of liver-relevant information while suppressing extraneous signals. Additionally, the proposed CAG incorporates residual connections, departing from conventional attention gate designs, to preserve decoder context, thereby enhancing segmentation robustness across diverse imaging conditions.

Beyond segmentation accuracy, explainability remains a critical concern in medical AI applications. The black-box nature of deep learning models limits their clinical adoption, as practitioners require transparency in decision-making. To address this, Explainable AI (XAI) [15] techniques, specifically Gradient-weighted Class Activation Mapping (Grad-CAM) [16], are integrated to visualize model attention during segmentation. Grad-CAM generates heatmaps highlighting critical regions, providing interpretable insights into feature importance and ensuring that the model prioritizes relevant liver structures while ignoring non-liver regions. This enhances trust and usability for clinical deployment.

This study introduces CAG-SwinUnet, a Transformer-based segmentation model designed to enhance both accuracy and interpretability in liver segmentation. The key contributions include:

- CAG in skip connections – a novel mechanism that selectively fuses encoder and decoder features,

improving boundary delineation and reducing feature contamination from irrelevant structures.

- Residual-Enhanced attention – unlike conventional attention gates, the CAG design incorporates residual connections, ensuring preservation of decoder context and refining segmentation across heterogeneous imaging conditions.

- State-of-the-Art performance – extensive evaluation on LiTS and SLIVER07 datasets demonstrates improved Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD), surpassing previous Transformer-based methods.

- Explainable AI with Grad-CAM – integration of Grad-CAM-based visualization to provide interpretable model predictions, enabling clinicians to validate model reliability and decision-making transparency.

- Ablation study on feature contributions – a detailed ablation study quantifies the impact of cross-attention, residual connections, and output projection, confirming their role in refining segmentation precision.

This paper is structured as follows: Section 2 reviews conventional and deep learning-based liver segmentation approaches, highlighting key challenges. Section 3 introduces the CAG-SwinUnet architecture, detailing the CAG and its role in skip connections. Section 4 presents experimental evaluations, including implementation details, quantitative and qualitative analyses, and an ablation study assessing CAG's contributions. Section 5 discusses the findings, comparing CAG-SwinUnet with existing methods and analyzing its clinical relevance and explainability. Finally, Section 6 summarizes key contributions and outlines potential future directions.

2 Related Work

Liver segmentation from CT images has seen significant progress through deep learning models, particularly CNNs, Transformers, Generative Adversarial Networks (GANs) [17], and U-Net-based architectures. Each of these approaches contributes to segmentation performance, but challenges persist in terms of computational complexity, boundary delineation, and feature representation.

Transformers have become a powerful tool for liver segmentation due to their ability to model long-range dependencies. High-Resolution Swin Transformer (HRSTNet) [18] replaces standard convolutional layers with transformer blocks, ensuring a continuous information flow across multi-resolution feature maps. UNeTr [19] follows a U-shaped architecture, integrating a Transformer-based encoder and decoder with skip connections for improved semantic segmentation. AD-DUNet [20] employs Axial Transformers to capture long-range dependencies while leveraging cascaded dilated convolutions for local feature extraction. ResTransUNet [21] further refines this hybrid approach by integrating CNNs with Transformers to mitigate feature loss during encoding. LGMA-Net [22] introduces an SNP convolutional Transformer block that balances

local feature extraction with global attention, enhancing segmentation accuracy in complex liver structures.

GANs have been explored for liver segmentation, particularly to address data scarcity and enhance feature representation. U-Net GAN [23] applies adversarial training to improve segmentation quality within a semi-supervised learning framework, enabling effective training even with limited labeled data. GAN Mask R-CNN [24] combines the strengths of GANs with Mask R-CNN, using anchor selection and k-means clustering to refine feature extraction and reduce segmentation errors in noisy CT images. While these approaches improve generalization, they often suffer from training instability and mode collapse, limiting their reliability in clinical applications.

Residual networks have been widely employed to improve feature propagation and ease optimization in deep segmentation models. ResU-Net [25] replaces traditional convolution modules with residual blocks, facilitating better gradient flow and feature reuse. Modified ResUNet [26] extends this approach by integrating deeper residual learning components. RDCTrans U-Net [27] combines ResNeXt50 with dilated convolutions, using residual learning while improving feature extraction efficiency. However, despite their advantages, residual networks often struggle to balance fine-grained feature preservation with global context modeling, impacting segmentation performance in cases with complex liver boundaries or low-contrast regions.

Attention mechanisms have been integrated into deep learning architectures to improve segmentation accuracy by focusing on the most relevant features. DRAUNet [28] employs a dual-effect attention module, combining spatial and channel attention to refine feature selection. GCHA-Net [29] introduces a hybrid attention network, incorporating a Global Attention Module for long-range dependency modeling and a Local Attention Module for preserving fine details. SAR-U-Net [30] integrates Squeeze-and-Excitation blocks to enhance channel-wise feature recalibration, allowing the model to suppress irrelevant information while emphasizing critical liver structures. While attention-based models significantly enhance segmentation quality, they also increase model complexity and can lead to overfitting when trained on limited datasets.

Several approaches have focused on multi-scale feature extraction to improve segmentation across varying liver sizes and shapes. Eres-UNet++ [31] employs deep supervision and nested residual blocks to capture hierarchical features effectively. LiM-Net [32] introduces a multi-level feature fusion mechanism using a Res2Net backbone, enhancing spatial detail preservation. SACU-Net [33] uses shape-aware skip pathways to bridge the semantic gap between encoder and decoder representations, leading to improved boundary delineation. DEMF-Net [34] incorporates a Multi-Scale Feature Fusion (MSFF) module within its skip connections, capturing both fine details and global structure simultaneously. However, multi-scale networks often introduce additional computational overhead and

struggle to maintain sharp boundary details in cases with small or irregular liver lesions.

Deep learning models have significantly improved liver segmentation, but several challenges remain. Feature representation inconsistencies lead to errors in boundary delineation, especially in cases with complex liver structures or low-contrast regions. Many models struggle to balance local detail preservation with global context, impacting segmentation robustness. Training stability, particularly in adversarial and attention-based methods, affects reliability, while hybrid architectures introduce feature misalignment, leading to segmentation inconsistencies. Additionally, the integration of multiple feature extraction mechanisms sometimes results in redundant or conflicting representations, reducing overall efficiency. Addressing these limitations requires improved feature fusion strategies, adaptive learning techniques, and enhanced generalization mechanisms for robust liver segmentation.

To address these challenges, CAG-SwinUnet builds upon SwinUnet [13] by introducing a CAG mechanism within the skip connections. Earlier studies have already explored the concept of Gated Attention [35], Cross-Attention [36] and Gated Cross-Attention [37] mechanisms for selective feature fusion, demonstrating its effectiveness in visual tasks. Building on these concepts, our proposed CAG-SwinUNet adapts and modifies cross-attention gating specifically for hierarchical Swin Transformer-based liver segmentation. Unlike conventional concatenation-based skip pathways that indiscriminately merge encoder and decoder features, CAG selectively refines feature fusion by dynamically weighing encoder features based on decoder queries. This ensures that relevant liver-specific information is emphasized while suppressing noise and irrelevant textures from surrounding organs. Additionally, the residual connection in CAG preserves contextual continuity, striking a balance between global context modeling and local detail retention.

3 Methodology

The proposed model introduces, CAG-SwinUnet a novel CAG mechanism within the skip connection pathway, replacing concatenation with a selective feature fusion process. The CAG utilizes cross-attention to dynamically weight encoder features based on decoder queries, emphasizing liver-specific information while suppressing extraneous data. Complementing this, the Swin Transformer blocks employ W-MSA and its shifted-window variant (SW-MSA) to model spatial relationships within feature maps, enhancing local and global feature extraction. Together, these dual attention mechanisms, W-MSA for intra-feature refinement and CAG for inter-feature fusion, improve segmentation precision and adapt to the liver's anatomical complexity.

3.1 CAG-SwinUnet Architecture

CAG-SwinUnet enhances the SwinUnet framework by preserving its encoder-decoder structure, which relies on

Swin Transformer blocks for hierarchical feature processing, while introducing a novel CAG mechanism to refine skip connections. This architecture is meticulously crafted to segment liver tissue from 2D medical images, transforming an input image $I \in R^{(H \times W \times C)}$ (where H and W denote height and width, and C is the number of channels), into a binary segmentation mask $O \in R^{(H \times W \times 1)}$, that assigns pixel-wise probabilities to distinguish liver regions from surrounding tissues. The process unfolds in a structured sequence: the encoder first converts the input image into a sequence of tokens, which are refined through a bottleneck stage, then upsampled and reconstructed by the decoder. At each decoder stage, CAG modules integrate features from the encoder's skip connections with precision, followed by additional Swin Transformer blocks that polish the representation further. The workflow concludes with a final layer that generates the segmentation output.

The proposed architecture provided in Fig. 1 illustrates the complete pipeline, starting with the input

image I entering the encoder, which comprises multiple stages, typically four, each featuring patch embedding, Swin Transformer blocks, and patch merging operations to progressively downsample the feature maps into a hierarchical representation. The deepest features are processed by the bottleneck, depicted as a compact set of Swin Transformer blocks. The decoder mirrors this structure in reverse, employing patch expanding layers to upsample the features, with CAG modules integrating skip connections from corresponding encoder stages, followed by additional Swin Transformer blocks for refinement. The final output layer, with sigmoid activation, generates the segmentation mask O . Arrows in the figure highlight the flow of skip connections from the encoder to the decoder via CAG modules, emphasizing their critical role in feature fusion, while the hierarchical nature of the Swin Transformer blocks is underscored by the multi-scale feature maps at each stage, visually capturing the transition from high-resolution local details to low-resolution global context and back.

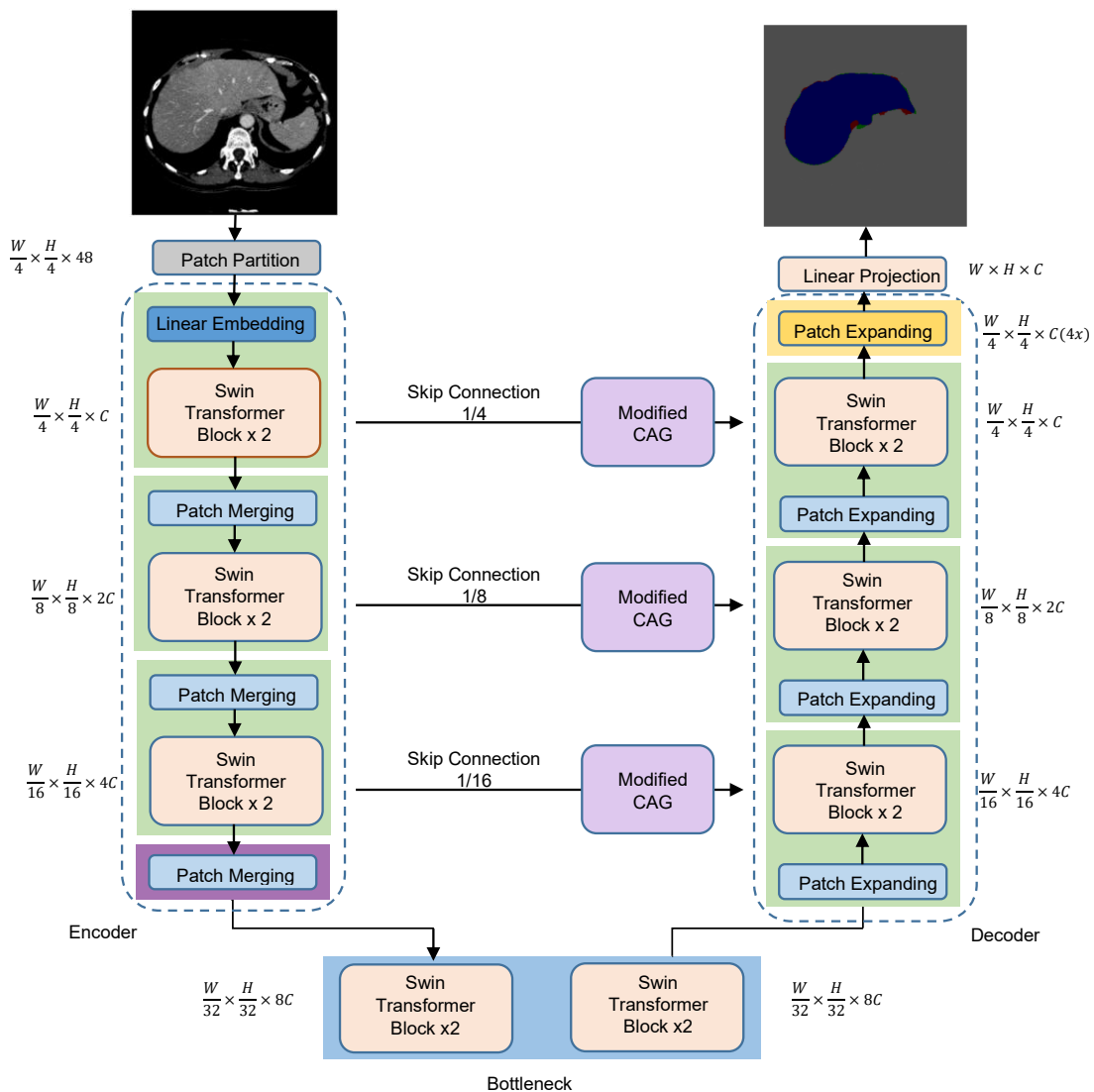


Fig. 1 Overview of the CAG-SwinUnet architecture utilizing Swin Transformer blocks for feature extraction and CAG modules for refined liver segmentation.

3.1.1 Encoder Structure

The encoder's primary function is to extract a multi-scale feature representation, capable of capturing both intricate local details, such as the branching patterns of hepatic veins, portal veins, or the heterogeneous textures of lesions like tumors or cysts and broader global anatomical context, including the liver's overall shape, its irregular contours, and its spatial relationships with adjacent organs such as the spleen, stomach, or diaphragm. This dual capability is essential given the liver's complex and highly variable anatomy, which can differ significantly across patients due to pathological conditions (e.g., cirrhosis, fatty liver disease, or hepatocellular carcinoma) or natural anatomical diversity. The feature extraction process begins with patch extraction, where the input image is divided into a grid of non-overlapping patches, each of size $P \times P$, with P being a configurable hyperparameter. Each patch encapsulates a localized region, potentially containing boundary edges, vessel segments, or lesion fragments and is flattened into a vector. The dimensionality of this vector is determined by the patch area (P^2 , the number of pixels in the patch) multiplied by the number of channels (C), resulting in an initial sequence of tokens denoted as X_0 . This tokenization reduces the spatial resolution from $H \times W$ to a coarser grid of $H/P \times W/P$, enabling efficient Transformer-based processing while preserving the structural integrity of liver features like sharp boundaries or delicate vascular networks.

These tokens undergo a patch embedding step, where a linear projection, maps each flattened patch vector to a fixed embedding dimension, augmented with positional encodings to retain critical spatial information lost during tokenization. This step is vital for Transformer-based models, which lack the convolutional inductive biases of CNNs that naturally preserve spatial relationships. By incorporating positional encodings, the model can distinguish between patches near the liver's periphery and those in its central regions, a distinction crucial for accurately segmenting the liver's irregular contours and internal structures. The embedding dimension is carefully selected to balance representational richness, necessary to capture the diverse features of liver tissue, with computational efficiency, ensuring practical training and inference times on modern hardware.

The encoder progresses through multiple stages, each comprising a series of Swin Transformer blocks followed by a patch merging operation to downsample the feature map progressively. Within each Swin Transformer block, token representations are refined through a detailed sequence of operations. LayerNorm (LN) normalizes the input features, stabilizing training by mitigating issues like vanishing or exploding gradients. This normalized input is processed by W-MSA, a hallmark of the Swin Transformer that computes self-attention within confined local windows of tokens rather than globally. The input is projected into query (Q_s), key (K_s), and value (V_s) matrices and the attention mechanism is computed within each window:

$$\text{Attention}(Q_s, K_s, V_s) = \text{SoftMax}(Q_s K_s^T / \sqrt{d} + B) V_s. \quad (1)$$

This localized attention excels at modeling fine-grained dependencies, such as the texture of hepatic veins or subtle boundary transitions. Outputs from all heads are concatenated and projected back using a linear layer.

In alternating blocks, SW-MSA shifts the window to enable cross-window interactions, allowing tokens in adjacent windows to attend to one another and capture longer-range dependencies, for instance, relating features between the liver's left and right lobes or across the liver-spleen boundary, enriching global anatomical understanding. The W-MSA or SW-MSA output is integrated via a residual connection given in Eq. (2):

$$\hat{z}^l = W_MSA(LN(z^{l-1})) + z^{l-1}, \quad (2)$$

followed by a two-layer MLP with GELU activation given in Eq. (3):

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l. \quad (3)$$

Next, z^l is processed by a multi-layer perceptron that utilizes GELU non-linearity, following a Layer Norm step. The output is then summed with \hat{z}^l , enhancing the information flow as described in Eq. (4):

$$\hat{z}^{l+1} = SW_MSA(LN(z^l)) + z^l. \quad (4)$$

For the second block, the output \hat{z}^{l+1} is obtained using the SW-MSA module applied to the Layer Norm-transformed z^l , along with adding the residual connection z^l as in Eq. (5):

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}. \quad (5)$$

Finally, \hat{z}^{l+1} is processed through another MLP following Layer Norm, and the resulting output is summed with \hat{z}^{l+1} as given in Eq. (5). The MLP typically expands the dimension then contracts it back, with residual connections ensuring training stability by facilitating gradient flow. These operations ensure robust capture of local details (e.g., vessel textures, lesion heterogeneity) and global context.

After each stage (except the deepest), patch merging downsamples the feature map by concatenating features from neighboring patches. A linear layer reduces this to a manageable size, creating a hierarchical representation transitioning from fine local details in early stages to coarse global context in deeper stages, optimizing efficiency while retaining liver-specific information.

3.1.2 Bottleneck

The bottleneck serves as a pivotal bridge between the encoder and decoder, comprising a small number, typically two, of consecutive Swin Transformer blocks operating on the deepest encoder features X_L , which have the lowest spatial resolution but the highest channel depth. These blocks apply W-MSA and SW-MSA, as in Eqs. (1–5), to refine the representation without altering resolution or dimension. This compact stage avoids

overfitting and convergence issues common in deep Transformer architectures, enhancing complex liver structures, such as lesions with heterogeneous intensities or intricate vascular networks, through W-MSA and SW-MSA's spatial modeling, preparing features for decoder upsampling.

3.1.3 Decoder Structure

The decoder reconstructs the segmentation mask from bottleneck features, employing patch expanding layers to upsample the spatial resolution while reducing the feature dimension. A linear layer adjusts the feature dimension, followed by a reshape that doubles the spatial resolution per stage, aligning decoder features with encoder stages for skip connection integration. This convolution-free approach avoids smoothing artifacts from convolutional upsampling, preserving fine details like edges. The CAG mechanism, detailed below, then integrates these features with encoder skip connections, followed by Swin Transformer blocks for further refinement.

3.1.4 Cross-Attention Gate (CAG)

The CAG mechanism integrates upsampled decoder features X'_{l+1} with encoder skip features S_l using cross-attention, offering a selective and efficient alternative to concatenation, adeptly handling the liver's anatomical variability. Figure 2 provides a visual breakdown of the CAG. The decoder features X'_{l+1} , representing a coarse liver reconstruction like approximate shape or major lobe outlines, are fed into a query projection, while encoder features S_l , rich with multi-scale details like boundary edges, and lesion textures, are split into key and value projections. The process unfolds as given in Eq. (6):

$$\begin{aligned} Q_c &= W_{qc} LN(X'_{l+1}), \\ K_c &= W_{kc} S_l, \\ V_c &= W_{vc} S_l, \end{aligned} \quad (6)$$

where $W_{qc}, W_{kc}, W_{vc} \in R^{D_l \times D_l}$ are learnable matrices tailored to the feature dimension D_l at the current stage, and c denotes cross-attention. Queries Q_c guide the attention process, seeking relevant encoder details, while keys K_c and values V_c supply these multi-scale features from the encoder. Attention scores are computed as follows in Eq. (7):

$$\begin{aligned} A &= SoftMax(Q_c K_c^T / \sqrt{D_l}), \\ A &\in R^{N_l \times N_l}, \end{aligned} \quad (7)$$

where $Q_c K_c^T \in R^{N_l \times N_l}$ is a similarity matrix, with each entry $(Q_c K_c^T)_{ij}$ measuring alignment between the i -th decoder token and the j -th encoder token, scaled by $\sqrt{D_l}$

for gradient stability. Softmax normalizes these into weights A_{ij} , focusing on liver-specific regions mainly boundary transitions at the liver-spleen interface or hepatic veins while suppressing irrelevant areas like stomach, diaphragm, or imaging artifacts like noise. The attention applied output given in Eq. (8) is obtained by enhancing X'_{l+1} with precise details from S_l , as depicted in Fig. 2's output flow, showing the refined feature map emerging from the cross-attention operation.

$$X_{attn} = AV_c, \quad X_{attn} \in R^{N_l \times D_l}. \quad (8)$$

A residual connection integrates this as given in Eq. (9):

$$X'_l = W_0(X_{attn} + X'_{l+1}), \quad W_0 \in R^{D_l \times D_l}, \quad (9)$$

where W_0 aligns dimensions and refines the fused features, balancing the decoder's global context with the encoder's local details for accurate segmentation. Unlike SwinUnet's concatenation, which doubles the feature dimension and requires additional projection, CAG maintains D_l , reducing overhead while improving focus on subtle transitions. Following CAG, Swin Transformer blocks refine the fused features using W-MSA and SW-MSA to capture local dependencies and long-range interactions between lobes, preserving fine details.

3.1.5 Output Layer

The final decoder output is reshaped to $H \times W$ and activated with sigmoid activation, X_0 is the final refined feature map, and the output provides pixel-wise probabilities. Sigmoid suits binary segmentation, allowing flexible thresholding for clinical needs.

CAG-SwinUnet's dual attention mechanisms ensure precise liver segmentation, where the encoder's hierarchical structure and Swin blocks capture multi-scale features, while CAG fuses them selectively. Its convolution-free design preserves boundary clarity, surpassing traditional methods across diverse liver conditions.

Although the model is built upon existing Transformer and attention components, its novelty lies in the design of a modified CAG integrated into the skip connections of SwinUNet. Unlike standard concatenation, the proposed CAG performs decoder-guided cross-attention to selectively emphasize liver-specific encoder features while suppressing irrelevant structures. The gate further incorporates a residual enhancement path and an output projection layer, ensuring stable feature fusion and consistent dimensionality without additional computational overhead. A stage-adaptive formulation is applied so that CAG behavior aligns with shallow, mid-level, and deep skip connections, improving the balance between fine-grained detail and global context. Combined with Grad-CAM-based explainability, these modifications distinguish CAG-SwinUNet from existing architectures and directly contribute to its improved segmentation accuracy.

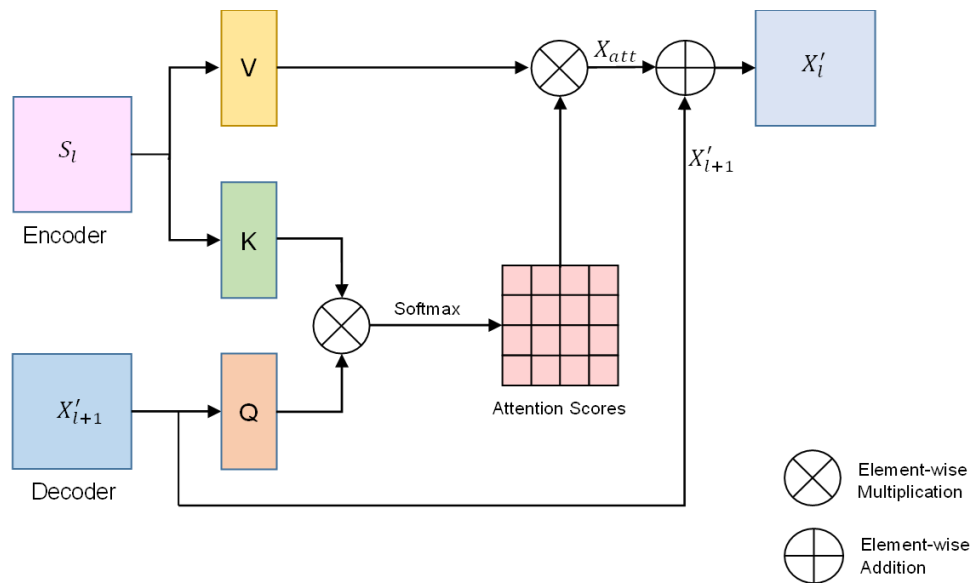


Fig. 2 Detailed structure of the Cross-Attention Gate employed for liver segmentation.

While attention mechanisms have been explored in segmentation architectures, the CAG module provides distinct enhancements tailored for liver segmentation within a Swin Transformer framework. Unlike Attention U-Net [35], which applies additive attention gates using decoder-derived grid gating signals to rescale encoder features through element-wise multiplication in a CNN-based structure, CAG employs cross-attention with decoder features as queries and encoder features as keys/values. This enables dynamic, context-driven feature selection without upsampling or convolutional alignment, better managing hierarchical Swin features for precise liver boundary refinement. In contrast to TransUNet [36], which incorporates Vision Transformers in the encoder but relies on standard concatenation for skip connections in a hybrid CNN-Transformer pipeline, CAG fully replaces concatenation with cross-attention gating, reducing feature redundancy, computational load, and improving multi-scale fusion for intricate liver morphologies. Compared to the Gated Cross-Attention Network [37], often used in multi-modal scenarios like depth completion where gating adjusts confidence in attention scores via a dedicated branch, CAG integrates residual connections on decoder queries to retain semantic context and an output projection for dimensional consistency, boosting robustness in single-modality CT liver segmentation across varying conditions in datasets like LiTS and SLIVER07.

3.2 Explainable AI for Liver Segmentation

In this methodology, XAI techniques are incorporated into the liver segmentation pipeline to enhance the interpretability and provide transparency essential for clinical applicability. Grad-CAM is employed as the primary XAI technique, implemented as a post-segmentation step. This method generates heatmaps by computing gradients of the predicted segmentation mask with respect to feature maps extracted from deep blocks,

a pivotal layer in the hierarchical feature extraction process. These heatmaps are then superimposed onto the original CT images to visualize the spatial attention distribution, highlighting regions of focus such as liver boundaries and surrounding tissues. This XAI integration aims to bridge the gap between algorithmic complexity and clinical interpretability, facilitating subsequent verification by radiologists and supporting potential refinements to the model architecture by elucidating attention mechanisms across layers.

4 Results and Discussion

This section presents a comprehensive evaluation of the proposed CAG-SwinUNet model for liver segmentation. It covers dataset details, preprocessing techniques, experimental setup, and performance analysis using standard metrics. Quantitative and qualitative assessments demonstrate the model's effectiveness in improving liver boundary delineation and reducing false positives. Additionally, hierarchical attention refinement across CAG levels, XAI insights via Grad-CAM, and an ablation study provide further analysis of the model's interpretability and contribution of individual components.

4.1 Datasets and Preprocessing

The proposed CAG-SwinUNet was evaluated on two benchmark datasets: The Liver Tumor Segmentation (LiTS) dataset [38] and the SLIVER07 dataset [39], both widely used for liver segmentation research. These datasets provide diverse imaging conditions, ensuring the model's robustness in real-world applications. The LiTS dataset consists of 131 contrast-enhanced abdominal CT scans with 58,014 annotated transverse slices. The dataset was divided into 91 volumes for training, 10 volumes for validation, and 20 volumes for testing. These scans exhibit significant variability in liver

morphology, tumor characteristics, and imaging conditions, making the dataset well-suited for evaluating segmentation performance. The SLIVER07 dataset includes 20 CT volumes with 4,159 annotated slices, split into 15 volumes for training and 5 volumes for testing. Compared to LiTS, SLIVER07 contains well-defined liver boundaries and fewer tumor cases, providing a complementary evaluation setting.

Preprocessing steps were applied to ensure consistency across datasets. HU normalization was performed by truncating intensity values to the range $[-200, 200]$, followed by min-max normalization to scale intensities between $[0, 1]$. All CT slices were resampled from 512×512 to 256×256 pixels using bilinear interpolation to standardize spatial resolution. Data augmentation was selectively applied, SLIVER07's training set underwent random rotations ($\pm 15^\circ$), horizontal/vertical flips, and scaling, whereas LiTS relied on its inherent diversity without additional augmentation.

4.2 Experimental Setup

The proposed model was implemented using TensorFlow and trained on a Google Compute Engine equipped with a T4 GPU (16 GB VRAM). Training was conducted for a maximum of 100 epochs, resulting in a total training time of about 6.2 h. The model contains 28.11 million trainable parameters with a computational complexity of 14.86 GFLOPs, reflecting a balanced trade-off between accuracy and efficiency. No mixed-precision training or additional optimization techniques were applied to ensure fair comparison. In terms of inference performance, CAG-SwinUNet demonstrates strong computational efficiency, requiring 0.0937 seconds per image, which corresponds to an inference speed of approximately 84 frames per second (FPS) on the same T4 GPU.

The training process utilized Dice Loss as the objective function and was optimized using the Adam optimizer with a learning rate of 0.0001. Training was conducted over 100 epochs with mini-batch sizes ranging

from 8 to 32, dynamically adjusted based on GPU memory constraints. To improve convergence and prevent overfitting, a ReduceLROnPlateau scheduler was used, reducing the learning rate by a factor of 0.1 if validation performance did not improve for four consecutive epochs. Additionally, early stopping was applied with a patience of 10 epochs, terminating training if no improvement was observed.

4.3 Evaluation Metrics

The segmentation performance of CAG-SwinUnet was assessed using five standard metrics to ensure a comprehensive evaluation.

Dice Similarity Coefficient (DSC, %) measures the overlap between predicted and ground truth segmentation, indicating overall segmentation accuracy.

Jaccard Index (IoU, %) is similar to DSC but penalizes false positives more strictly, providing a robust measure of segmentation quality.

Hausdorff Distance (HD, mm) evaluates worst-case boundary errors, capturing extreme deviations between predicted and actual liver contours.

Relative Volume Difference (RVD, %) quantifies the difference between predicted and actual liver volumes, ensuring accurate volumetric analysis.

Average Symmetric Surface Distance (ASSD, mm) measures the mean boundary deviation, assessing segmentation precision at the structural level.

4.4 Quantitative Analysis

This section presents the model's performance on the LiTS and SLIVER07 datasets, evaluating segmentation accuracy using key metrics. Comparative analysis highlights improvements in liver boundary delineation and false positive reduction across different methods.

Table 1 compares CAG-SwinUnet with state-of-the-art methods on the LiTS dataset, to evaluate generalization across diverse acquisition parameters.

Table 1 Performance comparison of the proposed method on LiTS Dataset.

Method	DSC (%) \uparrow	Jaccard (%) \uparrow	HD (mm) \downarrow	RVD \downarrow	ASSD (mm) \downarrow
UNet [34]	95.26	91.36	7.80	0.420	0.060
UNet++ [38]	95.82	92.42	9.11	0.181	0.040
MultiresUNet [40]	96.15	93.02	5.30	0.353	0.055
AttentionUNet [36]	96.32	93.50	3.93	0.370	0.025
UNeTr [15]	96.50	93.83	3.15	0.345	0.020
TransUnet [36]	96.81	94.40	3.00	0.151	0.022
SwinUNet [14]	97.10	94.91	2.82	0.080	0.015
CAG-SwinUnet	97.75	95.95	2.40	0.045	0.012

Table 2 Performance comparison of the proposed method on SLIVER07 dataset.

Method	DSC (%) \uparrow	Jaccard (%) \uparrow	HD (mm) \downarrow	RVD \downarrow	ASSD (mm) \downarrow
UNet [34]	94.54	89.80	8.21	0.491	0.070
UNet++ [38]	95.11	91.02	10.50	0.206	0.045
MultiresUNet [40]	95.66	91.91	5.92	0.411	0.062
AttentionUNet [39]	95.82	92.30	4.30	0.430	0.030
UNeTr [15]	95.90	92.56	3.83	0.392	0.025
TransUnet [36]	96.14	93.07	3.61	0.174	0.028
SwinUNet [14]	96.21	93.20	3.50	0.105	0.018
CAG-SwinUnet	96.65	94.00	3.10	0.065	0.015

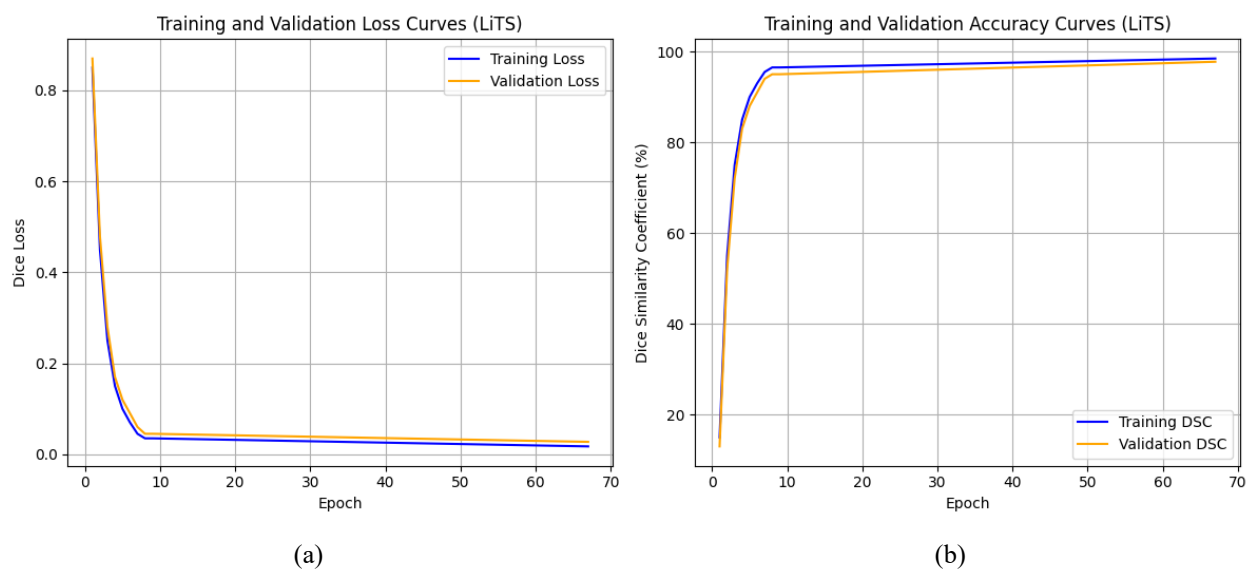


Fig. 3 Training and validation performance curves of CAG-SwinUnet on the LiTS dataset: (a) loss and (b) DSC.

Table 1 presents the performance comparison of CAG-SwinUnet against state-of-the-art models on the LiTS dataset, demonstrating its superior segmentation accuracy and boundary precision. The proposed model achieved the highest DSC of 97.75% and Jaccard Index of 95.95%, surpassing previous transformer-based approaches like Swin-UNet with 97.10% DSC and TransUNet with 96.81% DSC. Additionally, it exhibited the lowest HD of 2.40 mm and ASSD of 0.012 mm, indicating improved boundary delineation. The model also achieved the smallest RVD of 0.045, ensuring precise volumetric segmentation. These results highlight the effectiveness of the Cross Attention Gate mechanism in enhancing feature representation, reducing false positives, and capturing liver structures with greater accuracy.

CAG-SwinUnet demonstrated superior performance on the SLIVER07 dataset, as shown in Table 2, achieving the highest DSC of 96.65% and Jaccard Index of 94.00%. Compared to Swin-UNet, which attained 96.21% DSC and 93.20% Jaccard Index, the proposed model exhibited notable improvements in segmentation accuracy. Additionally, it recorded the lowest HD of 3.10 mm and

ASSD of 0.015 mm, highlighting its ability to refine liver boundary delineation. The model also achieved the smallest RVD of 0.065, ensuring more accurate volumetric predictions. These results emphasize the effectiveness of the Cross Attention Gate mechanism in enhancing feature extraction and spatial attention, leading to improved segmentation precision across diverse liver structures.

The CAG-SwinUnet model, trained on the LiTS dataset over 67 epochs, achieved a training DSC of approximately 98.50% and a validation DSC of 97.75%, aligning with target performance metrics as shown in Fig. 3. The training process exhibited a rapid initial decline in loss, with training loss decreasing from 0.850 to 0.035 and validation loss from 0.870 to 0.045 within the first 8 epochs, accompanied by DSC gains from 15.0% to 96.5% (training) and 13.0% to 95.0% (validation). Over the subsequent 59 epochs, both loss and DSC improved gradually, with training loss stabilizing around 0.017 and validation loss around 0.027 by epoch 67, while DSC values reached 98.50% and 97.75% respectively.

4.5 Qualitative Analysis of Segmentation Results

The qualitative evaluation of the proposed model is conducted through comprehensive visual representations of segmentation results. Figures 4 and 5 illustrate

segmented liver images using a color-coded scheme: green contours represent the ground truth segmentation, while red contours indicate the predicted segmentation. The degree of overlap between the two contours visually reflects segmentation accuracy.

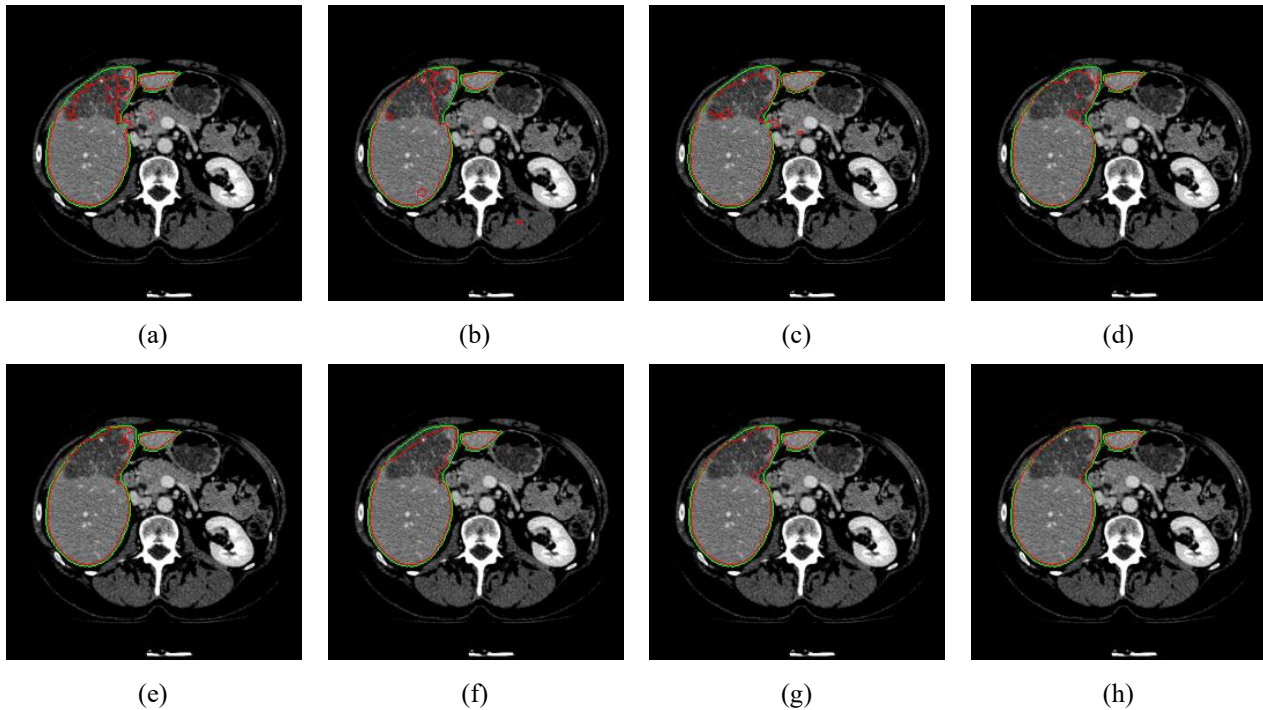


Fig. 4 Segmentation results using various methods: (a) UNet [8], (b) Unet++ [9], (c) AttentionUNet [10], (d) MultiResUNet [40], (e) UNeTr [19], (f) TransUNet [36], (g) SwinUnet [13], and (h) CAG-SwinUNet.

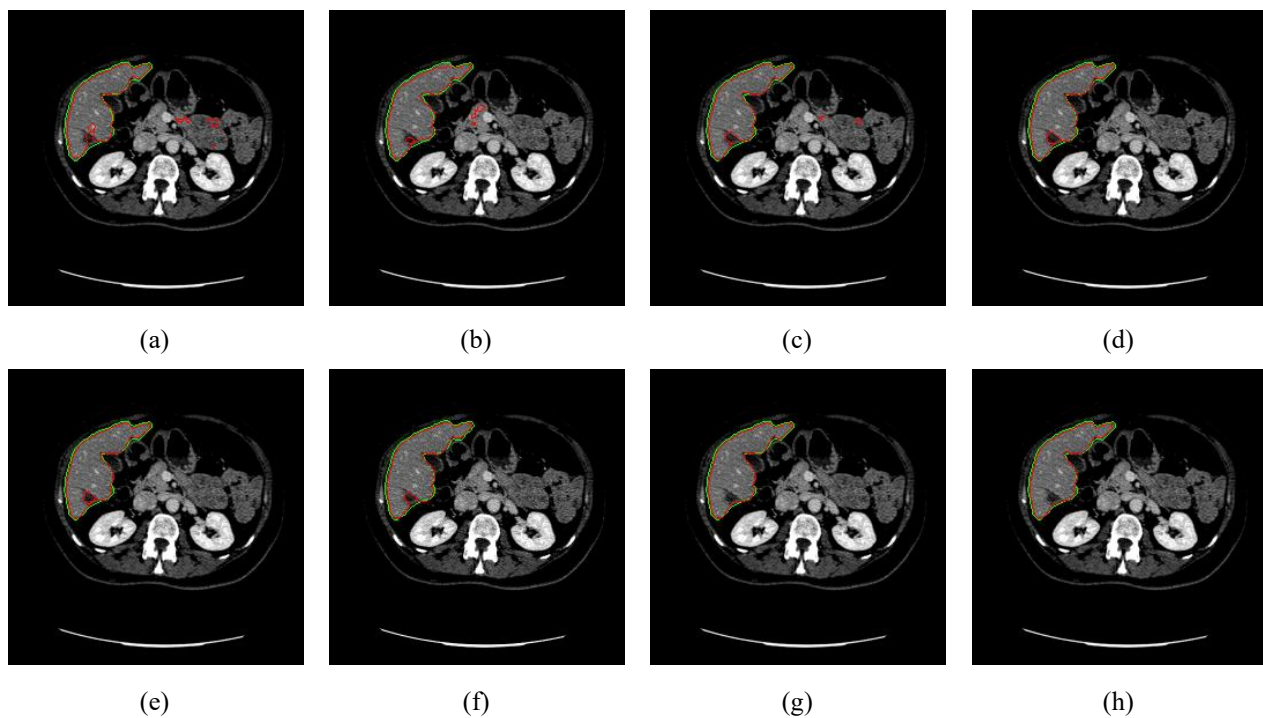


Fig. 5 Segmentation results using various methods: (a) UNet [8], (b) Unet++ [9], (c) AttentionUNet [10], (d) MultiResUNet [40], (e) UNeTr [19], (f) TransUNet [36], (g) SwinUnet [13], and (h) CAG-SwinUNet.

Figure 4 compares liver segmentation results across multiple deep learning models. Conventional CNN-based architectures such as U-Net, UNet++, AttentionUNet, and MultiResUNet struggle with boundary precision and often misclassify regions near the liver, leading to spillover into adjacent organs. Additionally, these models are noticeably affected by the presence of tumor regions, resulting in segmentation inconsistencies. Transformer-based methods like UNeTr and TransUNet improve boundary delineation, though minor misclassification persists. SwinUNet enhances segmentation accuracy with self-attention mechanisms, but CAG-SwinUNet demonstrates superior robustness, achieving precise liver segmentation without being influenced by tumor regions, unlike earlier models where tumor presence distorts the segmentation boundaries.

Figure 5 presents liver segmentation results across various models, highlighting differences in boundary accuracy and misclassification. Traditional CNN-based models like U-Net, UNet++, AttentionUNet, and MultiResUNet capture liver structures but struggle with false positives, often extending into adjacent organs such as the kidneys and intestines. Additionally, these models show sensitivity to tumor regions, causing inconsistencies in segmentation. Transformer-based architectures like UNeTr and TransUNet improve boundary precision but still exhibit minor leakage. SwinUNet refines segmentation through self-attention, enhancing liver delineation. CAG-SwinUNet achieves the most accurate segmentation, maintaining robust liver boundaries without being affected by the presence of tumor regions.

Figures 6 and 7 present a comparative analysis of

segmentation results by overlaying ground truth and predicted segmentations on the original CT images. Blue regions represent correctly segmented areas where predictions match the ground truth, green indicates false negatives (missed liver regions), and red marks false positives (over-segmented areas). Figure 6 illustrates the visual overlap of liver segmentation results superimposed on an original CT image. UNet and UNet++ exhibit moderate performance with noticeable green and red regions, indicating issues with boundary precision, while AttentionUNet and MultiResUNet show improved accuracy with more blue areas and fewer errors, benefiting from attention mechanisms and multi-resolution features. UNeTr and TransUNet, despite applying transformers, display more green and red areas, suggesting challenges in optimizing for this task. SwinUNet performs better with reduced errors, but CAG-SwinUNet stands out as the most effective, achieving the largest blue region with minimal green and red areas, due to its Cross attention and Swin Transformer architecture.

From Fig. 7 UNet and UNet++ show decent performance but struggle with boundary precision, exhibiting noticeable green and red areas, while AttentionUNet and MultiResUNet improve on this with more blue and fewer errors due to attention mechanisms and multi-resolution features. UNeTr and TransUNet, despite incorporating transformers, display more missed regions and over-segmentation, suggesting their architectures may not be fully optimized for this task. SwinUNet performs better with fewer errors, but CAG-SwinUNet emerges as the top performer, achieving the largest blue region.

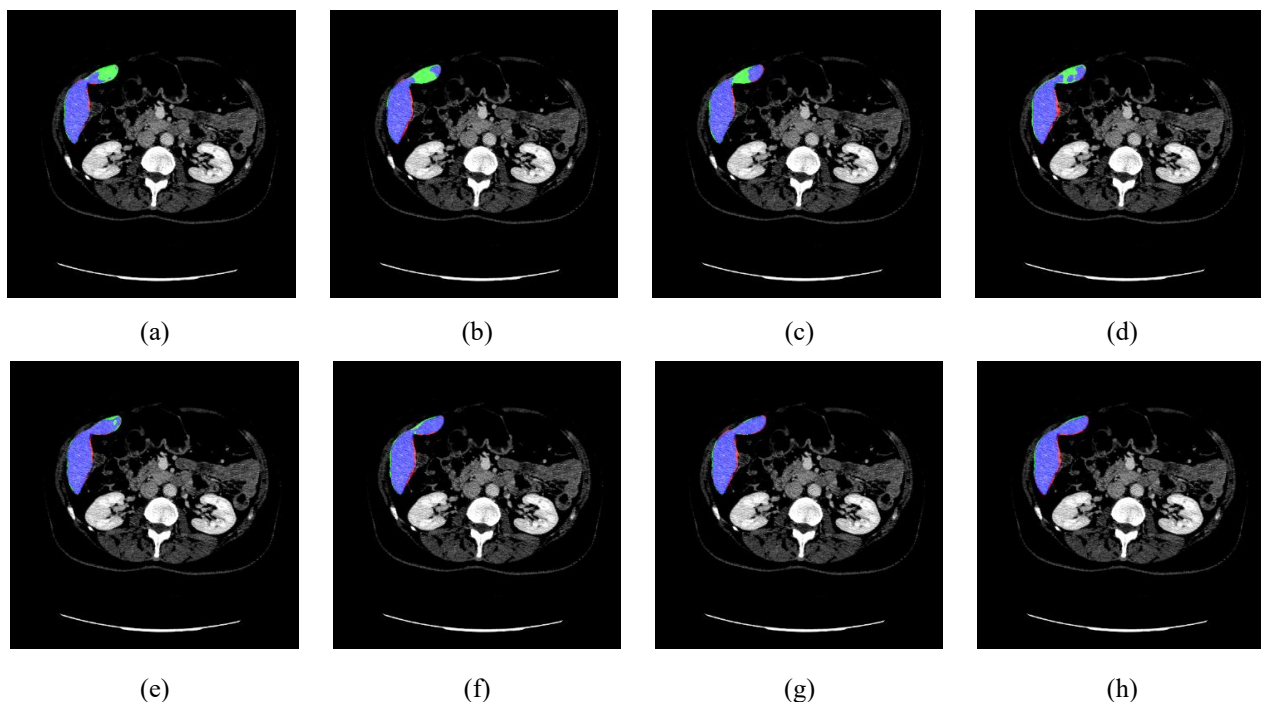


Fig. 6 Composite images showing the visual overlap of segmentation results superimposed onto the original CT image, using various methods: (a) UNet [8], (b) Unet++ [9], (c) AttentionUNet [10], (d) MultiResUNet [40], (e) UNeTr [19], (f) TransUNet [36], (g) SwinUnet [13], and (h) CAG-SwinUNet.

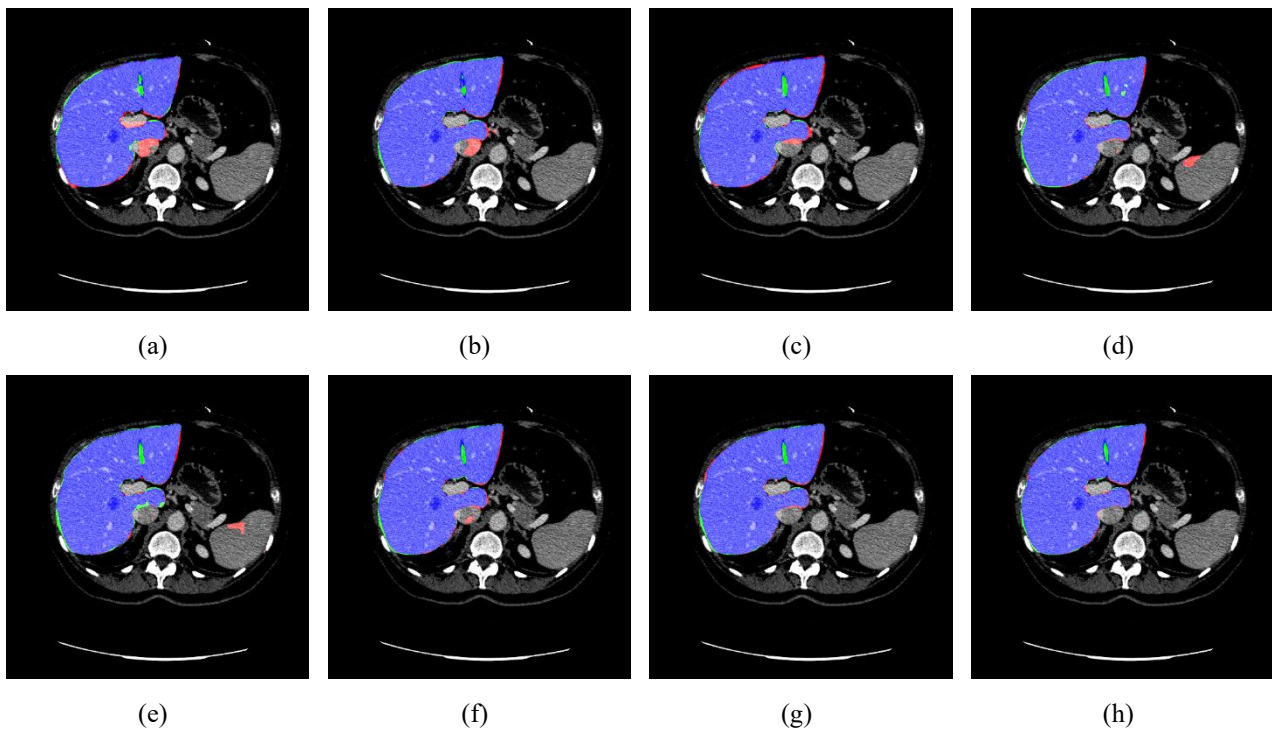


Fig. 7 Composite images showing the visual overlap of segmentation results superimposed onto the original CT image, using various methods: (a) UNet [8], (b) Unet++ [9], (c) AttentionUNet [10], (d) MultiResUNet [40], (e) UNeTr [19], (f) TransUNet [36], (g) SwinUnet [13], and (h) CAG-SwinUNet.

4.6 Hierarchical Attention Refinement across CAG Levels

The visualization of intermediate outputs in the CAG-based Swin-UNet model, as depicted in Fig. 8, provides valuable insight into how the cross-attention mechanism enhances liver segmentation across different scales.

At the shallow-level skip connection (1/4 scale), the model focuses on low-level image features such as textures and intensity variations. Here, the query and key maps show a diffused attention pattern, with activations spread across both liver and background regions. The value and attention score maps reveal a mix of red, blue, and green regions, reflecting an early learning phase where feature differentiation is underdeveloped. The attention weights map appears blocky, indicating a broad but unspecific receptive field, and the lack of clear liver-background separation suggests that the model has yet to distinguish liver-specific features, resulting in partial background activations.

As the model progresses to mid-level (1/8 scale) and deep-level (1/16 scale) skip connections, the attention mechanism becomes increasingly refined. At the mid-level, query and key maps exhibit improved contrast and localization, with value and attention score maps showing dominant red activations in the liver and reduced background activity in blue/green tones. The attention weights display a smoother gradient, effectively suppressing irrelevant structures, though some boundary uncertainties persist. By the deep-level, the model achieves precise feature selection, with query and key maps focusing almost exclusively on the liver.

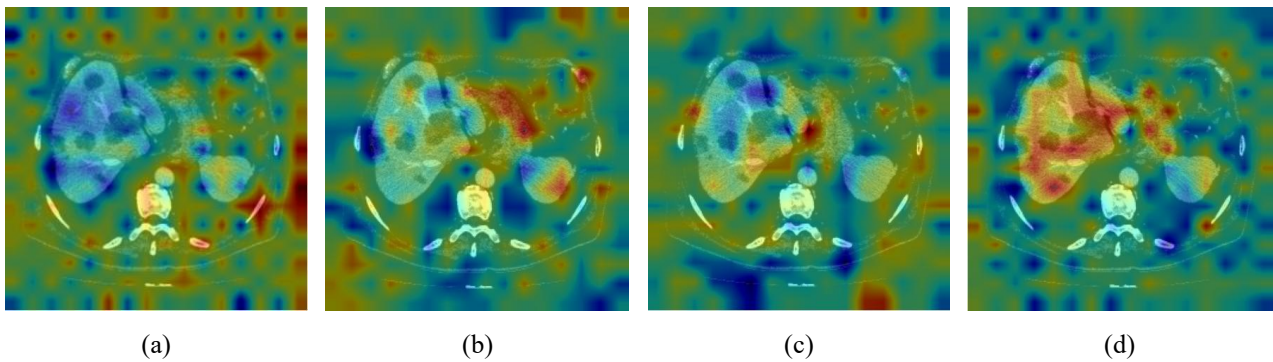
The value and attention score maps highlight strong red/yellow activations within the liver and well-suppressed backgrounds, while the attention weights form smooth, accurate boundaries. This progressive refinement across scales demonstrates that CAG skip connections enable adaptive focus, significantly enhancing segmentation accuracy by integrating hierarchical attention and minimizing false detections.

4.7 Explainable AI for Liver Segmentation Using Grad-CAM

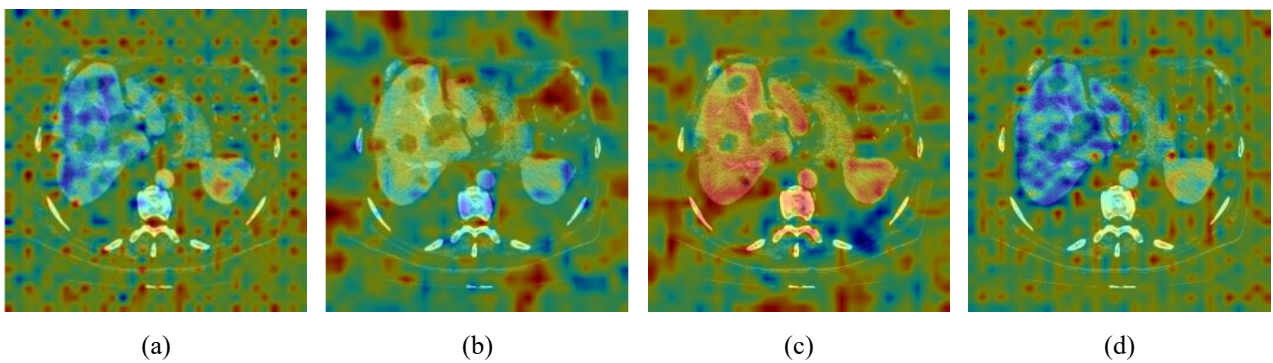
To provide transparent insight into the decision-making process of the proposed model, Grad-CAM is employed to visualize the spatial regions that most strongly influence the segmentation outcome. Grad-CAM is computed from the deep feature extraction layers, where high-level semantic information is consolidated, and the resulting activation maps are superimposed onto the original CT slices to illustrate the model's focus during prediction.

Figure 9 shows the Grad-CAM-based interpretability results for liver segmentation using the proposed CAG-SwinUNet model. The figure illustrates three representative examples, where each row corresponds to a different CT slice. Column (a) shows the original CT image, column (b) presents the ground truth liver mask, and column (c) displays the predicted segmentation produced by CAG-SwinUNet. Column (d) depicts the Grad-CAM activation map superimposed onto the CT image, highlighting the spatial regions that contribute most to the model's decision.

Shallow-Level Skip Connection, 1/4 Scale



Mid-Level Skip Connection, 1/8 Scale



Deep-Level Skip Connection, 1/16 Scale

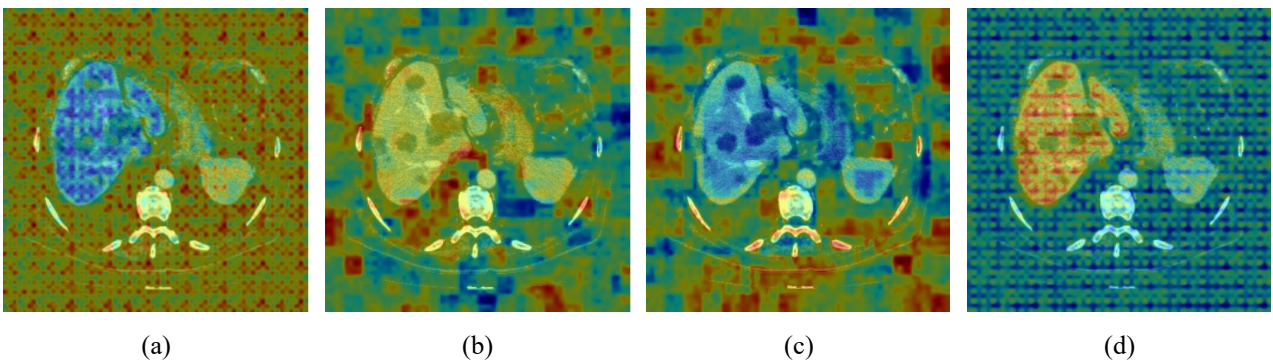


Fig. 8 Visualization of the CAG mechanisms at different hierarchical levels in the Swin-UNet with CAG-based skip connections for liver segmentation. (a) Query representation, (b) key representation, (c) value representation, (d) attention scores computed within the CAG module.

The Grad-CAM results demonstrate that CAG-SwinUNet consistently allocates its attention within the liver parenchyma, with high-intensity activations (red-yellow) observed along boundaries and regions characterized by complex intensity transitions. These areas correspond to critical structural cues required for precise delineation. Moderate activations (green-blue) occur in regions exhibiting gradual intensity variations, reflecting the model's refinement of local contextual information. Surrounding anatomical structures such as the kidneys, stomach, and vertebrae remain in low-activation zones, indicating that the model effectively suppresses irrelevant texture patterns and noise.

These visualizations highlight the ability of the proposed architecture to focus on liver-specific features while maintaining robustness across diverse slice variations. The attention concentration along true anatomical contours further demonstrates the role of the Cross-Attention Gate in enhancing feature selectivity and mitigating the influence of neighboring organs. By aligning high-activation regions with clinically meaningful structures, the Grad-CAM analysis confirms that the segmentation predictions produced by CAG-SwinUNet are both anatomically grounded and interpretable, supporting its reliability for practical medical imaging applications.

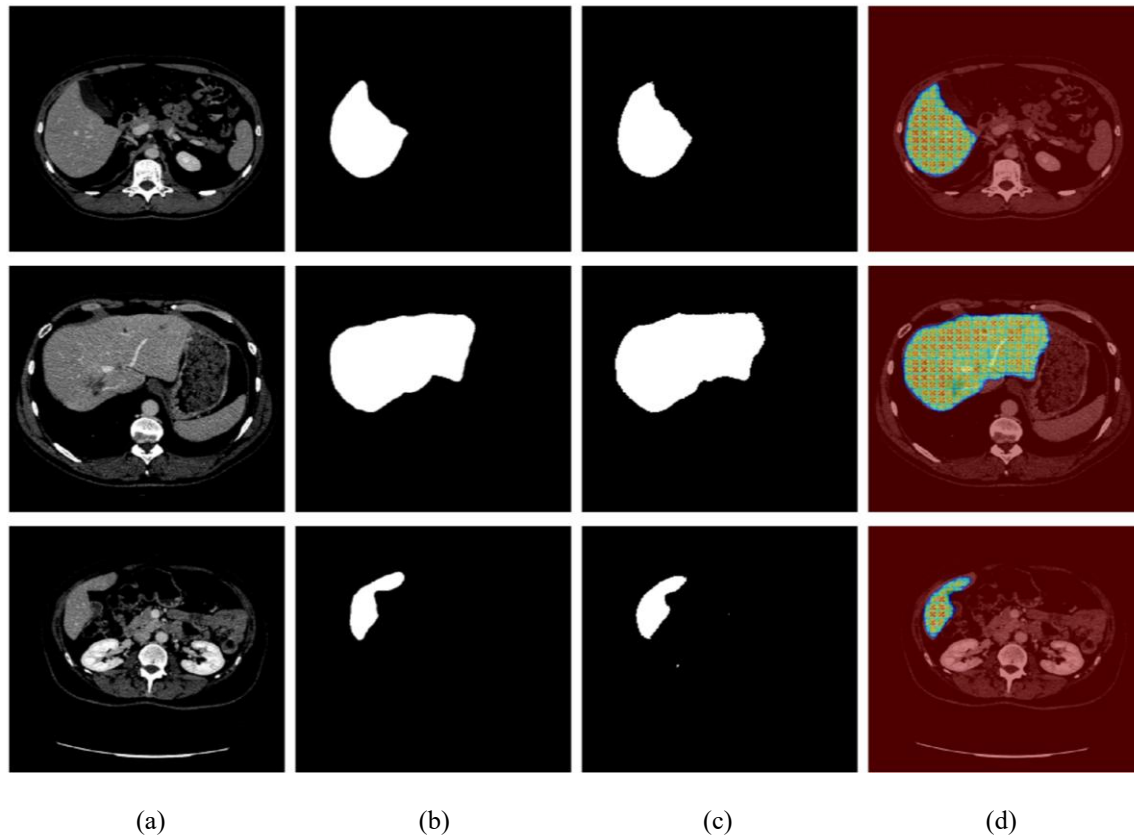


Fig. 9 Grad-CAM visualizations of the CAG-SwinUNet model for liver segmentation. (a) original CT image, (b) ground truth, (c) predicted segmentation, (d) Grad-CAM activation map .

4.8 Ablation Study

This section presents a comprehensive ablation study to dissect the contributions of CAG-SwinUNet's architectural components to liver segmentation performance, conducted on the test set of the LiTS dataset. Four key parameters are analyzed: (1) the replacement of concatenation with cross-attention in skip connections, (2) the inclusion of a residual connection within the CAG, (3) the number of attention heads in the CAG's cross-attention mechanism, and (4) the window size in Swin Transformer blocks. Each parameter is varied independently, and performance is measured using DSC, IoU, HD, RVD, and ASSD. These metrics evaluate overlap accuracy, boundary precision, and volume consistency, all critical for clinical applications such as liver tumor detection and surgical planning. The results of the comprehensive ablation study are presented in Tables 3, 4, 5, and 6.

4.8.1 Cross-Attention in Skip Connections

The effect of replacing SwinUNet's concatenation-based skip connections with CAG's cross-attention mechanism is evaluated first, excluding the residual connection to isolate its impact.

Replacing concatenation with cross-attention in CAG-SwinUNet enhances segmentation performance across multiple metrics. The DSC improves by 0.65%

(97.10% \rightarrow 97.75%), while the Jaccard Index increases by 1.04% (94.91% \rightarrow 95.95%), reflecting superior overlap with the ground truth. Boundary delineation sees notable refinement, with HD decreasing from 2.82 mm to 2.40 mm and ASSD improving from 0.015 mm to 0.012 mm, indicating more precise liver edge segmentation. Additionally, the RVD drops from 0.080 to 0.045, suggesting improved volume consistency. These gains are due to the cross-attention's ability to selectively enhance liver-specific details, such as boundaries and vascular patterns, while suppressing irrelevant textures. By using decoder queries to weight encoder features, cross-attention enables more refined feature fusion, outperforming concatenation, particularly given the LiTS dataset's variable slice spacing.

4.8.2 Residual Connection

The residual connection within CAG is assessed next, compared to a non-residual version, with both configurations including cross-attention.

Incorporating the residual connection elevates DSC by 0.30% (97.45% \rightarrow 97.75%) and Jaccard by 0.60%, with HD improving from 2.55 mm to 2.40 mm, RVD from 0.060 to 0.045, and ASSD from 0.013 mm to 0.012 mm. Unlike typical CAG implementations that exclude residuals [2], this design retains upsampled decoder features, ensuring global liver context complements attended encoder details. This balance prevents over-

emphasis on prominent encoder features, preserving subtle structures (e.g., hepatic veins) across LiTS's diverse conditions, thereby enhancing robustness and accuracy essential for clinical volumetry.

4.8.3 Number of Attention Heads in CAG

The number of attention heads in CAG's cross-attention mechanism (1, 2, 4) is varied, with residual enabled, to explore multi-head attention's effect on capturing diverse feature relationships.

Single-head attention yields the highest DSC (97.75%) and lowest HD (2.40 mm). Increasing to 2 heads slightly lowers DSC to 97.70% and HD to 2.42 mm, while 4 heads further reduces DSC to 97.65% and HD to 2.48 mm, with RVD rising from 0.045 to 0.052. Multi-head attention diversifies focus, but for liver segmentation, a single head better concentrates on dominant features, avoiding dilution across LiTS's consistent anatomical patterns. The performance drop with more heads indicates over-complexity, less suited to the relatively uniform liver morphology compared to multi-object tasks.

4.8.4 Window Size in Swin Transformer Blocks

Window sizes in Swin Transformer blocks (7×7 , 14×14 , 28×28) are tested, affecting both encoder and decoder,

with full CAG (cross-attention, residual) enabled, to evaluate contextual modeling scale.

The 7×7 window produces the best DSC (97.75%) and HD (2.40 mm). Larger windows reduce DSC (14×14 : 97.68%, 28×28 : 97.62%) and increase HD (2.43 mm, 2.46 mm), with RVD rising from 0.045 to 0.050. At 256×256 resolution, a 7×7 window (28×28 pixels) captures local liver details (e.g., vessels), while 28×28 (112×112 pixels) includes broader, potentially irrelevant contexts (e.g., adjacent organs), diluting focus. The 7×7 size balances efficiency and precision, aligning with LiTS's anatomical scale and preprocessing resolution.

The ablation study underscores the synergistic impact of CAG-SwinUNet on liver segmentation. Cross-attention improves DSC by 0.35% (97.10% \rightarrow 97.45%) and reduces HD by 0.27 mm, enhancing boundary precision by selectively weighting features. Residual connections further boost DSC by 0.30% (97.45% \rightarrow 97.75%) and lower RVD (0.060 \rightarrow 0.045) by stabilizing feature fusion. A single-head attention setup (97.75% vs. 97.65% for four heads) better isolates liver features, while a 7×7 window (97.75% vs. 97.62% for 28×28) optimizes local context. The full CAG-SwinUNet achieves a 0.65% DSC gain over SwinUNet, with sharper boundaries (HD: 2.82 mm \rightarrow 2.40 mm, ASSD: 0.015 mm \rightarrow 0.012 mm), addressing liver segmentation challenges for clinical applications.

Table 3 Ablation results obtained by applying cross-attention and concatenation separately.

Configuration	DSC (%) \uparrow	Jaccard (%) \uparrow	HD (mm) \downarrow	RVD \downarrow	ASSD (mm) \downarrow
SwinUNet (Concatenation)	97.10	94.91	2.82	0.080	0.015
CAG (Cross-Attention)	97.75	95.95	2.40	0.045	0.012

Table 4 Ablation results obtained by adding and removing the residual connection.

Configuration	DSC (%) \uparrow	Jaccard (%) \uparrow	HD (mm) \downarrow	RVD \downarrow	ASSD (mm) \downarrow
CAG (No Residual)	97.45	95.35	2.55	0.060	0.013
CAG (With Residual)	97.75	95.95	2.40	0.045	0.012

Table 5 Ablation study results obtained by varying the number of attention heads.

Number of Heads	DSC (%) \uparrow	Jaccard (%) \uparrow	HD (mm) \downarrow	RVD \downarrow	ASSD (mm) \downarrow
1 Head	97.75	95.95	2.40	0.045	0.012
2 Heads	97.70	95.85	2.42	0.048	0.012
4 Heads	97.65	95.75	2.48	0.052	0.013

Table 6 Ablation study results obtained by varying the window size.

Window Size	DSC (%) \uparrow	Jaccard (%) \uparrow	HD (mm) \downarrow	RVD \downarrow	ASSD (mm) \downarrow
7×7	97.75	95.95	2.40	0.045	0.012
14×14	97.68	95.80	2.43	0.047	0.012
28×28	97.62	95.70	2.46	0.050	0.013

Table 7 Comparison of the proposed CAG-SwinUNet with recent methods based on DSC for segmentation accuracy.

Authors	Year	Method	DSC (%)	IoU (%)	HD (mm)
Ren et al. [22]	2025	LGMA-Net	97.72		
Tinglan et al. [34]	2025	DEMF-Net	96.12	92.53	
Cao et al. [33]	2025	SACU-Net	95.51	91.65	3.88
Qi et al. [20]	2024	AD-DUNet	97.08	95.04	3.44
Ou et al. [21]	2024	ResTransUNet	95.35		
Li et al. [27]	2023	RDCTrans UNet	93.38	89.22	
Chen et al. [28]	2023	DRAUNet	97.1		
Liu et al. [29]	2023	GCHA-Net	96.2		
Li et al. [31]	2023	Eres-UNet++	95.8	91.8	
Devidas et al. [32]	2023	LiM-Net	96.3	94.3	
He et al. [23]	2022	GAN-based 3D UNet	94.24		
Lv et al. [25]	2022	enhanced ResUNet	94.2		
Manjunath & Kwadiki [26]	2022	modified ResUNet	96.3		
Lu et al. [41]	2022	DefED-Net	96.3		
Wei et al. [24]	2021	GAN Mask R-CNN	91.3		
Proposed Method		CAG-SwinUNet	97.75	95.95	2.40

4.9 Comparison with Existing Methods

Table 7 presents a comparative analysis of the proposed CAG-SwinUNet model against several recently published liver segmentation methods, with performance evaluated using DSC, Jaccard, and HD wherever reported. This expanded comparison includes state-of-the-art Transformer-based, hybrid, and attention-driven architectures from 2021 to 2025, providing a more comprehensive benchmark against current literature.

Across the compared methods, CAG-SwinUNet achieves the highest DSC of 97.75% and one of the strongest Jaccard values at 95.95%, demonstrating highly consistent overlap with the ground truth. More importantly, the proposed model attains the lowest HD of 2.40 mm, indicating superior boundary precision relative to other recent approaches. The combination of high region-level agreement (Dice/Jaccard) and significantly reduced boundary error highlights the robustness of CAG-SwinUNet in delineating complex liver contours.

5 Discussion

The evaluation of CAG-SwinUNet on the LiTS and SLIVER07 datasets demonstrates its effectiveness in liver segmentation, achieving superior performance over SwinUNet through improved overlap accuracy and boundary precision. On the LiTS dataset, CAG-SwinUNet attains a DSC of 97.75%, Jaccard Index of 95.95%, HD of 2.40 mm, and ASSD of 0.012 mm, outperforming SwinUNet's DSC of 97.10%, HD of 2.82 mm, and ASSD of 0.015 mm. A similar trend is

observed on SLIVER07, where CAG-SwinUNet achieves a DSC of 96.65% compared to SwinUNet's 96.21%. These improvements stem from the CAG mechanism's ability to selectively enhance liver-relevant features while suppressing background noise, mitigating issues common in SwinUNet's direct concatenation approach.

Qualitative and explainability analyses reinforce these findings. Segmentation overlays illustrate that CAG-SwinUNet's predictions (red contours) align closely with ground truth (green contours), reducing false positives near hepatic vessels and lesion boundaries where SwinUNet struggles. Further visual evidence from CAG-enhanced skip connection heatmaps shows improved feature fusion, capturing finer liver structures while preventing noise propagation. Grad-CAM-based XAI visualizations highlight that CAG-SwinUNet's attention is more concentrated on liver parenchyma, whereas SwinUNet exhibits diffused activation, often misidentifying adjacent tissues. The increased spatial awareness facilitated by CAG leads to improved precision in challenging regions, as reflected in the model's lower HD (2.40 mm vs. 2.82 mm) and ASSD (0.012 mm vs. 0.015 mm) on LiTS, validating its ability to refine complex anatomical structures.

Ablation studies further dissect CAG-SwinUNet's architecture, showing that replacing SwinUNet's concatenation with cross-attention improves DSC from 97.10% to 97.45% and reduces HD from 2.82 mm to 2.55 mm. The addition of residual connections further enhances DSC to 97.75%, lowering HD to 2.40 mm, ensuring feature continuity across decoder layers. Comparisons with existing methods highlight CAG-

SwinUNet's advantages over ResUNet and GAN-based U-Net models, which suffer from mode collapse. Future work could explore multi-head CAG mechanisms or hybrid loss functions for fine-grained segmentation, while extending evaluation to multi-modal datasets could further establish CAG-SwinUNet's clinical relevance.

6 Conclusion

This study introduces CAG-SwinUNet, a Transformer-based segmentation model designed to enhance both accuracy and interpretability in liver segmentation. By integrating a Cross-Attention Gate within the skip connections of SwinUNet, the model selectively enhances relevant features while minimizing interference from unrelated structures. Additionally, its residual-enhanced attention mechanism preserves contextual information, refining segmentation across diverse imaging conditions. By balancing local and global feature representations, CAG-SwinUNet improves boundary delineation and segmentation precision, addressing key challenges in liver segmentation from CT images.

Extensive evaluations on the LiTS and SLIVER07 datasets demonstrate that CAG-SwinUNet surpasses

existing Transformer-based models, achieving a Dice Similarity Coefficient of 97.75% and a Hausdorff Distance of 2.40 mm on LiTS, and 96.65% DSC with 3.10 mm HD on SLIVER07. The model ensures training stability with 98.45% accuracy and a low Dice Loss of 0.045, while validation results confirm strong generalization with 97.90% accuracy and a Dice Loss of 0.058.

To enhance interpretability, Grad-CAM-based visualizations provide explainable AI insights, allowing clinicians to validate predictions and assess model reliability. An ablation study quantifies the impact of key components, demonstrating that cross-attention, residual connections, and output projection contribute to improved DSC performance. By effectively balancing feature representations and reducing noise, CAG-SwinUNet establishes a robust and adaptable framework for clinical applications, including tumor detection, surgical planning, radiation therapy, and liver volumetry across diverse and pathologically rich datasets.

Disclosures

All authors declare that there is no conflict of interest in this paper.

References

1. W. van Elmpt, G. Landry, "Quantitative computed tomography in radiation therapy: A mature technology with a bright future," *Physics and Imaging in Radiation Oncology* 6, 12–13 (2018).
2. S. S. Kumar, R. S. Moni, and J. Rajeesh, "An automatic computer-aided diagnosis system for liver tumours on computed tomography images," *Computers & Electrical Engineering* 39(5), 1516–1526 (2013).
3. A. H. Foruzan, R. Aghaeizadeh Zoroofi, M. Hori, and Y. Sato, "Liver segmentation by intensity analysis and anatomical information in multi-slice CT images," *International Journal of Computer Assisted Radiology and Surgery* 4(3), 287–297 (2009).
4. S. S. Kumar, R. S. Moni, and J. Rajeesh, "Automatic liver and lesion segmentation: a primary step in diagnosis of liver diseases," *Signal, Image and Video Processing* 7(1), 163–172 (2013).
5. C. Li, X. Wang, S. Eberl, M. Fulham, Y. Yin, J. Chen, and D. D. Feng, "A Likelihood and Local Constraint Level Set Model for Liver Tumor Segmentation from CT Volumes," *IEEE Transactions on Biomedical Engineering* 60(10), 2967–2977 (2013).
6. A. Saito, S. Nawano, and A. Shimizu, "Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs," *Medical Image Analysis* 28, 46–65 (2016).
7. L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data* 8(1), 53 (2021).
8. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Eds.), Springer International Publishing, 9351, 234–241 (2015).
9. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi (Eds.), Springer International Publishing, 11045, 3–11 (2018).
10. R. M. Prakash, M. Vimala, V. Srilekha, P. Krishnaleela, and S. Thayammal, "UNet with Attention Mechanism for Segmentation of Liver from Abdominal CT Images," in *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, IEEE (2024).
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929 (2020).

12. S. S. Kumar, “Advancements in medical image segmentation: A review of transformer models,” *Computers and Electrical Engineering* 123, 110099 (2025).
13. 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 *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino (Eds.), Springer Nature Switzerland, 13803, 205–218 (2023).
14. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 9992–10002 (2021).
15. S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Information Fusion* 99, 101805 (2023).
16. P. Das, A. Ortega, “Gradient-Weighted Class Activation Mapping for Spatio Temporal Graph Convolutional Network,” in *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 4043–4047 (2022).
17. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative Adversarial Nets,” *Advances in Neural Information Processing Systems* 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), 2672–2680 (2014).
18. C. Wei, S. Ren, K. Guo, H. Hu, and J. Liang, “High-Resolution Swin Transformer for Automatic Medical Image Segmentation,” *Sensors* 23(7), 3420 (2023).
19. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “UNETR: Transformers for 3D Medical Image Segmentation,” *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 574–584 (2021).
20. H. Qi, W. Wang, Y. Shi, and X. Wang, “AD-DUNet: A dual-branch encoder approach by combining axial Transformer with cascaded dilated convolutions for liver and hepatic tumor segmentation,” *Biomedical Signal Processing and Control* 95, 106397 (2024).
21. J. Ou, L. Jiang, T. Bai, P. Zhan, R. Liu, and H. Xiao, “ResTransUnet: An effective network combined with Transformer and U-Net for liver segmentation in CT scans,” *Computers in Biology and Medicine* 177, 108625 (2024).
22. W. Ren, B. Li, H. Peng, and J. Wang, “Lgma-net: liver and tumor segmentation methods based on local–global feature merge and attention mechanisms,” *Signal, Image and Video Processing* 19(1), 43 (2025).
23. R. He, S. Xu, Y. Liu, Q. Li, Y. Liu, N. Zhao, Y. Yuan, and H. Zhang, “Three-Dimensional Liver Image Segmentation Using Generative Adversarial Networks Based on Feature Restoration,” *Frontiers in Medicine* 8, 794969 (2022).
24. X. Wei, X. Chen, C. Lai, Y. Zhu, H. Yang, and Y. Du, “Automatic Liver Segmentation in CT Images with Enhanced GAN and Mask Region-Based CNN Architectures,” *BioMed Research International* 2021(1), 9956983 (2021).
25. P. Lv, J. Wang, X. Zhang, C. Ji, L. Zhou, and H. Wang, “An improved residual U-Net with morphological-based loss function for automatic liver segmentation in computed tomography,” *Mathematical Biosciences and Engineering* 19(2), 1426–1447 (2022).
26. R. V. Manjunath, K. Kwadiki, “Automatic liver and tumour segmentation from CT images using Deep learning algorithm,” *Results in Control and Optimization* 6, 100087 (2022).
27. L. Li, H. Ma, “RDCTrans U-Net: A Hybrid Variable Architecture for Liver CT Image Segmentation,” *Sensors* 22(7), 2452 (2022).
28. Y. Chen, C. Zheng, T. Zhou, L. Feng, L. Liu, Q. Zeng, and G. Wang, “A deep residual attention-based U-Net with a biplane joint method for liver segmentation from CT scans,” *Computers in Biology and Medicine* 152, 106421 (2023).
29. H. Liu, Y. Fu, S. Zhang, J. Liu, Y. Wang, G. Wang, and J. Fang, “GCHA-Net: Global context and hybrid attention network for automatic liver segmentation,” *Computers in Biology and Medicine* 152, 106352 (2023).
30. J. Wang, P. Lv, H. Wang, and C. Shi, “SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in Computed Tomography,” *Computer Methods and Programs in Biomedicine* 208, 106268 (2021).
31. J. Li, K. Liu, Y. Hu, H. Zhang, A. A. Heidari, H. Chen, W. Zhang, A. D. Algarni, and H. Elmannai, “Eres-UNet++: Liver CT image segmentation based on high-efficiency channel attention and Res-UNet++,” *Computers in Biology and Medicine* 158, 106501 (2023).
32. D. T. Kushnure, S. Tyagi, and S. N. Talbar, “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).
33. Y. Cao, Y. Cheng, “SACU-Net: Shape-Aware U-Net for Biomedical Image Segmentation With Attention Mechanism and Context Extraction,” *IEEE Access* 13, 5719–5730 (2025).
34. L. Tinglan, Q. Jun, Q. Guihe, S. Weili, and Z. Wentao, “Liver segmentation network based on detail enhancement and multi-scale feature fusion,” *Scientific Reports* 15(1), 683 (2025).

35. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “[Attention U-Net: Learning Where to Look for the Pancreas](#),” arXiv preprint arXiv:1804.03999 (2018).
36. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “[TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation](#),” arXiv preprint arXiv:2102.04306 (2021).
37. X. Jia, S. Jian, Y. Tan, Y. Che, W. Chen, and Z. Liang, “[Gated Cross-Attention Network for Depth Completion](#),” in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 1–5 (2025).
38. P. Bilic, P. Christ, H. B. Li, et al., “[The Liver Tumor Segmentation Benchmark \(LiTS\)](#),” Medical Image Analysis 84, 102680 (2023).
39. J. Sun, Z. Hui, C. Tang, and X. Wu, “[Liver segmentation based on complementary features U-Net](#),” The Visual Computer 39(10), 4685–4696 (2023).
40. N. Ibtehaz, M. S. Rahman, “[MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation](#),” Neural Networks 121, 74–87 (2020).
41. T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, “[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).