Final Report

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

EN 601.656 Computer Integrated Surgery II

Andy Ding & Jessica Soong
5-7-2021

# Table of Contents

## Clinical Motivation

Operating in the temporal bone and lateral skull base is technically challenging. This region contains a complex geometry of nerves, arteries, veins, the end-organs for both hearing and balance, as well as the cranial nerves responsible for speech and swallowing.[1] To access this region, surgeons drill through varying densities of bone to identify surgical landmarks. In addition to the limited visibility of the surgical field and complex anatomical geometry in this space, critical anatomical structures are often within millimeters of each other.

Due to these conditions, temporal bone surgery poses a high risk of accidental damage to surrounding structures during free-hand procedures. For example, after cochlear implantation, cochlear implantation, 45% of patients experience changes in taste, with 20% of those patients having unresolved symptoms by the end of their follow-up period.[2] In more rare cases, patients also are at risk for facial paralysis due to accidental damage to the facial nerve.[3] Accidental damage to the brain or to the membrane surrounding the brain (dura) can lead to CSF leakage. Damage to the sigmoid sinus, which drains blood from the brain to the jugular vein, can lead to abnormal closure or even clotting of the sinus itself.[4]

One possible solution in mitigating accidental damage to surrounding structures is using a cooperative control robot intraoperatively. Previously, the Laboratory of Computational Sensing and Robotics (LCSR) has developed such a robot that holds on to the surgical drill, which the surgeon can freely control.[5] Robot-assisted surgery has the potential to reduce hand tremor and limit movement around sensitive structures, thereby increasing patient safety and improving long-term outcomes. However, a key dependency for realizing this technology in the operating room is providing meaningful information about patient anatomy so that the robot can safely guide the surgeon throughout the procedure. Effectively, this means highlighting important structures on patient CT imaging that can be registered to a robotic system.

## Prior Work

Previous work in the LCSR has focused on segmenting CTs through registration methods (Figure 1).[6] With a manually segmented template CT, deformable registration methods can map or propagate template segmentations to target CTs that have not been segmented before. These segmentations can then be locally optimized to produce a final segmentation for the target CT. This segmentation propagation method achieves submillimeter accuracy for segmenting inner and middle ear structures, with an average surface distance of < 0.2 mm and almost 90% overlap with ground truth segmentations.

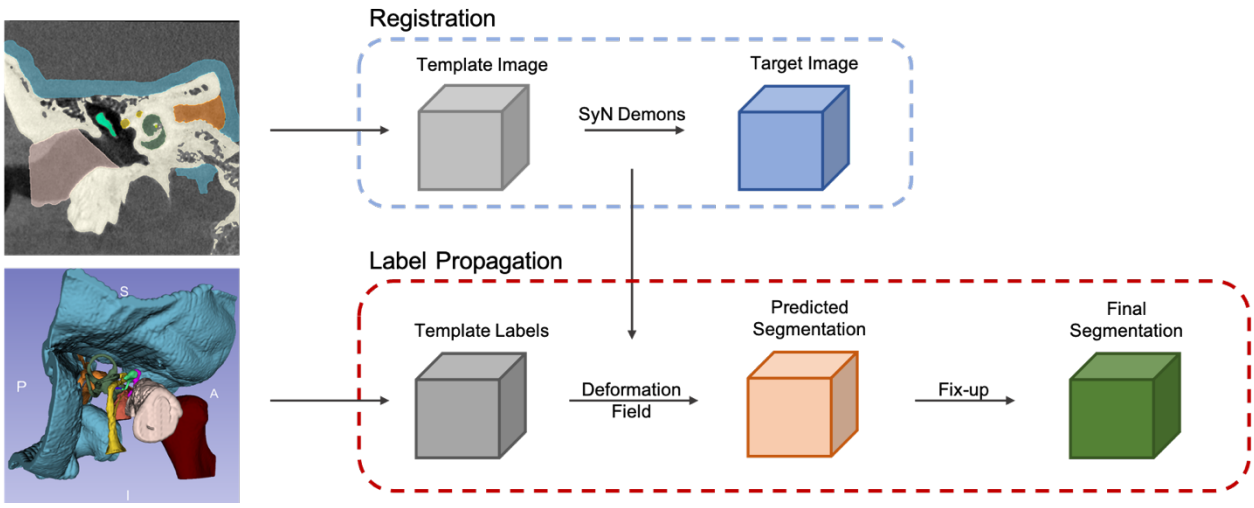


*Figure 1. Pipeline of the segmentation propagation method for temporal bone CT segmentation*

# Goals & Significance

## Significance:

Successful completion of this project will allow for more complete virtual safety barriers for robot-assisted temporal bone surgery. It can also be used to generate patient-specific segmentations as learning cases for junior otologists. Finally, this project has the potential to create the most complete dataset for model training and research.

Aside from the NIH OpenEar dataset, the dataset used for this project has the most complete segmentations of the temporal bone compared to any other group that has previously published in this area. In terms of anatomical boundaries of a mastoidectomy, which is the first step in virtually all temporal bone procedures, previous groups have only labeled one: the sigmoid sinus. This project's datasets not only label the sigmoid sinus, but also label the surrounding brain and the external auditory canal, which are the remaining mastoidectomy boundaries. By segmenting these areas, a cooperative control robotic system can then be able to apply virtual safety barriers to each of these boundaries, thereby providing for safe drilling throughout the procedure.

## Broad Goals

- To evaluate state state-of-the-art deep learning models for semantic segmentation of the temporal bone.
- To build the largest comprehensively annotated temporal bone CT database to date.

# General Experimental Setup

Our dataset consists of 21 manually segmented high resolution head CTs which have a voxel size of 0.1mm. The dimensions of the axial CT slices are 512x512 voxels$^2$ with an average z-stack of 494 images. The variability in the number of slices is due to differing patient anatomy and cropping parameters. All CT images are either of left or right temporal bone CTs with minor or no pathology and no prior temporal surgical procedures. The data has 16 labels: temporal bone, malleus, incus, stapes, vestibule

and cochlea, vestibulocochlear nerve, superior vestibular nerve, inferior vestibular nerve, cochlear nerve, facial nerve, chorda tympani, ICA, sinus and dura, vestibular aqueduct, TMJ, and EAC.

For running experiments, we used a local 3090 GPU workstation as well as a local 1080Ti GPU workstation, both of which have specifications laid out in Table 1. The key takeaway from the setups is that both have state-of-the art GPUs with high amounts of VRAM suitable for training on 3D images. SSH and VNC access were set up on the local Baltimore setup so that team members could always access the computer, even when not nearby physically.

*Table 1: Training rigs.*

| | Local California Setup | Local Baltimore Setup |
|---|---|---|
| GPU | GeForce GTX 1080Ti | GeForce RTX 3090 |
| VRAM | 11 GB DDR5 / 32 GB DDR5 | 24 GB DDR6 |
| OS | Ubuntu | Ubuntu |
| RAM | 32 GB | 32 GB |

# Technical Approach

Since the dataset used is small for deep learning approaches, other methods must be explored to train the model on less data. The method being explored is SSM based deformation, where we create an SSM of different deformation fields which were generated from registering a template image with over 40 other images not in the training, validation, or test set.

## Deep Learning Model

nnU-Net is a new benchmarking pipeline developed to standardize medical imaging.[7] It has top 33 leaderboard results for 53 different datasets, and effectively is a black box. It can be used to quickly establish a benchmark and there are 2D and 3D approaches available. The images are rigidly registered to a single template before the model training begins, and each image is intensity-normalized during pre-processing.

## Data Augmentation

Although data augmentation is built into nnU-Net, it is limited to rigid transformations. To provide more robust data augmentation, we have built statistical shape models (SSMs) of our temporal bone database using deformation fields from diffeomorphic image registration techniques (Figure 2). For each SSM, a dataset is designated as a template image. The remaining datasets are then deformably registered to the template image to generate inverse deformation fields for each registration process. As long as the template image is consistent between registration processes, the inverse deformation fields will have the same size and shape, which is requisite for SSM generation with principal component analysis (PCA). Once an SSM is generated, new deformation fields can be created by changing the mode weights of the model. These deformation fields can then be applied back to the template image and its corresponding segmentations to generate new training data for our nnU-Net model.

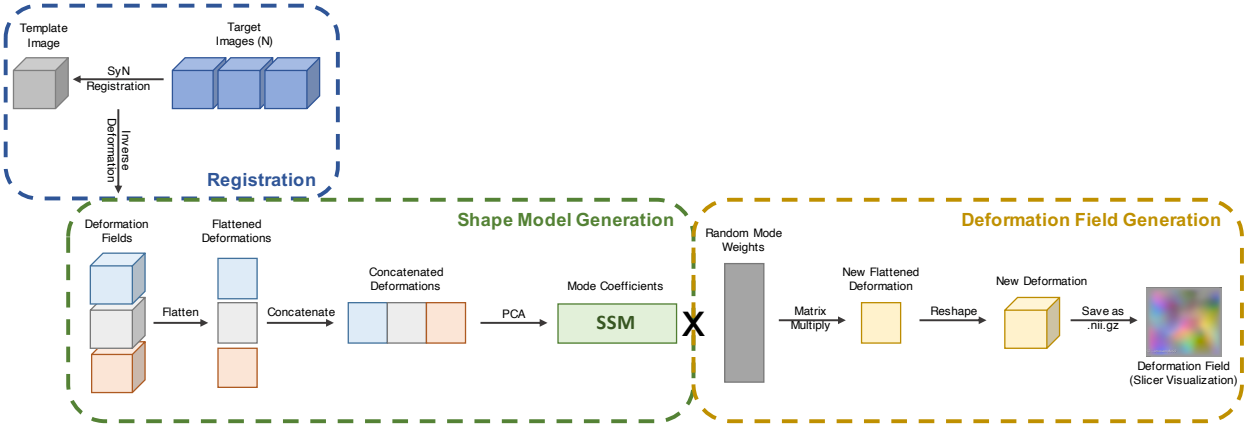


*Figure 2. Deformation field SSM workflow.*

## Generative Adversarial Network (GAN) Label Refinement

Using nnU-Net as a standard black box approach can be very effective, as Isensee et. al. has shown. However, results can always be improved upon and by using a clever adversarial set up, the labels have the potential to be refined further.

Generally speaking, GANs work by having a generator and a discriminator. The generator creates "fake" images, while the discriminator tries to detect the real image when presented with the two. The losses are combined and with careful consideration to keep the backpropagation graphs connected, this often times increases performance of the generator network. In this case, a pre-trained nnU-Net model can serve as the generator, with the ground truth label map and the original volume as the input. Then, the output prediction from the generator will be the fake image put into the discriminator, along with the original ground truth label map. There will be a loss term from both the generator and adversary, and they will be combined as shown in (Figure 3) to get a combined loss.

To preserve the backpropagation graph on the output predictions, which can be interpreted as a probability map of labels, the ground truth label is made to look like a probability map. This is done by one-hot encoding the ground truth label, then adding random noise to the ground truth, putting the resulting sum through a soft max across the relevant dimension, and then ensuring that the maximum probability per voxel of the "real" probability map corresponds to the correct ground truth label.

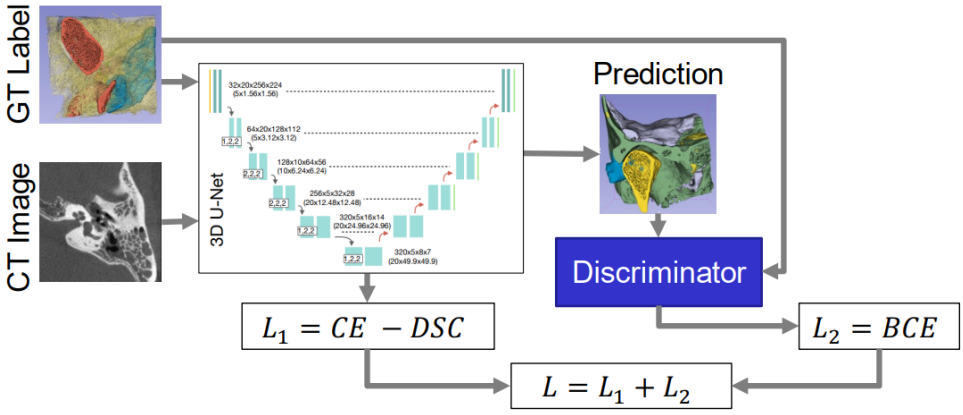


*Figure 3. GAN label refinement workflow.*

## Results

The out of the box nnU-Net implementation is referenced as the "Vanilla" results, which have 12 training volumes, and 3 validation volumes. There is a test dataset of 6 volumes but since the project is not yet complete, we cannot evaluate the test set yet without introducing bias. The SSM generated model ("SSM Gen.") contains the original template images for the SSM, as well as 10 additional generated datasets made from deformations produced from the SSM. Both were trained for 300 epochs, with an initial learning rate of 0.01. Both mean validation dice scores and mean validation modified Hausdorff distances are shown below in Table 1, and the training results are shown in Figure 4 and Figure 5.

*Table 2. nnU-Net validation accuracy metrics.*

| Class | Mean Val DSC | | Mean Val HD (mm) | |
|---|---|---|---|---|
| | Vanilla | SSM Gen. | Vanilla | SSM Gen. |
| **Bone** | .95 ± .01 | .95 ± .03 | **.001 ± .000** | 0.016 ± .020 |
| **Malleus** | **.93 ± .02** | .85 ± .04 | **.003 ± .001** | 0.010 ± .000 |
| **Incus** | **.93 ± .02** | .88 ± .02 | **.003 ± .000** | 0.037 ± .010 |
| **Stapes** | **.59 ± .16** | .46 ± .10 | **.023 ± .019** | 0.085 ± .087 |
| **Bony Labyrinth** | **.96 ± .01** | .94 ± .02 | .003 ± .002 | 0.003 ± .001 |
| **Internal Auditory Canal** | **.93 ± .02** | .84 ± .05 | **.015 ± .007** | 0.092 ± .034 |
| **Superior Vestibular Nerve** | .62 ± .10 | **.76 ± .03** | .099 ± .058 | **0.018 ± .005** |
| **Inferior Vestibular Nerve** | .53 ± .32 | **.71 ± .04** | .479 ± .766 | **0.046 ± .030** |
| **Cochlear Nerve** | .79 ± .07 | **.82 ± .02** | .131 ± .138 | **0.034 ± .026** |
| **Facial Nerve** | .85 ± .04 | **.86 ± .02** | **.027 ± .016** | 0.038 ± .012 |
| **Chorda Tympani** | **.72 ± .09** | .52 ± .02 | **.143 ± .085** | 0.598 ± .473 |
| **Internal Carotid Artery** | .93 ± .02 | .93 ± .03 | .061 ± .067 | **0.037 ± .033** |
| **Sigmoid Sinus + Dura** | .80 ± .04 | .80 ± .01 | .263 ± .366 | **0.204 ± .105** |
| **Vestibular Aqueduct** | **.67 ± .06** | .51 ± .16 | **.095 ± .098** | 0.274 ± .228 |
| **Mandible** | .94 ± .02 | **.96 ± .01** | .002 ± .002 | **0.001 ± .000** |
| **External Auditory Canal** | **.84 ± .02** | .81 ± .05 | **.130 ± .060** | 0.351 ± .271 |

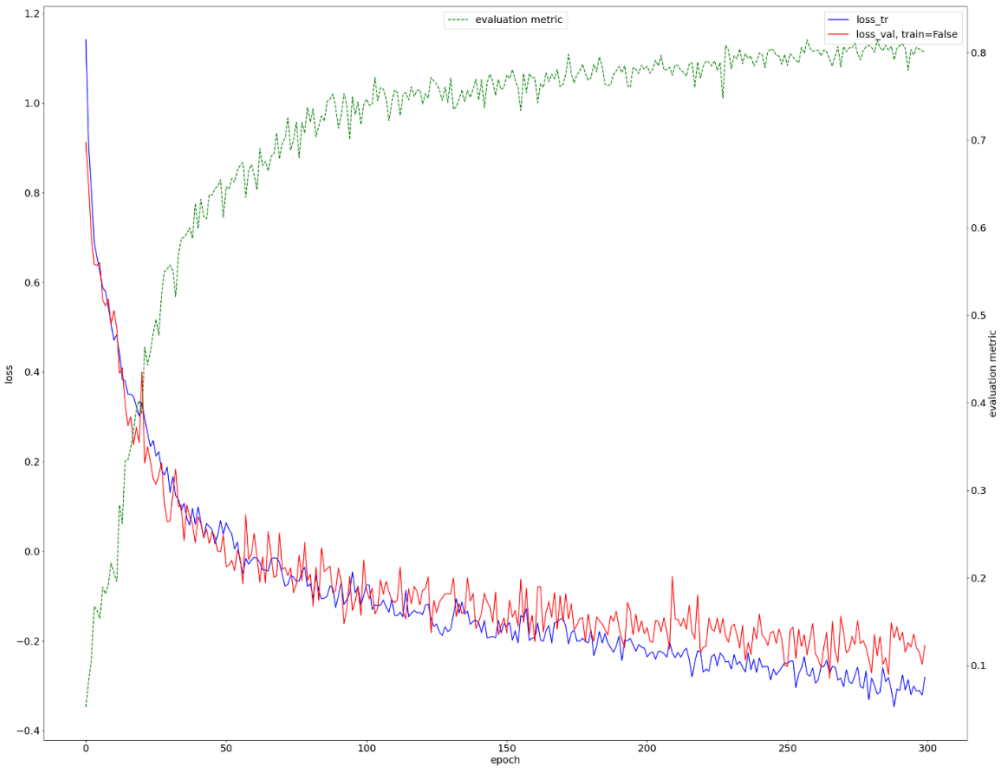


*Figure 4. Vanilla nnU-Net training progress*

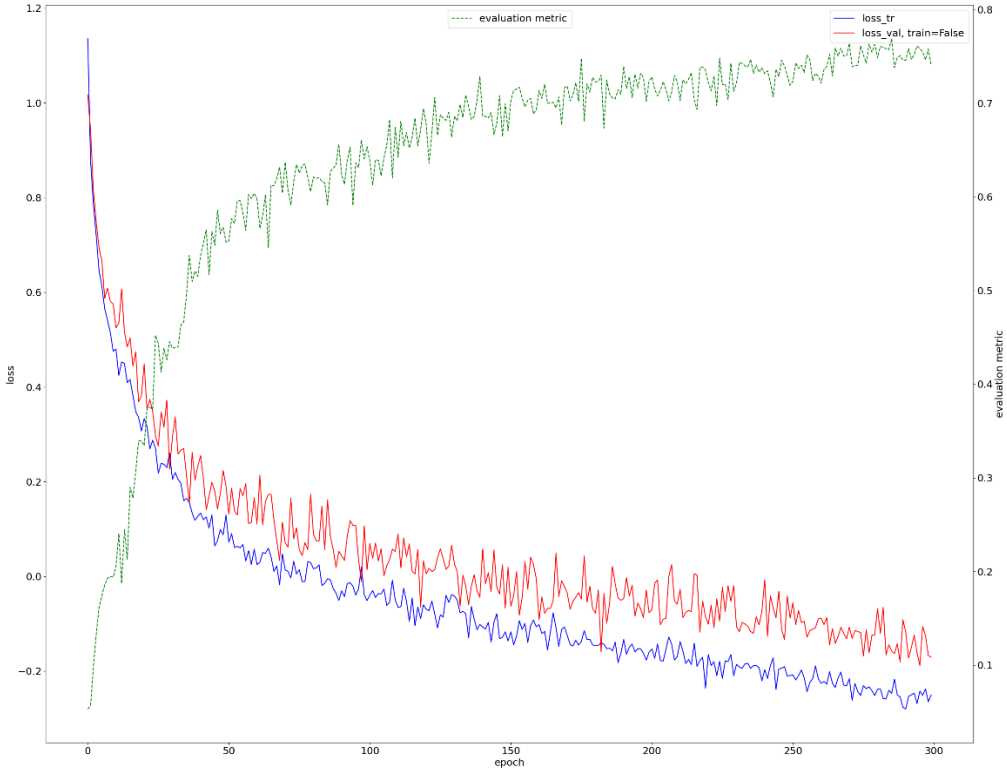


*Figure 5. SSM Gen. nnU-Net training progress*

## Discussion

As shown in our results, there is potential to squeeze more performance out of the SSM Gen. nnU-Net by potentially training for longer. While there is some separation between the training and validation curves, it does not look like divergence since they still are trending downwards, and the evaluation metric is trending upwards. More experiments will be conducted to confirm or deny this hypothesis. For the actual results, we can see that there is a mix of performance between the Vanilla and SSM Gen. model performances. Further analysis is needed to determine whether some differences are statistically significant, but for the most part it appears there may be trade-offs in the models, depending on what structures are important. This is expected since a single template, even with an infinite number of potential deformations, cannot capture all the variability in temporal bone CTs that is natural in human anatomy. Regardless, both results satisfy the requirement established at the beginning of the project of Hausdorff distances less than 0.6 mm and reach state-of-the art level dice scores, which is impressive especially considering the level of class imbalance and oddly shaped structures some of the labels have.

## Progress Evaluation

### Key Activities and Deliverables

The key activities and deliverables can be found in Table 3. They changed significantly throughout the course of the project as results developed and the original plan received feedback. The minimum, expected, and maximum activities and corresponding deliverables are laid out, and a Gantt Chart including the exact details and timeline is found in **Error! Reference source not found.**. The original key deliverables can be found in the appendix, although a summary of the changes can also be found below.

*Table 3: Activities and corresponding deliverables*

| | Activity | Deliverable | Status |
|---|---|---|---|
| Minimum | Synthesize deformed temporal bone CTs with labels to augment training dataset. | Statistical shape model of temporal bone CTs. | ✓ |
| Expected | Implementing nnU-Net. | Fully functioning model for CT segmentation with documentation. | ✓ |
| | Training model, then validating nnU-Net results on validation and test data. | Internal validation report with ground truth segmentations. | ⧗ |
| | Application of nnU-Net to external dataset. | External validation reports with Western University's dataset.[8] | ⧗ |
| Maximum | Implementing GAN label refinement into nnU-Net. | GAN label refinement model for CT segmentation with documentation. | ⧗ |
| | Final manuscript preparation. | Submittable manuscript. | ✖ |
| | Application of segmentation model to unlabeled dataset. | High quality segmented temporal bone CT dataset using our segmentation models. | ✖ |

Originally implementing nnU-Net was a minimum deliverable. This is because we underestimated the difficulty of augmenting our dataset, which became the new minimum deliverable for the project. The expected deliverables were then a functioning model for CT segmentation with documentation, as well as internal validation with ground truth segmentations. This is almost accomplished, but we cannot evaluate on the test set without completing all planned studies. The last expected deliverable of nnU-Net applied to an external dataset has not been met yet, since the collaborator at Western University recently had a heart attack and is not yet back to work. We will continue this work in the summer, trying to collaborate potentially with Vanderbilt University, or with manually registered NIH data. The GAN label refinement has been implemented but needs to be run and tuned; all training has been delayed due to unmet dependencies and contingencies that fell through. The final manuscript preparation will be completed in the summer as the other studies complete, as with the application of the segmentation mode to the unlabeled dataset.

## Dependencies

The project is mainly virtual so there are few physical dependencies. The dependencies are listed in Table 3. The only dependency met on time was the supervision agreement with Dr. Unberath, which was resolved in early February. Late in the project we realized that the labels were not consistent between two of the annotators. This meant that even though we originally thought the first dependency of finalizing labels was completed on time, it was not, and we had to retrain all the models we had trained before April. Furthermore, the workstation was delayed in arrival to 4/15, almost 1.5 months from the original date, and the contingency plans fell through. With MARCC, the most available GPU, the K80, was not suited for the 3D convolutions and mixed precision computing used in the implementation of nnU-Net. In short, it would take approximately 10 days to complete 300 epochs of a single training instance on MARCC with that GPU. Shortly after discovering this, we attempted to train on Google Cloud, but quickly burned through the free credits from the class due to the amount of training 3D segmentation problems need (approximately 28 hours per model). In the end this caused some large delays in the number of studies we were able to push out, but with the work continuing in the summer with all dependencies met, there should be no more major road bumps.