

Article

DA-TransResUNet: Residual U-Net Liver Segmentation Model Integrating Dual Attention of Spatial and Channel with Transformer

Kunzhan Wang ¹, Xinyue Lu ¹, Jing Li ^{2,*} and Yang Lu ^{1,*}

¹ School of Computer Science, Jilin Normal University, Siping 136000, China; wangkunzhan@mails.jlnu.edu.cn (K.W.); luxinyue@mails.jlnu.edu.cn (X.L.)

² School of Educational Science, Jilin Normal University, Siping 136000, China

* Correspondence: lij964@jlnu.edu.cn (J.L.); luyang@jlnu.edu.cn (Y.L.)

Abstract

Precise medical image segmentation plays a vital role in disease diagnosis and clinical treatment. Although U-Net-based architectures and their Transformer-enhanced variants have achieved remarkable progress in automatic segmentation tasks, they still face challenges in complex medical imaging scenarios, particularly around simultaneously modeling fine-grained local details and capturing long-range global contextual information, which limits segmentation accuracy and structural consistency. To address these challenges, this paper proposes a novel medical image segmentation framework termed DA-TransResUNet. Built upon a ResUNet backbone, the proposed network integrates residual learning, Transformer-based encoding, and a dual-attention (DA) mechanism in a unified manner. Residual blocks facilitate stable optimization and progressive feature refinement in deep networks, while the Transformer module effectively models long-range dependencies to enhance global context representation. Meanwhile, the proposed DA-Block jointly exploits local and global features as well as spatial and channel-wise dependencies, leading to more discriminative feature representations. Furthermore, embedding DA-Blocks into both the feature embedding stage and skip connections strengthens information interaction between the encoder and decoder, thereby improving overall segmentation performance. Experimental results on the LiTS2017 dataset and Sliver07 dataset demonstrate that the proposed method achieves incremental improvement in liver segmentation. In particular, on the LiTS2017 dataset, DA-TransResUNet achieves a Dice score of 97.39%, a VOE of 5.08%, and an RVD of −0.74%, validating its effectiveness for liver segmentation.

Keywords: liver image segmentation; Transformer; dual attention block; spatial pyramid pooling; residual structure



Academic Editor: Jonathan Blackledge

Received: 23 December 2025

Revised: 27 January 2026

Accepted: 3 February 2026

Published: 5 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

MSC: 68U10

1. Introduction

Liver segmentation is a critical task in medical image processing, aiming to accurately delineate the boundaries of the liver. This process is indispensable for disease diagnosis, shaping treatment strategies, and conducting follow-up assessments postoperatively [1,2]. Compared with manual delineation, automatic segmentation not only improves efficiency but also reduces labor costs and inter-observer variability without compromising accu-

racy [3]. Therefore, there is an increasing demand for precise and reliable automatic liver segmentation techniques in clinical practice [1,4].

Convolutional neural networks (CNNs) have demonstrated strong capability in capturing local features and have been widely adopted in medical image segmentation. Among them, the U-Net architecture has become a cornerstone due to its encoder–decoder design and skip connections, which effectively preserve spatial information. Numerous U-Net variants have achieved remarkable success in various segmentation tasks [5]. Inspired by residual learning, Res U-Net was proposed to facilitate deeper feature learning and improve gradient propagation [6]. DA-ResUNet further incorporates dual-attention blocks (DA-Blocks) and residual connections into the U-Net framework, enhancing feature representation. However, CNN-based methods still suffer from inherent limitations, including restricted receptive fields and strong inductive biases of convolution operations. As a result, their ability to capture long-range dependencies and global contextual information remains limited, which can negatively affect segmentation accuracy, especially in complex anatomical structures such as the liver.

The Transformer architecture was originally proposed for sequence-to-sequence modeling in natural language processing and has demonstrated outstanding capabilities in capturing long-range dependencies [7]. Recently, Vision Transformer (ViT) has been successfully introduced into computer vision, including medical image segmentation, leading to notable performance improvements. Building upon ViT [8], TransUNet, proposed by Chen et al. in 2021, integrates Transformer encoders into the U-Net framework, combining the global modeling capability of Transformers with the localization strength of CNN-based decoders [9]. This hybrid design significantly improves segmentation accuracy by leveraging global contextual information. However, TransUNet does not fully exploit image-specific characteristics, particularly spatial and channel-wise feature dependencies. To further enhance hierarchical feature modeling, Swin-UNet incorporates Swin Transformer modules into a U-shaped architecture and achieves promising results [10]. Nevertheless, blindly deepening Transformer-based architectures does not always result in consistent improvements in segmentation performance. Consequently, developing a more effective integration strategy between Transformers and U-Net architectures for accurate medical image segmentation remains an open research challenge.

To address the aforementioned challenges, we propose an automated liver segmentation framework termed DA-TransResUNet, which is built upon a Res U-Net backbone rather than the conventional U-Net architecture. The main contributions of this work are summarized as follows:

- (1) We propose DA-TransResUNet, a novel liver segmentation network that adopts Res U-Net as the baseline architecture and integrates residual learning, dual-attention (DA) blocks, and Transformer encoders in a unified framework. Instead of a simple aggregation of existing modules, each component is carefully embedded into the network to enhance feature representation at different semantic levels, thereby improving segmentation accuracy.
- (2) We design a task-oriented hybrid encoder in which DA blocks are employed to explicitly model spatial and channel-wise dependencies in convolutional feature maps, while the Transformer encoder focuses on capturing global contextual information. This complementary design enables effective interaction between local detail preservation and global dependency modeling. In addition, Atrous Spatial Pyramid Pooling (ASPP) is introduced between the encoder and decoder to further enhance multi-scale feature representation. Moreover, each skip connection is refined by a DA block to reduce feature redundancy and alleviate the semantic gap between encoder and decoder features.

- (3) Unlike existing approaches such as TransUNet, DA-TransUNet, and SAR-U-Net, which mainly emphasize either global modeling or attention-enhanced convolutional features, DA-TransResUNet explicitly explores the synergistic collaboration among residual learning, dual attention, and Transformer-based global modeling within a Res U-Net framework. Experimental results on two public benchmark datasets show that DA-TransResUNet delivers competitive performance and further improves segmentation accuracy compared with several existing methods.

2. Related Work

2.1. Residual Structure

The residual structure was first introduced by He et al. [11] to address the degradation problem that arises when network depth increases. By introducing identity shortcut connections, residual learning enables deep networks to be trained more effectively and stably. In the field of medical image segmentation, residual architectures have demonstrated strong performance. For instance, Bi et al. [6] proposed a deep residual network for liver lesion segmentation and achieved fourth place in the LiTS2017 liver segmentation challenge.

As illustrated in Figure 1, a typical residual block consists of two convolutional layers followed by a shortcut connection. Instead of directly learning a complete mapping, the residual block learns a residual function, which is then added to the input feature map through element-wise addition. This design allows the network to preserve low-level features while progressively refining higher-level representations. As a result, residual structures facilitate deeper network construction and improve feature propagation in segmentation tasks.

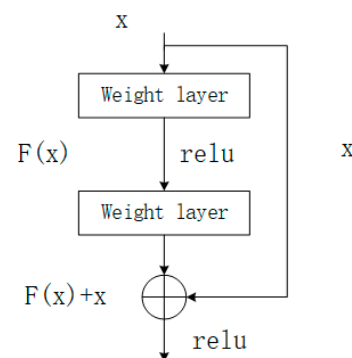


Figure 1. The diagram of the Residual Block.

2.2. Dilated Convolution

Dilated convolution was initially proposed by Yu et al. [12], is a variant of the standard convolution operation that inserts predefined gaps between kernel elements, thereby enlarging the effective receptive field without increasing the number of parameters. As illustrated in Figure 2, this mechanism allows the convolutional layer to capture broader contextual information while maintaining the original spatial resolution of the feature maps. In contrast, conventional convolution often relies on downsampling to expand the receptive field, which can lead to a loss of spatial detail. By preserving resolution, dilated convolution enhances the network's ability to model long-range dependencies and improves its representational capacity, which is particularly important in dense prediction tasks such as image segmentation [13]. For instance, Chen et al. [14] incorporated dilated convolution into the DeepLab V1 framework, demonstrating its effectiveness in capturing both local and global contextual information. Since then, this technique has become a foundational component in many semantic and medical image segmentation models,

effectively addressing challenges related to spatial resolution preservation, contextual modeling, and overfitting.

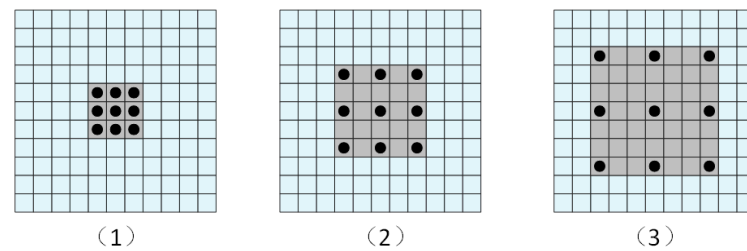


Figure 2. (1) Ordinary convolution; (2,3) dilated convolutions. The black dots denote the sampling points of the convolution kernel, and the shaded region represents the effective receptive field.

2.3. Transformer Encoder

The Transformer architecture, originally proposed for sequence modeling in natural language processing, has been widely adopted due to its powerful self-attention mechanism. In recent years, it has been successfully extended to computer vision tasks, giving rise to Vision Transformer (ViT)-based models [8].

As illustrated in Figure 3, a typical Transformer encoder consists of three main components: patch embedding, multi-head self-attention (MHSA), and a feed-forward multilayer perceptron (MLP). For visual inputs, the input image or feature map is first divided into a set of non-overlapping patches. Each patch is flattened and projected into a latent embedding space, and positional embeddings are added to preserve spatial information.

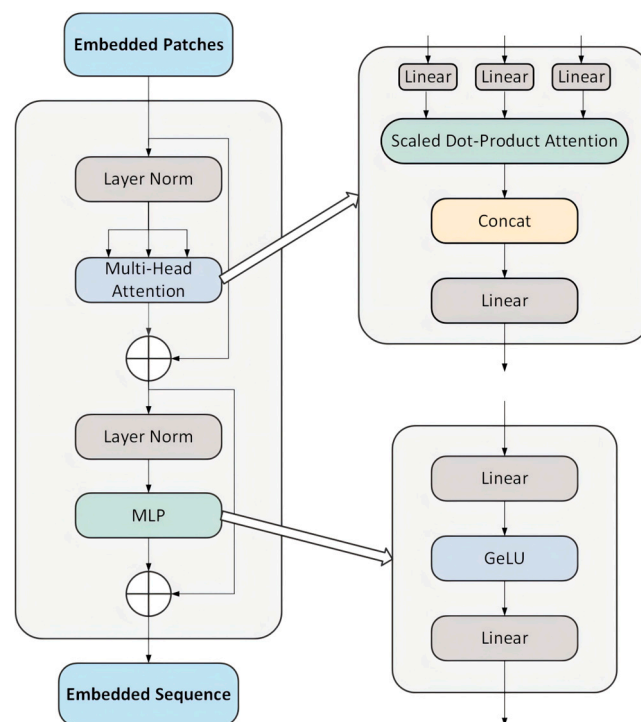


Figure 3. Diagram of the Transformer Encoder Block.

The embedded tokens are then processed by stacked Transformer encoder layers. Each encoder layer comprises an MHSA module and an MLP block, with residual connections and layer normalization applied to stabilize training. The self-attention mechanism explicitly models pairwise interactions among all tokens, enabling the capture of long-range dependencies and global contextual information. Moreover, multi-head attention allows the

model to attend to information from different representation subspaces, thereby enhancing feature diversity.

Owing to its ability to model global context and long-range relationships, the Transformer encoder has been widely integrated into medical image segmentation frameworks, often in combination with convolutional neural networks, to complement local feature extraction and improve segmentation performance.

2.4. Squeeze-and-Excitation Blocks

The Squeeze-and-Excitation (SE) block is a classical channel attention mechanism designed to enhance the representational capability of convolutional neural networks (CNNs). By explicitly modeling channel-wise feature dependencies, the SE block enables adaptive recalibration of feature responses.

As illustrated in Figure 4, the SE block consists of a squeeze operation followed by an excitation operation. In the squeeze stage, global average pooling is applied to the input feature map to generate a channel-wise descriptor that captures global contextual information. In the excitation stage, this descriptor is passed through two fully connected layers with a ReLU activation in between, followed by a sigmoid function to produce channel-wise weighting coefficients. These weights are then multiplied with the original feature map in a channel-wise manner to emphasize informative features and suppress less relevant ones.

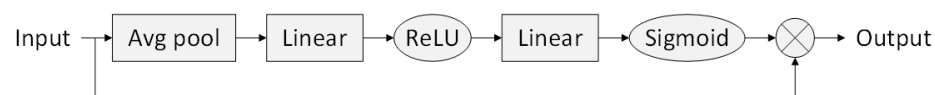


Figure 4. Diagram of the SE block.

Due to its simple structure and effectiveness, the SE block has been widely adopted in CNN-based architectures and has shown consistent improvements in various computer vision tasks, including medical image segmentation.

3. Methodology

3.1. Architecture

The proposed DA-TransResUNet is built upon the Res U-Net backbone and follows a symmetric U-shaped encoder–decoder architecture, as illustrated in Figure 5. The network processes input images with a resolution of $512 \times 512 \times 1$ and produces binary segmentation outputs of the same size. Based on the Res U-Net framework, residual blocks, squeeze-and-excitation (SE) blocks, dual-attention (DA) blocks, and Transformer-based encoding layers are integrated to enhance both local feature extraction and global contextual modeling.

Since Res U-Net serves as the baseline, residual blocks are employed throughout both the encoder and decoder to facilitate deeper representation learning while alleviating degradation issues. Each residual block incorporates shortcut connections to improve gradient flow. Batch normalization is applied after each convolution to stabilize training and provide implicit regularization, while the ReLU activation function mitigates vanishing gradient problems.

The DA blocks are strategically positioned after each residual block to refine the extracted features by emphasizing both spatial and channel-wise information. To alleviate the reduction in spatial resolution caused by successive downsampling, the network employs Atrous Spatial Pyramid Pooling (ASPP) as a transition layer. The ASPP module captures contextual information at multiple scales, enriching the feature maps with multi-scale semantic cues. Notably, the ASPP module is also incorporated at the decoder output to

further enhance segmentation accuracy. As illustrated in Figure 6, ASPP samples convolutional features at different dilation rates, effectively integrating multi-scale contextual information and facilitating the preservation of both fine details and global structure in the reconstructed feature maps.

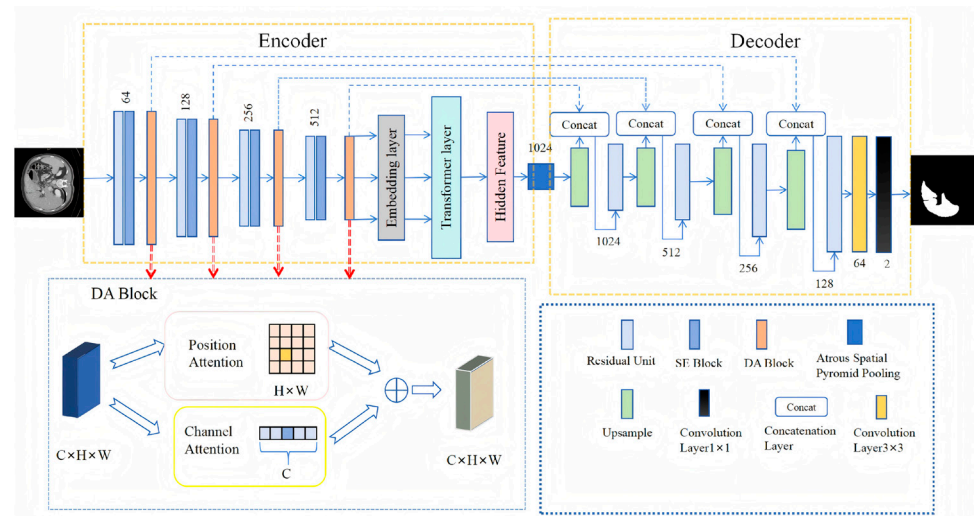


Figure 5. The architecture of DA-TransResUNet.

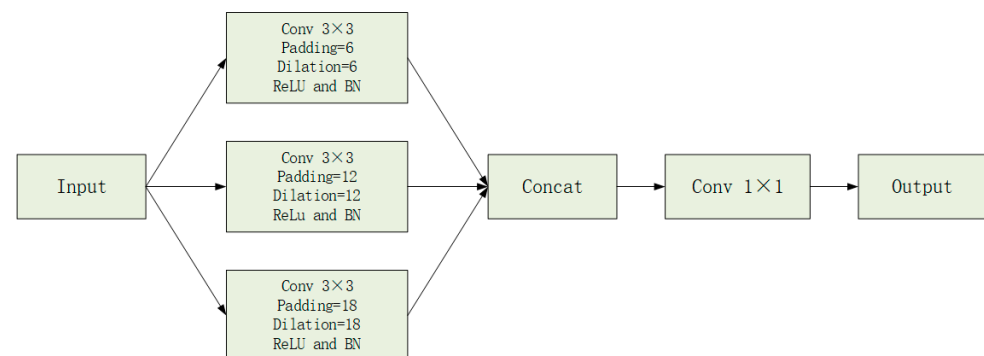


Figure 6. The diagram of the Atrous Spatial Pyramid Pooling model.

Through this combination of residual learning, dual attention, Transformer-based global modeling, and multi-scale context aggregation, the DA-TransResUNet effectively balances local detail preservation, long-range dependency modeling, and semantic richness, leading to improved segmentation performance for challenging medical images such as liver CT scans.

3.2. Encoder Combining Transformer, Residual Blocks and Dual-Attention Modules

As highlighted in Figure 5, the proposed encoder is composed of five main components: residual blocks, dual-attention (DA) blocks, squeeze-and-excitation (SE) blocks, Transformer layers, and an ASPP module. This architecture integrates hierarchical convolutional feature extraction with attention-based refinement and global contextual modeling, drawing inspiration from prior works on residual learning, attention mechanisms, and Transformer-based vision models [15]. The encoder begins with four residual blocks, which progressively expand the receptive field while preserving stable gradient propagation. Each stage performs spatial downsampling by a factor of two while doubling the channel dimension, thereby balancing feature representation capacity across hierarchical stages. These residual blocks extract multi-scale local features and establish a strong convolutional foundation for subsequent processing. Following convolutional feature extraction,

dual-attention blocks are introduced to enhance the discriminative quality of intermediate feature representations. Positioned prior to the Transformer layers, the DA blocks refine features by emphasizing informative spatial regions and salient feature channels, enabling more structured and semantically meaningful inputs for subsequent global modeling. An embedding layer then transforms the refined feature maps into a sequence representation with appropriate dimensional alignment, serving as a bridge between convolutional and Transformer-based feature processing. The Transformer encoder follows a standard multi-head self-attention architecture, configured with an embedding dimension of 512 ($d_{\text{model}} = 512$), 4 attention heads ($n_{\text{heads}} = 4$), a feed-forward hidden dimension of 32 ($d_{\text{ff}} = 32$), and 4 stacked Transformer layers ($L = 4$). This light weight configuration enables effective long-range dependency modeling while maintaining manageable computational complexity, making it suitable for high-resolution medical image segmentation tasks.

During forward propagation, the input image is processed through the residual blocks to extract hierarchical features, which are subsequently refined by the DA blocks. The enhanced representations are embedded and fed into the Transformer encoder to capture global contextual relationships beyond the limitations of conventional CNNs. Finally, the Transformer outputs are reshaped into spatial feature maps and forwarded to the intermediate decoding stage for multi-scale fusion and segmentation refinement.

3.3. Dual Attention Block

The Dual Attention (DA) module is designed to enhance feature representation by jointly modeling spatial dependencies and channel-wise relationships, enabling more discriminative and task-relevant feature extraction. This capability is particularly important for liver segmentation in CT images, where challenges such as low tissue contrast, ambiguous organ boundaries, heterogeneous intensity distribution, and large anatomical variability across slices often hinder accurate segmentation. Conventional convolutional operations primarily capture local patterns and may struggle to effectively distinguish liver regions from surrounding organs with similar visual characteristics. Although Transformer-based architectures are proficient at modeling global contextual dependencies, they are less specialized in emphasizing fine-grained spatial structures and channel-specific cues that are critical for medical image segmentation. To address this limitation, the DA module introduces complementary attention mechanisms that refine feature maps by selectively emphasizing informative spatial locations and discriminative feature channels.

By integrating position attention and channel attention, the DA block strengthens the network's ability to focus on anatomically meaningful liver regions while suppressing irrelevant background responses. This dual mechanism improves both boundary delineation and semantic separability, leading to more accurate segmentation of liver contours and internal structures. Furthermore, the DA module enhances feature consistency across encoder–decoder pathways, helping preserve fine details during feature propagation and reconstruction.

As illustrated in Figure 7, the DA block consists of two synergistic components: a Position Attention Module (PAM) and a Channel Attention Module (CAM), inspired by prior dual-attention frameworks for scene segmentation [16,17]. The PAM captures long-range spatial dependencies to emphasize structurally important regions, whereas the CAM models inter-channel correlations to prioritize feature responses that are most relevant to liver tissue. The detailed mechanisms of these two components are described in Sections 3.3.1 and 3.3.2, respectively.

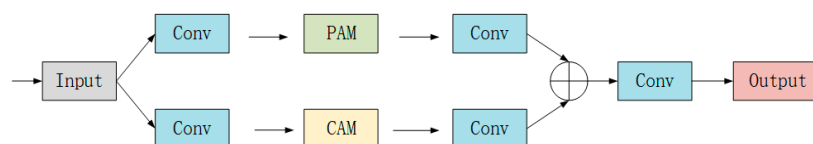


Figure 7. The diagram of the dual-attention block.

3.3.1. Position Attention Module

The primary role of the Position Attention Module (PAM) is to identify the spatial dependencies among different locations in the feature map. It achieves this by calculating a weighted sum of the features from all positions to refine specific features. The weights are assigned based on the similarity of features between two distinct positions. As a result, PAM is highly effective at extracting important spatial features.

As illustrated in Figure 8, a local feature $A \in \mathbb{R}^{C \times H \times W}$ is initially fed into a convolutional layer, producing two new feature maps, B and C, where $\{B, C\} \in \mathbb{R}^{C \times H \times W}$. Subsequently, B and C are reshaped into the format $\mathbb{R}^{C \times N}$, with $N = H \times W$ indicating the total number of pixels. Ultimately, the transposes of B and C undergo matrix multiplication, and the resulting product is processed through the softmax function to derive the spatial attention map $S \in \mathbb{R}^{N \times N}$. The calculation is represented in Equation (1).

$$s_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^N \exp(B_i \cdot C_j)} \quad (1)$$

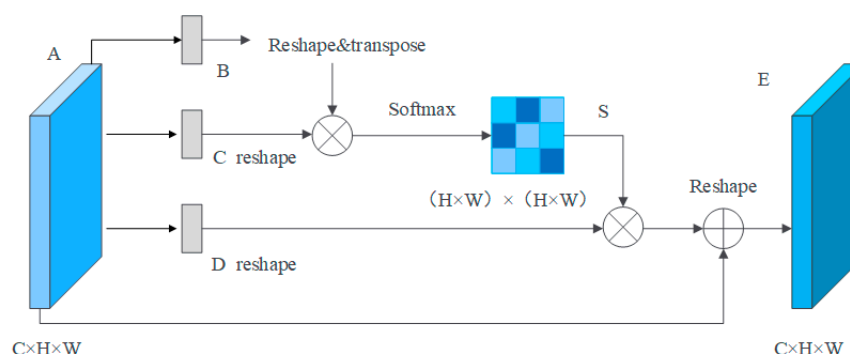


Figure 8. The diagram of the Position Attention Module.

In this context, s_{ji} denotes the effect of position i on position j . Next, feature map A is fed into the convolutional layer, resulting in a new feature map $D \in \mathbb{R}^{C \times H \times W}$. Subsequently, D is transformed into a matrix of dimensions $\mathbb{R}^{C \times N}$. Following that, a matrix multiplication is performed between D and the transpose of S, and the outcome is reshaped back into dimensions $\mathbb{R}^{C \times H \times W}$. Ultimately, this result is scaled by the parameter α , and an element-wise addition is conducted with the original feature map A, leading to the final output $E \in \mathbb{R}^{C \times H \times W}$. The calculation is represented in Equation (2).

$$E_j = \alpha \sum_{i=1}^N (s_{ji} D_i) + A_j \quad (2)$$

The parameter α starts off at zero and is gradually adjusted over time. As shown in Equation (2), the feature E_j at any given position represents a weighted accumulation of features from all positions, along with the original feature itself. This allows PAM to capture a comprehensive understanding of the context, selectively integrating information according to the spatial attention map. Consequently, it adeptly retrieves positional details while preserving the broader contextual information.

3.3.2. Channel Attention Module

The Channel Attention Module primarily aims to optimize the efficiency of channel dimension information within feature maps in neural networks. In the context of a Convolutional Neural Network, feature maps encompass details across both channel and spatial dimensions. The role of CAM is to dynamically adjust the significance of channel features, enabling the network to focus more on channels that hold greater relevance for the task at hand. As illustrated in Figure 9, this module's architecture allows for the highlighting of interdependent feature mappings, thereby enhancing the representation of specific semantic information. Ultimately, the feature output for each channel is derived from a weighted sum of all channels' features combined with the original feature. This approach mimics the long-term semantic relationships among feature maps and bolsters the distinctiveness of the features.

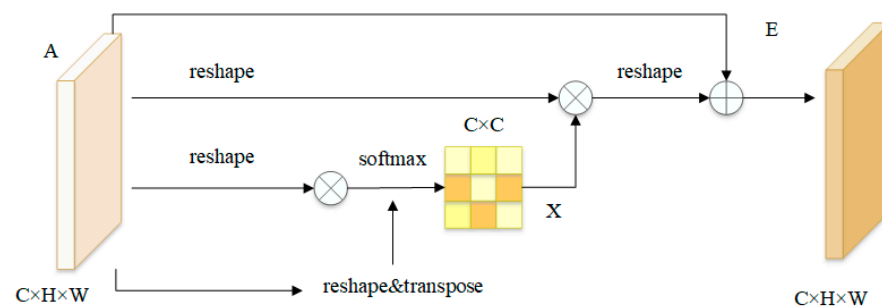


Figure 9. The diagram of the Channel Attention Module.

In contrast to the Position Attention Module, the Channel Attention Module (CAM) computes the channel attention map $X \in \mathbb{R}^{C \times C}$ directly from the original feature tensor $A \in \mathbb{R}^{C \times H \times W}$. To achieve this, the feature tensor A is first reshaped into $\mathbb{R}^{C \times N}$, followed by a matrix multiplication of A with its transpose. The resulting channel attention map $X \in \mathbb{R}^{C \times C}$ is then derived by applying the softmax function. The formula for this calculation is presented in Equation (3).

$$x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^C \exp(A_i \cdot A_j)} \quad (3)$$

Among these, x_{ji} quantifies the impact of the i -th channel on the j -th channel. Additionally, matrix multiplication is executed between X and the transpose of A , and the resulting product is reshaped into dimensions of $\mathbb{R}^{C \times H \times W}$. Following this, the outcome is scaled by a parameter β , after which an element-wise summation with A is performed to yield the final output $E \in \mathbb{R}^{C \times H \times W}$.

$$E_j = \beta \sum_{i=1}^C (x_{ji} A_i) + A_j \quad (4)$$

Among these, β progressively acquires the value starting from zero. As illustrated in Equation (4), the ultimate characteristic of each channel is derived from the weighted aggregation of all channel features alongside the original feature. This approach mimics the enduring semantic relationships within feature mappings, thereby enhancing the distinctiveness of the features.

3.4. Decoder

The decoder, illustrated on the right side of Figure 5, is responsible for reconstructing high-resolution feature maps from the encoded representations. It achieves this by inte-

grating features from the encoder and leveraging skip connections to retain fine-grained spatial information.

The decoder consists of three main components: feature fusion, upsampling convolutional blocks, and the segmentation head. During feature fusion, the decoder combines intermediate feature maps from the encoder (via skip connections) with the upsampled features from previous stages. This integration ensures that both high-level semantic information and low-level spatial details are preserved. The upsampling convolutional blocks progressively double the spatial dimensions of the feature maps while reducing the number of channels, effectively restoring the resolution at each stage. Finally, the segmentation head generates the output feature maps with the same spatial dimensions as the original input, providing pixel-wise predictions for segmentation.

The overall workflow of the decoder can be summarized as follows: at each stage, feature maps are upsampled and fused with the corresponding encoder features through skip connections. This process is repeated across multiple stages, systematically reconstructing the feature maps to their original resolution. By combining multi-scale information and refining spatial details, the decoder facilitates precise reconstruction of the input structure and enables accurate segmentation. This design allows the network to fully exploit both high-level semantic cues and low-level positional information, ensuring effective feature recovery and high-quality segmentation outputs.

4. Experiment

4.1. Dataset and Experimental Environment

Two datasets were used in this study: the LiTS2017 dataset and the Sliver07 dataset.

4.1.1. LiTS2017 Dataset

The LiTS2017 dataset is a public benchmark for the Liver Tumor Segmentation Challenge (ISBI2017 and MICCAI2017) and is widely used in liver and tumor segmentation studies. It comprises 131 abdominal CT scans with varying slice numbers (42–1026) and axial resolution of 512×512 pixels. The inter-slice spacing ranges from 0.45 mm to 6.0 mm. In our experiments, datasets 0–9 were used for validation, datasets 10–120 for training, and datasets 121–131 for testing.

4.1.2. Sliver07 Dataset

The Sliver07 dataset contains liver CT scans and corresponding segmentation labels for 20 patients, publicly available through the MICCAI2007 Liver Segmentation Challenge. Each image has a slice resolution of 512×512 pixels, with the number of slices per patient ranging from 64 to 502. The inter-slice spacing is from 0.7 mm to 5 mm, and the intra-slice spacing is from $0.56 \text{ mm} \times 0.56 \text{ mm}$ to $0.86 \text{ mm} \times 0.86 \text{ mm}$.

4.1.3. Data Split Strategy and Overfitting Analysis

Although the number of validation and test cases in LiTS2017 is relatively limited, each 3D CT volume contains a large number of 2D slices. After slice-wise decomposition, the effective number of training samples increases substantially, resulting in tens of thousands of training slices and providing sufficient data diversity for stable optimization.

The data splitting strategy was designed to balance training efficiency, validation reliability, and fair testing. A larger proportion of cases was allocated to training to improve model robustness, while independent validation and test subsets were preserved to prevent data leakage and ensure unbiased evaluation.

To mitigate the risk of overfitting, we applied data augmentation, intensity normalization, and regularization during training. Furthermore, the trained model was evaluated on

the independent Sliver07 dataset as an external benchmark, providing additional evidence of its generalization capability across heterogeneous datasets.

While k-fold cross-validation can provide additional statistical robustness, performing full cross-validation on slice-based CT data derived from high-resolution volumetric scans would significantly increase computational cost and training time. Instead, we prioritized maintaining a sufficiently large training set and assessed generalization using an independent external dataset (Sliver07). Future work will explore more extensive cross-validation protocols as computational resources permit.

4.2. Experimental Environment and Parameters

The models were trained using the Adam optimizer with an initial learning rate (*init_lr*) of 0.001. The learning rate was decayed according to Equation (5):

$$lr = init_lr \times \gamma^{\frac{epoch}{step_size}} \quad (5)$$

where $\gamma = 0.5$ and $step_size = 4$. Each model was trained for 60 epochs with a batch size of 4. All experiments were conducted on a GRID V100S-16C GPU with 16 GB of video memory, and a system equipped with 64 GB RAM. The implementation was based on Python 3.8 and the PyTorch 1.12 deep learning framework.

4.3. Image Preprocessing

To facilitate effective network training, the original liver CT images were preprocessed using several steps. First, a windowing technique with Hounsfield unit values ranging from -200 to 200 was applied to suppress irrelevant tissues and enhance the contrast between the liver and surrounding structures. Next, histogram equalization was performed to further improve image contrast and visibility. To standardize the data, all scans were resampled to 1 mm along the z-axis, and image intensity values were normalized to the range $[0, 1]$. As shown in Figure 10, these preprocessing steps sharpen the liver's texture and boundaries, improving the visibility of anatomical structures and facilitating subsequent segmentation tasks.

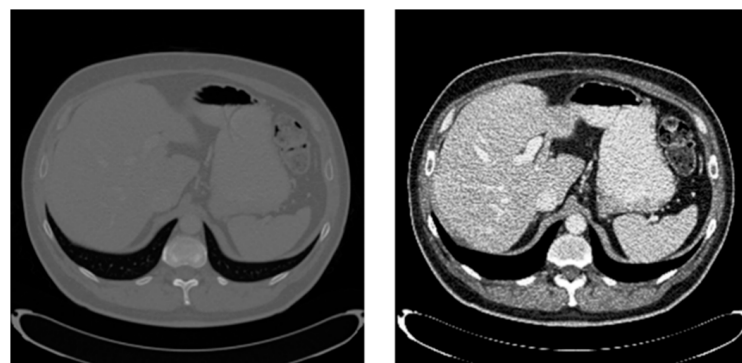


Figure 10. Image slices before and after preprocessing.

4.4. Loss Function

Cross-entropy is widely used in the training of various models by measuring the similarity between the predicted distribution and the actual situation. Its formula is as follows:

$$L = -\frac{1}{N} \sum_{i=1}^N [\hat{P}_i \log P_i + (1 - \hat{P}_i) \log(1 - P_i)] \quad (6)$$

In this formulation, $\hat{P}_i$ denotes the model's prediction and P_i denotes the gold standard. Additionally, N denotes the total count of samples. However, when it comes to

liver segmentation, the uneven distribution between the foreground (liver) and the background (non-liver) can significantly diminish the accuracy in smaller liver regions. To address this issue, a class-balancing strategy is adopted. An extra weight factor, denoted as ω_i^{class} , is incorporated into the loss function outlined in Equation (7), with its calculation formula being

$$\omega_i^{class} = \frac{N - n_i}{n_i} \quad (7)$$

Here, ω_i^{class} does not function as a regularization term; instead, it serves as a class-balancing coefficient that adjusts the contribution of each category to the loss, helping mitigate pixel-level class imbalance. n_i signifies the aggregate count of pixels that fall within the i -th category, while “ N ” encompasses the grand total of pixels across all categories. Essentially, the ratio is calculated by dividing the sum of pixels not in the class (N minus n_i) by the total pixels across all classes. Consequently, as the pixel count within the i -th category increases, the weight coefficient shrinks. Given that there is an abundance of background pixels in liver CT scans compared to liver-specific pixels, this weight factor adeptly addresses the imbalance between the two, ultimately enhancing the segmentation’s precision.

In this work, weighted cross-entropy is employed as the loss function, as it provides stable optimization, smooth gradient propagation, and effective handling of class imbalance, which is particularly important given that liver pixels occupy a small fraction of the CT slices relative to background. This weighting helps the network focus on underrepresented liver regions and improves segmentation accuracy.

We acknowledge, however, that weighted cross-entropy has certain limitations. Unlike Dice-based losses, it does not directly optimize overlap metrics such as Dice coefficient, which may limit its performance in maximizing foreground-background overlap. Additionally, weighted cross-entropy may be less sensitive to small structural errors at boundaries compared with hybrid Dice-cross-entropy losses. Exploring hybrid or Dice-based losses in future work could potentially further enhance segmentation performance, especially for fine-grained liver structures.

4.5. Evaluation Metrics

For liver segmentation evaluation, a quintet of standard metrics are utilized, including the Dice coefficient, volume overlap error, relative volume error, average symmetric surface distance, and maximum surface distance. Let A denote the segmented liver and B represent the ground truth. The descriptions of these five metrics are outlined below.

Dice coefficient (Dice): It is a metric used to measure the similarity between two samples, and it is especially commonly used to evaluate the degree of overlap between the segmentation result and the ground truth.

$$DC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (8)$$

Volume Overlap Error (VOE): It is a commonly used quantitative evaluation metric that measures segmentation accuracy by comparing voxel overlap between the segmentation result and the ground truth (Ground Truth), usually calculated based on the overlapping voxels and total volumes of both

$$VOE(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (9)$$

Relative Volume Error (RVD): It is used to measure the difference between the volume obtained by segmentation and the true (standard) volume. The closer the value is to 0, the higher the segmentation accuracy will be.

$$\text{RVD}(A, B) = \frac{|B| - |A|}{|A|} \quad (10)$$

Average Symmetric Surface Distance (ASD): It is a distance-based metric that measures the discrepancy between the segmentation result and the ground truth. ASD computes the average shortest Euclidean distance from each voxel on one surface to the corresponding surface voxels of the other. It serves as a quantitative indicator for evaluating segmentation model performance.

$$\text{ASD}(A, B) = \frac{1}{|S(A)| + |S(B)|} \left(\sum_{p \in S(A)} d(p, S(B)) + \sum_{q \in S(B)} d(q, S(A)) \right) \quad (11)$$

Maximum Surface Distance (MSD): This indicates the greatest surface distance between segmentation outcome A and the ground truth B. The smaller the MSD value is, the more concentrated the data distribution will be, with less volatility, and it will better meet the requirements of stability and accuracy.

$$\text{MSD}(A, B) = \max \left\{ \max_{p \in S(A)} d(p, S(B)), \max_{q \in S(B)} d(q, S(A)) \right\} \quad (12)$$

5. Experimental Results and Analysis

To evaluate the effectiveness of the proposed method, experiments were conducted using samples from Cases 121–130 of the LiTS2017 dataset and the Sliver07 dataset.

For consistent interpretation of the visualization results, the color scheme and evaluation indicators are defined as follows: The blue contours correspond to the ground truth, whereas the green contours represent the predicted segmentation results. Greater overlap between the predicted boundaries and the ground truth indicates superior segmentation performance. The Dice values shown in the visual results reflect slice-level accuracy and may differ from the mean volumetric Dice reported in the quantitative evaluation tables.

5.1. Ablation Experiment

To verify the effectiveness of the proposed network architecture and analyze the contribution of each component, a series of ablation experiments were performed. Res U-Net was adopted as the baseline model, and additional modules were incrementally integrated to assess their influence on segmentation performance. First, all convolutional blocks in both the encoder and decoder of the baseline U-Net were replaced with residual units, forming the Res U-Net architecture. Subsequently, modifications were applied to the encoder structure: SE blocks and Transformer blocks were individually introduced after the residual blocks to enhance channel attention and global contextual modeling, respectively. In addition, DA blocks were incorporated into both the residual blocks and the skip connections to further improve feature discrimination and multi-scale information fusion. Furthermore, based on the Res U-Net backbone, we conducted additional experiments by pairwise combining the DA, SE, and Transformer modules, in order to further validate the effectiveness and complementarity of each component in the proposed model. Finally, based on the Res U-Net backbone, all aforementioned modules were combined to construct the complete DA-TransResUNet model.

All models were trained on the LiTS2017 dataset and evaluated on test sets from both LiTS2017 and Sliver07. As shown in Tables 1 and 2, incorporating individual modules such

as SE blocks, DA blocks, and Transformer components into the Res U-Net framework results in consistent performance improvements. These findings indicate that each proposed module contributes positively to segmentation accuracy, and their combined integration achieves superior and more stable segmentation performance compared with individual module configurations.

Table 1. The results of the ablation experiment on the LiTS2017 dataset.

Model	DICE (%)	VOE (%)	RVD (%)	ASD (mm)	MSD (mm)
Res U-Net	96.93	9.84	−2.78	8.06	46.11
ResU-Net + DA	97.39	7.86	1.31	8.10	38.49
ResU-Net + SE	97.59	7.7	−1.66	8.11	39.83
ResU-Net + Transformer	97.64	8.99	−1.88	8.03	38.12
ResU-Net + DA + SE	97.67	7.62	−0.95	8.02	38.70
ResU-Net + DA + Transformer	97.74	7.55	−1.10	7.99	37.90
ResU-Net + SE + Transformer	97.76	7.48	−1.25	7.98	37.80
DA-TransResUNet	97.79	7.52	−1.02	7.97	37.58

Table 2. The results of the ablation experiment on the Sliver07 dataset.

Model	DICE (%)	VOE (%)	RVD (%)	ASD (mm)	MSD (mm)
Res U-Net	83.53	24.97	−7.61	13.88	155.50
ResU-Net + SE	83.87	19.13	5.19	11.36	149.04
ResU-Net + Transformer	86.19	17.79	4.65	9.73	138.61
ResU-Net + DA	86.87	16.72	3.16	9.69	101.43
ResU-Net + DA + SE	87.35	16.10	2.45	9.42	100.10
ResU-Net + SE + Transformer	88.30	16.35	2.10	9.05	105.40
ResU-Net + DA + Transformer	89.62	15.55	1.85	8.92	99.20
DA-TransResUNet	91.79	15.17	1.10	8.45	98.51

As shown in Table 1, each additional module contributes to performance improvements over the Res U-Net baseline. For instance, the Dice score increases from 96.93% to 97.64% with the integration of the Transformer block, while the MSD decreases from 46.11 mm to 38.12 mm, reflecting enhanced global contextual understanding. Moreover, multi-module configurations such as ResU-Net + SE + Transformer and DA-TransResUNet achieve more consistent improvements across Dice, VOE, and surface distance metrics, indicating that combining attention mechanisms and global modeling yields more robust segmentation behavior. From the results, while adding more modules generally improves segmentation performance, the marginal gains gradually decrease as additional modules are stacked. This trend indicates that each module provides complementary benefits, and the full integration achieves superior and stable performance without redundancy. These observations clarify the individual contributions and combined effectiveness of the proposed modules.

Figure 11 presents a qualitative comparison between the proposed method and other models listed in Table 1 on the LiTS2017 dataset. As modules are progressively incorporated into the baseline model, the segmentation results exhibit more precise boundary delineation and improved preservation of fine structural details. In particular, the complete model demonstrates more consistent contour alignment in complex regions, suggesting enhanced robustness in handling challenging anatomical structures. These qualitative observations align with the quantitative results, indicating that the integration of multiple feature enhancement mechanisms contributes to more accurate and reliable segmentation performance.

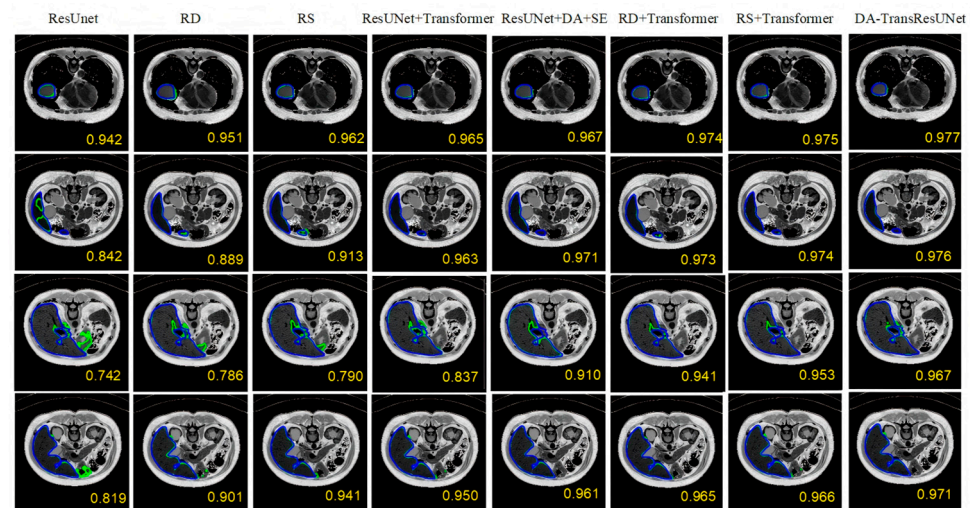


Figure 11. Prediction result graph on the LiTS2017 dataset. The blue bounding boxes represent the ground-truth annotations, while the green contours indicate the segmentation results, and the Dice value displayed in the bottom-right corner represents the slice-level segmentation performance (Note: ResU-Net + DA and ResU-Net + SE are abbreviated as RD and RS, respectively).

To further evaluate the contribution of each module under more challenging conditions, we conducted ablation experiments on the Sliver07 dataset using the same training settings and evaluation protocols. The quantitative results are summarized in Table 2.

Table 2 shows that augmenting the Res U-Net baseline with SE, Transformer, and DA modules yields consistent performance improvements on the Sliver07 dataset. The combined model demonstrates balanced gains across overlap and distance-based metrics, indicating enhanced robustness and segmentation reliability under challenging conditions. While adding additional modules generally improves segmentation performance, the marginal gains tend to decrease as more modules are stacked. This trend indicates that each module provides complementary benefits, and the full integration achieves robust and stable segmentation performance without redundancy. These observations clarify the individual contributions and combined effectiveness of the proposed modules on the Sliver07 dataset.

The qualitative results on the Sliver07 dataset indicate that the baseline Res U-Net model exhibits limitations in accurately delineating liver boundaries, particularly in regions with complex structures. Common issues include boundary inaccuracies and the mis-segmentation of non-liver tissues, as illustrated in Figure 12. As additional modules are progressively integrated into the network, these segmentation errors are noticeably reduced. The enhanced model demonstrates improved boundary precision, stronger discrimination of fine anatomical details, and a lower tendency to misclassify irrelevant regions as liver tissue. These improvements suggest that the proposed architectural refinements contribute to more reliable liver region identification and more accurate segmentation under challenging conditions.

In summary, the ablation experiments demonstrate that each proposed module contributes positively to the segmentation performance. The integration of DA, SE, and Transformer components consistently enhances liver region delineation, boundary accuracy, and the reduction in mis-segmentation errors across both the LiTS2017 and Sliver07 datasets. These results indicate that the modular enhancements provide complementary benefits, improving the model's reliability and robustness, particularly in handling complex anatomical structures.

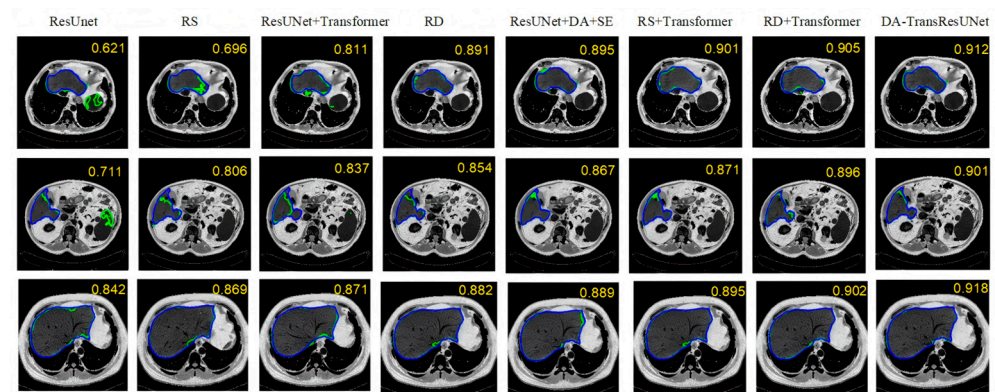


Figure 12. Prediction result graph on the Sliver07 dataset. The blue bounding boxes represent the ground-truth annotations, while the green contours indicate the segmentation results, and the Dice value displayed in the upper-right corner represents the slice-level segmentation performance (Note: ResU-Net + DA and ResU-Net + SE are abbreviated as RD and RS, respectively).

5.2. Comparative Experiment

To evaluate the effectiveness of the proposed method, we conducted comparative experiments against eight representative segmentation models, including UKAN, KM-UNet, FCN, U-Net, Attention U-Net, RMAU-Net, GCHA-Net, and SAR-U-Net, using the LiTS2017 dataset. All models except RMAU-Net and GCHA-Net were re-implemented and trained in our environment under identical experimental settings, utilizing 2D axial slices as input. The results of RMAU-Net and GCHA-Net are taken from other literature. During training, liver and lesion annotations were merged to focus on liver region segmentation. The quantitative results of the nine models on the LiTS2017 test set are summarized in Table 3.

Table 3. The results of the comparative experiment on the LiTS2017 dataset (results of RMAU-Net and GCHA-Net are reported from other literature; all other results are obtained in our environment).

Model	DICE (%)	VOE (%)	RVD (%)	ASD (mm)	MSD (mm)
UKAN	94.63	9.81	−2.30	8.26	53.15
KM-UNet	94.88	9.22	−1.95	8.20	51.97
FCN	91.83	14.64	−6.91	9.03	58.59
U-Net	92.67	13.20	−5.41	8.30	55.81
Attention U-Net	94.12	10.93	−3.58	8.21	54.47
RMAU-Net [18]	95.21	8.19	−0.40	-	-
GCHA-Net [19]	92.68	11.85	−1.50	-	-
SAR-U-Net	94.98	9.42	−2.15	8.08	52.61
DA-TransResUNet	95.22	7.52	−1.02	7.97	37.58

As shown in Table 3, the proposed DA-TransResUNet consistently achieves competitive performance across all evaluation metrics. Specifically, it obtains the highest Dice score of 95.22%, surpassing UKAN and KM-UNet by 0.59% and 0.34%, respectively, and outperforms SAR-U-Net by 0.24%. In terms of volumetric overlap error (VOE) and relative volume difference (RVD), DA-TransResUNet reduces errors by 1.90% and 1.13% compared with SAR-U-Net models, indicating more accurate liver region coverage. Moreover, improvements in average surface distance (ASD) and maximum surface distance (MSD) demonstrate enhanced contour delineation and boundary consistency, with DA-TransResUNet achieving 7.97 mm ASD and 37.58 mm MSD, which are the lowest among all evaluated models. These results collectively highlight the robust segmentation capability of the proposed method and its ability to provide precise and structurally consistent liver delineation compared with representative baselines.

Figure 13 presents a qualitative comparison of segmentation results produced by the models listed in Table 3 on the LiTS2017 dataset. Existing methods, including UKAN, KM-

UNet, FCN, U-Net, Attention U-Net, RMAU-Net, GCHA-Net, and SAR-U-Net, exhibit limitations in accurately delineating liver boundaries, often leading to blurred contours and confusion between the liver and adjacent anatomical structures. In contrast, the proposed model demonstrates more precise boundary delineation and improved contour continuity. Benefiting from high-resolution input and enhanced feature extraction, it achieves clearer separation between liver tissue and surrounding regions. The reduced occurrence of boundary misclassification suggests improved segmentation reliability, which may contribute to more consistent downstream medical analysis and clinical decision-making.

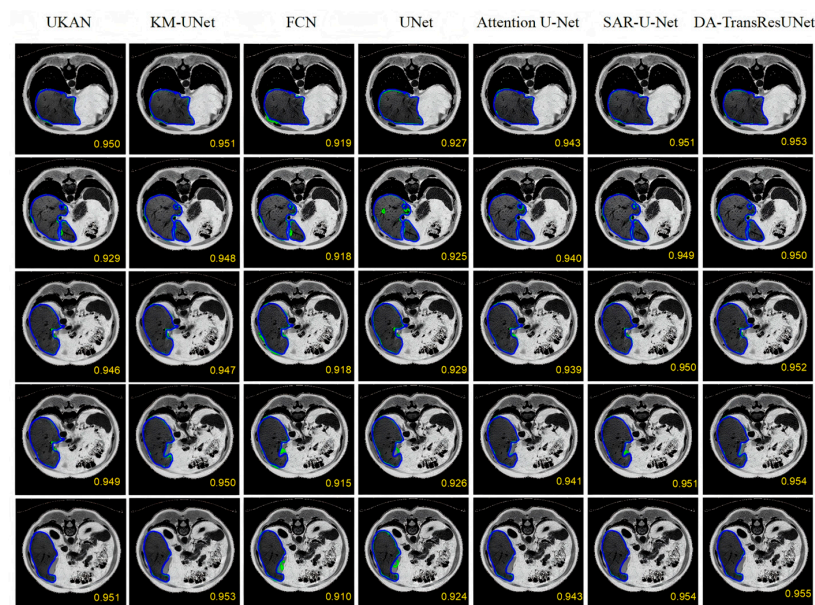


Figure 13. Comparison chart of predicted outcomes against other models on the LiTS2017 dataset. The blue bounding boxes represent the ground-truth annotations, while the green contours indicate the segmentation results, and the Dice value displayed in the bottom-right corner represents the slice-level segmentation performance.

Similarly, comparative experiments were conducted on the Sliver07 dataset to further evaluate the generalization ability of the proposed method. The quantitative results are summarized in Table 4.

Table 4. The findings from the comparative study using the Sliver07 dataset.

Model	DICE (%)	VOE (%)	RVD (%)	ASD (mm)	MSD (mm)
UKAN	89.18	18.31	−3.19	9.90	145.13
KM-UNet	90.05	17.44	−2.11	9.06	135.47
FCN	78.57	35.29	−12.36	15.54	174.04
U-Net	82.69	27.66	−8.93	12.51	167.20
Attention U-Net	88.81	20.12	−4.09	10.04	147.75
SAR-U-Net	90.51	17.11	−1.51	8.98	129.14
DA-ResTransUNet	91.79	15.17	1.10	8.45	98.51

As shown in Table 4, the proposed DA-ResTransUNet achieves the highest Dice score (91.79%) and the lowest ASD (8.45 mm) and MSD (98.51 mm) among all compared models, indicating improved overlap accuracy and boundary precision. Compared with other representative methods such as UKAN, KM-UNet, and SAR-U-Net, the proposed model demonstrates more consistent performance across both region-based and distance-based metrics. These findings further confirm that the proposed model maintains strong performance and reliable segmentation on the Sliver07 dataset.

Figure 14 presents a qualitative comparison of the segmentation results on the Sliver07 dataset for the models listed in Table 4. The results indicate that UKAN and KM-UNet, which do not incorporate explicit liver-specific anatomical constraints, have difficulty modeling complex hepatic boundaries, leading to segmentation errors in fine structures. FCN and U-Net occasionally misclassify regions outside the liver, reflecting insufficient sensitivity to background and surrounding anatomical structures. Attention U-Net and SAR-U-Net exhibit some challenges in accurately delineating liver boundaries, resulting in blurred contours between the liver and adjacent tissues.

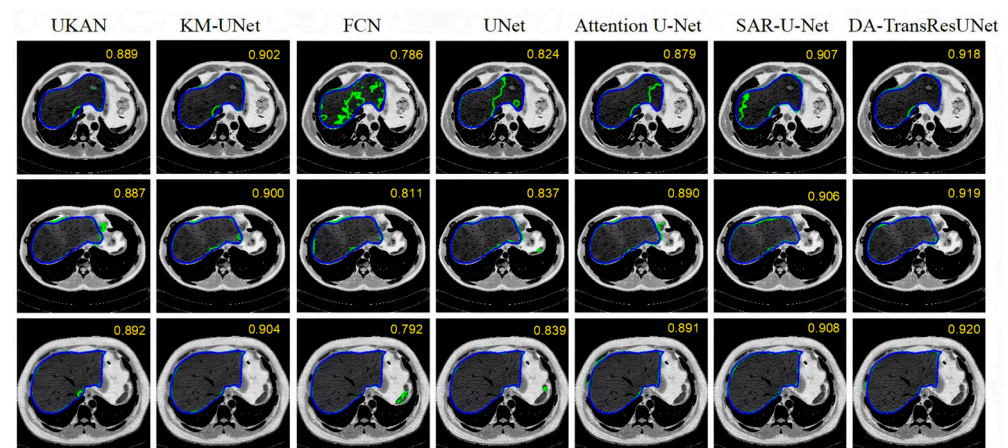


Figure 14. Comparison chart of prediction results with other models on the Sliver07 dataset. The blue bounding boxes represent the ground-truth annotations, while the green contours indicate the segmentation results, and the Dice value displayed in the upper-right corner represents the slice-level segmentation performance.

In contrast, the proposed DA-ResTransUNet demonstrates improved boundary delineation and more precise liver region identification. This improvement can be attributed to enhanced feature extraction and optimized training strategies. Overall, these observations suggest that the proposed method reduces boundary misclassification and improves segmentation reliability, which may support more accurate downstream medical analysis and clinical decision-making.

6. Conclusions

In this study, we proposed DA-TransResUNet, an automated liver segmentation framework for CT images built upon a Res U-Net backbone. The framework systematically integrates a dual-attention (DA) block, Transformer-based contextual modeling, residual connections, and ASPP-based multi-scale feature extraction. This combination enables the model to effectively capture both local spatial details and long-range contextual dependencies, enhancing feature representation and segmentation accuracy.

Liver segmentation remains challenging due to unclear boundaries and inaccurate delineation of fine structures. The proposed DA-TransResUNet addresses these challenges by adaptively aggregating long-range contextual information through the DA block and efficiently extracting multi-scale features with the ASPP module, resulting in improved segmentation accuracy and structural consistency. Experimental results and qualitative analyses demonstrate that the integration of the DA block and Transformer encoder provides clear advantages in capturing global context, particularly for complex anatomical structures, while the residual connections and multi-scale feature extraction further enhance feature discrimination and training stability.

Despite these promising results, the current framework processes 3D CT data as 2D slices, which limits exploitation of inter-slice contextual information along the z-axis.

Future work will focus on extending the model to 3D or hybrid 2.5D architectures to fully leverage volumetric information and further improve segmentation performance in practical clinical settings. Overall, DA-TransResUNet provides a robust and accurate approach to liver segmentation, highlighting the importance of contextual modeling, long-range dependency capture, and multi-scale feature integration in enhancing medical image segmentation performance.

Author Contributions: Conceptualization, Y.L.; Methodology, X.L.; Software, K.W.; Investigation, J.L.; Resources, J.L. and Y.L.; Writing—original draft, X.L.; Writing—review and editing, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Jilin Province (No. YDZJ202301ZYTS157).

Data Availability Statement: The data presented in this study are openly available in [LiTS2017] at [<https://sliver07.grand-challenge.org/>] (accessed on 21 March 2025)].

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

Abbreviations

The following abbreviations are used in this manuscript:

DA	Dual Attention
PAM	Position Attention Module
CAM	Channel Attention Module
CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
ViT	Vision Transformer
SAR	Spatial Attention Residual
SE	Squeeze-and-Excitation

References

1. Sattari, M.A.; Zonouri, S.A.; Salimi, A.; Izadi, S.; Rezaei, A.R.; Ghezelbash, Z.; Hayati, M.; Seifi, M.; Ekhteraei, M. Liver margin segmentation in abdominal CT images using U-Net and Detectron2: Annotated dataset for deep learning models. *Sci. Rep.* **2025**, *15*, 8721. [[CrossRef](#)] [[PubMed](#)]
2. Jaitner, N.; Ludwig, J.; Meyer, T.; Boehm, O.; Anders, M.; Huang, B.; Jordan, J.; Schaeffter, T.; Sack, I.; Reiter, R. Automated liver and spleen segmentation for MR elastography maps using U-Nets. *Sci. Rep.* **2025**, *15*, 10762. [[CrossRef](#)] [[PubMed](#)]
3. Nayantara, P.V.; Kamath, S.; Kadavigere, R.; Manjunath, K.N. Automatic liver segmentation from multiphase CT using modified SegNet and ASPP module. *SN Comput. Sci.* **2024**, *5*, 377. [[CrossRef](#)]
4. Delmoral, J.C.; Tavares, J.M.R.S. Semantic segmentation of CT liver structures: A systematic review of recent trends and bibliometric analysis: Neural network-based methods for liver semantic segmentation. *J. Med. Syst.* **2024**, *48*, 97. [[CrossRef](#)] [[PubMed](#)]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [[CrossRef](#)]
6. Diakogiannis, F.I.; 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. [[CrossRef](#)]
7. Vaswani, A. Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Available online: <https://arxiv.org/abs/1706.03762> (accessed on 1 April 2025).
8. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [[CrossRef](#)]
9. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [[CrossRef](#)]

10. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the Computer Vision—ECCV 2022 Workshops, Tel Aviv, Israel, 23–27 October 2022*; Proceedings, Part III; Springer Nature: Cham, Switzerland, 2022; pp. 205–218. [CrossRef]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; pp. 770–778. Available online: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (accessed on 1 April 2025).
12. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; pp. 472–480. Available online: https://openaccess.thecvf.com/content_cvpr_2017/html/Yu_Dilated_Residual_Networks_CVPR_2017_paper.html (accessed on 2 April 2025).
13. Guo, X.; Wang, Z.; Wu, P.; Li, Y.; Alsaadi, F.E.; Zeng, N. ELTS-Net: An enhanced liver tumor segmentation network with augmented receptive field and global contextual information. *Comput. Biol. Med.* **2024**, *169*, 107879. [CrossRef] [PubMed]
14. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
15. Wang, J.; Lv, P.; Wang, H.; Shi, C. SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in Computed Tomography. *Comput. Methods Programs Biomed.* **2021**, *208*, 106268. [CrossRef] [PubMed]
16. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; pp. 3146–3154. Available online: https://openaccess.thecvf.com/content_CVPR_2019/html/Fu_Dual_Attention_Network_for_Scene_Segmentation_CVPR_2019_paper.html (accessed on 4 April 2025).
17. Sun, G.; Pan, Y.; Kong, W.; Xu, Z.; Ma, J.; Racharak, T.; Nguyen, L.-M.; Xin, J. DA-TransUNet: Integrating spatial and channel dual attention with transformer U-net for medical image segmentation. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1398237. [CrossRef] [PubMed]
18. Jiang, L.; Ou, J.; Liu, R.; Zou, Y.; Xie, T.; Xiao, H.; Bai, T. RMAU-Net: Residual multi-scale attention U-Net for liver and tumor segmentation in CT images. *Comput. Biol. Med.* **2023**, *158*, 106838. [CrossRef] [PubMed]
19. Liu, H.; Fu, Y.; Zhang, S.; Liu, J.; Wang, Y.; Wang, G.; Fang, J. GCHA-Net: Global context and hybrid attention network for automatic liver segmentation. *Comput. Biol. Med.* **2023**, *152*, 106443. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.