Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.083345

ARTICLE

ECANet: Enhanced Convolutional Attention Network for Liver Segmentation

Yuyan Ning[1,2], Haiyun Huang[1], Legend Zhang[3], Wei Wei[4], Hao Quan[5] and Bo Yang[1,*]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2]Glasgow College, University of Electronic Science and Technology of China, Chengdu, China
[3]School of Artificial Intelligence, Guangzhou Huashang University, Guangzhou, China
[4]School of Mechanical and Material Engineering, Xi'an University, Xi'an, China
[5]Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy

*Corresponding Author: Bo Yang. Email: boyang@uestc.edu.cn

Received: 02 April 2026; Accepted: 12 May 2026; Published: 30 June 2026

ABSTRACT: Hybrid CNN-Transformer models are widely used in medical image segmentation because they combine CNN-based local feature extraction with Transformer-based global context modeling. Despite their popularity, these models face several challenges, including computational complexity, noise blurring, and information loss. This paper proposes an enhanced convolutional attention network (ECANet) for liver segmentation. ECANet uses a U-shaped architecture with efficient channel-attention-based skip connections. Both the encoder and decoder are constructed using enhanced convolutional Transformer (ECT) blocks, where group convolution is integrated into the convolutional attention module for efficient Token embedding and channel disentanglement, and a Token-wise multi-layer perceptron (MLP) branch is incorporated into the wide-focus module to improve feature representation across channels. Deep supervision and a hybrid of Binary Cross-Entropy (BCE) and Dice loss are used to improve boundary accuracy. We evaluate the proposed model on the publicly available LiTS17 dataset. Experiments show that ECANet outperforms the compared CNN-based and CNN-Transformer baseline models on both quantitative and qualitative measures.

KEYWORDS: Medical image segmentation; convolutional transformer; group convolution; efficient channel attention; deep supervision

1 Introduction

The liver, as one of the largest solid organs in the human body, is particularly prone to primary and metastatic malignant tumors. Therefore, accurately delineating the liver structure and conducting reliable volume assessment are of vital importance for surgical planning and treatment decisions. Current clinical imaging techniques include CT, ultrasound, and MRI, which provide noninvasive and high-resolution views of liver anatomy. However, the resulting images often suffer from poor soft-tissue contrast. They may also contain acquisition noise and ambiguous organ boundaries. In routine clinical practice, radiologists still rely on manual layer-by-layer annotation, which is not only labor-intensive but also prone to significant inter-observer variations. This is mainly because the quality of the annotations largely depends on the individual's clinical experience. Traditional segmentation approaches fall into two categories: image-based [1,2] and machine-learning-based. Machine learning techniques, such as support vector machines, have demonstrated greater accuracy than conventional methods. However, they often require large annotated datasets and computationally intensive hyperparameter optimization. Recent advancements in deep learning have driven the development of fully automatic segmentation systems. These systems can enhance the efficiency and

accuracy of tumor delineation and provide reliable volume measurement data for clinicians' diagnostic and treatment decisions.

Convolutional neural networks (CNNs) have revolutionized image segmentation thanks to their exceptional ability to capture spatial features. Two classical CNN architectures, the fully convolutional network (FCN) [3] and the U-Net [4], demonstrate strong performance in feature learning and representation. The U-Net architecture features a contracting path for context extraction, a symmetric expanding path for precise localization, and skip connections to enhance feature propagation. Several U-shaped variants have been developed to improve medical image segmentation further. U-Net++ [5] enhances feature fusion by using nested dense connections and deep supervision. ResUNet [6] introduces residual learning to improve gradient flow and stabilize network optimization. Attention U-Net [7] further incorporates attention gates, which help the model dynamically highlight salient regions and suppress irrelevant background information. Beyond these architectural improvements, nnU-Net [8] provides a self-configuring segmentation framework. It automatically adapts preprocessing, network topology, and training strategies according to the characteristics of each dataset. Because of this strong adaptability, nnU-Net has become a widely used baseline in medical image segmentation benchmarks.

Despite their success, CNNs are fundamentally constrained by their limited receptive field, which hinders their ability to model long-range spatial dependencies. Recently, Transformers were introduced to computer vision [9] to address this limitation. They are good at capturing global contexts through parallel self-attention mechanisms. In medical image segmentation, Swin-Unet [10] exemplifies this approach as a Transformer-based U-shaped network that uses Swin Transformer blocks to build the encoder-decoder structure. Other notable pure-Transformer architectures have also been proposed for medical image segmentation. MISSFormer [11] improves long-range dependency modeling through an Enhanced Mix-FFN and a context bridge, which strengthens multi-scale feature correlation. NNFormer [12] combines convolution and self-attention in an interleaved manner, and introduces local-global volume-based attention for 3D segmentation. DAE-Former [13] further reformulates self-attention to capture both spatial and channel relationships, while using cross-attention-based skip connections to improve feature fusion. More recently, State Space Models (SSMs) have emerged as a promising alternative to Transformers. Compared with standard self-attention, SSMs can model long-range dependencies with linear computational complexity. Representative SSM-based segmentation methods include U-Mamba [14] and VM-UNet [15]. U-Mamba integrates Mamba blocks into a U-shaped encoder to enhance long-range feature modeling, while VM-UNet adopts Visual State Space (VSS) blocks to build a pure Mamba-based encoder-decoder architecture. Despite these advances, Transformer-based methods still have certain limitations. In many cases, images need to be down-sampled into patches before tokenization. This process may weaken critical low-level features and remove fine local anatomical details. To address this issue, hybrid CNN-Transformer architectures have been developed. These models combine the local feature extraction ability of CNNs with the global contextual modeling ability of Transformers.

Hybrid CNN-Transformer architectures have achieved remarkable success in medical image segmentation by combining local feature extraction with global context modeling. Existing studies have explored several representative structural designs. One common strategy is to embed Transformer modules into CNN-based encoder-decoder frameworks, either at the bottleneck stage [16] or through skip connections [17]. Another strategy is to use Transformer-based encoders together with CNN decoders, as adopted in methods such as Swin UNETR and Focal UNETR [18–20]. In addition, dual-branch architectures have also been proposed, where CNN branches focus on local patterns and Transformer branches capture long-range dependencies in parallel [21,22]. UCTransNet [17] rethinks skip connections from a channel-wise perspective, replacing conventional concatenation with channel Transformer modules to fuse multi-scale

features selectively. PHTrans [23] parallelly hybridizes Transformer and CNN in the main building blocks to produce hierarchical representations and adaptively aggregates global and local features. To improve feature representation, these hybrid models often incorporate advanced attention mechanisms within their fusion modules. These mechanisms include channel attention [24] for feature recalibration, spatial attention [25] for region emphasis, and cross attention [26] for inter-modality interaction. TransAttUnet [27] further advances the hybrid paradigm. It introduces multi-level guided attention and multi-scale skip connections to enhance semantic segmentation performance jointly. Recent developments in hybrid CNN-Transformer architecture have introduced novel modules that combine complementary feature extraction paradigms synergistically. The ACMix module [28], for example, combines convolution and self-attention modules via a shared first stage that uses 1×1 convolutions for feature projection, and a second stage that is composed of two parallel paths: a shift-and-sum-based convolutional path and an attention-weight-and-aggregation-based Transformer path. Another innovative approach, the fully convolutional Transformer (FCT) [29], surpasses traditional self-attention by replacing linear projections with depthwise separable convolutions, and incorporating multi-ratio parallel dilated convolutions for multi-scale spatial context capture. CoTr [30] introduces deformable self-attention into a CNN-Transformer hybrid architecture, enabling efficient multi-scale attention computation by attending to a sparse set of key sampling points across feature levels. Despite these advances, several challenges remain. Medical image segmentation requires accurate boundary delineation, especially for complex anatomical structures. At the same time, the model must capture sufficient global contextual information while maintaining computational efficiency. Achieving a balanced trade-off among these factors remains difficult. To address these challenges, we build upon recent advances in convolutional Transformer architectures [29,31]. We propose an enhanced convolutional attention network, named ECANet, for precise liver segmentation. ECANet follows a U-shaped framework. Both its encoder and decoder are constructed using stacked enhanced convolutional Transformer (ECT) blocks. The main contributions of this work are summarized as follows:

1. ECT block: we integrate group convolutions into the convolutional attention module to improve the efficiency of token embedding and incorporate a parallel point-wise multi-layer perceptron (MLP) pathway to strengthen fine-grained (point-level) feature aggregation within the wide-focus module.
2. Efficient channel attention (ECA) for skip connections: We incorporate the ECA mechanism into the skip connections of the U-Net framework. This allows the network to dynamically recalibrate channel-wise features without spatial downsampling, thereby improving feature propagation across different network levels.

The proposed network is validated on the publicly available LiTS17 dataset. Experiments demonstrate the superiority of our network, which achieves highly competitive performance compared to state-of-the-art (SOTA) methods. Ablation studies further validate the effectiveness of each proposed component.

2 Methods

2.1 Overall Structure

The overall structure of the proposed ECANet is illustrated in Fig. 1. ECANet adopts a U-shaped architecture consisting of an encoder and a decoder. The encoder extracts multi-scale contextual features from input images, while the decoder restores spatial resolution level by level to produce accurate segmentation maps. Skip connections bridge corresponding encoder-decoder layers to preserve fine-grained details and improve gradient flow during training.

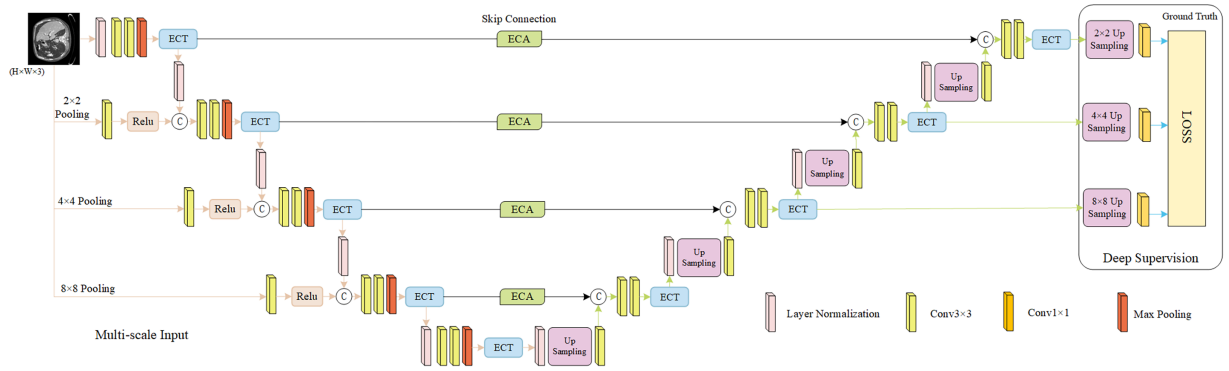

Figure 1: ECANet network structure.

The encoder consists of five levels of modules. In addition to receiving features from the previous level, the three intermediate levels incorporate multi-scale input images through progressively larger pooling operations of 2 × 2, 4 × 4, and 8 × 8. This design enhances the model's ability to capture diverse regions of interest (ROIs) across different resolutions. At each level, the multi-scale input is first processed by a 3 × 3 convolution followed by ReLU activation. The resulting feature maps are then concatenated with the batch-normalized output from the preceding level and passed into the ECT block. Two successive 3 × 3 convolutions and a max-pooling operation follow, after which the down-sampled features enter another ECT block for deeper feature extraction. These ECT blocks enhance feature expressiveness by capturing local and long-range dependencies before passing the refined features to the bottleneck stage.

The decoder mirrors the encoder in structure, which progressively reconstructs the segmentation map from the bottleneck representation through upsampling and refinement operations. The output of the bottleneck ECT block first undergoes layer normalization (LN), followed by upsampling and a 3 × 3 convolution. The feature maps are then concatenated with the corresponding skip-connection output from the encoder, which the ECA module has enhanced to emphasize informative channels [32]. The concatenated features are further refined by two consecutive 3 × 3 convolutional layers before being passed into the ECT block in the decoder. The above processes are repeated at each decoder level, each of which has the same resolution as its corresponding encoder level. Finally, the outputs of the last three decoder ECT blocks are upsampled by scales of 2×, 4×, and 8×, respectively. Each upsampled output is followed by a 1 × 1 convolution to generate prediction maps, which are compared with the ground truth under the deep supervision scheme. The loss is computed using a hybrid BCE-Dice loss function, specifically designed in this work to enhance boundary accuracy and class balance.

The deep supervision strategy enables the penultimate and antepenultimate ECT blocks to output intermediate segmentation maps, injecting additional supervisory information to enhance the model's predictive accuracy. The strategy is not applied at the lowest resolution scale, because small ROIs in medical images often become indistinguishable at this level (28 × 28). Low-scale supervision may introduce biases that cause the model to misclassify certain ROIs as background.

2.2 ECT Block

The vanilla multi-head self-attention (MHSA) mechanism suffers from high computational costs due to the high-dimensional linear transformations required for query, key, and value projections. To adapt the MHSA to high-dimensional image data, Vision Transformer (ViT) [9] reduces the number of Tokens by patching. However, in the image segmentation task, the patching process reduces the spatial resolution

of the features and causes a chunking effect on the segmentation map. In [31], a new vision Transformer architecture, called the convolutional vision Transformer (CvT), is proposed, which introduces convolution operations into the ViT model to improve the efficiency of MHSA. Tragakis et al. [29] further design a fully convolutional Transformer block based on the CvT, and use it as a basic component to construct a U-shaped segmentation network. In this work, we propose an enhanced convolutional Transformer (ECT) block by introducing group convolution to optimize token embedding and adding a point-wise MLP branch in the wide-focus stage to optimize the aggregation of MHSA output features. The workflow of the ECT block is shown in Fig. 2. Each ECT block consists of two modules connected in series, convolutional attention module and wide-focus module.

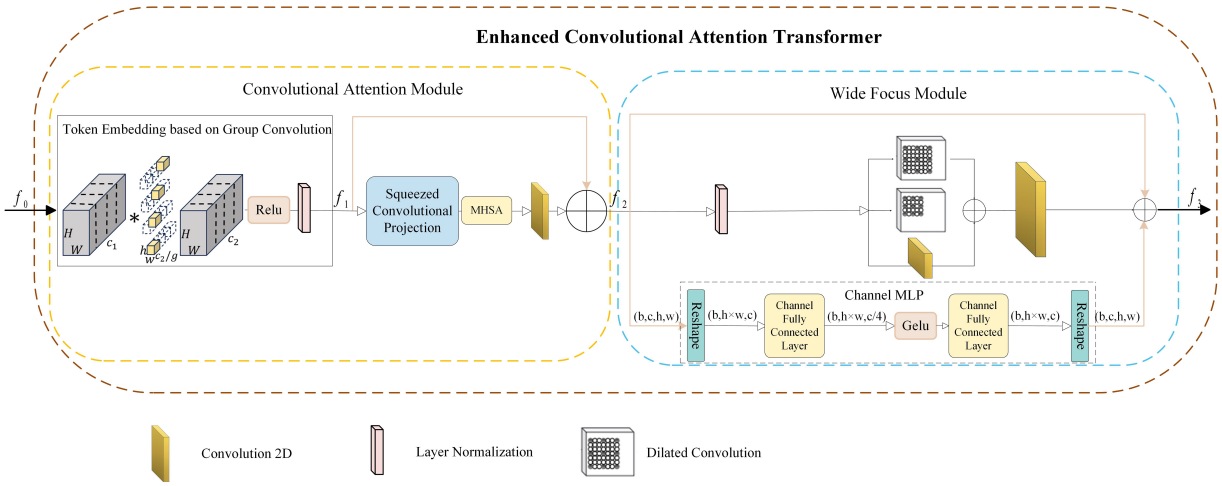

Figure 2: Workflow of enhanced convolutional transformer (ECT) block.

2.2.1 Convolutional Attention Module

The convolutional attention module consists of two stages: group-convolution-based Token embedding and squeeze convolutional projection (SCP) followed by multi-head self-attention (MHSA).

Token Embedding via Group Convolution

Compared with standard convolutional layers, group convolution reduces the number of trainable parameters while enabling channel disentanglement, i.e., each group learns distinct feature representations that are later aggregated.

Let $f_0 \in \mathbb{R}^{B\times c_1\times H\times W}$ denote the input features of the ECT block, where B is the batch size, c_1 is the number of input channels, and $H \times W$ is the spatial resolution. The c_1 channels are partitioned into g equal groups:

$$f_0 = \left[f_0^{(1)}, f_0^{(2)}, \ldots, f_0^{(g)}\right], \quad f_0^{(i)} \in \mathbb{R}^{B\times(c_1/g)\times H\times W}, \tag{1}$$

where $f_0^{(i)}$ is the i-th group of feature maps. Each group is independently processed by a $k \times k$ convolution with a kernel $W^{(i)} \in \mathbb{R}^{(c_2/g)\times(c_1/g)\times k\times k}$, where c_2 is the number of output channels and k is the kernel size.

The per-group output is $\hat{f}^{(i)} = W^{(i)} * f_0^{(i)}$. All group outputs are concatenated along the channel dimension:

$$\hat{f} = \left[\hat{f}^{(1)} \,\|\, \cdots \,\|\, \hat{f}^{(g)}\right] \in \mathbb{R}^{B\times c_2\times H\times W}, \tag{2}$$

where $\|$ denotes concatenation. The Token embedding is then obtained as

$$f_1 = \mathrm{LN}\left(\mathrm{GELU}\left(\hat{f}\right)\right) \in \mathbb{R}^{B\times c_2\times H\times W}, \tag{3}$$

where GELU is the Gaussian error linear unit activation function and LN denotes layer normalization.

The number of trainable parameters in the group convolution is $k^2c_1c_2/g$, compared to $k^2c_1c_2$ for a standard convolution, yielding a g-fold reduction. In this work, we set $k = 3$ and $g = 4$, which reduces the Token-embedding parameters by 75% while preserving the full c_2-dimensional output space. Although each group is convolved independently, the concatenated output is subsequently processed by attention, point-wise/channel mixing operations, and convolutional layers, allowing cross-group information exchange in later stages.

Squeeze Convolutional Projection and MHSA

To reduce computational cost while preserving spatial information, the standard linear projections used to compute Query (Q), Key (K), and Value (V) in MHSA are replaced with a squeeze convolutional projection (SCP) layer inspired by [31]. The SCP layer uses depth-wise separable convolutions to construct Q, K, and V:

$$Q,\ K,\ V = \mathtt{Flatten}(\mathtt{DSConv}_{Q,K,V}(f_1)), \tag{4}$$

where $\mathtt{DSConv}_{Q,K,V}$ denotes the depth-wise separable convolution for each of Q, K, and V, respectively. Each DSConv is implemented as Depth-wise Conv2d → LN → Depth-wise Conv2d [31], and Flatten reshapes the 2-D feature maps into a sequence of Token vectors for the subsequent attention computation.

Compared to a standard convolutional projection, SCP reduces the computational complexity from $\mathcal{O}(s^2c_2^2HW)$ to $\mathcal{O}(s^2c_2HW)$, where s is the depth-wise convolutional kernel size, by decoupling channel mixing from spatial filtering. Furthermore, the strides of the K and V branches are set to 2, while the stride of the Q branch is kept at 1 to maintain spatial alignment with the input. This reduces the number of key/value Tokens from HW to $HW/4$, thereby lowering the MHSA complexity from $\mathcal{O}(d(HW)^2)$ to $\mathcal{O}(d(HW)^2/4)$, where d denotes the embedding dimension per head.

The total per-block complexity of the convolutional attention module is:

$$\mathcal{O}\left(\frac{k^2c_1c_2\,HW}{g} + s^2c_2\,HW + \frac{d\,(HW)^2}{4}\right), \tag{5}$$

where the three terms correspond to group-convolution Token embedding, depth-wise separable QKV projection, and strided MHSA, respectively.

The output of the convolutional attention module is computed as

$$f_2 = f_1 + \mathtt{Conv}(\mathtt{MHSA}(Q,\ K,\ V)), \tag{6}$$

where MHSA denotes the multi-head self-attention computation, and Conv is a 3 × 3 convolution that maps the attention output back to the original spatial dimensions. The residual connection adds the attention output to f_1, forming f_2, which then serves as input to the wide-focus module.

2.2.2 Wide-Focus Module

The output of the convolutional attention module f_2 is fed to the wide-focus module for fine-grained information processing without loss of spatial context. The original wide-focus module employs multiple

branches of conventional and dilated convolutions with different dilation rates to capture hierarchical features via convolutional receptive fields at different scales. To further enhance the point-level feature representation and cross-channel feature aggregation, we add a parallel channel MLP branch, as shown in Fig. 2. It enhances each token's representational capacity by applying non-linear channel-wise transformations, without altering the spatial resolution of the feature map.

Our channel MLP branch is implemented by

$$\texttt{cMLP}\left(f_2\right) = \texttt{Reshape}\left(\texttt{FC}\left(\texttt{GELU}\left(\texttt{FC}\left(\texttt{Flatten}(f_2)\right)\right)\right)\right). \tag{7}$$

where `Flatten` first reshapes the input features of shape (b, c, h, w) to $(b, h \times w, c)$, allowing subsequent token-wise `FC` (full connection) processing. The first `FC` layer reduces the channel dimension to $c/4$, followed by a `GELU` activation to introduce non-linear elements. The second `FC` restores the channel dimension to c, and then the output is reshaped to the original dimensions (b, c, h, w). This design provides a good trade-off between computational efficiency and representational power, allowing the MLP to model non-linear channel-wise transformations within each token.

The final output of the wide-focus module is written as,

$$f_3 = f_2 + \texttt{Conv}\left(\sum_{d\in\{1,2,4\}} \texttt{DConv}_d\left(\texttt{LN}(f_2)\right)\right) + \texttt{cMLP}(f_2) \tag{8}$$

where $\texttt{DConv}_d$ denotes the dilation convolution of rate d.

2.3 ECA Skip Connections

In the proposed ECANet, we employ the efficient channel attention (ECA) module [32] in the skip connections to recalibrate channel-wise feature responses before encoder–decoder concatenation. The structure of the ECA module is shown in Fig. 3.

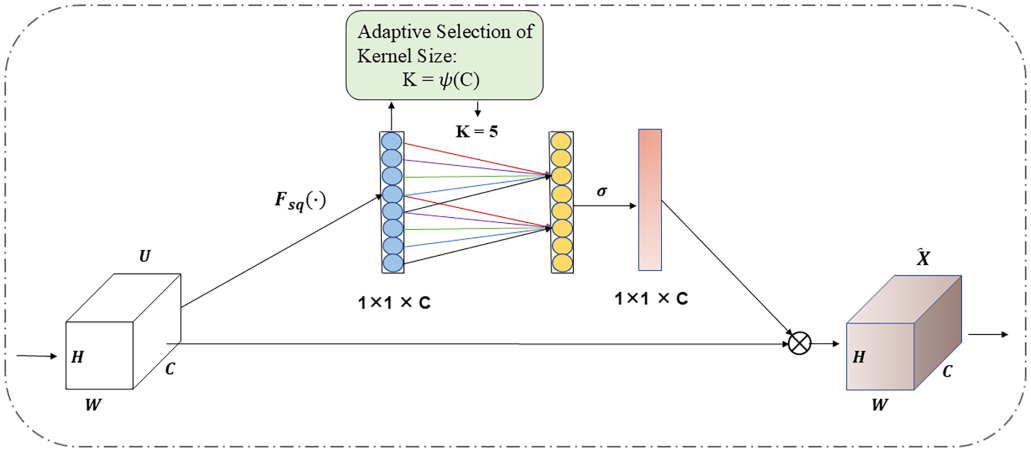

Figure 3: Structure of the ECA module used in skip connections.

The squeeze-and-excitation (SE) block [24] is a widely used channel attention mechanism that computes channel weights through global average pooling followed by two fully connected (FC) layers with a reduction ratio r, incurring $2C^2/r$ parameters. However, the intermediate dimensionality reduction ($C \to C/r \to C$) can break the direct correspondence between individual channels and their attention weights. ECA addresses this limitation by replacing both FC layers with a single 1-D convolution of kernel size k, which operates directly

on the full channel descriptor and models local cross-channel interactions without information compression. This design reduces the parameter count from $2C^2/r$ to just k (typically 3 or 5), regardless of the channel dimension C.

Specifically, given input features $f_{in} \in \mathbb{R}^{H\times W\times C}$, global average pooling first produces a channel descriptor $z_0 \in \mathbb{R}^C$, whose elements are computed as

$$z_0(c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_{in}(i, j, c), \tag{9}$$

where $\{i, j, c\}$ index the height, width, and channel dimensions, respectively. A 1-D convolution of kernel size k then infers the importance of each channel from z_0. After a Sigmoid activation (σ), a channel attention map $z_1 \in \mathbb{R}^C$ is obtained:

$$z_1 = \sigma(\texttt{Conv1D}_k(z_0)). \tag{10}$$

Following [32], the kernel size k is adaptively determined by the channel dimension to control the coverage of local cross-channel interaction:

$$k = \left| \frac{\log_2(C)}{2} + \frac{1}{2} \right|_{\text{odd}}. \tag{11}$$

In our implementation, $k = 3$ is used uniformly across all encoder levels, since the channel dimensions (8, 16, 32, 64) are small and the formula yields $k = 3$ for $C \leq 64$. The output features are obtained by channel-wise multiplication: $f_{out} = z_1 \cdot f_{in} \in \mathbb{R}^{H\times W\times C}$.

In the context of liver segmentation, the encoder's skip-connection features carry mixed semantic content: shallow layers encode fine edge textures, while deeper layers capture high-level structural context. By applying ECA before the encoder–decoder concatenation, diagnostically informative channels (e.g., those responding to organ boundaries or lesion contrast) are selectively amplified, while noise-dominated channels are suppressed. This channel-level recalibration is performed independently at each encoder level, preserving multi-scale feature diversity. Moreover, the skip connections provide shortcut paths for gradient propagation during back-propagation, which, combined with the deep supervision strategy described in Section 2.4, effectively shortens the gradient path and mitigates the vanishing gradient problem. The ablation study in Section 3.4 confirms that adding ECA alone yields a 0.779% Dice gain on liver segmentation and a 0.369% gain on liver tumor segmentation, with further synergistic improvements when combined with the ECT-block enhancements.

2.4 Deep Supervision

The deep supervision (DS) strategy accelerates model convergence by introducing intermediate supervisory signals at multiple decoder resolutions, which shortens the effective gradient path to early layers and mitigates the vanishing gradient problem. Without DS, gradients must propagate through all decoder blocks before reaching the encoder; with DS, each intermediate output provides an additional, shorter pathway, so the encoder and mid-level decoder layers receive more frequent and stronger gradient updates.

The strategy is deliberately *not* applied to the lowest-resolution level (28 × 28, Block 5). On the LiTS17 dataset, the smallest tumor lesions occupy as few as 1–3 pixels at this resolution, making reliable supervision infeasible; forcing a loss at this scale biases the model to classify small ROIs as background. This threshold

was verified empirically: applying DS at the 28×28 level degraded tumor Dice by approximately 0.4% in preliminary experiments.

The total loss function for training ECANet is defined as

$$\text{Loss}_{\text{total}} = \sum_{k=1}^{3} \alpha_k \, \text{Loss}_k, \tag{12}$$

where Loss_k is the loss at the k-th decoder output (ordered from highest to lowest resolution). Furthermore, α_k is the corresponding weight. We set $\alpha_{\{1,2,3\}} = \{0.5, 0.3, 0.2\}$ to prioritise higher-resolution outputs, which carry more spatial detail; this weighting was selected by grid search over $\{(0.6, 0.3, 0.1), (0.5, 0.3, 0.2), (0.4, 0.4, 0.2)\}$ and $(0.5, 0.3, 0.2)$ achieved the best liver Dice on a held-out validation fold. To ensure consistent loss calculation at all levels, bilinear interpolation upsampling and 1×1 convolution are used to scale up the outputs at levels $k = 2, 3$ to match the ground-truth resolution.

A composite loss is used at each level, combining BCE and Dice loss. The BCE Loss provides pixel-wise supervision that is sensitive to fine boundary details, while the Dice loss directly optimizes region-overlap and is robust to the severe class imbalance present in liver CT data (background pixels outnumber liver pixels by roughly 5:1 and tumor pixels by up to 50:1). BCE alone tends to predict the background to minimize loss; Dice alone can average out point-level boundary errors. Their combination jointly optimizes both pixel-level accuracy and global overlap, which is critical for segmenting small tumor lesions with complex boundaries.

Let $y_{i,j,c} \in \{0, 1\}$ and $\hat{y}^{(k)}_{i,j,c} \in (0, 1)$ denote the ground-truth label and predicted probability at position $\{i, j, c\}$ for level k. The composite loss at level k is

$$\text{Loss}_k = \alpha_{\text{BCE}} \, \text{BCE}^{(k)}_{\text{loss}} + \text{DICE}^{(k)}_{\text{loss}}, \tag{13}$$

where $\alpha_{\text{BCE}} = 0.5$ balances the two terms (selected by cross-validation) and

$$\text{BCE}^{(k)}_{\text{loss}} = -\frac{1}{N} \sum_{i,j,c} \left(y_{i,j,c} \log \hat{y}^{(k)}_{i,j,c} + (1 - y_{i,j,c}) \log(1 - \hat{y}^{(k)}_{i,j,c}) \right), \tag{14}$$

$$\text{DICE}^{(k)}_{\text{loss}} = 1 - \frac{2 \sum_{i,j,c} y_{i,j,c} \, \hat{y}^{(k)}_{i,j,c}}{\sum_{i,j,c} y_{i,j,c} + \sum_{i,j,c} \hat{y}^{(k)}_{i,j,c}}, \tag{15}$$

with N being the total number of elements of the ground truth.

3 Experiments

3.1 Dataset and Experimental Setup

We use the LiTS17 dataset, which was released as part of the 2017 MICCAI Liver Tumor Segmentation Challenge. LiTS17 comprises abdominal CT scans from 131 patients, along with expert-annotated labels for the liver and tumors, stored in NIfTI (.nii) format.

We applied a series of preprocessing steps to prepare the dataset for training. Firstly, we improved the visibility of the liver boundaries by limiting the voxel intensities within clinically relevant numerical ranges using the intensity window width technique. Additionally, we employed threshold segmentation to suppress irrelevant background regions. Next, we performed histogram equalization to adjust the image contrast, reduce noise, and further emphasize liver structures. Finally, we obtained 19,211 labeled 2D slices from 131 CT volumes, resizing each slice to 448×448 pixels. We randomly divided these slices into training and testing sets, allocating 70% for training and 30% for testing.

We implemented all methods using Python 3.10 and PyTorch 2.1.2 on an NVIDIA RTX 4060 GPU with 16 GB of memory. Table 1 summarises the hyperparameters used for training all models. The batch size was set to 8, which is the largest value that fits in the 16 GB GPU memory for the 448×448 input resolution. Stochastic gradient descent (SGD) with Nesterov momentum was adopted as the optimiser, with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 3×10^{-4}. These values were selected through a grid search over learning rates $\{10^{-2}, 10^{-3}, 10^{-4}\}$, weight decay values $\{10^{-3}, 3 \times 10^{-4}, 10^{-4}\}$, and optimisers {Adam, SGD}; the reported combination achieved the highest liver Dice on a held-out validation fold. Each model was trained for a maximum of 500 epochs with an early-stopping patience of 50 epochs (i.e., training is terminated if the validation loss does not decrease for 50 consecutive epochs), which prevents overfitting while ensuring convergence. The input images consist of 3 channels (the windowed CT slice concatenated with its histogram-equalised version and the original slice). No data augmentation was applied during training, so that all performance differences stem solely from architectural design. The composite BCE + Dice loss described in Section 2.4 was used, with the BCE weighting coefficient $\alpha_{\text{BCE}} = 0.5$ and deep-supervision weights $\alpha_{\{1,2,3\}} = \{0.5, 0.3, 0.2\}$ as detailed therein. All baseline models were trained with identical hyperparameters and data splits for a fair comparison.

Table 1: Hyperparameter settings for all experiments.

Hyperparameter	Value	Selection Method
Input resolution	448×448	Standard for LiTS17
Input channels	3	Fixed by preprocessing
Batch size	8	GPU memory constraint
Optimiser	SGD (Nesterov)	Grid search
Initial learning rate	0.01	Grid search
Momentum	0.9	Common default
Weight decay	3×10^{-4}	Grid search
Max epochs	500	Convergence guarantee
Early stopping patience	50 epochs	Empirical
Loss function	BCE + Dice	Task-specific design
BCE weight α_{BCE}	0.5	Cross-validation
DS weights $\alpha_{\{1,2,3\}}$	{0.5, 0.3, 0.2}	Grid search
Dropout rate	0.3	Common default

To evaluate the performance of the proposed model, we conducted extensive comparative experiments against representative CNN-based and CNN-Transformer-based segmentation models, including U-Net [4], U-Net++ [5], Attention U-Net [7], TransUNet [16], and FCT [29]. We used the Dice coefficient, intersection over union (IoU), volumetric overlap error (VOE), average symmetric surface distance (ASSD), and root mean square distance (RMSD) to quantitatively evaluate the segmentation accuracy. In addition to segmentation accuracy, we further report the number of parameters, floating-point operations (FLOPs), and inference time to analyse the computational cost of each model under the same input resolution and hardware setting. Additionally, ablation studies are performed to analyze the impact of each innovative component.

To enhance the reliability of the experimental evaluation, each model was independently trained and evaluated ten times using different random seeds. Therefore, the final quantitative results are presented in the form of the mean and standard deviation of key segmentation metrics (including the Dice coefficient and IoU) over multiple repeated runs. This setup can reduce the influence of random initialization and

data shuffling, thereby enabling a more stable assessment of the model performance. U-Net++ [5] reduces the semantic gap by using nested dense skip connections and deep supervision. However, this design also introduces considerable parameter redundancy and memory overhead. Attention U-Net [7] uses spatial attention gates in the skip connections to suppress irrelevant regions. Nevertheless, it does not perform channel-wise feature recalibration. It is also still constrained by the local receptive field of CNNs. FCT [29] expands the receptive field by replacing linear projections with depthwise separable convolutions. It also introduces multi-scale dilated convolutions to reduce part of the computational overhead. However, its wide-focus module lacks cross-channel point-wise aggregation during runtime. In addition, its skip connections do not include explicit channel recalibration.

The proposed ECANet addresses these limitations from multiple architectural perspectives. In the convolutional attention module, group convolution is used to reduce the cost of token embedding compared with dense convolution. The subsequent squeeze convolutional projection (SCP) further reduces computational complexity by separating channel and spatial operations. To improve fine-grained feature mixing, a Token-wise MLP branch is added after the wide-focus module. This branch provides point-level cross-channel aggregation, which is not present in FCT [29]. For the skip connections, the ECA mechanism introduces lightweight channel reweighting. This design avoids the dimensionality bottleneck commonly found in SE-based modules. Overall, these improvements allow ECANet to improve feature representation while maintaining a practical balance between segmentation accuracy and computational cost, rather than simply pursuing the lowest parameter count or fastest inference speed.

3.2 Comparative Results

Fig. 4 shows a comparison of the segmentation results from the five baseline models and the proposed model for several test samples. While all baseline models demonstrate basic capability in identifying the liver's general structure and approximating tumor locations, they have significant limitations in precise boundary delineation and tumor localization. In contrast, our ECANet consistently produces superior segmentation outcomes, providing more accurate anatomical boundaries and improved consistency in tumor region identification. For example, in Sample 4, the upper right region of the liver is extruded by other tissue, forming a more complex C-shaped contour. The contours segmented by all baseline models in this region are deformed, whereas the contours output by ECANet are well restored to the ground truth.

The advantages of our ECANet are even more pronounced in the segmentation of liver tumors with small tissue structures. For Sample 1, only U-Net and ECANet correctly identified all three tumor regions, while all other models missed some detections, and ECANet is significantly superior to U-Net in terms of tumor contours. For Sample 2, ECANet, benefiting from the enhanced convolutional Transformer block, is the only model that segments all tumor regions, including very small punctate tumor regions. For Samples 3 and 4, all models detect the correct number of tumor regions. However, our ECANet produces tumor shapes significantly closer to the ground truth. These results strongly illustrate our model's advantages in segmenting fine edges, especially those of small regions of interest.

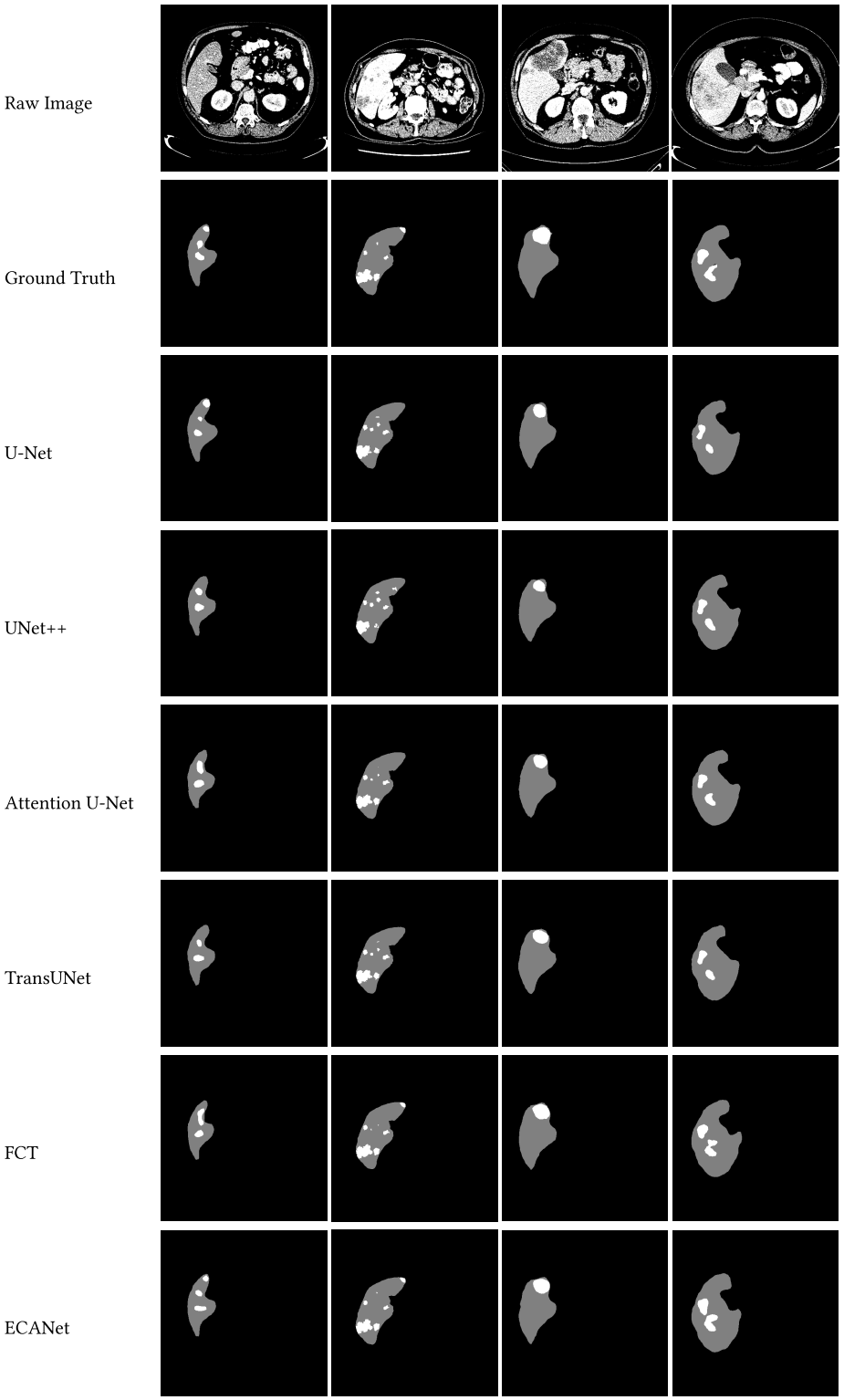

Figure 4: Segmentation results obtained by different models.

The quantitative evaluation results for the evaluated models for liver and liver tumor segmentation are given in Tables 2 and 3, respectively. Consistent with the visualization results in Fig. 4, the proposed ECANet outperforms the baseline models with significant improvements in liver and liver tumor Dice coefficients. Compared to CNN-based models (U-Net, U-Net++, and Attention U-Net), ECANet's liver Dice increased by 3.249%, 2.547%, and 1.953%, respectively. Compared to the CNN-Transformer hybrid baselines (TransUNet and FCT), ECANet's liver Dice increased by 1.948% and 1.461%, respectively. The advantage of ECANet is even more evident in liver tumor segmentation with fine tissue structures. Compared to CNN-based models (U-Net, U-Net++, and Attention U-Net), ECANet's tumor Dice increased by 5.877%, 4.789%, and 3.715%, respectively. Compared to the CNN-Transformer hybrid baselines (TransUNet and FCT), ECANet's tumor Dice increased by 2.606% and 1.591%, respectively. ECANet also achieves the best results in other quantitative metrics, IoU, VOE, and two surface-based metrics (ASSD and RMSD).

Table 2: Quantitative results on the liver dataset (mean ± std over 10 runs).

Model	Dice↑ (%)	mIoU↑	VOE↓	ASSD↓ (mm)	RMSD↓ (mm)
U-Net [4]	93.021 ± 0.132	0.870 ± 0.003	0.130 ± 0.003	2.251 ± 0.048	5.376 ± 0.089
UNet++ [5]	93.723 ± 0.118	0.882 ± 0.003	0.118 ± 0.003	2.113 ± 0.042	4.873 ± 0.076
Att. U-Net [7]	94.317 ± 0.105	0.892 ± 0.002	0.108 ± 0.002	2.046 ± 0.039	4.692 ± 0.071
TransUNet [16]	94.322 ± 0.121	0.893 ± 0.003	0.107 ± 0.003	2.037 ± 0.044	4.375 ± 0.082
FCT [29]	94.809 ± 0.097	0.901 ± 0.002	0.099 ± 0.002	2.016 ± 0.035	4.235 ± 0.063
Ours	**96.270 ± 0.085**	**0.928 ± 0.002**	**0.072 ± 0.002**	**1.678 ± 0.032**	**3.603 ± 0.055**

Note: Best results are shown in **bold**, and second-best results are underlined.

Table 3: Quantitative results on the liver tumor dataset (mean ± std over 10 runs).

Model	Dice↑ (%)	mIoU↑	VOE↓	ASSD↓ (mm)	RMSD↓ (mm)
U-Net [4]	69.741 ± 0.291	0.535 ± 0.004	0.465 ± 0.004	12.439 ± 0.263	20.176 ± 0.354
UNet++ [5]	70.829 ± 0.273	0.548 ± 0.003	0.452 ± 0.003	12.021 ± 0.226	21.948 ± 0.347
Att. U-Net [7]	71.903 ± 0.278	0.561 ± 0.004	0.439 ± 0.004	13.156 ± 0.251	20.810 ± 0.318
TransUNet [16]	73.012 ± 0.268	0.575 ± 0.003	0.425 ± 0.003	11.621 ± 0.234	19.802 ± 0.308
FCT [29]	74.027 ± 0.243	0.588 ± 0.003	0.412 ± 0.003	9.802 ± 0.222	17.968 ± 0.298
Ours	**75.618 ± 0.254**	**0.608 ± 0.004**	**0.392 ± 0.004**	**7.348 ± 0.165**	**13.277 ± 0.296**

Note: Best results are shown in **bold**, and second-best results are underlined.

To ensure the reliability and statistical significance of our experimental results, all models were trained and evaluated 10 times independently with different random seeds. Tables 2 and 3 report the mean and standard deviation (mean ± std) over 10 independent runs. Across all metrics, the standard deviations for ECANet remain low, pointing to statistically reliable gains rather than artifacts of training randomness. In particular, ECANet records the lowest Dice standard deviation for both liver (0.085%) and liver tumor (0.254%) segmentation. These results suggest that ECANet combines top segmentation accuracy with strong run-to-run consistency.

3.3 Computational Complexity Analysis

To further evaluate the computational characteristics of different models, we compare the number of trainable parameters, FLOPs, and inference time in Table 4. All models are evaluated under the same input resolution and hardware environment. The purpose of this comparison is not to claim that ECANet has the

lowest overall computational cost, but to clarify the relationship between architectural modification, model complexity, and segmentation performance.

Table 4: Comparison of model complexity and inference time.

Model	Params (M)	FLOPs (G)	Inference Time (ms)
U-Net [4]	31.39	171.47	6.91
UNet++ [5]	36.63	424.66	17.76
Attention U-Net [7]	34.88	204.07	8.21
TransUNet [16]	105.73	98.91	11.63
FCT [29]	1.95	3.76	39.05
Ours	**2.20**	**3.44**	**39.34**

Note: Best results are shown in **bold**.

As shown in Table 4, ECANet has 2.20M parameters and 3.44G FLOPs. Compared with FCT, the proposed model introduces a slightly larger number of parameters, increasing from 1.95 to 2.20 M. This increase mainly comes from the additional token-wise MLP branch and ECA-enhanced skip connections, which are designed to strengthen channel-wise feature interaction and multi-scale feature propagation. Meanwhile, the group-convolution-based token embedding reduces the parameter cost of the corresponding convolutional embedding operation. Therefore, the proposed architectural design does not simply reduce the total parameter count; instead, it redistributes the computational budget by reducing the cost of token embedding while enhancing the feature learning ability of other modules.

It is worth noting that the FLOPs of ECANet are slightly lower than those of FCT, decreasing from 3.76 to 3.44 G, while the inference time remains at a comparable level. This indicates that the proposed modifications do not introduce a substantial additional computational burden. More importantly, under a similar overall complexity level, ECANet achieves clear improvements in segmentation performance, especially for liver tumor segmentation, as reported in Tables 2 and 3. These results support the motivation of the proposed design: ECANet is not intended to be a purely lightweight model, but rather to achieve a better trade-off between segmentation accuracy and computational cost.

3.4 Ablation Study

This section evaluates the individual and combined contributions of the proposed innovative components based on the five-stage U-shaped architecture derived from the FCT model. Three components are evaluated: group convolution ('+Group') and the MLP branch ('+MLP'), added to the convolutional attention module and wide-focus module in the ECT block, and the ECA module ('+ECA'), applied to the skip connections. We start from the baseline FCT ('Base') and add each component one at a time to isolate its individual contribution. Tables 5 and 6 report the quantitative results for liver and liver tumor segmentation, respectively. In the liver segmentation task, adding group convolution brings a 0.903% Dice gain over the base model. Adding the MLP module leads to consistent improvements across all metrics. Incorporating the ECA module further improves the model performance. This module introduces lightweight channel-wise attention and selectively emphasizes informative features while suppressing less relevant responses. At the same time, it maintains computational efficiency.

Table 5: Ablation results on the liver dataset (mean ± std over 10 runs).

Model	Dice↑ (%)	mIoU↑	VOE↓	ASSD↓ (mm)	RMSD↓ (mm)
base	94.809 ± 0.098	0.901 ± 0.002	0.099 ± 0.002	2.016 ± 0.038	4.235 ± 0.067
+mlp	95.473 ± 0.087	0.913 ± 0.002	0.087 ± 0.002	1.911 ± 0.034	4.137 ± 0.058
+group	95.712 ± 0.093	0.918 ± 0.002	0.082 ± 0.002	1.872 ± 0.041	3.982 ± 0.063
+eca	95.588 ± 0.079	0.915 ± 0.002	0.085 ± 0.002	1.893 ± 0.031	4.025 ± 0.054
+mlp+group	95.927 ± 0.082	0.922 ± 0.002	0.078 ± 0.002	1.841 ± 0.036	3.726 ± 0.061
+mlp+eca	95.801 ± 0.091	0.919 ± 0.002	0.081 ± 0.002	1.876 ± 0.039	3.809 ± 0.057
+group+eca	96.019 ± 0.076	0.923 ± 0.002	0.077 ± 0.002	1.795 ± 0.029	3.627 ± 0.052
+mlp+group+eca	**96.270 ± 0.085**	**0.928 ± 0.002**	**0.072 ± 0.002**	**1.678 ± 0.032**	**3.603 ± 0.055**

Note: Best results are shown in **bold**, and second-best results are underlined.

Table 6: Ablation results on the liver tumor dataset (mean ± std over 10 runs).

Model	Dice↑ (%)	mIoU↑	VOE↓	ASSD↓ (mm)	RMSD↓ (mm)
base	74.027 ± 0.243	0.588 ± 0.004	0.412 ± 0.004	9.802 ± 0.218	17.968 ± 0.312
+mlp	74.721 ± 0.228	0.596 ± 0.003	0.404 ± 0.003	9.177 ± 0.196	17.028 ± 0.287
+group	74.219 ± 0.261	0.590 ± 0.004	0.410 ± 0.004	9.226 ± 0.237	16.745 ± 0.326
+eca	74.396 ± 0.219	0.592 ± 0.003	0.408 ± 0.003	9.218 ± 0.184	16.226 ± 0.269
+mlp+group	75.124 ± 0.237	0.602 ± 0.004	0.398 ± 0.004	8.386 ± 0.211	15.904 ± 0.301
+mlp+eca	75.415 ± 0.208	0.605 ± 0.003	0.395 ± 0.003	8.109 ± 0.175	14.026 ± 0.258
+group+eca	75.275 ± 0.249	0.604 ± 0.004	0.396 ± 0.004	8.125 ± 0.223	14.469 ± 0.317
+mlp+group+eca	**75.618 ± 0.254**	**0.608 ± 0.004**	**0.392 ± 0.004**	**7.348 ± 0.165**	**13.277 ± 0.296**

Note: Best results are shown in **bold**, and second-best results are underlined.

In the more challenging task of liver tumor segmentation, the effects of each component become more pronounced. These results validate the effectiveness of each component. Their combination produces a synergistic effect, leading to consistent performance improvements in both liver and liver tumor segmentation tasks. The ablation study shows that in the more difficult liver tumor segmentation task (see Table 6), each proposed component has made a significant contribution to performance improvement. These experimental results provide clear empirical verification of the effectiveness of each architectural innovation. More importantly, the combination of these components shows a strong synergistic effect. Each component contributes complementary improvements to the segmentation process. Together, they lead to consistent performance gains in both liver and liver tumor segmentation tasks.

Similarly to the comparative experiments, all ablation configurations were trained and evaluated over 10 independent runs with different random seeds. The standard deviations in Tables 5 and 6 further confirm the stability of our results. As more components are gradually integrated, the standard deviations remain consistently low. This indicates that the observed performance gains are robust and reproducible. It also suggests that these gains are not artifacts of random initialization.

4 Conclusions

This paper presents an enhanced convolutional attention network, named ECANet, for liver segmentation. The model follows a U-shaped architecture, which introduces efficient channel attention into the skip connections. This design strengthens multi-scale feature transfer and improves gradient flow across the network. In both the encoder and decoder, enhanced convolutional Transformer (ECT) blocks combine

group convolution—for lightweight token embedding and channel disentanglement—with a token-wise MLP branch that enriches cross-channel feature interaction. Training is guided by a deep supervision strategy with a hybrid loss of binary cross-entropy and Dice loss, improving boundary delineation in particular. We evaluate ECANet extensively on the public LiTS17 dataset. The results show that it surpasses existing CNN-Transformer hybrids on both quantitative metrics and visual quality. Complexity analysis further shows that ECANet achieves these improvements under a computational cost comparable to FCT, indicating that the proposed design improves segmentation accuracy without introducing a substantial additional computational burden. ECANet draws on the complementary strengths of convolutional attention and Transformer modules, yielding a model that is both accurate and robust. This makes it a practical choice for clinical settings like liver disease diagnosis, in which segmentation precision has a direct impact on downstream decisions. Future work may extend the architecture to other organ segmentation tasks or incorporate larger and more diverse clinical datasets.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by Chengdu Science and Technology Program (2026-YF08-00034-GX).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Yuyan Ning and Bo Yang; methodology, Yuyan Ning, Haiyun Huang and Bo Yang; software, Yuyan Ning; validation, Yuyan Ning, Haiyun Huang and Legend Zhang; formal analysis, Yuyan Ning; investigation, Yuyan Ning and Haiyun Huang; resources, Bo Yang and Wei Wei; data curation, Yuyan Ning and Legend Zhang; writing—original draft preparation, Yuyan Ning; writing—review and editing, Haiyun Huang, Legend Zhang, Wei Wei, Hao Quan and Bo Yang; visualization, Yuyan Ning; supervision, Bo Yang; project administration, Bo Yang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This study used the publicly available Liver Tumor Segmentation Challenge 2017 (LiTS17) dataset. The dataset supporting the findings of this study is available from the LiTS Challenge website. The experimental results generated and analyzed during the current study are included in this published article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shenouda M, Gudmundsson E, Li F, Straus CM, Kindler HL, Dudek AZ, et al. Convolutional neural networks for segmentation of pleural mesothelioma: analysis of probability map thresholds (CALGB 30901, alliance). J Imag Inform Med. 2025;38(2):967–78. doi:10.1007/s10278-024-01092-z.
2. Zhao J, Zhou Z, Wang X, Zhang H, Duan Z, Wang S, et al. Hematoma segmentation of spontaneous intracerebral hemorrhage based on watershed and region-growing algorithm. J Sichuan Univ. 2022;53(3):511–6.
3. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 3431–40.
4. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015. Berlin/Heidelberg, Germany: Springer; 2015. p. 234–41.
5. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Berlin/Heidelberg, Germany: Springer; 2018. p. 3–11.
6. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens. 2020;162:94–114.

7. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. arXiv:1804.03999. 2018.
8. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.
9. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
10. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. In: Computer Vision—ECCV 2022 Workshops. Berlin/Heidelberg, Germany: Springer; 2023. p. 205–18.
11. Huang X, Deng Z, Li D, Yuan X, Fu Y. MISSFormer: an effective transformer for 2D medical image segmentation. IEEE Trans Med Imag. 2023;42(5):1484–94.
12. Zhou HY, Guo J, Zhang Y, Han X, Yu L, Wang L, et al. nnFormer: volumetric medical image segmentation via a 3D transformer. IEEE Trans Image Process. 2023;32:4036–45.
13. Azad R, Arimond R, Aghdam EK, Kazerouni A, Merhof D. DAE-former: dual attention-guided efficient transformer for medical image segmentation. In: Predictive intelligence in medicine. Cham: Springer Nature Switzerland; 2023. p. 83–95.
14. Ma J, Li F, Wang B. U-Mamba: enhancing long-range dependency for biomedical image segmentation. arXiv: 2401.04722. 2024.
15. Ruan J, Li J, Xiang S. VM-UNet: vision mamba UNet for medical image segmentation. arXiv:2402.02491. 2024.
16. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal. 2024;97(2):103280. doi:10.1016/j.media.2024.103280.
17. Wang H, Cao P, Wang J, Zaiane OR. UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer. Proc AAAI Conf Artif Intell. 2022;36(3):2441–9. doi:10.1609/aaai.v36i3.20144.
18. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: transformers for 3D medical image segmentation. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA. p. 1748–58.
19. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Berlin/Heidelberg, Germany: Springer; 2022. p. 272–84.
20. Li C, Qiang Y, Sultan RI, Bagher-Ebadian H, Khanduri P, Chetty IJ, et al. FocalUNETR: a focal transformer for boundary-aware prostate segmentation using CT images. In: Medical image computing and computer assisted intervention—MICCAI 2023. Berlin/Heidelberg, Germany: Springer; 2023. p. 592–602.
21. Heidari M, Kazerouni A, Soltany M, Azad R, Aghdam EK, Cohen-Adad J, et al. HiFormer: hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 6202–12.
22. Zhang Y, Liu H, Hu Q. TransFuse: fusing transformers and CNNs for medical image segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2021. Berlin/Heidelberg, Germany: Springer; 2021. p. 14–24.
23. Liu W, Tian T, Xu W, Yang H, Pan X, Yan S, et al. PHTrans: parallelly aggregating global and local representations for medical image segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2022. Berlin/Heidelberg, Germany: Springer; 2022. p. 235–44.
24. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7132–41.
25. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Computer vision—ECCV 2018. Berlin/Heidelberg, Germany: Springer; 2018. p. 3–19.
26. Chen CR, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 347–56.

27. Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. TransAttUnet: multi-level attention-guided U-Net with transformer for medical image segmentation. IEEE Trans Emerg Top Comput Intell. 2024;8(1):55–68.
28. Pan X, Ge C, Lu R, Song S, Chen G, Huang Z, et al. On the integration of self-attention and convolution. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 805–15.
29. Tragakis A, Kaul C, Murray-Smith R, Husmeier D. The fully convolutional transformer for medical image segmentation. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 3649–58.
30. Xie Y, Zhang J, Shen C, Xia Y. CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2021. Berlin/Heidelberg, Germany: Springer; 2021. p. 171–80.
31. Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, et al. CvT: introducing convolutions to vision transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 22–31.
32. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11531–9.