Automatic Liver Tumor Segmentation

Dang Tran, Bobby Bona, Dima Kazlouski, Wanchen Hong
{rjbona, dimak, dangtran, wghong}@bu.edu

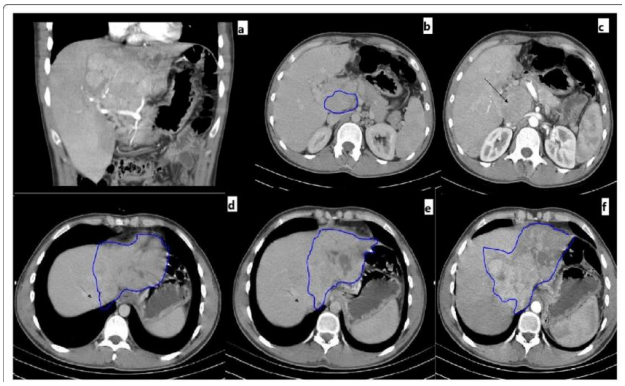

Figure 1: Example CT Images of liver tissue to be analyzed for the detection of liver cancer.

1. Task

Our task is to build a model that will allow clinicians to determine whether a tumor in a patient's liver is that of liver cancer or not by analyzing CT images of the tissue (like those in Figure 1) taken from a biopsy of the tumor. For tumor detection, the main difficulty lies in the accurate diagnosis of the exact type of tumor as well as the lack of structured cancer-related health data which makes it more difficult to validate our model on larger datasets. We have decided that the best way to solve these issues is instead of focusing on the effect of using only one pre-trained model and comparing it to other models, we will build a model that combines several convolutional neural network models. Furthermore, in addition to leveraging a model based on UNet++, this project will also explore the use of a transformer model for automatic liver tumor segmentation. The goal is to independently apply and evaluate a transformer-based approach to determine its efficacy in distinguishing liver tumors from CT images. This way, a comparative analysis can be drawn between the UNet++ model and the transformer model in terms of performance, accuracy, and efficiency in liver tumor segmentation.

2. Related Work

The journey toward robust liver tumor segmentation has seen remarkable strides with the advent of U-Net and its nested variant, U-Net++, both of which have been rigorously evaluated across multiple medical imaging tasks, including liver segmentation [1].

In the endeavor to further refine segmentation accuracy, the Feature Pyramid Network (FPN) and attention mechanisms like Squeeze-and-Excitation (SE) and Convolutional Block Attention Module (CBAM) have been explored. Particularly, an improved U-Net model demonstrated a promising approach by integrating attention mechanisms within the network's skip connections, resulting in enhanced localization and segmentation of liver tumors, thus addressing the challenges of oversegmentation inherent to liver tumor segmentation [2][3][4].

Moreover, the importance of loss functions in training robust segmentation models cannot be overstated. Binary Cross Entropy, Dice, and Jaccard loss have been recognized as effective loss functions in handling the class imbalance often encountered in medical image segmentation tasks, thereby potentially improving the segmentation performance [5][6].

The state of the art models often suffer from oversegmentation and are sensitive to class imbalance, which can adversely affect the segmentation performance, especially in medical imaging tasks where precise segmentation is crucial. Moreover, many models heavily rely on well-annotated datasets, which pose a significant bottleneck in real-world clinical settings. Our project, inspired by these advancements, aims to address these flaws by integrating the strengths of U-Net/U-Net++, FPN, and attention mechanisms, alongside employing Binary Cross Entropy, Dice, and Jaccard loss functions to enhance liver tumor classification. Through rigorous comparative analyses and validations on larger datasets, we aim to bridge the gap between academic advancements and real-world clinical diagnostic needs. Additionally, Transformer models, particularly those inspired by the architecture proposed in TransUNet, have shown promise in medical image segmentation tasks [10]. TransUNet combines the strengths of transformers and U-Net to deliver improved segmentation performance, which can be particularly advantageous in handling the challenges associated with liver tumor segmentation. The inclusion of transformer models can potentially address the problems of oversegmentation and class imbalance, which are common in medical image segmentation tasks.

3. Datasets

We use the LiTS17 (Liver Tumor Segmentation Challenge 2017) dataset organized by P. Bilic, et al. The training data set contains 114 CT scans in which we use 90 scans for the training set and 24 scans for the validation set, and the test data set 70 CT scans.

In addition, we also consider a smaller Kvasir SEG dataset which contains 1000 polyp JPEG images and their corresponding ground truth to verify the validity of our transUNet model.

4. Approach

4.1. UNet_ASPP

UNet is known for fast and precise segmentation of images, but UNet++ outperforms UNet in terms of medical image segmentation [1]. There are two types of architectures that can be used in encoder: DenseNet and ResNeXt, but DenseNet is prioritized due to its capability of capturing complex structures and help in transfer learning. We implement a modified UNet++ architecture with the addition of Atrious Spatial Pyramid Pooling (ASPP) model at the end going through the output of $x_{0,4}$ (Figure 2) which consists of a dilated convolutional block to enrich the fields of view and capture image context at multiple scales, thus improving the accuracy of capturing tumor boundaries and tumors with complex structures.

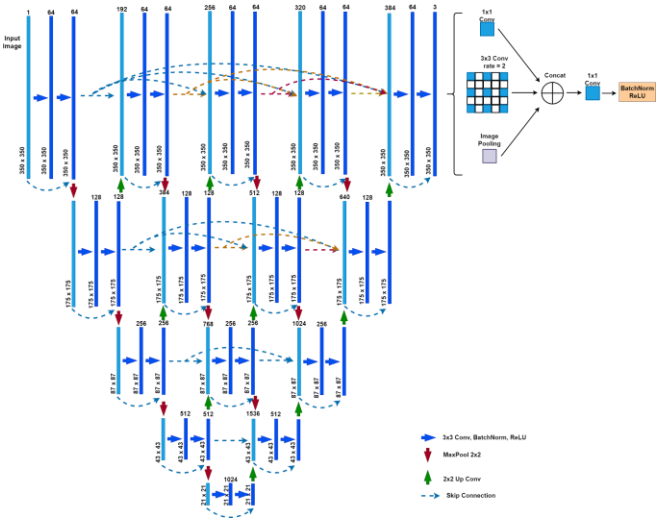

Figure 2: UNet_ASPP module.

Each block consists of two convolutional layers and each is followed by a batch normalization with ReLu activation function. In addition, there is a residual convolutional layer to avoid vanishing gradients after passing through multiple blocks.

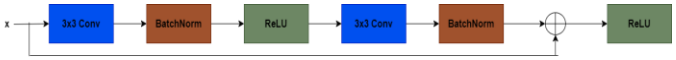

Figure 3: Convolution block of the model

Assume that H(x) is the convolutional operation and B(x) is the batch normalization operation:

$$y = \max(0, B(H(\max(0, B(H(x))))) + H(x))$$

We include a skip connection from each block to the following blocks of the same level which is concatenated with the upsampling from a lower-level block. Furthermore, for lower-level layers beginning from the second layer, for the model to also gain information from higher-level layers, we have a max pooling operation from higher-level to lower-level layers. By doing this, there is another modified UNet++ architecture inside the outer modified UNet++. This architecture helps the model learn semantic information in multi-label segmentation. $H(x_{i,j})$ is the convolution operation including activation function at $x_{i,j}$, and $U(x_{i,j})$ is the upsampling from $x_{i,j}$. At top-level layer from $x_{0,0}$ to $x_{0,4}$ we have the output $y_{i,j}$ of node $x_{i,j}$, as:

$$y_{0,j} = H(x_{0,j}) \quad , \text{if } j = 0$$
$$= H([\sum_{k=0}^{j-1} x_{0,k}, U(x_{1,j-1})]) \quad , \text{otherwise}$$

After that, the output $x_{0,4}$ goes to the ASPP model to expand the field of view of the feature maps and capture image context, which are liver and tumor boundaries for precise segmentation. Assume d is the dilation rate, R(x,r) is atrous convolution operation with dilation rate r from x, P(x) is average pooling operation, then we have:

$$y = R(x,r_2) + H(x) + H(P(x))$$

4.2 TransUNet with Attention

Additionally, a separate pipeline is established to implement a transformer model inspired by TransUNet for the task of liver tumor segmentation. This model employs a transformer as a robust encoder to extract features from CT images, which are then fed into a U-Net-like decoder for segmentation.

One of the novelties we introduce is the integration of spatial channels attention blocks within the decoder. For each upsampling stage in the decoder, a spatial channels attention block will be introduced. The primary aim of this block is to recalibrate the channel-wise feature responses by taking into account the spatial dependencies within the image. By doing so, we aim to refine the segmentation outputs by emphasizing important spatial structures and suppressing less relevant areas. This approach can help in better capturing the intricate details of the liver tumor regions and potentially improve the segmentation accuracy,

especially in cases where the tumor structures are complex.

Furthermore, skip-connections between the encoder and decoder will be implemented for retaining high-resolution features crucial for accurate segmentation, akin to the TransUNet architecture. The transformer's ability to capture long-range dependencies and global context from the input images, coupled with the attention mechanism, could prove beneficial in accurately segmenting liver tumors.

To evaluate the effectiveness of this model architecture, we will also compare the results of a simple transUNet without the attention blocks to the results of our implementation of transUNet with the attention blocks.

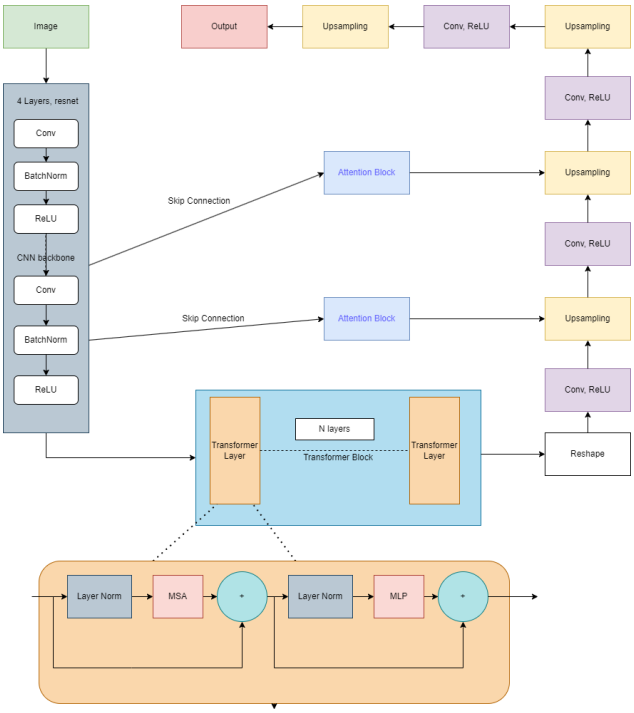

Figure 4: TransUNet with attention blocks.

5. Loss Functions

Dice loss is being used to measure the overlap of the predictions and ground-truth images.

$$L = 1 - \frac{2 * (p*g) + s}{p+g+s}$$

in which p is the prediction, g is the ground-truth tensor and s is the smooth variable (0.00001) to avoid dividing by 0.

For multi-label segmentation, common Dice loss alone is not enough to separate tumor from liver. We modified the binary Dice loss into multi-label Dice loss for three labels: 0 - background, 1- liver, 2 - tumor.

$$L_D = \frac{1}{N}\sum_{i=1}^{C} \left(1 - \frac{2 * (p*g) + s}{p+g+s}\right)$$

in which C is the number of classes, and N is the number of labels.

In addition, Cross-entropy loss is added to track pixel-by-pixel predictions to further improve the accuracy of each label prediction.

$$L_E = - \frac{1}{N}\sum_{i=1}^{N} y_i * \log(p_i)$$

in which p_i is the prediction, and y_i is the ground truth of label i with N labels.

Finally, our loss functions is concluded as:

$$L = 0.3 * L_D + 0.7 * L_E$$

We want the model to mainly focus on Cross-entropy loss, since we want it to check the label of each pixel in the input image.

6. Evaluation Metrics

We use Intersection-over-Union (IoU) which is defined as (Area of overlap)/(Area of union) [8] followed by a precision-recall (PR) curve as one of the metrics for our model, and we want the IoU to be close to 0.9 without overfitting the model.

Our goal was to find the mean IoU across all classes of segmentation. This is concluded as:

$$\text{IoU} = \left|\frac{y_i \cap p_i}{y_i \cup p_i}\right|$$

in which p_i is the prediction and y_i is the ground truth of label y. C is number of classes

7. Results

7.1. UNet_ASPP

Both UNet_ASPP and the baseline model UNet++ are trained with 12 epochs, batch size of 4 with the same loss functions and datasets.

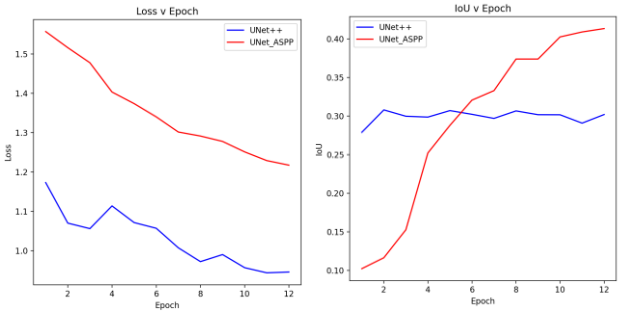

Figure 5: Loss and IoU values of UNet_ASPP and UNet++.

From Figure 5, our model needs more time to convert, while UNet++ starts at lower loss hence it converts faster as we can see that at epoch 12 their losses are

1.2175 and 0.9457 respectively, but our model is more stable during the training process. Apart from the loss, we also check if the IoU metric increases as the loss decreases. Although our model's convergence is slow, its IoU value gradually grows (0.4135) compared to the almost unchanged IoU value of baseline UNet++ (0.3021). Below we have a comparison between our model's predictions and UNet++'s. While UNet++ struggles to identify the liver and tumor, our model can give a prediction which is very close to the ground truth. Segmented region is prediction after going through softmax activation function to get the labels with the highest probability.

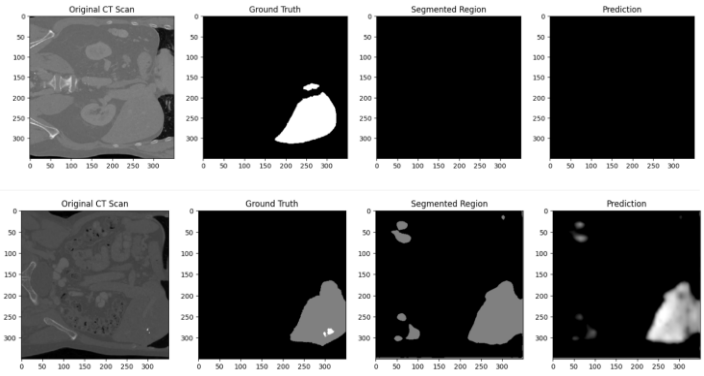

Figure 6: Segmented regions of UNet++ (above) and UNet_ASPP (below).

From Figure 5 and 6, our model outperforms UNet++ in multi-label segmentation tasks. Despite having higher loss than UNet++ does, our model gives clearer and more accurate predictions. Through using the ASPP model to expand the field of view, we are able to precisely detect the liver's region. Although we can see different regions with different labels in the prediction image, that information is lost along with the softmax activation function. It is still challenging for the model to accurately separate tumors from liver.

7.2. TransUNet with Attention

For the TransUNet model, after a few attempts, we decided to use a modified version of the ResNet as the backbone CNN structure for the architecture. The ResNet backbone is chosen because of its simplicity and saves enough memories for the transformer.

The transUNet models are trained using 15 epochs, batch size of 1, and 6 layers of transformer using the same loss function and dataset. The batch size and transformer layer size are chosen because any larger size would cost more memory than the machine allows.

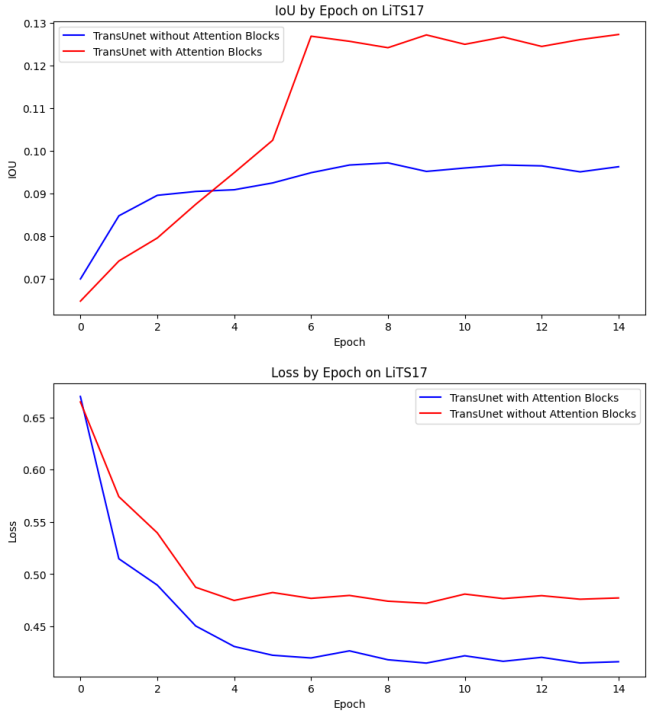

Figure 7: IoU (top) and Loss (bottom) results of TransUNet with and without attention blocks on LiTS17 dataset.

The results in Fig. 7 show that both implementations are converging, in which the implementation with attention blocks has better performance in both IoU (0.1283 to 0.0951) and loss (0.4103 to 0.4725). However, the low IoU values suggest that both models are not learning effectively. To verify this, the segmented regions are also plotted (Fig. 8).

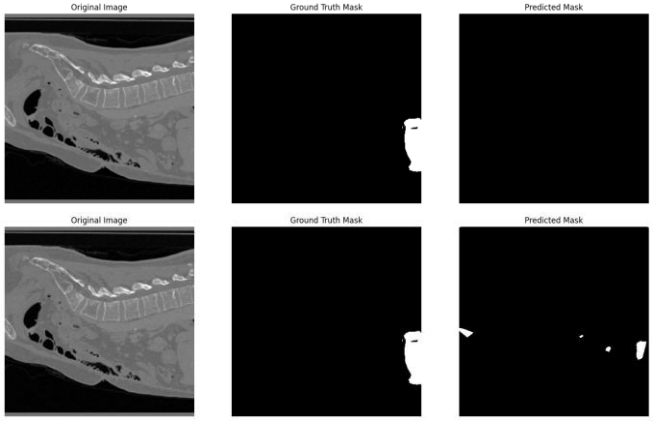

Figure 8: The same segmented region processed by TransUNet without attention blocks (top) and with attention blocks (bottom) on LiTS17 dataset.

The results of the segmented regions suggest that the implementation with attention blocks is learning, however, not very much compared to the implementation without attention blocks. It is worth noting that most of the results from the implementation with attention blocks also have the same segmentation result as in Fig. 8 top, which does not show any signs of

learning. Combining the result of the loss and IoU curves (Fig. 7), we very likely ran into a training plateau.

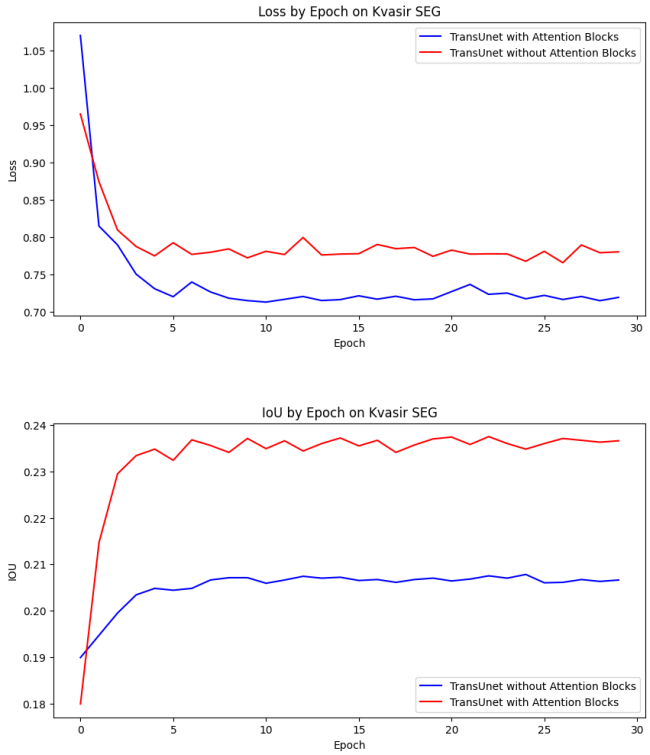

Figure 9: IoU (top) and Loss (bottom) results of TransUNet with and without attention blocks on Kvasir SEG dataset.

To further test the validity of our implementation, we would use the Kvasir SEG dataset, which is a much simpler dataset with 2D images in JPEG. For this dataset, the models are trained using 30 epochs, batch size of 16, and 12 layers of transformer with the same loss function. The results tell a similar story. As shown in Figure 9, the loss and IoU for both implementations are converging, but only to a very low limit, especially for IoU, only able to reach 0.2057 for without the attention blocks and 0.2370 for with attention blocks. Furthermore, the early convergence at around epoch 5 once again suggests a training plateau, which is verified by the results of the segmented regions (Fig. 10), that the relatively accurate results is not capturing the main region of interest, but rather including a lot of noise. This is even more the case for the less accurate result. Together, these results suggest that the model is not robust enough for these datasets.

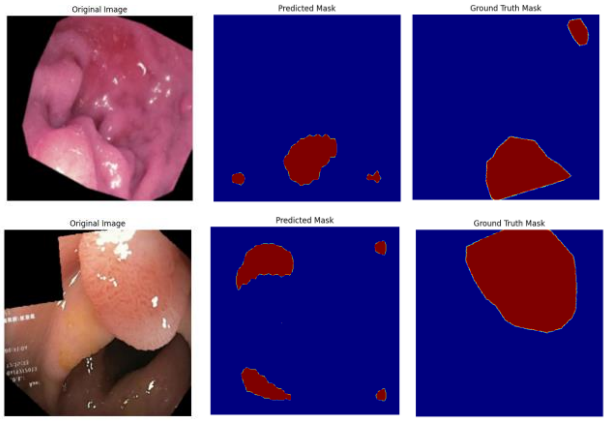

Figure 10: The segmented regions processed by TransUNet with attention blocks, of relatively accurate (top) and inaccurate (bottom) results on Kvasir SEG dataset.

7.3. Models Comparison

UNet ASPP:

Metrics	UNet++	UNet_ASPP
Loss	0.9457	1.0863
IoU	0.3021	0.4641

TransUNet, LiTS 17:

Metrics	Without Attention Block	With Attention Block
Loss	0.4725	0.4103
IoU	0.0951	0.1283

TransUNet, Kvasir SEG:

Metrics	Without Attention Block	With Attention Block
Loss	0.7656	0.7129
IoU	0.2057	0.2370

8. Conclusion and Future Work

Throughout our project we have made advancements in liver tumor segmentation by integrating and evaluating various convolutional neural network models and transformer-based approaches. Our primary goal was to distinguish liver cancer tumors from CT images with enhanced accuracy and efficiency. The incorporation of multiple models, including UNet++, TransUNet, and

attention mechanisms, has led to improvements in segmentation performance.

Our approach using UNet_ASPP demonstrated a slower convergence compared to UNet++ but achieved more stable training and superior segmentation accuracy, as evidenced by the higher Intersection-over-Union (IoU) values. This underscores the efficacy of Atrious Spatial Pyramid Pooling (ASPP) in capturing intricate tumor structures.

The implementation of the TransUNet model, both with and without attention blocks, revealed that the attention blocks enhance performance, albeit marginally, in both the LiTS17 and Kvasir SEG datasets. This suggests that while the attention mechanism adds value, further optimization is required to fully leverage its potential in medical image segmentation.

Key lessons learned from this project include the importance of model architecture optimization for specific medical imaging tasks, and the need to balance complexity with performance and resource constraints. The challenge of training plateau encountered in the TransUNet models highlights the necessity for ongoing model tuning and evaluation.

For future improvements, we propose exploring alternative attention mechanisms and transformer architectures to address the limitations observed in the TransUNet models. Additionally, augmenting our dataset or employing more sophisticated data augmentation techniques might improve model robustness and performance. Exploring advanced loss functions tailored to medical imaging could also enhance our model's ability to differentiate between tumor and non-tumor regions with greater precision. Lastly, real-world clinical validation of our models would be an essential step to bridge the gap between academic research and clinical application, ensuring that our models not only perform well in controlled environments but also deliver reliable results in diverse clinical settings.

9. Team Contributions

Task	Dang	Bobby	Dima	Wanchen
Data preprocessing	X		X	X
Implement baseline model	X		X	X
Implement UNet_ASPP	X			
Implement TransUNet			X	X
Loss functions	X	X		
Training and validate models	X	X	X	X
Make improvements for UNet_ASPP	X			
Make improvements for TransUNet			X	X
Testing and compare results	X	X	X	X

10. Code

GITHUB: <https://github.com/DangTranQL/Liver-Tumor_Segmentation>

References

1) Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018). 2018 Sep;11045:3-11. doi: 10.1007/978-3-030-00889-5_1. Epub 2018 Sep 20. PMID: 32613207; PMCID: PMC7329239.
2) Li, H.; Liang, B. Liver Tumor Computed Tomography Image Segmentation Based on an Improved U-Net Model. *Appl. Sci.* 2023, *13*, 11283. https://doi.org/10.3390/app132011283
3) Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017, April 19). *Feature Pyramid Networks for Object Detection*. arXiv.org. https://arxiv.org/abs/1612.03144
4) Woo, S. *et al.* (2018) *CBAM: Convolutional Block Attention Module*, *arXiv.org*. Available at: https://arxiv.org/abs/1807.06521 (Accessed: 10 October 2023).
5) Yeung, M. *et al.* (2021) *Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation*, *arXiv.org*. Available at: https://arxiv.org/abs/2102.04525 (Accessed: 12 October 2023).
6) Bertels, J. *et al.* (2019) *Optimizing the DICE score and Jaccard Index for Medical Image Segmentation: Theory & Practice*, *arXiv.org*. Available at: https://arxiv.org/abs/1911.01685 (Accessed: 12 October 2023).
7) Bilic, P. *et al.* (2022) *The Liver Tumor Segmentation Benchmark (lits)*, *arXiv.org*. Available at: https://arxiv.org/abs/1901.04056 (Accessed: 12 October 2023).
8) Hofesmann, E. (2021) *IOU A Better Detection Evaluation Metric*, *Medium*. Available at: https://towardsdatascience.com/iou-a-better-detection-evaluation-metric-45a511185be1 (Accessed: 15 October 2023).
9) Q. Huang, Y. Zhou and L. Tao, "Dual-Term Loss Function For Shape-Aware Medical Image

Segmentation," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Nice, France, 2021, pp. 1798-1802, doi: 10.1109/ISBI48211.2021.9433775

10) Chen, J. *et al.* (2021) *TransUNet: Transformers make strong encoders for medical image segmentation*, *arXiv.org*. Available at: https://arxiv.org/abs/2102.04306 (Accessed: 01 November 2023).