Liver and Hepatic Vessel Segmentation Using Attention Residual UNet Architecture

Abdallah Abouabdallah
Fachbereich Medizintechnik und Technomathematik
Medical Informatics laboratory
abdallah.abouabdallah@rwth-aachen.de

***Abstract*—In this paper, we use the Attention Residual UNet architecture to segment liver and hepatic vessels from CT images. The method uses two linked models, one aimed at the liver and another for the vessels. We train these models with the Medical Decathlon data set and their performance is evaluated using the Dice coefficient. This project's aim is to give insight toward the real-world use of medical images segmentation in health care, particularly in improving diagnostic accuracy and treatment planning.**

***Index Terms*—Attention Residual UNet, Liver Segmentation, Hepatic Vessel Segmentation, CT Imaging, Medical Image Segmentation, Deep Learning**

I. Introduction

The topic of medical image segmentation has been popular for a long time due to its usefulness in diagnosis and various healthcare applications. Historically, the segmentation of medical images followed a more classic approach, often relying on manual or semi-automated methods, which required significant human intervention. However, with the emergence of machine learning, there has been a shift towards more automated and data-driven approaches that can provide faster and more accurate results.

The liver, like many other organs in the human body, can reveal a lot about its health and condition simply by examining it through medical imaging. Using imaging techniques such as CT scans and MRI scans, doctors are able to assess various conditions affecting the liver. Segmentation through machine learning significantly enhances the diagnostic process, making it faster, more reliable, and less dependent on manual intervention.

Many machine learning methods can be implemented for this task, such as classical Convolutional Neural Networks (CNNs) and the well-known UNet architecture. In this paper, we will focus on the Attention Residual UNet architecture and develop a model to segment not only the liver from CT images but also the hepatic vessels. By doing so, we aim to provide a more comprehensive tool for liver analysis, which is critical for both diagnosis and treatment planning.

Abbreviations and Acronyms

CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
UNet	U-shaped Network
ResNet–	Residual Network
ARUNet–	Attention Residual UNet
ROI	Region of Interest
Dice	Dice Similarity Coefficient
IoU	Intersection over Union
ML	Machine Learning
DL	Deep Learning
HV	Hepatic Vessels
TP	True Positive
FP	False Positive
FN	False Negative

II Related Work

Traditional Approaches

We will first look into traditional approaches preceding machine learning area, the methods that exist, their concept, usage and performance. This pre-ML look gives us a view into the steps taken in data science and hardware capabilities, which now let us use AI and machine learning in difficult tasks. These new methods have touched areas like medicine—where sharp split is key—as well as in engineering, self-driving cars, and even new teaching methodologies. These traditional techniques included Thresholding, region-growing, and active contour model. [1] As an example Thresholding is a traditional approach that makes use of the varying intensity of the pixels in an image, marking the desired region based on a specific intensity threshold and segmenting that part accordingly [2]. This method is commonly used in images where the target structure has consistent intensity values. However, this consistency is not always retained, making segmentation via thresholding problematic. The latter becomes considerably more difficult when attempting to separate the liver in CT images, or any medical image in general, from other kinds of tissue, as these may lay anywhere across the intensity spectrum.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are powerful tools in image analysis, particularly for segmentation, due to their ability to capture complex features and anatomical structures. Through convolutional layers that work like feature detectors, and with the help of techniques such as max pooling, convolutions, and activation functions, CNNs automatically learn spatial hierarchies in images—a task that was challenging to achieve with traditional segmentation methods. [3] This foundational method serves as the building block for more complex approaches, such as the UNet architecture.

U-net: Convolutional networks

As the field progressed, more specialized models emerged to improve segmentation accuracy, with one key model being UNet. The UNet architecture builds on the basic structure of CNNs but includes specific adjustments for medical image segmentation. This architecture utilizes a U-shaped design, where an encoder path captures features through successive convolutions and pooling layers, and a decoder path upsamples these features, reconstructing spatial details [4].

UNet's design bridges lower-resolution features with higher-resolution context, which helps to retain fine details—essential for precise segmentation in medical images, such as the liver and hepatic vessels in CT scans. This model has become foundational in medical imaging, as it enables accurate and contextually aware segmentation even when data is limited, which is often the case in clinical applications.

Residual Networks

The next advancement builds upon CNN and UNet foundations, integrating the concept of residual learning. Residual Networks (ResNets) introduce "shortcut connections" that allow the model to bypass certain layers, effectively helping to avoid the degradation problem seen in deeper networks [5]. These connections make it easier for the model to learn complex features by preserving information across layers, which is especially useful in medical segmentation tasks that demand high precision and depth.

In the context of liver and hepatic vessel segmentation, residual connections help maintain details through the network's depth, which is key when segmenting fine structures. The addition of residual connections within UNet, forming an attention residual UNet, merges both attention mechanisms and residual learning to improve focus on critical regions and stabilize learning in deeper networks.

Attention Mechanisms

Attention mechanisms are methods designed to improve the focus of a network by highlighting the most relevant parts of an image while reducing emphasis on less critical areas. In this method the input data undergoes convolution to extract features and attention modules then highlight these features according to their significance, for the task at hand. Spatial attention plays a role in filtering out areas, which is particularly important, in segmentation tasks that require precise distinctions to be made. Furthermore temporal attention components are created to sweep through the input sequence and pinpoint frames or instances that have the most impact in accurate segmentation. By showing a response to areas anticipated as part of the category the attention mechanism guarantees that the model stays attentive to regions, with high responses enhancing the overall reliability and precision of segmentation. [6]

Cascaded Liver Segmentation

One relevant paper to this project is "Cascaded Fully Convolutional Neural Networks for Automatic Liver and Tumor Segmentation" by Li et al. [7]. Their approach was to divide the segmentation task into stages. Initially, the first network performs the segmentation of the liver region, capturing the general shape and boundaries. This initial output is then passed to a second network focused on refining the segmentation by detecting finer details, including tumors or other abnormalities within the liver. In their experiments, Li et al. demonstrated that the cascaded approach significantly improves segmentation accuracy compared to single-stage networks. However, the study also notes that the cascaded approach can be computationally intensive, as it requires running two networks in sequence.

III Methodology

Overview of the Method

In this work, we utilize an advanced architecture known as the Attention Residual UNet to perform liver and hepatic vessel segmentation from CT images. This architecture integrates three core components: the UNet structure, residual connections, and attention mechanisms. Each of these components contributes uniquely to enhance segmentation accuracy and model performance. The UNet framework provides a foundation for feature extraction and reconstruction, residual connections allow for deeper networks by stabilizing gradient flow, and attention mechanisms ensure the model focuses on relevant areas within the images. Below, we dive into each component and discuss their technical implementations in detail.

UNet Architecture

The UNet architecture is the foundation of our model, comprising an encoder path, a decoder path, and skip connections that link them. The encoder and decoder paths are symmetrical, with each level in the encoder capturing spatial hierarchies and each level in the decoder reconstructing these features to output a segmented image.

1 *Encoder Path:* The encoder path is composed of multiple convolutional blocks, each consisting of two convolutional layers followed by a ReLU activation function and batch normalization. Each convolutional layer in the encoder applies a set of filters that slide over the input, capturing features at different scales. After each convolutional block, a max-pooling layer is used to reduce the spatial dimensions by a factor of 2, effectively downsampling the image. This downsampling enables the model to learn more abstract, high-level features as it progresses deeper into the encoder.

2 *Skip Connections:* To ensure that high-resolution features from the encoder are not lost, UNet introduces skip connections that transfer information directly from each encoder block to the corresponding decoder block. This setup preserves spatial details that are critical for accurate segmentation, especially for small structures like hepatic vessels. Mathematically, let E_i represent the output from the i^{th}

encoder block and D_i represent the corresponding decoder block. The skip connection can be represented as:

$$D_i = f(E_i) + g(D_{i+1})$$

where f and g are transformation functions, often implemented as convolutions, that adjust the dimensions before concatenation.

3 Decoder Path: The decoder path mirrors the encoder in structure, but it performs upsampling to reconstruct the original image dimensions. Each upsampling layer is followed by a convolutional block that combines features from the previous layer and the corresponding encoder skip connection. The upsampling is typically implemented through transposed convolutions, which increase spatial resolution by learning to reverse the downsampling effect. This setup allows the decoder to produce a detailed segmentation map that aligns closely with the input image.

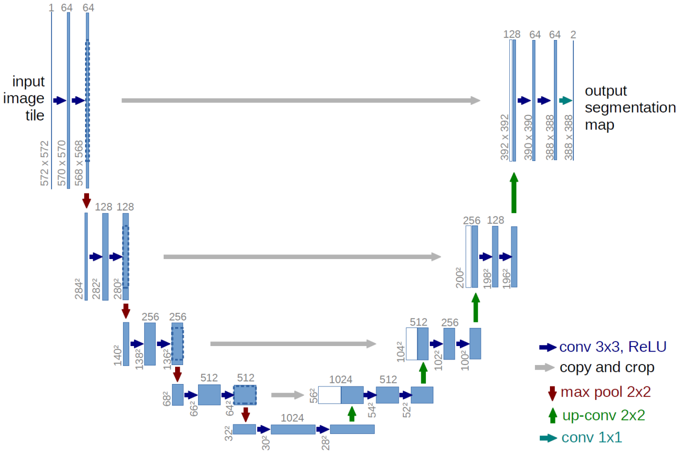


Fig. 1. Diagram of the UNet architecture, illustrating the encoder, decoder, and skip connections. Source: [4].

Residual Connections

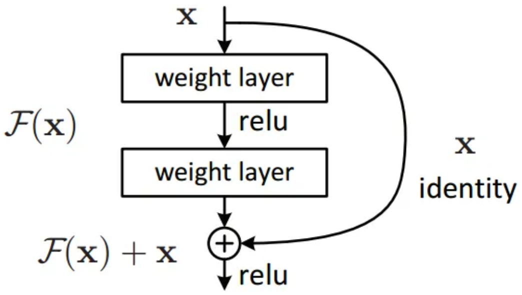
Residual connections, originally proposed in ResNet, are incorporated into the UNet architecture to enhance model depth without suffering from the vanishing gradient problem. These connections work by adding the input of a layer directly to its output, allowing the model to learn ”residuals” instead of absolute mappings.

1 Residual Block Structure: In our model, residual connections replace standard convolutional blocks within the encoder and decoder paths. A residual block consists of two convolutional layers with batch normalization and ReLU activation, followed by an addition operation that combines the block's input and output. Formally, for an input x, a residual block computes:

$$y = x + F(x, \{W_i\})$$

where $F(x, \{W_i\})$ represents the operations within the block (convolutions, activations), and the addition operation allows the gradient to flow through the shortcut, stabilizing training.

2 Placement of Residual Blocks in UNet: Residual blocks are strategically placed in each level of the encoder and decoder paths, replacing traditional convolutional blocks. This modification enables the model to learn more complex features with reduced risk of gradient vanishing, particularly in deeper layers. The residual connections, by allowing information to bypass several layers, also help preserve fine details crucial for segmentation accuracy.



x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 2. Diagram of the Residual Connection architecture, illustrating the skip pathways that allow information to bypass layers. Source: [5].

Attention Mechanisms

Attention mechanisms are added to improve the model's focus on relevant regions, enhancing segmentation performance by highlighting areas of interest (such as the liver and vessels) while suppressing irrelevant regions. In our model, spatial attention layers are applied at the decoder stages to refine feature maps based on spatial importance.

1 Implementation of Attention in UNet: Attention layers are placed just before each skip connection in the decoder path, allowing the network to weigh features from the encoder based on their relevance to the target region. The attention mechanism calculates an attention map, α, which scales feature maps, F, as follows:

$$F_{\text{att}} = \alpha \odot F$$

where $\odot$ denotes element-wise multiplication. The attention map α is computed by applying a sigmoid activation to the output of a learned function that considers spatial dependencies across the feature map.

2 Mathematics of Attention: The attention mechanism in our model follows a scaled dot-product approach. Given input features Q (query), K (key), and V (value), the attention map is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

where d_k is the dimensionality of K. This formula applies a softmax function to normalize the dot products of Q and K, ensuring that high values correspond to areas of high attention. The output is then used to modulate V, effectively focusing the network's resources on the most informative regions.

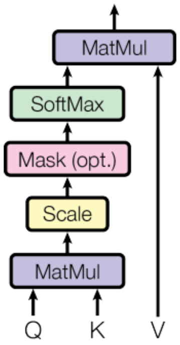


Fig. 3. Diagram of the Attention Mechanism, showing how it highlights important regions while suppressing irrelevant features. Source: [6].

Combined Architecture: Attention Residual UNet

The Attention Residual UNet combines the strengths of UNet's encoder-decoder structure, residual connections, and attention mechanisms to achieve accurate segmentation of liver and hepatic vessels from CT images. Each component uniquely enhances feature extraction, information flow, and focus on critical regions.

The UNet encoder-decoder framework serves as the foundation, with the encoder progressively downsampling the input to capture features at multiple levels, and the decoder upsampling these features to reconstruct a detailed segmentation map. **Skip connections** link each encoder and decoder layer, preserving spatial details essential for segmenting fine structures like hepatic vessels.

Residual connections replace standard convolutional blocks within the encoder and decoder paths, allowing gradients to bypass layers, which stabilizes training and improves feature learning in deeper networks. These connections enable the model to capture subtle anatomical details by learning ”residuals,” thus reducing the risk of vanishing gradients.

Attention mechanisms are applied to the skip connections, selectively focusing on relevant features while suppressing irrelevant ones. By assigning higher weights to important regions (e.g., liver boundaries), attention layers help the model concentrate on critical areas. The attention mechanism computes an attention map, α, using a scaled dot-product approach with query, key, and value matrices, enhancing segmentation precision by dynamically highlighting areas of interest.

Overall, the integration of residual and attention mechanisms within the UNet framework creates a robust architecture optimized for medical image segmentation, ensuring high accuracy even in complex and detailed anatomical structures.

IV Data and Training Setup

Dataset and Data Split

For the task of liver segmentation, we use a subset of the Medical Segmentation Decathlon dataset [8], comprising 54 contrast-enhanced, labeled CT images. To ensure robust model evaluation and avoid data leakage, the dataset is divided into a distinct Training Set and Validation Set with no image overlap. The Training Set consists of 43 images, while the Validation Set includes 11 images, ensuring a clear separation between training and validation data. Each image contains between 70 and 300 slices.

This subset was selected to include only images with fewer than 300 slices to reduce computational demands. Images with a high slice count significantly increase training time, yet provide limited advantage over smaller images for this specific task. This choice allows for more efficient training while preserving the model's segmentation accuracy.

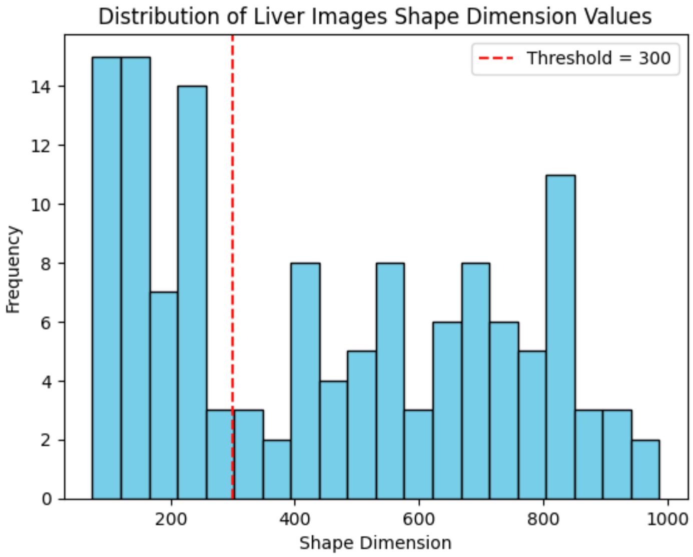


Fig. 4. Distribution of liver images by shape dimension values. Images with more than 300 slices are excluded to reduce computational demands. The red dashed line indicates the 300-slice threshold.

The histogram in Figure 4 shows the distribution of shape dimension values (number of slices) for the liver images in our dataset. As illustrated, a large number of images have fewer than 300 slices, with the rest of images extending beyond this threshold. To manage computational resources effectively, only images with fewer than 300 slices were selected for training and validation, as indicated by the red dashed line. This threshold ensures that training time remains manageable without compromising the quality of the segmentation results.

The same approach was applied to the hepatic vessel segmentation task. For this task, we used a subset of the Hepatic Vessel Dataset, which initially comprised 303 labeled CT images. To control training time while maintaining segmentation quality, we selected images with fewer than 80 slices, as indicated by the red dashed line in Figure 6. This threshold allows the model to focus on relevant image dimensions without the excessive computational demands posed by larger images.

The resulting subset includes 216 images, split into distinct Training and Validation Sets. The Training Set consists of 172 images, while the Validation Set includes 44 images. This clear separation helps ensure reliable model evaluation and minimizes data leakage.

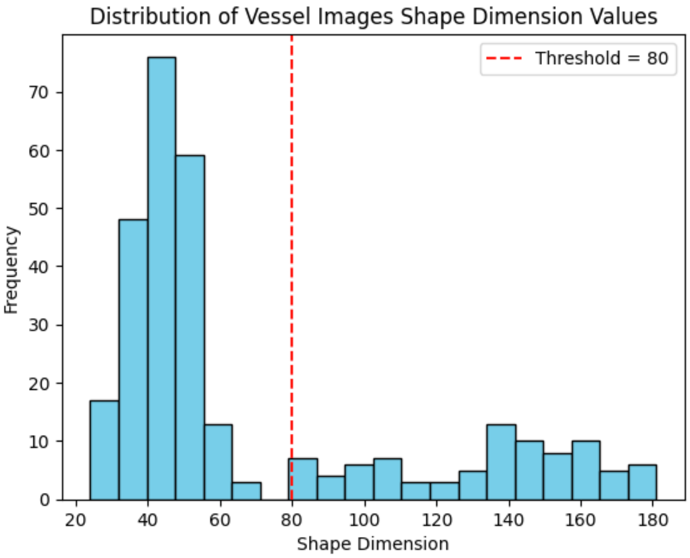


Fig. 5. Distribution of vessel images by shape dimension values. Images with more than 80 slices are excluded to optimize training efficiency. The red dashed line indicates the 80-slice threshold.

Training Configuration

Our training configuration is designed to optimize the model performance while maintaining computational efficiency. Each epoch is run with a batch size of 32 and the model is trained over an indefinite amount of epochs initially. The training only stops once the Halt mechanism is activated. The Halt mechanism makes sure the training continues unless the validation loss value starts increasing, meaning the model has reached it's peak and cannot learn anymore from the current data and training setup. The Halt mechanism is a simple but very effective method, it has a patience equal to 5. That means the condition for it to activate is to have the validation loss not improve for 5 consecutive epochs, essentially avoiding having occasional bumps in the validation loss stop the training but also making sure the model doesn't start overfitting. The Liver segmentation task was trained for 25 epochs before automatically stopping and the Hepatic vessel segmentation task was trained for 18 epochs. The final model is chosen as the best model with the smallest loss validation value read. The initial learning rate is set to 1×10^{-4}, which is adjusted using a learning rate scheduler. The scheduler reduces the learning rate by a factor of 0.5 if the validation performance does not improve for 3 consecutive epochs (patience = 3). Additionally Adam Optimizer is used. This adaptive approach allows the model to converge effectively while avoiding unnecessary adjustments to the learning rate.

Loss Function and Dice Score

To train and evaluate the model, we use a Dice-Cross Entropy (DiceCE) loss function, which combines the Dice similarity coefficient and cross-entropy loss to balance foreground and background accuracy. In addition, we compute the Dice Score during validation to assess segmentation accuracy. The Dice Score, a common metric for evaluating segmentation tasks, measures the overlap between predicted and ground truth regions. It is defined as follows:

$$\text{Dice} = \frac{2 \times |P \cap T|}{|P| + |T|}$$

where P represents the predicted segmentation mask, T is the ground truth mask, and $|P \cap T|$ is the intersection of predicted and actual segmented regions. The Dice Score ranges from 0 to 1, with a higher score indicating better overlap and therefore better segmentation accuracy.

For each batch, the Dice Score is computed by applying a threshold to convert the predicted probabilities into binary values, where values above 0.5 indicate a positive prediction. The intersection of the predicted mask and ground truth mask is calculated as the sum of element-wise multiplications across spatial dimensions, representing the correctly predicted pixels. Similarly, the union is calculated as the sum of pixels in both the predicted and ground truth masks. A smoothing term is included in both the numerator and denominator to avoid division by zero. The Dice Score for each image in the batch is averaged to provide an overall measure of the model's performance.

Model Details

The Attention Residual UNet used for the Liver segmentation task has 8 features in its base layer. While the model used for the Hepatic Vessels segmentation has 16 features in its base layer. this difference stems from the difference in size of the two datasets. Making it possible to use more features for the second task regarding the Hepatic Vessels because more memory is available. The data used in the Liver segmentation task is significantly larger than the data used in Hepatic Vessel task, as shown and described in the Data and Training setup section. These features serve as the initial filters in the network and are doubled with each downsampling layer in the encoder. This progressive increase in features enhances the model's ability to capture fine details in the CT images, contributing to precise segmentation of the liver and hepatic vessels.

Results

The results of the Liver and tumor segmentation task showed a dice coefficient of 0.88 for the liver and 0.73 for the tumor. While The second task of the Hepatic vessels and tumor segmentation showed dice coefficient of 0.73 for the Hepatic vessels and 0.66 for the Tumor. We can see that these dice coefficients are lower than the optimal or acceptable value, making the segmentation mostly unreliable. Although the training and validation loss had reached substantially low values, the dice coefficient for both the segmentation tasks remains unsatisfying.

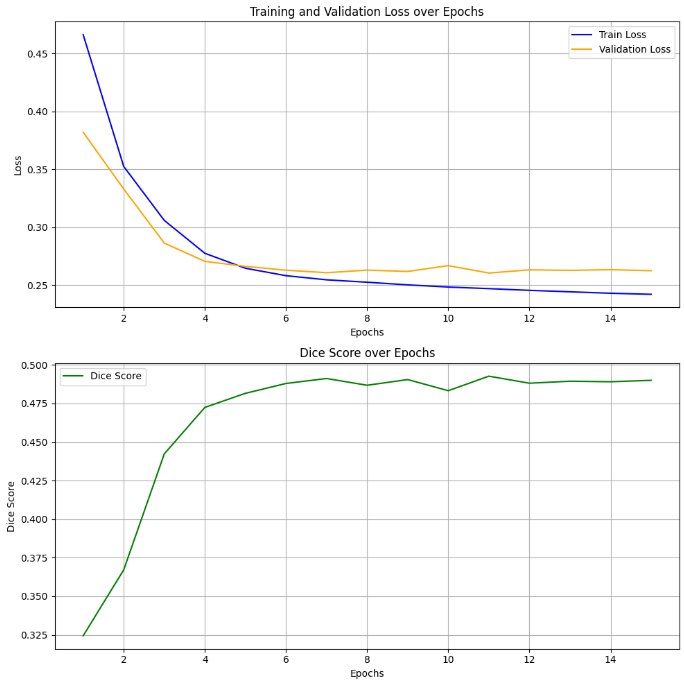


Fig. 6. Training and Validation Loss, and Dice Score Progression over Epochs.

The top plot shows the training and validation loss decreasing over epochs, with validation loss stabilizing around epoch 8 and reaching it's lowest around epoch 14. The bottom plot displays the Dice score, which improves steadily before leveling off, indicating the model's segmentation accuracy reaches a consistent performance after initial epochs. It is to be noted that the dice score shown in the plot is the combined score of all classes, which also includes background labels. That means this dice score is unreliable as it is inflated. The dice scores to be taken into consideration are the ones calculated for each class independently of the other.

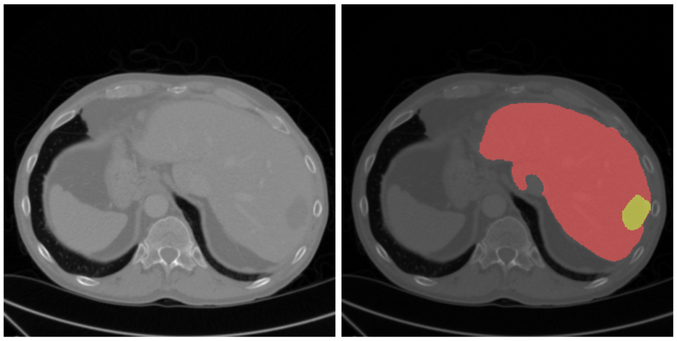


Fig. 7. Visualization of Liver and Tumor Segmentation Results.

The left image shows an original CT scan slice, while the right image displays the segmentation output produced by the Attention Residual UNet model. The red overlay represents the segmented liver region, and the yellow overlay highlights the segmented tumor region within the liver.

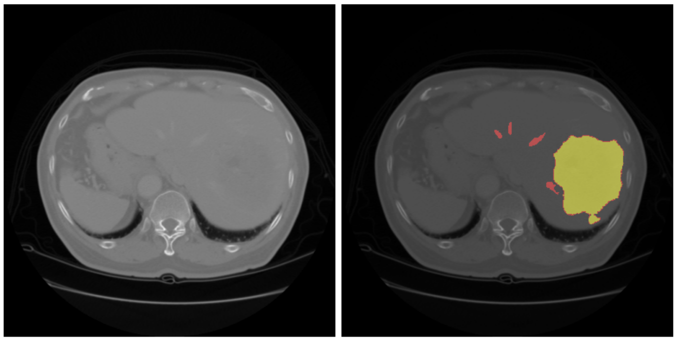


Fig. 8. Visualization of Hepatic Vessel and Tumor Segmentation Results.

The left image shows an original CT scan slice, while the right image displays the segmentation output produced by the Attention Residual UNet model. The yellow overlay represents the segmented tumor region, and the red overlays highlight the segmented hepatic vessels.

V Discussion

The Dice scores obtained for liver (0.48) and hepatic vessel (0.43) segmentation indicate moderate accuracy in segmenting these larger anatomical structures, which is reflected in the clear boundaries visible in the segmentation overlays. However, the lower Dice scores for tumors—0.13 for the liver-tumor segmentation and 0.36 for tumor segmentation within hepatic vessels—suggest challenges in accurately capturing smaller, irregularly shaped tumor regions. This discrepancy highlights the model's effectiveness in segmenting larger, well-defined structures, while pointing to limitations in handling finer details in tumor boundaries.

One possible reason for this disparity could be the choice of loss function. The low training and validation loss values indicate that the model is not overfitting, which suggests that the issue may not lie in the model's capacity but rather in how it is penalized for its mistakes. The current loss function may not adequately address the imbalanced class distribution between the labels, because the background labeled pixel account for a large portion of the image in comparison to for example the tumors, which only occupy a minuscule area in the whole 3D CT image.

A loss function that accounts for class distribution differences—such as a weighted Dice loss or a focal loss—could potentially improve performance by penalizing errors in underrepresented classes more heavily. Adjusting the loss function to better handle these disparities may lead to improved segmentation accuracy. This change could enhance the model's ability to focus on these challenging areas, ultimately resulting in a more balanced and accurate segmentation outcome across all targeted regions.

VI Conclusion

In this project, we used an Attention Residual UNet model to perform liver and hepatic vessel segmentation from CT images. The model was able to achieve moderate segmentation accuracy for larger anatomical structures, such as the liver and hepatic vessels, while facing challenges in precisely capturing smaller, less defined tumor regions. The low Dice scores suggest further improvements may be needed, potentially through a more specialized loss function that accounts for class imbalances.

Figures

List of Figures

References

[1] M. Trimpl, S. Primakov, P. Lambin, E. Stride, K. Vallis, and M. Gooding, “Beyond automatic medical image segmentation - the spectrum between fully manual and fully automatic delineation,” *Physics in Medicine and Biology*, vol. 67, no. 12, Jun. 2022, funding Information: This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.766276. KAV acknowledges support by CRUK (Grant Number A28736). Publisher Copyright: © 2022 The Author(s). Published on behalf of Institute of Physics and Engineering in Medicine by IOP Publishing Ltd.

[2] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–165, 2004.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” vol. 9351, 10 2015, pp. 234–241.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 06 2017.

[7] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “Cascaded fully convolutional networks for automatic liver and tumor segmentation,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2018, pp. 346–358.

[8] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. Landman, G. Litjens, B. Menze, O. Ronneberger, R. Summers, B. Ginneken, M. Bilello, P. Bilic, P. Christ, R. Do, M. Gollub, S. Heckers, H. Huisman, W. Jarnagin, and M. J. Cardoso, “The medical segmentation decathlon,” *Nature Communications*, vol. 13, p. 4128, 07 2022.