A Liver Tumor Image Segmentation Method using Convolutional Attention

YuFu Li, *Ji Zhao**

***Abstract*—In the context of abdominal computed tomography (CT) imaging, precise identification and delineation of the liver and malignant tumors are critical for enabling accurate disease diagnosis and facilitating the development of advanced computer-assisted medical treatments. Given the varied forms of the affected regions and the indistinct edges,To achieve precise liver tumor segmentation in medical imaging, this study introduces a novel convolutional attention mechanism that enhances the accuracy of tumor boundary detection. The U-net network structure is employed, and the CHD convolutional attention module designed in this paper is used in the encoder to extract image features. Moreover, the SP operation is added to the skip-connections between the encoder and decoder to refine the shallow image feature information extracted by the encoder. The decoder is used to restore the image resolution and merge it layer by layer with shallow features, eventually obtaining the segmentation results. On the LiTS dataset, the Dice coefficients, VOE, and RVD results for the liver and tumors were respectively 96.76%, 8.29%, 2.53%, and 86.59%, 10.81%, 11.25%. Compared with other methods, this method has a certain advantage in liver tumor segmentation, which proves the effectiveness of this method.**

***Index Terms*—Liver tumors, Image Segmentation, Convolution Attention, U-net.**

I. Introduction

LIVER is the body's largest detoxification organ, and liver cancer, which originates in the liver cells, is one of the most common cancers worldwide. According to statistics, liver cancer ranks as the fifth leading cause of cancer-related deaths [1]. Common diagnostic methods for liver cancer include liver ultrasound, computed tomography (CT) scans, magnetic resonance imaging (MRI), and liver biopsies [2]. Among these, CT scans are widely used in clinical diagnosis due to their fast imaging speed and relatively low cost. CT scans provide three-dimensional images that reveal the liver's structure, the location, size, and shape of intrahepatic tumors, and extrahepatic metastases. Clinical physicians can use this information to develop appropriate treatment plans based on the lesions identified [3]. Nevertheless, the liver's grayscale contrast is nearly indistinguishable from the adjacent surrounding tissues and the tumor's neighboring regions, which significantly limits the ability to detect subtle changes in the hepatic tissue.In the context of computed tomography (CT) imaging, it is common to observe not only indistinct anatomical boundaries but also intricate internal configurations that often defy simple interpretation. Additionally, manual annotation is time-consuming and labour-intensive, with segmentation accuracy highly dependent on expert radiologists' subjective judgment. Therefore, leveraging computer-aided diagnosis to achieve real-time, accurate segmentation of liver tumors remains a highly challenging task. Early research on liver tumor image segmentation mainly attempted to use traditional image processing techniques [4]such as threshold-based segmentation, edge detection, and region growing methods [5]. However, in practical applications, traditional methods usually require manual feature selection, making algorithm parameter adjustment difficult, and they also demand high image quality [6]. With the advancement of computer technology, researchers began to explore machine learning-based methods for liver tumor segmentation, such as support vector machines, random forests, and k-means clustering. Nevertheless, due to the diversity and complexity of liver tumors, as well as the issue of data imbalance between normal tissues and tumors, machine learning-based methods often yield poor segmentation results for tumor regions and suffer from poor model generalization capability [7]. With the rise of artificial intelligence technology, deep learning-based methods have gradually become the mainstream approach for liver tumor image segmentation in recent years. In 2015, Ronneberger et al. [8] conducted a groundbreaking study that explored the potential applications of neural networks in medical diagnostics.In a groundbreaking approach, researchers introduced the U-Net architecture, which is based on an encoder-decoder framework.The architecture employs skip connections to integrate diverse feature representations derived from both the encoder and decoder components.This accomplishment highlights an exceptional capability within the domain of medical imaging analysis, where the system has consistently outperformed existing methodologies. The architecture of U-Net and its skip connections have profoundly impacted subsequent research. Researchers have proposed numerous variants of the U-Net framework to meet specific task requirements. For instance, U-Net++ [9] inserts multiple feature fusion modules between the encoder and decoder, each containing multiple feature maps at different levels. Res-Unet [10] and Dense-Unet [11] are inspired by residual connections and dense connections, respectively. They replace each submodule in the U-Net backbone with forms incorporating these connections, enhancing segmentation performance to varying degrees. Subsequent research on convolutional models has primarily focused on large-window convolutions, aiming to expand the receptive field. Recently, visual attention mechanisms such as Vision Transformers (VIT) [12] and Swin Transformers [13] have demonstrated superior performance in the image domain [14]. For instance, models like TransUnet [15] and LeViT-UNet [16] combine U-Net and Transformers by incorporating self-attention into the encoder. By embedding convolutional feature maps into

Manuscript received September 9, 2024; revised June 26, 2025. This work was supported Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

YuFu Li is a teaching assistant at the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: lyf22019@163.com).

Ji Zhao is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (Corresponding author, e-mail: 319973500069@ustl.edu.cn).

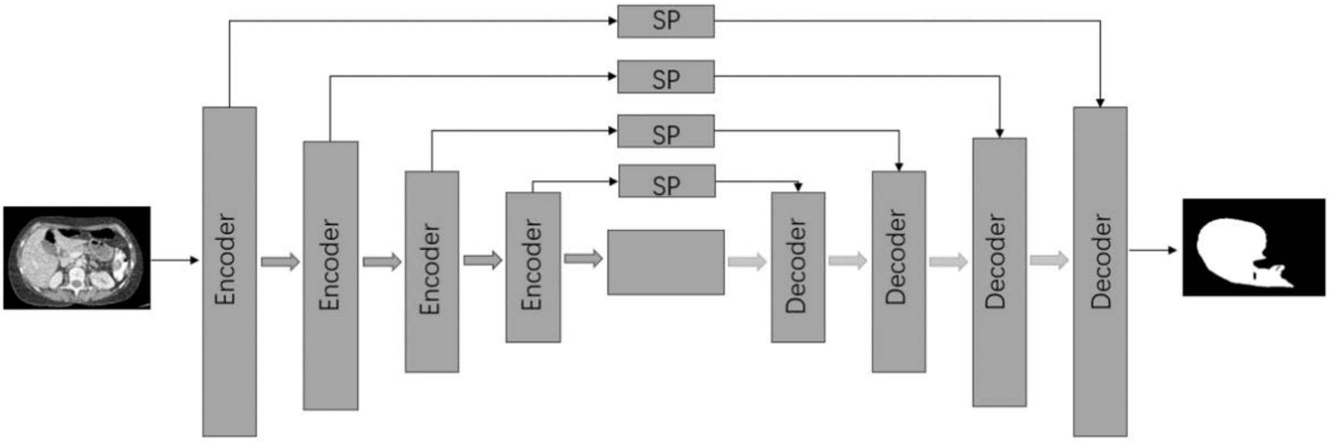

Fig. 1: Network structure

a tokenized representation of image patches, these models effectively capture global contextual information, which enables them to outperform existing approaches in multi-organ segmentation tasks. Swin-Unet [17] is restructured by integrating the entire encoder with a series of Swin Transformer blocks, which are designed to operate on shifted windows to capture contextual information effectively. It also designs a decoder with patch expansion layers in the Swin Transformer to perform upsampling and restore the spatial resolution of feature maps, outperforming TransUnet on multi-organ datasets. Similarly, the Swin Unetr [18] model, which shares this design philosophy, has achieved advanced performance on brain tumor datasets. These methods leverage Transformers for global long-range modelling of images, significantly enhancing image understanding capabilities. Nevertheless, the integration of attention mechanisms in models significantly elevates the parameter count, which in turn results in increased model intricacy, augmented hardware and memory demands, and complicates the training process [19]. The proposal of ConvNeXT [20] suggests that the superior performance of Transformers is due to their advanced network design philosophy. By altering the design of convolutional networks and adopting training strategies from Swin Transformers, the authors achieved performance surpassing that of Swin Transformers. Hou et al. [21] combined the design philosophies of ConvNeXT and Swin Transformer to propose a hierarchical convolutional network called Conv2Former. This network employs Hadamard [22] convolutional attention modulation instead of the self-attention mechanism. Doing so, reduces the number of parameters while achieving performance superior to mainstream convolutional network models and Transformer-based models. Due to the high computational cost associated with self-attention networks, despite their improved segmentation performance [23], this study is inspired by Conv2Former. We design the CHD convolutional attention module as the backbone network within the U-net framework. Additionally, we incorporate Strip Pooling [24] operations within the skip connections.This method enhances feature maps by systematically addressing both horizontal and vertical orientations, effectively eliminating non-essential areas while maintaining the critical edge details of the original image.Our findings indicate that the novel approach significantly improves the precision of liver and tumor segmentation in medical imaging.

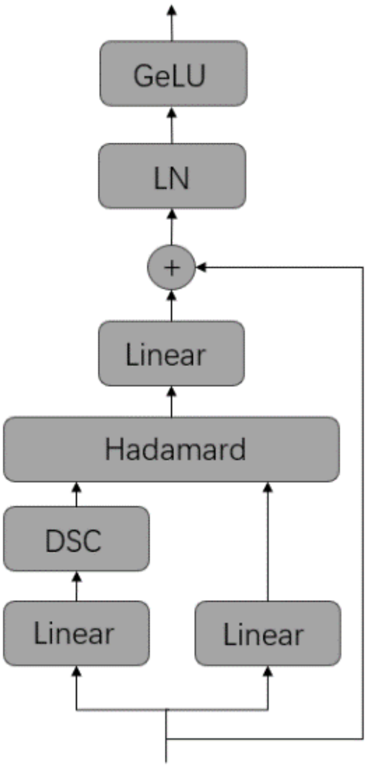

Fig. 2: CHD Convolutional Attention module

II. Methods

A. Framework Design

The U-Net architecture has had a significant impact on the development of medical image segmentation methodologies, primarily through its unique symmetric design and the implementation of skip connections that effectively bridge the encoder and decoder layers.To ensure that the output maintains high precision, retains all essential details, and remains resilient under varying conditions, the methodology incorporates advanced algorithms designed to refine accuracy, preserve information integrity, and improve system stability.In recent years, a significant number of leading-edge neural network architectures have emerged as derivative versions of the U-Net framework. Today, many high-performing network models are variants of U-net. Therefore, in this study, we adopt U-net as the fundamental design framework for the network model. The overall network structure is illustrated in Figure 1. The backbone network utilizes the CHD convolutional attention module designed in this paper, where the encoder is responsible for extracting image feature information, and the decoder performs upsampling operations to restore image resolution, followed by feature fusion with the extracted features from the decoder. Recognizing that the original U-net's skip connections directly fuse shallow image feature information extracted by the encoder with high-level semantic information obtained by the decoder, leading to information redundancy and affecting segmentation performance, we introduce Strip Pooling (SP) operations into the corresponding stage of skip connections between the encoder and decoder to optimize shallow features before fusing them with the high-level semantic information obtained by the decoder.

B. Convolutional block structure design

In the design of the convolutional block, drawing inspiration from the design principles of ConvNeXT and Conv2Former, each stage of the encoder and decoder consists of two CHD convolutional attention modules. The design structure of the CHD convolutional attention module is similar to that of the Transformer, consisting of linear layers, depth-wise separable convolutional layers [25], Hadamard modulation layers, layer normalization [26], and GeLU activation functions [27], as illustrated in Figure 2.

Original image $X \in R^{H \times W \times C}$ where H, W, and C represent the height, width, and number of channels of the image, respectively. The original image enters the CHD convolutional attention module through three paths, one of which serves as a residual connection to avoid the gradient vanishing and exploding problems caused by excessively deep networks. The other two paths serve as convolutional feature modulation values to simplify and replace self-attention. We use features extracted from depth-wise separable convolution and linear layers as weights for Hadamard convolutional modulation operations, replacing self-attention layers. The specific operations are as follows:

$$V = W_2 X \tag{1}$$

$$A = DSConv\left(W_1 X\right) \tag{2}$$

$$Z = A \odot V \tag{3}$$

Let X denote the input image, W_1 and W_2 are the weights of two linear layers, $DSConv(W_1 x)$ denotes the depth-wise separable convolution operation, e stands for the Hadamard product operation, and Z is the final output. Unlike traditional convolutional operations, depth-wise separable convolutions significantly reduce the number of parameters and lower computational demands without sacrificing their ability to enhance model performance. In contrast to the significantly increased parameter count of self-attention layers in the CHD convolutional attention module, the linear growth in the parameter count of the Hadamard convolutional modulation operation offers advantages.

C. Skip connection

In the skip connection stage, considering that the shallow-level features extracted by the encoder often contain low-level features of the image, while the high-level feature information of the decoder contains more semantic information, directly concatenating these two types of feature information will lead to information redundancy and reduce the model's generalization ability [28]. As a result, the Skip Connection is enhanced by incorporating a Strip Pooling (SP) module to refine the multi-scale feature information processed by the encoder, as shown in Figure 3. Unlike the SegNeXT model [29], which decomposes deep convolutions into two serial stripe convolutions, the SP module consists of two parallel paths. One path focuses on capturing distant context information of liver tumor images from horizontal and vertical dimensions, while avoiding unnecessary connections between distant positions, highlighting liver and important tumor information. The other path serves as a residual connection to avoid overfitting. Let $X \in R^{H \times W \times C}$ denote the input image, where C is the number of input channels. The dual parallel routes consist of a horizontal or vertical stripe pooling layer, which serves as the core component in this architectural design.Next, we introduce a single-layer convolutional architecture that employs a 3x3 kernel to process the input data.Previously, the model was designed to adjust the current location and its adjacent elements.Given the definitions of $y^h_{c,j} \in R^{C \times H}$ and $y^v_{c,j} \in R^{C \times W}$, Then $y \in R^{C \times H \times W}$ can be represented

$$y_{c,i,j} = y^h_{c,j} + y^v_{c,j} \tag{4}$$

The final output is represented as:

$$z = Scale\left(x, \sigma\left(f\left(y\right)\right)\right) \tag{5}$$

Where $Scale\left(.,.\right)$ represents element-wise multiplication, σ stands for the Sigmoid function, and f represents the 1x1 convolution.

III. Experiments and Analysis

A. Dataset

The dataset used is the MICCAI 2017 Liver Tumor Segmentation Challenge (LiTS) dataset, which includes CT scans from 201 patients. Among these, 131 CT scans are designated as the training dataset and 70 CT scans as the test dataset. This dataset, sourced from six medical centres worldwide, is currently the largest and most authoritative publicly available dataset in the liver tumor CT image segmentation field. All CT scans from the 201 patients were annotated by experienced radiology professionals. In abdominal CT scans, the Hounsfield Unit (HU) values for the liver and tumors can vary between 50 and 70. Therefore, a threshold is set to limit the HU values to the range of [-200, 200], with values outside this range being ignored. Subsequently, the pixel values are normalized (0, 1), with overflow values clamped to the boundary values. The training data comprises 131 medical CT images, each with a resolution of 512×512 pixels, and the number of slices in each image varies between 230 and 960.As a consequence of the processing steps, a total of 85,150 individual image slices were generated.

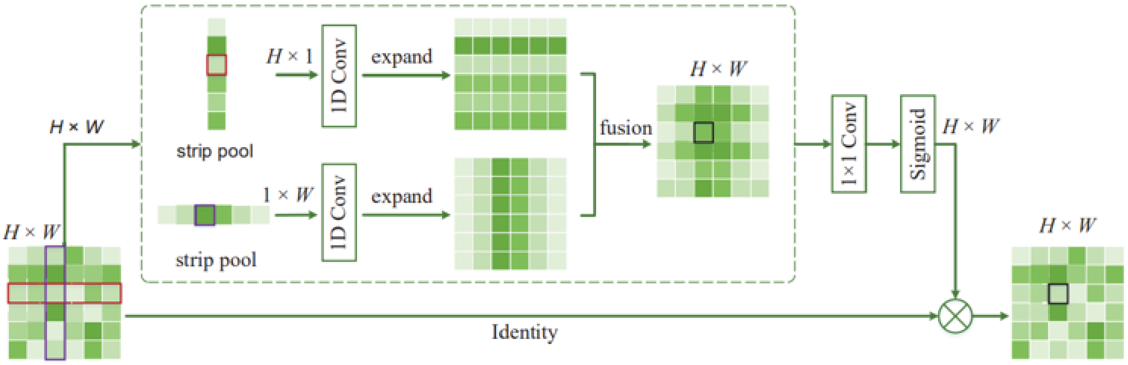

Fig. 3: Strip Pooling Moudle

B. Experimental environment

The experimental environment for the method described in this paper consists of a computing platform running the Ubuntu Linux operating system, equipped with an Intel Xeon E5 processor, 128GB of memory, and two Nvidia GTX Titan XP high-performance graphics cards. The development process is based on the Python programming language, and it leverages the PyTorch deep learning framework to build and train the model.The model training process is executed with the Adam optimization algorithm, employing its default parameters, which include an initial learning rate of 0.001.The model employs a first-moment decay coefficient of 0.9, which ensures rapid dissipation of initial energy, and a second-moment decay coefficient of 0.999, further enhancing the transient behavior of the system, while the stability parameter is set to 1e-8 to maintain numerical robustness during simulation.The total number of training iterations is set to 200, and the batch size used during each epoch is 8.

C. Evaluation Criteria

Three primary quantitative assessment tools were employed in this study to measure the effectiveness of the network model's segmentation process.The Dice Coefficient, the Volumetric Overlap Error (VOE), and the Relative Volume Difference (RVD) are three distinct metrics that provide complementary insights into the spatial alignment and geometric consistency of overlapping datasets. The formulas are as follows:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$VOE = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

$$RVD = \frac{|B| - |A|}{|A|} \tag{8}$$

Among these metrics, A denotes the anticipated outcome, while B corresponds to the actual value, TP denotes the number of instances where both the predicted and actual labels are positive, reflecting correctly identified positive cases.In this context, FN denotes the number of actual negative cases that were incorrectly classified as positive. However, the ground truth is positive, and FP represents the false positive count where the prediction is positive, but the ground truth is negative. The Dice Coefficient and the Jaccard Index are used to assess the similarity of image segmentation results. However, in liver tumor segmentation tasks, the liver occupies a much larger portion of the image compared to the tumor, leading to class imbalance issues. Compared to the Jaccard Index, the Dice Coefficient places more emphasis on matching positive samples, making it more suitable for medical image processing tasks.

D. Loss Function

In classification tasks, a loss function quantifies the disparity between the predicted outputs of a model and the actual target labels. The goal of neural network optimization is continuously updating parameters to find weights that minimize the loss function. The cross-entropy loss function is a commonly used loss function for classification tasks. It is defined as follows:

$$L_{BCE} = \frac{1}{N}\sum_{i=1}^{N} L_i = -\frac{1}{N}\sum_{i=1}^{N}\sum_{C=1}^{M} y_{ic}\log\left(p_{ic}\right) \tag{9}$$

where N is the number of pixels, C is the number of predicted classes,y_{ic} is the label for class c, which takes the value of 1 or 0, and p_{ic} is the predicted i probability of the sample belonging to class c. In the liver tumor image segmentation task, since the background and liver occupy most of the overall image and the tumor proportion is very small, the amount of negative samples far exceeds that of positive samples, leading to a class imbalance problem. As a result, the cross-entropy loss function inherently favors training outcomes that align more closely with samples of larger magnitude in the dataset.By decelerating the network's rate of integration, this approach inadvertently hampers the precision of data partitioning.In the field of medical image segmentation, the Dice loss function is frequently employed due to its ability to effectively manage class imbalance and address challenges posed by unclear or fuzzy boundaries in images.This concept refers to the ratio between twice the area of the overlapping region formed by two sets and the total combined size of both sets.The following explanation provides a detailed clarification:

$$L_{Dice} = 1 - \frac{2\left|X \bigcap Y\right|}{|X| + |Y|} = 1 - \frac{2TP}{FP + 2TP + FN} \tag{10}$$

where X represents the set of predicted values, and Y represents the ground truth set. This data set is considered

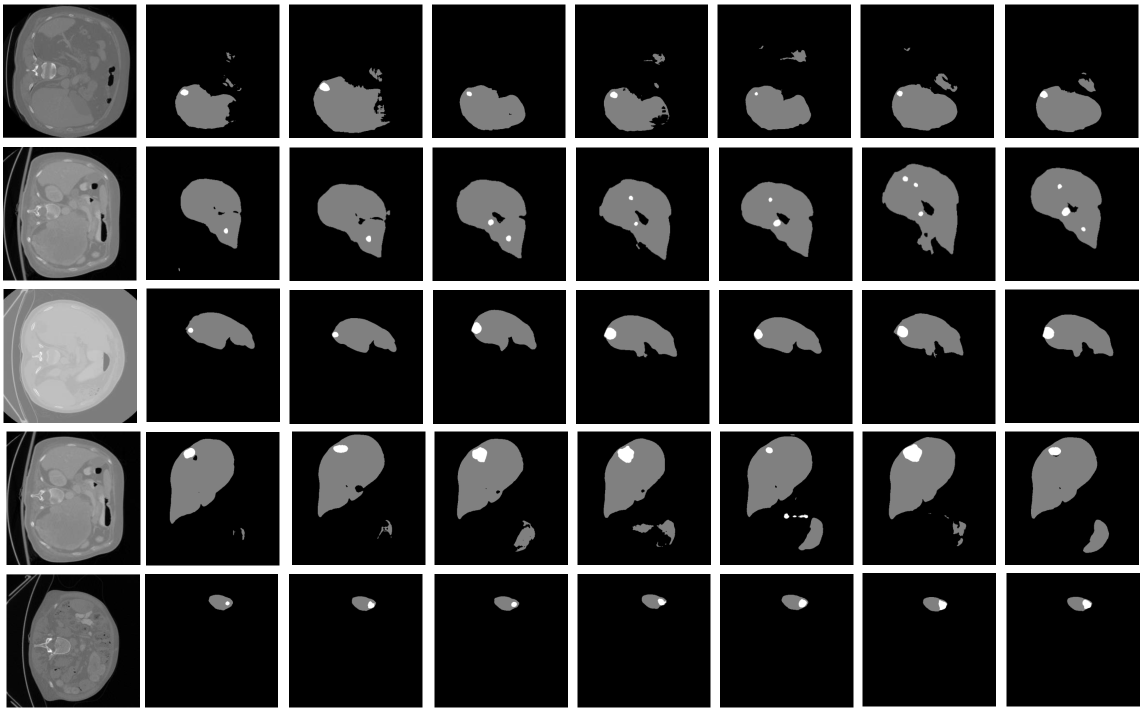

Fig. 4: Comparison of experimental results

to be the benchmark or reference standard for evaluating the accuracy of our model's predictions.Taking into account the distinct features of the loss functions applied in liver tumor segmentation,This study integrates the cross-entropy loss function with the Dice similarity coefficient to create a novel hybrid objective function that effectively balances both classification accuracy and structural similarity. The weight θ is manually assigned, and it is defined as follows:

$$L_{Loss} = \theta L_{BCE} + (1 - \theta) L_{Dice} \tag{11}$$

E. Comparative Experiments

To confirm the efficacy of the new approach, extensive experiments were conducted on the LiTS liver tumor public dataset, comparing it with six other classical and outstanding methods, including U-net, Res-Unet, TransUnet, and Swin-Unet, under the same dataset and experimental environment. To facilitate a clearer assessment of the segmentation efficacy across different models,A total of five distinct sets of patient computed tomography (CT) scans were randomly chosen from the test dataset to undergo visualization analysis, as detailed in Figure 4. In the figure, black represents the segmentation background, the grey shades denote the liver tissue that has been separated from the surrounding area, while the white regions indicate the tumor's boundaries.

As seen in Figure 4, U-net, Res-Unet, and Dense-Unet exhibit over-segmentation and under-segmentation issues due to directly merging shallow feature information extracted by the encoder with the high-level semantic information of the decoder without considering the impact of shallow layer noise. Additionally, the simplistic backbone network of U-net results in insufficient feature extraction by the encoder, making it challenging to segment smaller tumors accurately.

TABLE I: Results of liver segmentation by different methods on the LiTS dataset

Model	Dice(%)	$VOE(\%)$	$RVD(\%)$
U-net	94.24	12.14	16.81
U-net++	94.75	11.86	15.47
Res-Unet	95.43	10.09	6.59
Dense-Unet	95.36	10.54	6.76
TransUnet	96.29	8.53	3.48
Swin-Unet	96.72	8.07	2.91
Our	96.76	8.29	2.53

Despite its integration of the strengths of conventional convolutional and self-attention techniques, the segmentation outcomes are affected by data bias and overfitting, which are both exacerbated by the relatively small dataset size. The Swin-Unet model, based on the Swin Transformer, is even more dependent on data quality and diversity, making it more susceptible to overfitting, and both models demand high computational resources.

The proposed method significantly reduces the computational resource requirements by using Hadamard convolution modulation instead of the attention mechanism. The designed CHD convolution module demonstrates strong feature extraction capabilities, and the SP operation added in the skip connections effectively addresses the noise in shallow feature information, leading to better feature fusion with high-level semantic information. This technique enables precise delineation of the liver's anatomical edges and tumor margins,

TABLE II: Results of tumor segmentation by different methods on LiTS dataset

Model	Dice(%)	$VOE(\%)$	$RVD(\%)$
U-net	75.74	32.12	36.85
U-net++	79.27	30.98	33.06
Res-Unet	81.74	25.07	28.65
Dense-Unet	82.36	20.59	27.67
TransUnet	84.29	19.85	20.34
Swin-Unet	85.72	15.03	14.29
Our	86.59	10.81	11.25

even for minute lesions.This approach not only validates the efficacy of the proposed method but also confirms its robustness in liver tumor segmentation tasks, as demonstrated through rigorous experimental validation.

IV. Conclusion

Accurate delineation of liver tumors within abdominal CT imaging data is crucial for both clinical assessment and surgical strategy development, as it enables precise tumor localization and enhances treatment efficacy.This study introduces a novel liver tumor segmentation approach, which integrates a CHD convolutional block into the U-Net architecture to enhance feature extraction and improve segmentation accuracy.By integrating advanced architectural modifications into the encoder, we have successfully improved its ability to extract critical features from input data.The encoder-decoder architecture utilizes spatial pooling techniques during the skip connections to enhance shallow feature representation by reducing noise and maintaining critical visual elements.These intermediate outputs are subsequently combined with high-level semantic data extracted from the decoder to ultimately produce the final segmentation outcome.The findings from our experimental study indicate that this approach provides superior segmentation accuracy for liver and tumor regions within the LiTS dataset, surpassing existing techniques in both precision and recall metrics. Future work will optimize this method to improve segmentation accuracy and generalization capabilities. Additionally, research will explore combining weakly supervised and semi-supervised approaches to address the scarcity of liver tumor image datasets, thereby enhancing model performance while reducing the cost of creating liver tumor images.

References

[1] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver Tumor Segmentation Benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

[2] Z. Xu, S. Liu, D. Yuan, L. Wang, J. Chen, T. Lukasiewicz, Z. Fu, and R. Zhang, "ω-net: Dual Supervised Medical Image Segmentation with Multi-dimensional Self-attention and Diversely-connected Multi-scale Convolution," *Neurocomputing*, vol. 500, pp. 177–190, 2022.

[3] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep Semantic Segmentation of Natural and Medical Images: A Review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[4] H. Rahman, T. F. N. Bukht, A. Imran, J. Tariq, S. Tu, and A. Alzahrani, "A Deep Learning Approach for Liver and Tumor Segmentation in CT Images Using Resunet," *Bioengineering*, vol. 9, no. 8, p. 368, 2022.

[5] H. Liu, Y. Fu, S. Zhang, J. Liu, Y. Wang, G. Wang, and J. Fang, "Gcha-net: Global Context and Hybrid Attention Network for Automatic Liver Segmentation," *Computers in Biology and Medicine*, vol. 152, p. 106352, 2023.

[6] X.-d. HOU, Z.-y. LI, J.-y. NIU, and H.-p. LIU, "Retinal Vessel Segmentation Based on Attention Mechanism and Multi-path U-net," *Journal of Computer-Aided Design & Computer Graphics*, vol. 35, no. 1, pp. 55–65, 2023.

[7] Y. Zhang, C. Peng, L. Peng, H. Huang, R. Tong, L. Lin, J. Li, Y.-W. Chen, Q. Chen, H. Hu *et al.*, "Multi-phase Liver Tumor Segmentation with Spatial Aggregation and Uncertain Region Inpainting," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 68–77.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[9] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "U-net++: A Nested U-net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, MLCDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[10] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted Res-unet for High-quality Retina Vessel Segmentation," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 2018, pp. 327–331.

[11] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully Dense Net for 2-d Sparse Photoacoustic Tomography Artifact Removal," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 568–576, 2019.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, no. 2017, 2017.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[14] G. Castellano, P. De Marinis, and G. Vessio, "Weed Mapping in Multispectral Drone Imagery Using Lightweight Vision Transformers," *Neurocomputing*, vol. 562, p. 126914, 2023.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[16] G. Xu, X. Zhang, X. He, and X. Wu, "Levit-unet: Make Faster Encoders with Transformer for Medical Image Segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 42–53.

[17] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like Pure Transformer for Medical Image Segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 205–218.

[18] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin Unetr: Swin Transformers for Semantic Segmentation of Brain Tumors in Mri Images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[19] M. Y. Ansari, Y. Yang, S. Balakrishnan, J. Abinahed, A. Al-Ansari, M. Warfa, O. Almokdad, A. Barah, A. Omer, A. V. Singh *et al.*, "A Lightweight Neural Network with Multiscale Feature Enhancement for Liver CT Segmentation," *Scientific Reports*, vol. 12, no. 1, p. 14153, 2022.

[20] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A Convnet for The 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[21] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A Simple Transformer-style Convnet for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention U-net: Learning Where to Look for The Pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[23] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M. Cheng, and S. Hu, "Attention Mechanisms

in Computer Vision: A Survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[24] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip Pooling: Rethinking Spatial Pooling For Scene Sparsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4003–4012.

[25] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and Improving Layer Normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] J. Kang, R. Liu, Y. Li, Q. Liu, P. Wang, Q. Zhang, and D. Zhou, "An Improved 3d Human Pose Estimation Model Based on Temporal Convolution with Gaussian Error Linear Units," in *2022 8th International Conference on Virtual Reality (ICVR)*. IEEE, 2022, pp. 21–32.

[28] Y. Zhao, X. Ma, B. Hu, Q. Zhang, M. Ye, and G. Zhou, "A Large-scale Point Cloud Semantic Segmentation Network via Local Dual Features and Global Correlations," *Computers & Graphics*, vol. 111, pp. 133–144, 2023.

[29] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking Convolutional Attention Design for Semantic Segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.