Temporally Aware Segmentation of Tumor Ablation

Samin bin Karim
Illinois Institute of Technology

Abhinav Theramel Baiju
Illinois Institute of Technology

Jack Harrison Mohr
Illinois Institute of Technology

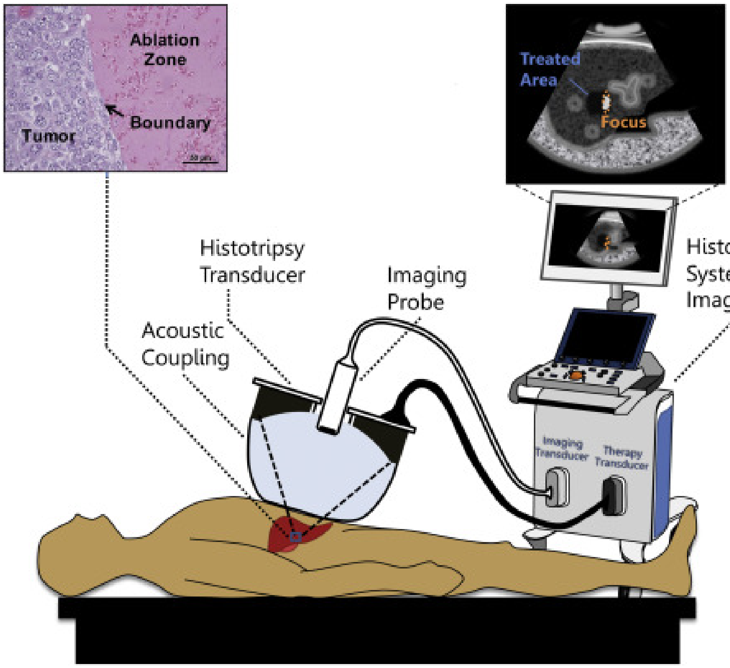


Figure 1: histotripsy procedure

1 INTRODUCTION

We developed an image segmentation model for use in a non-invasive ultrasound-based treatment for renal cell carcinoma being developed at the University of Chicago Department of Radiology by Dr. Kenneth Bader. The histotripsy procedure [2] ablates tumors using a high-powered ultrasound therapy and is intended to replace traditional invasive procedures. One challenge is that the real-time ultrasound images the physician sees while performing the procedure are not sufficient for determining the amount of ablation by visual inspection (see for figures 2 and 3, which depict an image from the dataset and its accompanying segmentation label). It is important to have accurate knowledge of how much tissue has been ablated so that treatment stops as soon as possible. Otherwise, risks of medical complications are increased. Thus, we develop a computer vision model to segment the ultrasound images of the tumor ablation areas. We implement and compare multiple architectures.

Our model uses an in vitro dataset produced by ablating a tissue-like agarose phantom with an ultrasound transducer in a lab. Our model is a proof-of-concept, demonstrating a useful level of performance. Given our positive result, Bader Lab can proceed to the next stage of research, gathering in vivo data and proceeding to porcine trials.

In addition to supporting the development of a novel cancer treatment, this project investigates multiple active areas of research in the field of medical computer vision. This project is one of the first to use 3D convolution and 3D transformer models for image segmentation of medical time-series image data.

To the best of our knowledge, there are no medical image segmentation models currently available that use 3D convolutions where the third dimension is time, not space (in contrast, there is an abundance of models using three spatial dimensions). Furthermore, there are no medical image segmentation models currently available that use transformers on 3D image input, where the third dimension is time. Such models are, however, increasingly common in non-medical domains, e.g., for image segmentation of drone footage or in quality control for industrial manufacturing. We will, therefore, apply state-of-the-art 3D convolution and transformer models to the segmentation of medical image time-series data.

We are using an in vitro dataset of tumor ablation (histotripsy) procedures in which each procedure is a time-series of 100 grayscale ultrasound images with corresponding binary labels. We have 10 unique procedures in the dataset, resulting in 1000 images in the dataset.The dataset was collected by Bader Lab at the University

CS577 Deep Learning, IIT, Chicago, IL

ACM ISBN

of Chicago Department of Radiology. Bader Lab is developing cancer treatments based on therapeutic ultrasound whereby renal tumors are liquified non-invasively by way of high-power ultrasound waves. A performant computer vision algorithm for this dataset will allow radiologists to monitor treatment progress in real time and reduce the risk of complications by reducing procedure time.

Our project makes multiple contributions. Firstly, we are contributing to the development of new cancer treatments by applying state-of-the-art image segmentation models to a novel dataset. Secondly, we will pioneer the use of 3D convolution and transformer models for image segmentation of medical time-series image data.

We will use a 2D UNET medical image segmentation model as a baseline. We will compare it to state-of-the-art 2D image segmentation models, both convolution and transformer-based. We will then compare the 2D approaches to 3D approaches that use multiple frames of the time-series as input. For 3D approaches, we will compare a 3D convolution to a 3D transformer model. Finally, we will modify the models to optimize performance given the unique requirements of our dataset. We hope that the modifications we make to the models will provide novel insights into adapting image segmentation models to the unique domain of medical time-series datasets.

We began by meticulously reviewing state-of-the-art 2D and 3D medical image segmentation literature. Based on this comprehensive survey, we selected a specific set of model architectures to implement. Our approach involved creating a 2x2 matrix, with the x-axis representing the choice between convolution and transformer-based architectures, and the y-axis distinguishing between 2D and time-series 3D models.

1.1 Project Overview

1.1.1 Data Handling. To facilitate our research, we developed a specialized data loader tailored to the unique requirements of our project. This data loader offers a level of complexity beyond the conventional ones, enabling the training of 3D time series-based models. These models necessitate the input of a user-defined number of sequential frames, and our data loader seamlessly accommodates this demand. Additionally, we implemented functions for visualizing test results to enhance our understanding of model performance.

1.1.2 Baseline Model Training. We conducted training and testing of a baseline UNET segmentation model, serving as a foundational reference point for our research.

1.1.3 . Enhanced Model Development Building upon the baseline UNET model, we proceeded to train and test a more advanced UNET++ model. Notably, this model exhibited superior performance compared to the baseline UNET. We then trained and compared a number of other models including UNET 3D and TransUnet.

1.1.4 Collaboration and Guidance. Our interactions with Dr. Bader, the Principal Investigator of the research project, have been invaluable. We have held multiple meetings at the University of Chicago, during which Dr. Bader provided essential guidance and advice, steering our project in the right direction.

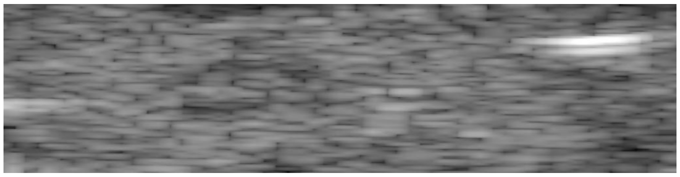


Figure 2: input cell image

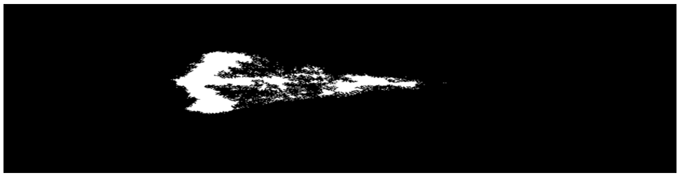


Figure 3: ablation segmentation

2 LITERATURE REVIEW

This literature review provides an overview of 2D and 3D Image segmentation models that can be used for medical image segmentation.

2D Medical Image Segmentation Models: UNet: UNet[6] is a popular medical image segmentation model that uses encoder-decoder structure with skip connections for more precise image segmentation with smaller datasets making it ideal as a foundational architecture for medical image segmentation.

UNET++: UNET++[7] builds the original UNET and enhances feature extraction by introducing nested and dense skip pathways. The dense skip pathways reduce the difference between the feature maps of the encoder and the decoder allowing for context awareness and simplifying the task for the decoder layers. The model achieved greater IoU scores than previous UNet models.

Transunet: Transunet[4] represents the integration of transformer-based attention mechanisms into the traditional UNet architecture. Leveraging transformers allows the model to capture long-range dependencies in medical images, enhancing its ability to understand complex spatial relationships.

SwinUnet: SwinUnet[3] is a pure transformer model that is structured like a Unet. It uses skip connections and bottleneck like UNet but all the blocks are transformer blocks with patch merging layers and patch expanding layers. It out performs TransUnet by a small margin in medical image segmentation tasks.

3D Medical Image Segmentation Models:

UNET3D: UNET3D[8] extends the UNet architecture to three dimensions. It replaces the 2D operations with equivalent 3D functions. The 3D dimension gives the model spatial awareness and the model useful to predict volumetric medical data using sparse volumetric annotations. 3D Unets perform well on tasks that are inherently trying to predict labels for 3D structures like MRI data.

Non-Medical Image Segmentation Models for Time-Series Data: STEm-Seg: STEm-Seg[1] models videos as a 3D spatio-temporal volume and proposes a new model that tracks object instances across space and time. Model main aim is to cluster a an a 3D object across time without breaking the task into multiple stages. Model can potentially be adapted segmenting medical video and 3D data as well.

TransUnet3D TransUnet3D[5] combines a 3D Unet architecture with transformer-based attention mechanisms that gives the model the ability to capture long range dependencies within a frame and across frames, where the frames can be related by time or space. Additionally TransUnet also add another decoder layer that includes learnable queries and uses novel loss function that combine pixel level loss and image level loss. TranUnet3D performs well on both 2D and 3D medical segmentation tasks.

3 METHODS

3.1 Establishing a baseline

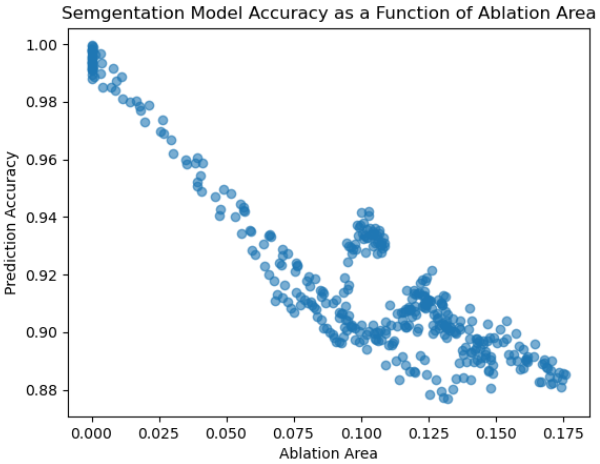


Figure 4: Prediction accuracy as a function of ablation area

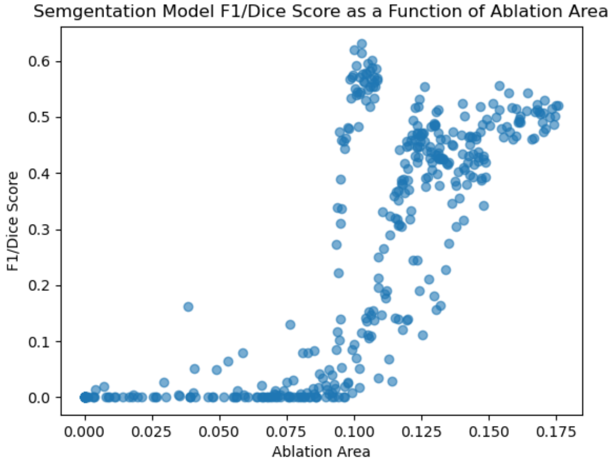


Figure 5: Ablation Area and F1

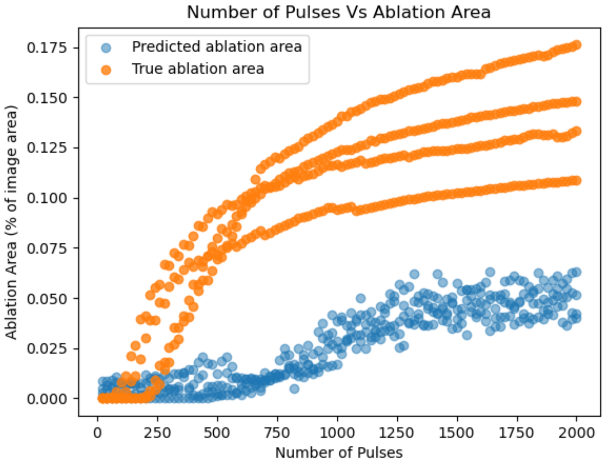


Figure 6: Prediction dice score as a function of ablation area

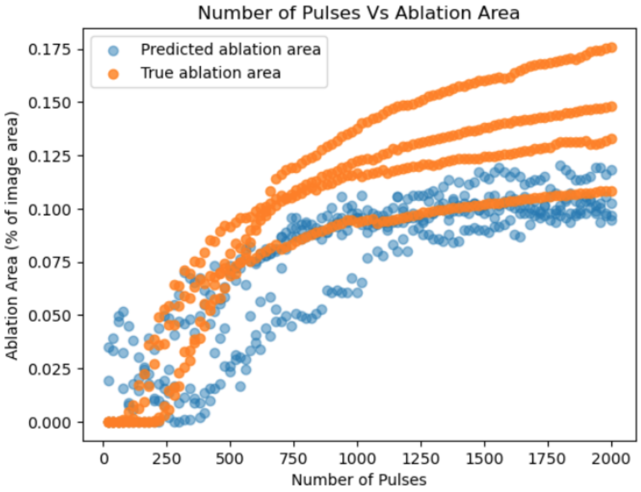


Figure 7: Dice Score as a function of ablation area for modified loss function

Our initial efforts to solve the segmentation task used a vanilla UNET model, which we refined over multiple iterations. During this early stage of the project we gained insights about the nature of our dataset that informed design decisions down the line. Without any modifications, the original UNET model achieved a dice score of 0.49 on the test set. After various optimizations the improved UNET model achieved a dice score of 0.67.

3.1.1 Improving UNET.

First, we created an image downsampling pipeline that sped up training time by multiple fold without a significant effect on model performance. We experimented with different downsampling parameters such as downsampling factor and the amount of gaussian blur to apply and whether to apply the blur before or after downsampling or both. Through trial-and-error we found that a 16x downsample factor yielded the excellent speedup without sacrificing model performance. Beyond 16x pixel reduction, model performance began to degrade. We originally had

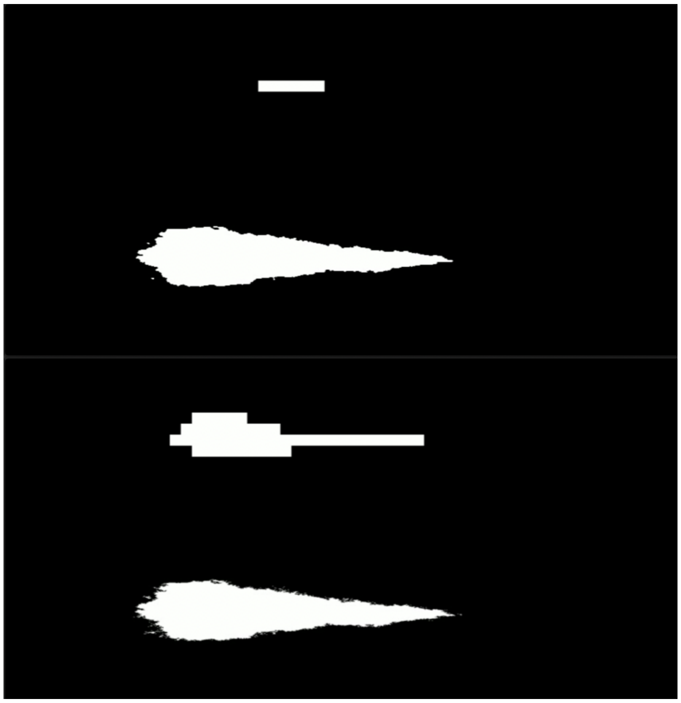


Figure 8: prediction with original loss function (top) vs weighted loss function (bottom)

the model create the downsampled dataset from the original dataset each time at runtime but eventually after we found appropriate settings we created a hard copy of the downsampled data and accessed it directly when training. We used this downsampled version of the data in subsequent models as well.

3.1.2 Lessons from first tests. The first key insight we gleaned from our initial tests is that the dataset has an unequal distribution of '0' and '1' pixels. There are far more '0' pixels than '1' pixels. This fact has multiple implications. Firstly, it means that 'accuracy' is an inappropriate performance metric for this task. Because most of the pixels are black, especially in the early phases of the treatment, a model that just predicts all black pixels will perform quite well. Figures 4 and 5 illustrate the misleading nature of using accuracy as a metric for this data. Accuracy here means the ratio of correctly predicted to incorrectly predicted pixels. Figure 4 shows the model accuracy as a function of the ablation area (keep in mind that the ablation area is increasing monotonically over the course of the treatment). The most accurate predictions are for images at the beginning of the treatment that are almost completely black (i.e. non-ablated) and as the treatment progresses the accuracy drops.

In contrast, figure 5 shows how the dice coefficient, which penalizes false positive and false negative predictions (equation 1), gives a more realistic depiction of the model's performance.

$$\text{Dice Score} = \frac{2 \times \text{TP}}{(\text{TP} + \text{FP}) + (\text{TP} + \text{FN})} \tag{1}$$

$$\text{F-beta Score} = \frac{(\beta^2 + 1) \times \text{TP}}{\beta^2(\text{FN} + \text{FP}) + (1 - \beta^2) \times \text{TP}} \tag{2}$$

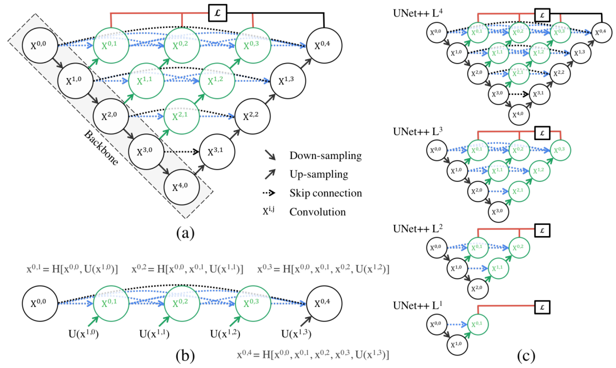
However, note that the dice score is extremely low at the early treatment phases. Here the dice score suggests an overly pessimistic view because the numerator in the dice score ratio is 2 * true positive predictions. So when there are almost no true positive pixels in an image, it is like finding a needle in the haystack event though the model is overwhelmingly correct in its prediction of black pixels. As the ablation area increases, the dice score tends to improve. Following these findings we chose to use dice score as our key performance metric because it tends to be more representative of the model's usefulness. That is to say, upon visual inspection of the predictions, a prediction with a good dice score tends to correlate with a prediction that is meaningfully similar to the true label much more than the accuracy score does.

A second consequence of the uneven distribution of '0' and '1' pixels is that the model becomes biased toward predicting '0'. Consequently, it chronically underestimates the total number of '1' pixels in an image. This is shown graphically in the 'number of pulses vs ablation area' graphs. The orange dots are the true ablation areas while the blue dots are the corresponding predicted areas. We partially corrected this problem by modifying the dice loss function in the training loop to a weighted dice loss, the so-called 'f-beta dice loss', which disproportionately penalizes false negative predictions, thus encouraging the model to predict more '1'-valued pixels. The second of the two graphs shows the blue predicted area dots more closely aligned with the orange true area dots. We also include an example of a prediction. In the attached figure the top two sections show the label and prediction using the model trained on the original dice loss function and the second two sections show the same but the prediction is from the model trained on the modified loss function. This shows how the original loss function led to severe underestimations of ablation area, especially in early stages of the treatment. Underestimations at early treatment stages is especially detrimental to the success of the histotripsy procedure because physicians need to be able to detect the inflection point at which the increase in ablation area begins to decelerate rapidly (see again the number of pulses vs ablation area graphs for the 'knee' around 500 pulses in). If physicians can stop treatment around this inflection point, they will have delivered most of the value of ablation without spending more time than necessary in treatment.

Finally, from visual inspection of the dataset we discovered that the ablation boundary is much more apparent visually when looking at multiple frames in succession than when looking at a single frame. Based on this observation, we created a pipeline to turn model predictions into short movies that show the segmentation for a whole treatment sequentially. This helped us gain intuition for how our model was performing. This also inspired us to try a 3D model that takes multiple sequential frames as input. The idea is that if a human can recognize the pattern better when seeing multiple frames, perhaps a machine can as well.

3.2 UNET++

3.2.1 Model Architecture Overview. U-Net++ is an extension of the original U-Net architecture designed for semantic segmentation tasks. It enhances the U-Net structure by introducing nested and



$\mathcal{L}$
$X^{0,0}$ $X^{0,1}$ $X^{0,2}$ $X^{0,3}$ $X^{0,4}$
$X^{1,0}$ $X^{1,1}$ $X^{1,2}$ $X^{1,3}$
$X^{2,0}$ $X^{2,1}$ $X^{2,2}$
$X^{3,0}$ $X^{3,1}$
$X^{4,0}$
Backbone
Down-sampling
Up-sampling
Skip connection
$X^{i,j}$ Convolution
(a)
$x^{0,1} = H[x^{0,0}, U(x^{1,0})]$
$x^{0,2} = H[x^{0,0}, x^{0,1}, U(x^{1,1})]$
$x^{0,3} = H[x^{0,0}, x^{0,1}, x^{0,2}, U(x^{1,2})]$
$U(x^{1,0})$ $U(x^{1,1})$ $U(x^{1,2})$ $U(x^{1,3})$
(b)
$x^{0,4} = H[x^{0,0}, x^{0,1}, x^{0,2}, x^{0,3}, U(x^{1,3})]$
UNet++ L^4
UNet++ L^3
UNet++ L^2
UNet++ L^1
(c)

Figure 9: Unet++

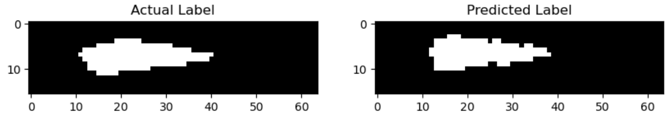


Figure 10: Unet++: prediction vs ground truth

dense skip pathways. The architecture includes multiple convolutional blocks with skip connections at different resolutions to capture both local and global contexts.UNet++ has been extensively used in medical image segmentation tasks such as identifying organs or anomalies in MRI or CT scans.UNet++ often achieves better segmentation accuracy compared to the original UNet due to its improved feature extraction capabilities.

The U-Net++ architecture incorporates nested skip pathways and dense connections, as proposed in the U-Net++ paper. These additions aim to enhance the model's ability to capture hierarchical features and improve information flow across different scales.Unet++ uses a pre-trained ResNet backbone, which provides a strong feature extractor. U-Net++ may utilize subnets such as ResNet or VGG as the backbone for its convolutional blocks. This choice is often driven by the ability of these subnets to capture intricate features and hierarchical representations.If VGG is used as a subnet in U-Net++, the paper might report fair performance. VGG is known for its simplicity and effectiveness in capturing features, which could contribute to improved segmentation results.

3.2.2 Results. Table 1 shows the dice scores obtained from training the UNet++ model.

Data Aug.	Training Dice Co.	Validation Dice Co.	Test Dice Co.
No	0.95	0.93	0.80

Table 1: Unet++: Dice coefficients

3.2.3 Shortcomings. The increased complexity due to nested skip connections and dense pathways results in a more intricate architecture. This complexity demands more computational resources, both in terms of memory and processing power

The expanded network structure might lead to overfitting, particularly when training data is limited. Its effectiveness could be influenced by factors such as dataset size, diversity, and the complexity of the segmentation task.

3.2.4 Potential Improvements. Instead of dense connections, selective or learned to skip connections could be explored to reduce redundancy and enhance information flow.

Integrating attention mechanisms could help the model focus on relevant features and regions, potentially improving segmentation accuracy.

3.3 UNET3D

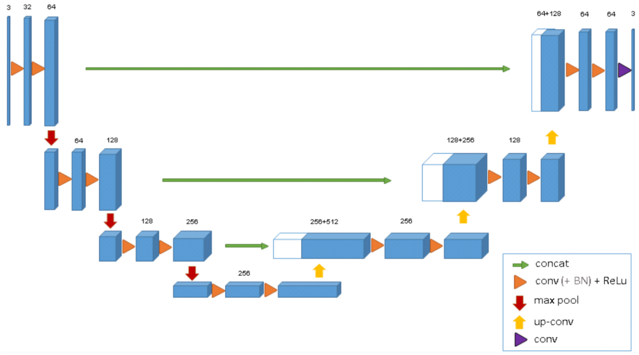


Figure 11: Unet3D

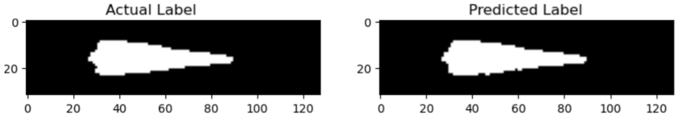


Figure 12: Unet3D: prediction vs ground truth

3.3.1 Model Architecture Overview. UNET3D is a three-dimensional extension of the U-Net architecture designed for volumetric data segmentation tasks. Like the original UNet, it also includes skip pathways and can train and perform well using smaller datasets. The contracting path with the bottleneck provides global information while the skip connections add high-resolution information back, giving the model better context to make decisions. Unet3D uses batch normalization, and in our experiments, retaining the original mean and variance produced better performance than using running mean and variance.

3.3.2 Modifications to the Original Model. Unet 3D was originally designed for 3D medical data segmentation like MRI images. The 3rd dimension in that case is a spatial dimension. In our problem the 3rd dimension is time. Images from previous 8 previous timesteps are stacked together and passed through the model. The model output is the same shape as the input, therefore 2 more CNN layers were added to flatten the output into a single mask prediction matching the label for the latest timestep in the input image.

3.3.3 Results. Unet3D model worked better on the test set than the other models while using using the batch normalization mean and variance from the original training set.

Increasing the number of timesteps in the model input to 16 frames results in similar training and validation dice scores while it reduces the testing dice score by 2 percent.

Augmenting the training data by blurring and adding noise to it decreased the training dice score but increased the testing dice score.

Data Aug.	Prev. Timesteps	Training Dice Co.	Validation Dice Co.	Test Dice Co.
No	8	0.98	0.97	0.82
No	16	0.98	0.97	0.81
Yes	8	0.97	0.97	0.84

Table 2: Unet3D: Dice coefficients

3.3.4 Shortcomings. While Unet3D performed better that the other models it still performs poorly on the test set with 0.84 dice score being the highest achieved.

The model is also not truly temporally aware and the 3rd dimension was originally meant as a spatial one. The only reason the model may be performing well is because the tumour cells have a 3D structure as they and being crushed as time progresses.

3.3.5 Potential Improvements. We can add more data augmentations and add them to our dataset to make it larger and also to make our model less prone to overfitting and potentially perform better on the test set.

We can add temporal understanding to the model by taking inspiration from LSTMs and give the model access to the previous outputs.

3.4 Transunet

3.4.1 Model Architecture Overview. Trans-Unet combines Unet model with the transformer based attention mechanism. It gives the model the ability to relate image patches that are further away while producing encodings, which should in theory allow it to encode better global data. The original paper used a VIT model for the transformer bottleneck with ResNet blocks for the Unet.

In our implementation, we used VGG blocks which is also a comparable alternative with simpler architecture.

3.4.2 Results.

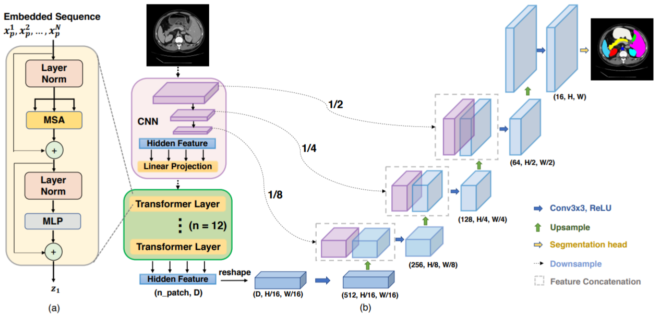


Figure 13: TransUNet

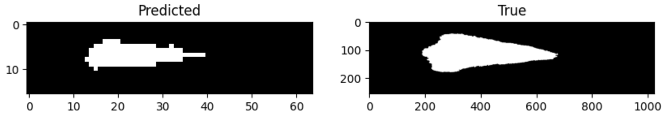


Figure 14: TransUNet: prediction vs ground truth

Data Aug.	Training Dice Co.	Validation Dice Co.	Test Dice Co.
No	0.96	0.89	0.65

Table 3: TransUNet: Dice coefficients

3.4.3 Shortcomings. The transformer layers in TransUnet do not have the advantage of CNNs that inherently assume as neighbourhood relation. Hence, the attention layers must be trained longer for them to understand how different image patches relate for a given downstream task. However, in our case our training dataset only consists of 880 images and the attention layers can learn a wide variety of invalid relations between patches and still produce good training accuracy.

3.4.4 Potential Improvements. We can use a pretrained ViT backbone to alleviate the problem of poorly fitted attention layers. Pretraining will allow the ViT attention layers to understand which patches to relate better, which could not be done well using our tiny dataset. Then the model can be retrained on our smaller data to perform well on the downstream task. Doing has the potential to drastically improve the dice score on the test set.

3.5 Summary of Results

Model Desc.	Data Aug.	Training Dice Co.	Validation Dice Co.	Test Dice Co.
UNet	No	0.79	0.78	0.67
UNet++	No	0.95	0.93	0.80
UNet3D	Yes	0.97	0.97	0.84
TransUNet	No	0.96	0.89	0.65

Table 4: Results summary

4 CONCLUSION AND FUTURE WORK

In conclusion, we successfully presented a proof-of-concept for image segmentation histotripsy ultrasound images. We likewise demonstrated that a 3D time-series segmentation model can achieve comparable results to a 2D model on a medical image segmentation task. These findings signify a promising opportunity for improving image segmentation for videos of medical procedures.

To further advance this research, future work should focus on enhancing the overall performance of the models. This could involve fine-tuning hyperparameters, exploring advanced architectures, or incorporating additional data augmentation techniques to improve generalization.

Moreover, transitioning towards real-time implementation is crucial for practical use by physicians. Conducting thorough latency testing and optimizing the models for speed will be essential steps in ensuring the feasibility of deploying these segmentation models in real-world scenarios.

In terms of methodological advancements, more rigorous testing is warranted to comprehensively compare the performance of timeseries image segmentation models against their 2D counterparts. Future experiments should be carefully designed to address questions such as the feasibility and advantages of using a 3D spatial model in the context of a 3D timeseries model. Additionally, a detailed comparison between 3D and 2D models needs to be conducted, considering various performance metrics and real-world applications.

In summary, the success of this proof-of-concept lays the foundation for future research endeavors aimed at improving model performance, implementing real-time systems, and conducting more extensive comparative studies. These efforts will contribute to the continued evolution of machine learning applications in the field of medical image segmentation, ultimately enhancing diagnostic capabilities and advancing healthcare technologies.

REFERENCES

[1] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. 2020. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 158–177.

[2] Kenneth B. Bader, Eli Vlaisavljevich, and Adam D. Maxwell. 2019. For Whom the Bubble Grows: Physical Principles of Bubble Nucleation and Dynamics in Histotripsy Ultrasound Therapy. *Ultrasound in Medicine Biology* 45, 5 (2019), 1056–1080. https://doi.org/10.1016/j.ultrasmedbio.2018.10.035

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*. Springer, 205–218.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).

[5] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. 2023. 3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers. *arXiv preprint arXiv:2310.07781* (2023).

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV]

[7] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: 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, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 3–11.

[8] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. arXiv:1606.06650 [cs.CV]