# Automated Segmentation of Temporal Bone CT Imaging for Robot-Assisted Microsurgery

Paper Seminar Report

*EN 601.656 Computer Integrated Surgery II*

Jessica Soong

jsoong1@jhu.edu

# Project Summary

The ear has complex anatomy and for robotic surgery this means extremely high resolution data is needed for safe trajectory planning. Even in high resolution computer tomography (CT) scans, many of the structures may only be 1-2 voxels in size. Furthermore, due to the nature of otologic surgery, many important structures near the ossicles cannot be hit, for example, the facial nerve. This project is motivated by this need to have high quality segmentations of patient anatomy from temporal bone CTs for safe robotic surgery.

## **Paper**

One of the goals of the project is to compare and contrast the performance of different deep neural networks on our dataset. The relevance of "PWD-3DNet: A Deep Learning-Based Fully-Automated Segmentation of Multiple Structures on Temporal Bone CT Scans" is rather clear, in that in order to accomplish our goal, we need to find state-of-the-art implementations to test with our dataset. This paper is one of the first methods to develop a 3D approach to temporal bone segmentation from CT images, and has already done extensive experiments to find optimal hyperparameters on a similar dataset.

### Summary and Key Results

This paper implements a 3D patch-based approach to temporal bone segmentation with a novel sampling method that the authors credit to its high performance when compared to other state-of-the-art methods. This was the first fully automated 3D pipeline designed for temporal bone segmentation. The model has good performance, measured by a high dice similarity score and low Hausdorff distances for the temporal bone structures.

### Background

Patient specific segmentations of temporal bone organs from CT scans are useful to improve the feasibility of robotic surgery planning. Manual labeling is very time consuming and takes anywhere from one hour to one day, depending on the image and the skill of the clinician. Furthermore, since segmentations are used to train novice surgeons, often times skilled surgical otolaryngologists are making these segmentations, taking them away from the operating room.  Previous algorithms implemented for this purpose use 2D U-Nets ensembled in some way [2]. This is inefficient since 2D ensembles will have redundant convolutions that would be captured with 3D convolutions, and 2D convolutions lose spatial context in the slice-level interpretation. Another common 3D segmentation approach is to use down-sampling to fit the entire image on a GPU. Since some structures are extremely small but still vital, this method cannot be used without completely losing the small structures like the chorda tympani.

# Dataset

39 adult cadaveric temporal bone specimens were used. The scans were labeled with region growing and manual correction. Each subject has a micro-CT and a clinical scan. A micro-CT is done by scanning each slice of the cadaver for extremely high resolution, while a clinical scan is done by putting the

cadaver in a CT scanner. In total there are 78 micro-CT and clinical scans each from the left and right ears. In the training, validation, and test set, there are 126, 14, and 18 scans respectively.

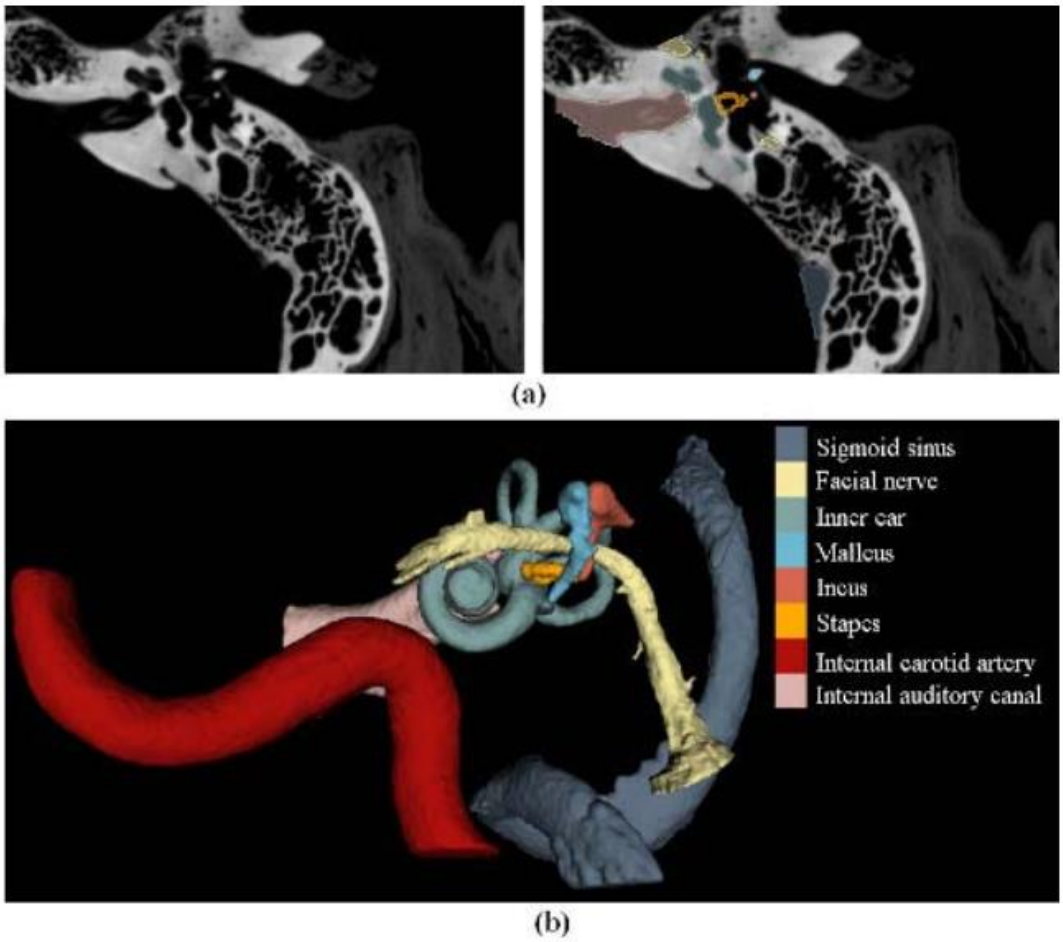


*Figure 1: CT and Manual Labels*

## Model Architecture/Workflow

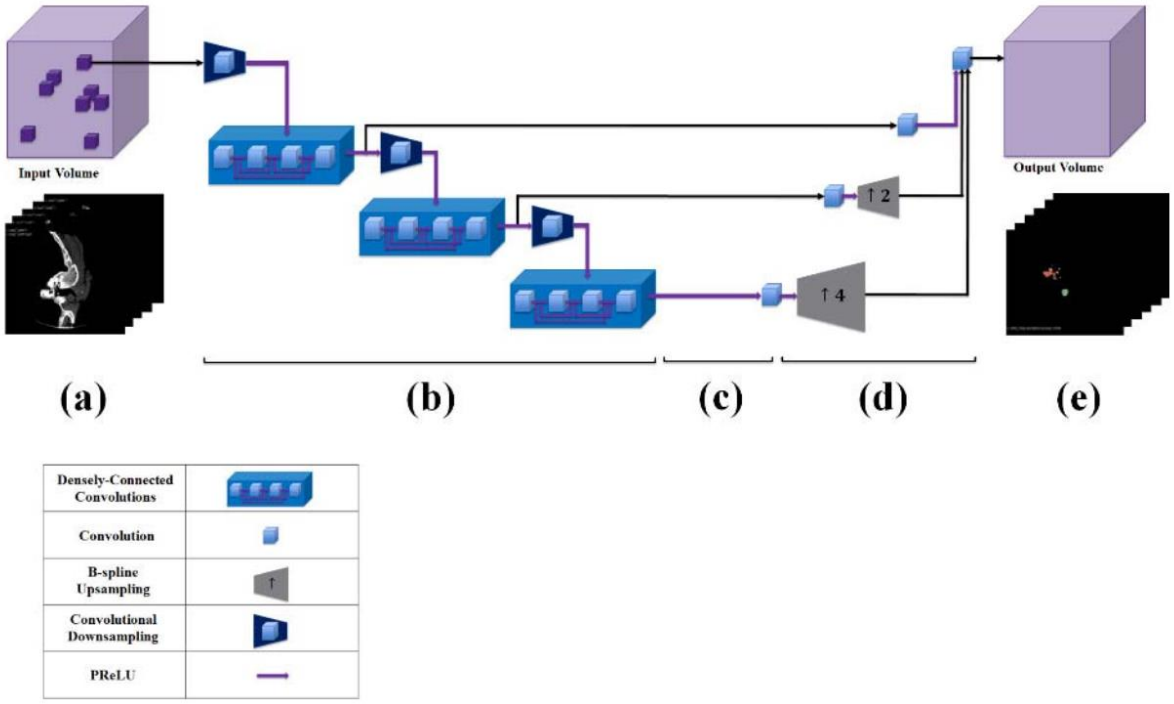


*Figure 2: PWD-3D Net Architecture*

The input volume is first sub-sampled using a balanced patch sampler. Then, each patch will be sent through the network, which has a similar architecture to DenseVNet. The sub-volumes go through a series of deeply connected residual layers and are unsampled at each level before being concatenated and passing through the final layer to form the output volume. The use of B-spline upsampling gradually decreases the compression which hypothetically reduces the number of artifacts in the system. Then, the loss is calculated between the predictions and the labels, using a dice score sensitive to outliers:

$$DSS(V,\hat{V}) = \frac{1}{|M|}\sum_{m\in M}\frac{2\sum_j v_m^j\,\hat{v}_m^j}{\sum_j\left(v_m^j\right)^2+\sum_j\left(\hat{v}_m^j\right)^2}$$

where $V$ is the ground truth volume and $\hat{V}$ is the predicted volume, and $v_m^j$ and $\hat{v}_m^j$ denote the j-th voxels of class m for the ground truth and predicted volume respectively. The inclusion of the square in the denominator is important to this implementation since it effectively acts as a class regularizer—harshly punishing outliers.

## Sampling Methods

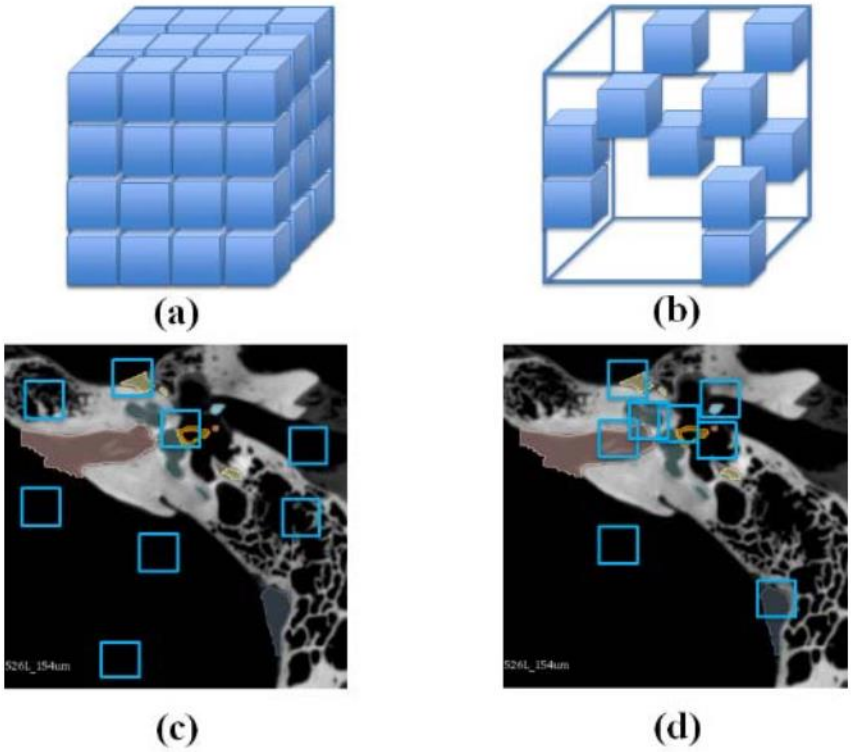


*Figure 3: Comparison of Sampling Methods. (a) Grid Window Sampling (b) Uniform Window Sampling (c) Background/Foreground Sampling, (d) Balanced Window Sampling.*

Some common types of sampling methods are shown in Figure 3. Figure 3a shows a method called grid window sampling, which subsamples patches derived from a sliding window and a given stride. This has many redundant calculations depending on the stride, and for training is non-optimal due to the redundancy and also the large resulting class imbalance. In an image of mostly background, a sliding window approach will result in mostly background. Since neural networks are sensitive to the class representation, this would yield unsatisfactory results. Figure 3b shows uniform window sampling, where windows are randomly sampled from all window configurations. This reduces redundant calculations since there is less overlap, but the method is still slow since al feasible locations of windows must be calculated and then randomly sampled. Furthermore, there is no guarantee against class imbalance. Figure 3c shows Background/Foreground sampling, which samples the dataset in a 50-50 split of background and foreground. This method still has some imbalance for multi-class classification tasks. The proposed sampling method, balanced window sampling, shown in Figure 3d, uses ground truth labels as sampler weights so each class is sampled with equal probability. This solves the redundancy problem as well as the class imbalance problem.

## Data Synthesis + Augmentation

Since the dataset is still small relative to other image datasets, augmentation was used to simulate a larger dataset. Three main methods were used, global blurring, resampling, and an addition of clinical CTs.

*Blurred Micro-CT:* Depending on the CT scanner, the image can be optimized for bony structures or soft tissue. *A g*lobal gaussian filter was applied to the micro-CT images to allow the resulting model to segment various types of CT scans optimized for different tissues.

*Resampled Micro-CT*: Images were downsampled to larger slice thickness, then unsampled to original voxel size using B-splines to simulate CTs taken with a lower resolution.

*Addition of clinical CTs*: Micro-CTs are not performed in human patients, only cadavers. Cadavers were scanned before Micro-CTs were performed to increase the dataset size.

On the fly augmentation was also performed. Histogram standardization and whitening normalization was performed on every input to the network, with the parameters for these transformations derived soley from the training set. Random rotation in the range of -10° to 10° of each orthogonal plane was applied as well as spatial rescaling in the range of 0.9 to 1.1 times the original image size.

## Inference

Balanced patch based sampling is for training only and requires labels to sample. For inference an unseen image is used and patches are sampled using the grid window sampling method, with amount of overlap as a hyperparameter. More overlap means a slower calculation, whereas less overlap may mean a decreased accuracy. In the experiments done, this was shown to be true up to a patch size of 32x32x32. Post-processing needed since the patch-based method results in noise and disconnected components (Fig 4).


*Figure 4: Inference Post-Processing. (a) Model Output. (b) After post-processing.*

## Evaluation/Results

The overall results from the model are shown in Table II. For all large structures there is a dice score of greater than 0.8, but the dice score for smaller structures like the facial nerve and stapes fall slightly short of that. If we look deeper into the types of data in Figure 6, we can see that the clinical CT results are mixed, with a high amount of variation in dice score for the results.

TABLE II

AVERAGE DICE SIMILARITY SCORE (DSS), HAUSDORFF DISTANCES (MILLIMETER) AND JACCARD SCORE (JS) OF THE PROPOSED SEGMENTATION ALGORITHM (AUGMENTED NETWORK) FOR EIGHT TEMPORAL BONE STRUCTURES. STRUCTURES ARE SIGMOID SINUS (SS), FACIAL NERVE (FN), INNER EAR (IE), MALLEUS (M), INCUS (I), STAPES (S), INTERNAL CAROTID ARTERY (ICA) AND INTERNAL AUDITORY CANAL (IAC)

| | SS | FN | IE | M | I | S | ICA | IAC |
|---|---|---|---|---|---|---|---|---|
| **DSS** | 0.86 | 0.74 | 0.90 | 0.84 | 0.85 | 0.77 | 0.81 | 0.89 |
| **HD** | 1.91 | 1.23 | 0.27 | 0.26 | 0.28 | 0.28 | 1.96 | 0.62 |
| **JS** | 0.75 | 0.59 | 0.82 | 0.72 | 0.74 | 0.63 | 0.68 | 0.80 |

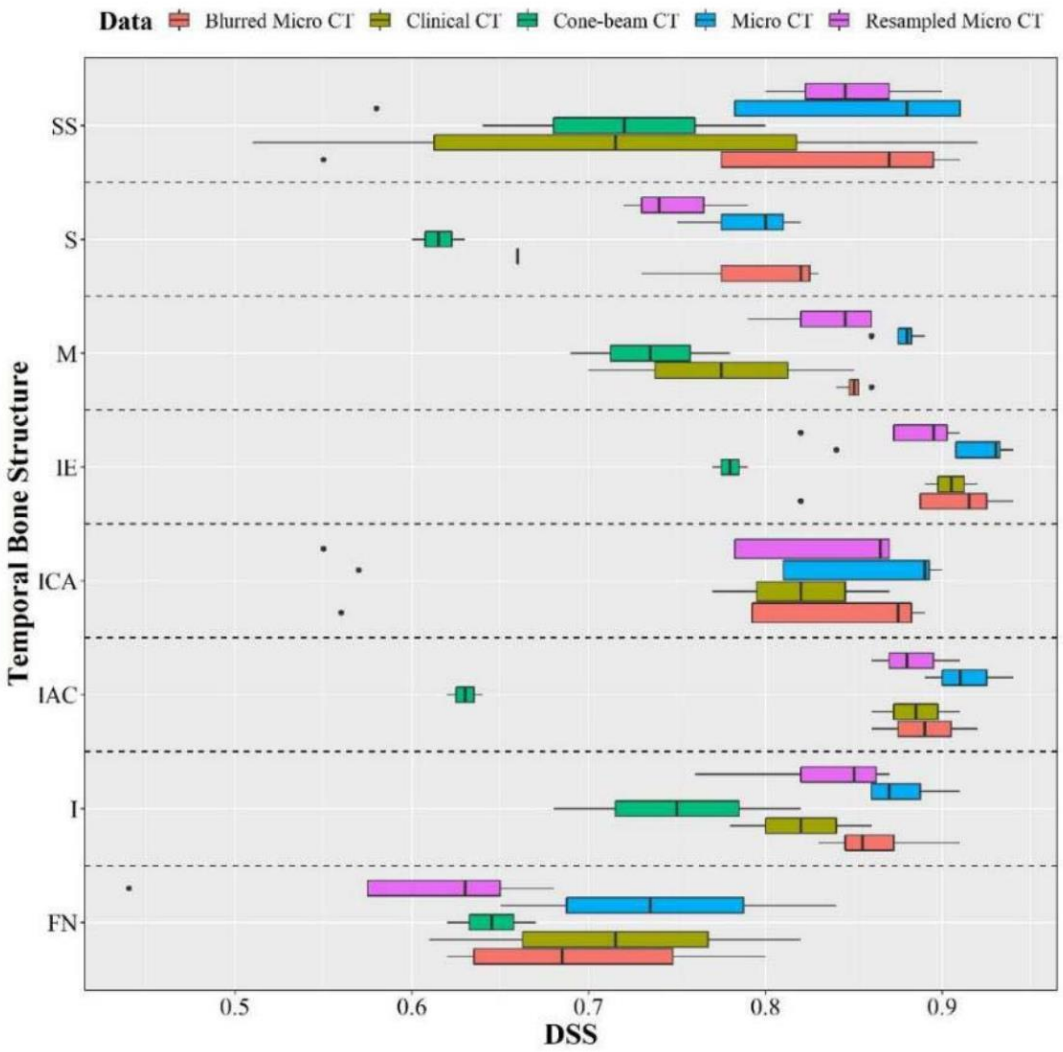


*Figure 5: DSS of the proposed segmentation algorithm for eight temporal bone structures, sigmoid sinus (SS), facial nerve (FN), inner ear (IE), malleus (M),*

## Conclusions

The experiments done show that the patch-size of 144x144x144 is optimal, and experiments done on the volume of overlap for inference show that 32x32x32 is optimal. Balanced window sampling is the most appropriate sampling technique for temporal bone segmentation due to the large class imbalances. The authors credit the outperformance of other state of the art models due to this. When tested on a held-out dataset from another institution, an average dice score of 0.64 was obtained. The authors claim that this score suggests the model is generalizable and synthesis/augmentation of data is key to making the data generalizable on a small dataset.

## Critiques

This paper was very thorough in its write-up and in its experiments. The paper's training methods were well explained, and the model architecture was explained with limited introduction of mathematical notation, which is generally appreciated by clinicians. The code is available on github with a docker file, and data is available through written request. The rigor of the experiments include comparisions of

performance on augmented and non-augmented datasets, as well as experiments to find the optimal patch size, volume of overlap for inference, and a comparison with two other state-of-the-art methods for medical image segmentation problems, one of which uses an alternative sampling method. Although there were many differences between the alternate network and sampling scheme, it served as a loose control for the proposed sampling method.

While the code is publicly available, it is also poorly documented. Furthermore, the inference method was not clearly explained. For example, there is no mention of what was done with the overlapping subsampled pixels with different predicted labels. Given the numerous experiments run, this seems like an important detail that was left out. Furthermore, the dataset after augmentation was primarily micro-CTs, which are not done on living subjects. For increasing feasibility of robotic surgery, it would be better if the study could be done as closely as possible to the same workflow in a clinical setting, using mostly clinical CTs. Furthermore, augmentation of a flat gaussian blur is not a good method for simulating another CT data optimized for softer tissues. If anything, since the image is already labeled, blurring the image based on the ground truth in areas of soft tissues may result in something closer to a different clinical scan. Finally, taking a 0.64 dice score on external validation as sign that the algorithm can generalize well is generous. While there is a page limit for journals and it is easy to recognize that the authors had much to write, this is a claim that needs to have more justification as a dice score of 0.64 for a supervised method is not particularly great.

## References


1. S. Nikan *et al*., "PWD-3DNet: A Deep Learning-Based Fully-Automated Segmentation of Multiple Structures on Temporal Bone CT Scans," in *IEEE Transactions on Image Processing*, vol. 30, pp. 739-753, 2021, doi: 10.1109/TIP.2020.3038363.
2. Fauser, J., et al. (2019). "Toward an automatic preoperative pipeline for image-guided temporal bone surgery." International Journal of Computer Assisted Radiology and Surgery **14**(6): 967-976.