2020

# PWD-3DNet: A deep learning-based fully-automated segmentation of multiple structures on temporal bone CT scans

Western University

Western University

Western University

Western University

Western University

*See next page for additional authors*



Citation of this paper:

Nikan S, Van Osch K, Bartling M, Allen DG, Rohani SA, Connors B, Agrawal SK, Ladak HM. PWD-3DNet: A Deep Learning-Based Fully-Automated Segmentation of Multiple Structures on Temporal Bone CT Scans. IEEE Trans Image Process. 2021;30:739-753. doi: 10.1109/TIP.2020.3038363. Epub 2020 Dec 4. PMID: 33226942.

# PWD-3DNet: A deep learning-based fully-automated segmentation of multiple structures on temporal bone CT scans

Soodeh Nikan, Kylen Van Osch, Mandolin Bartling, Daniel G. Allen, S. Alireza Rohani, Ben Connors, Sumit K. Agrawal, and Hanif M. Ladak

***Abstract*—The temporal bone is a part of the lateral skull surface that contains organs responsible for hearing and balance. Mastering surgery of the temporal bone is challenging because of this complex and microscopic three-dimensional anatomy. Segmentation of intra-temporal anatomy based on computed tomography (CT) images is necessary for applications such as surgical training and rehearsal, amongst others. However, temporal bone segmentation is challenging due to the similar intensities and complicated anatomical relationships among critical structures, undetectable small structures on standard clinical CT, and the amount of time required for manual segmentation. This paper describes a single multi-class deep learning-based pipeline as the first fully automated algorithm for segmenting multiple temporal bone structures from CT volumes, including the sigmoid sinus, facial nerve, inner ear, malleus, incus, stapes, internal carotid artery and internal auditory canal. The proposed fully convolutional network, PWD-3DNet, is a patch-wise densely connected (PWD) three-dimensional (3D) network. The accuracy and speed of the proposed algorithm was shown to surpass current manual and semi-automated segmentation techniques. The experimental results yielded significantly high Dice similarity scores and low Hausdorff distances for all temporal bone structures with an average of 86% and 0.755 millimeter (mm), respectively. We illustrated that overlapping in the inference sub-volumes improves the segmentation performance. Moreover, we proposed augmentation layers by using samples with various transformations and image artefacts to increase the robustness of PWD-3DNet against image acquisition protocols, such as smoothing caused by soft tissue scanner settings and larger voxel sizes used for radiation reduction. The proposed algorithm was tested on low-resolution CTs acquired by another center with different scanner parameters than the ones used to create the algorithm and shows potential for application beyond the particular training data used in the study.**

***Index Terms*— Temporal bone CT, fully-automated segmentation, multi-structure, patch-wise, balanced-weighting, PWD-3DNet.**

## I. Introduction

Medical image segmentation is a classification task for labeling image pixels as background or regions of interest corresponding to specific anatomical structures. Image segmentation creates a simplified representation of anatomy for further clinical investigations. The temporal bone, a major part of the lateral skull surface, contains organs responsible for hearing and balance [1]. This small volume contains the critical anatomical structures including the middle ear, inner ear, nerves and vessels. Medical image segmentation of intra-temporal anatomy is challenging due the complicated anatomical relationships among the various structures. Creating three-dimensional (3D) patient-specific representations of temporal bone organs from computed tomography (CT) scans is useful to improve the feasibility of virtual reality surgical simulations for non-invasive surgery rehearsal, robotic surgery planning, and the design of effective artificial hearing implants [2].

Current approaches to segmentation represent regions of interest in medical images by labeling image voxels based on the variations in their intensity levels, or Hounsfield values. However, segmentation of temporal bone anatomy, where the Hounsfield values of critical structures are very similar, is a challenging task [3], [4]. Segmentation approaches can be classified as manual, semi-automatic and automatic. Repeated manual segmentation of a series of two-dimensional (2D) slices is impractical, labor intensive and time consuming. Furthermore, manual segmentation is vulnerable to inter/intra-operator variabilities leading to inconsistent subjective interpretations due to the intensity-level variations in different organs. These challenges in manual segmentation provided the impetus for pursuing semi/fully-automated segmentation techniques with consistent and quantitative analyses.

Semi-automated segmentation of temporal bone anatomy requires user input to initialize and/or guide an automatic algorithmic component. A supervised learning approach was described by Lu et al. [5], [6] which refines the resolution of CT scans, using the learned mapping obtained from the manual segmentations. Although such input increases reliability [7], the amount of user time required varies from one hour to one day depending on the size of input data and level of software

Submitted on March 17, 2020.

S. Nikan is with the Schulich School of Medicine & Dentistry, Western University, London, Ontario N6A 5C1, Canada (e-mail: snikan@uwo.ca).

K. Van Osch, and M. Bartling are with the Schulich School of Medicine & Dentistry, Western University.

D. G. Allen, and B. Connors are with the Department of Electrical and Computer Engineering, Western University.

S. K. Agrawal, and S. A. Rohani are with the Department of Otolaryngology – Head and Neck Surgery, Western University.

H. M. Ladak is with the Department of Electrical and Computer Engineering and Department of Medical Biophysics, Western University.

proficiency [4]. As ear surgeons are often the ones performing temporal bone segmentations for their trainees, this time requirement makes it impractical in a clinical setting. Statistical shape modeling, region growing, registration-free techniques and atlas-based methods have been previously considered to automate medical image segmentation. These techniques are robust to inter-subject variability. However, they are sensitive to noise and topological changes because of assumptions like diffeomorphic transformation priors [8].

Recent advances in machine learning have introduced fully automated segmentation strategies. Significant progress in computational tools have enabled machine/deep-learning-based segmentation algorithms that can process many scans in a short period of time. Souadih et al. [9] proposed a fully automatic approach using fuzzy c-mean, morphology operations and a dilated residual convolutional auto-encoder to segment the sphenoid sinus. Computer vision research has been focused on the effectiveness of fully convolutional neural networks (FCNs) to automatically learn distinctive representations of the training data and label them. In the field of medical imaging, automatic segmentation has progressed significantly in the past few years, with applications in cardiac images [10], cardiac MRI [11]-[15], skin lesion [16], [17], brain lesions [18] and neuropathologies [19], using 2D and 3D convolution kernels. Fauser et al. [20] proposed a shape regularized deep learning-based segmentation of temporal bone anatomy and trajectory planning. The authors used a majority voting approach with slice-by-slice predictions from individually trained 2D UNets [19] to initialize the active shape regularization. Although 2D techniques are memory efficient, they are associated with a loss in spatial context. Conversely, volumetric segmentations using 3D representations take the whole volume content into account and are more desirable in medical image analysis due to their relevance to the majority of interventional imaging modalities used in clinical practice such as volumetric CT, ultrasound and magnetic resonance imaging (MRI) scans [21]. Recent improvements in modern imaging techniques have created the possibility to capture detailed anatomical information from volumetric scans [22]. Li et al. [23] proposed a 3D deep supervised densely network (3D-DSD Net) using a dense Unet with a multi-pooling feature fusion encoding and a supervised mechanism based decoding to segment the anatomical organs of the temporal bone. The network takes the resampled volume as input.

The current work presents a patch-wise densely connected network (PWD-3DNet) and its application to multi-structure segmentation of the temporal bone from CT scans, with improved accuracy compared to other state-of-the-art techniques. The main contributions of this work can be summarized as follows:

1) This is the first study to use 3D convolutional neural networks in a fully automated pipeline for multi-structure segmentation of temporal bone CT scans (simultaneous segmentation of critical small and large scale organs). Our implementation is available at https://github.com/Auditory-Biophysics-Lab/Slicer-ABLTemporalBoneSegmentation.

2) In order to train and evaluate the network, a comprehensive set of images were acquired and manually segmented by domain experts. Such data are important for cross-site replication and domain transfer studies. The data are available through written request to HML (hladak@uwo.ca).

## II. Related works

Semi-automated segmentation techniques, such as Fast GrowCut, develop the label volume automatically based on the voxels pre-marked by an expert for the regions of interest [24]. Unfortunately, these techniques require an extensive manual corrections unless the structures are very clearly demarcated. Powell et al. [25] presented an atlas-based segmentation to develop the gold-standard atlas and regions of interest. Intensity-based segmentation was then used to segment critical structures in the temporal bone anatomy. In atlas-based segmentation methods, the anatomy to be segmented must be similar to the atlas. Also, due to inter-subject variability in the appearances of different anatomical structures, determining the correspondences between images from different subjects is challenging [3]. Therefore, the level of reliability of these methods varies by the correctness of atlases and statistical priors [7]. In multi-atlas label fusion (MALF) methods, a training dataset is registered to each new image and the propagated reference segmentations are combined to generate new segmentations [26]. However, this method is computationally expensive and prone to inaccurate fusion due to registration errors [27]. Statistical shape modeling co-registers training images to investigate the anatomical correspondences between images from different subjects and construct a shape/appearance-based model of the training data. Segmentation of the new test samples is performed by fitting the corresponding model to the new image [28]. Noble et al. [29]-[31] used a combination of a model, created by statistical a-priori intensity and shape information of the temporal bone structure, and atlas-based registration to align the model to the CT scan. However, the variability in organ shape and appearance, soft tissue deformation, and registration accuracy can significantly affect these techniques.

Convolutional Neural Networks (CNNs) are capable of learning distinctive visual representations of the training data automatically and can then classify them similar to a human expert. As the inputs are processed through the network layers, the level of abstraction of the resulting features increases. Shallower layers grasp local information, while deeper layers use filters whose receptive fields are much broader, thus capturing global information. Applying CNNs to medical image segmentation was first considered for knee cartilage segmentation in MRI scans [32]. After this, many researchers adopted deep learning for segmentation of other anatomies, including brain tissue [33], neuroanatomical structures [34], prostate [35], [36], heart [37]-[39], bone [8], [40], and tumors [41], [42]. Recently, researchers showed that utilizing multi-level features through FCN with skip connections across layers has a great impact on the performance of medical image segmentation [43]. Multi-organ strategies lead to a more accurate representation of the complex spatial and

physiological inter-organ relationships in human anatomy [44]. Deep learning approaches can learn the global context characteristics and inter-organ interactions from examples automatically [44], [40]. FCNs can be applied to segment multiple structures concurrently with acceptable accuracy. In the shape regularized preoperative pipeline proposed by Fauser et al. [20], multiple structures of the temporal bone from CT slices were segmented using 2D sub-UNets. The subnetworks were subdivided into multi-class and binary segmentations to deal with the class label imbalance in the training dataset. However, their algorithm was not fully automated and the segmentation results were used as the input to the active shape models by restricting the segmentation to the trained shape space [20]. Deep residual networks, in which bypassing paths are used to prevent the vanishing gradient problem by propagating information among network blocks, have shown nice convergence in largescale applications. A voxel-wise residual network (VoxResNet) was proposed by Chen et al. [21], which combined the shape features with low-level appearance and high-level context information for improving the segmentation predictions. Gibson et al. [45] created a framework with the application to segment multiple abdomen organs on CT scans using DenseVNet with a hinge loss, resulting in a similarity score of 0.82 which is better than that of existing MALF and FCN methods. The authors added a low-resolution map of trainable parameters as a spatial prior to the network's output likelihood. However, the trained parameters may misguide the output probability. Also, their proposed dense training strategy increases the risk of losing discriminative information of small regions of interest. However, including the full-size image as in VoxResNet [21] and DenseVNet [45] is not memory efficient since the size of 3D medical scans is relatively large. Li et al. proposed 3D-DSD Net [23] to segment the inner ear and some organs of the middle ear. However, very small and hardly visible structures (stapes and facial nerve), which are challenging in clinical CT scans, were not segmented.

Rather than including the whole image area which causes redundant calculations of background regions in the learning process, patch-based or region-specific strategies have recently been introduced in medical image segmentation applications to reduce the complexity [44]. Li et al. [34] proposed a high-resolution, compact CNN with residual connections (HighRes3DNet), which used randomly sampled sub-volumes from a brain MRI for the volumetric segmentation of fine structures with an average Dice similarity score of 0.84. Another successful technique in brain tissue segmentation was the DeepMedic algorithm by Kamnitsas et al. [33]. This method used the patch-based sampling strategy and two CNN pathways for capturing both global and local features at multiple scales, as well as a 3D fully-connected conditional random field for postprocessing at the final level of the network. Although this technique achieved a Dice similarity score of 0.85 on 110 test cases from the BRATS 2015 challenge [46], it has a dual-path architecture with many layers of 3D convolutions which is not computationally efficient. In addition, it uses uniform or background/foreground window sampling which takes patches from random locations in the volume or equal number of random patches from background and foreground regions, which may cause an imbalance in patch distributions. Therefore, in multi-class segmentation, there is no guarantee that samples from all regions of interest exist in each training iteration of learning. In particular, some classes of labels occupy very small regions of the volume, and the network may become stuck in a local minimum of the objective function. This would yield partially detected results since the predictions are biased towards larger regions [35].

In addition to the architecture of the network which improves the segmentation performance, different cost functions have been investigated to deal with class imbalance challenges between the foreground and multi-class background voxels. Milletari et al. proposed an objective function based on the gradient of Dice in a V-shape compression/ decompression network for binary segmentation of the prostate in MRI volumes from the full-size image volume [35]. Boutillon et al. [40], proposed additional shape priors and adversarial regularization terms in the loss function which assessed the global similarity between predicted and ground truth delineations for multi-structure bone segmentation on scarce heterogeneous pediatric imaging. Kamnitsas et al. also proposed an unsupervised domain adaptation segmentation of brain lesions using adversarial loss and a domain classifier invariant to differences in imaging protocols, which improves the generalization of the segmentation models [47].

Furthermore, another strategy to tackle the class imbalance problem in multi-organ segmentation is the patch sampling, dividing the original image into smaller sub-volumes, by either using a sliding window approach (grid sampling) or distributing patches to the background and foreground regions randomly or equally, as in Gibson et al. [48] and Rossello [49], respectively.

The current work proposes the first single multi-class fully automated deep-learning-based method for the segmentation of multiple critical structures of the temporal bone, with various size of very small and hardly visible structures and large scale organs simultaneously, from CT scans. One of the important contributions in this work was to augment the training set by including different types of CT scans to account for common variabilities due to scanner models and acquisition protocols. Therefore, the algorithm can potentially generate accurate predictions for inference images with different modalities, resolutions and scanner parameters. Also, a weighted patch-wise strategy is proposed to create balanced learning from multiple labels with various sizes in a temporal bone volume.

## III. Materials

### *A. Micro-CT and clinical CT*

Thirty-nine adult cadaveric temporal bone specimens were used. All specimens were obtained with approval from the body bequeathal program at Western University, London, Ontario, Canada in accordance with the Anatomy Act of Ontario and Western's Committee for Cadaveric Use in Research (Approval number #19062014).

To fit within the micro-CT scanner bore, a cylindrical drill

bit with a diameter of 40 mm and a height of 60 mm was used to cut out the region of interest from the temporal bone. The cadaveric temporal bone specimens were scanned using the GE Healthcare eXplore Locus micro-CT scanner (GE Healthcare, Chicago, IL). The scanner was set at a voltage of 80 kV and a working current of 0.45 mA. Approximately 900 views were captured, with an incremental angle of 0.4 degrees. Images were reconstructed into thirty-nine 3D volumes with an isometric voxel size of 154 micrometers (μm). Fig. 1 shows a lateral view of a micro-CT scan of a temporal bone in the dataset with the manual segmentation labels and 3D model of the label map.

All of the specimens were also scanned at a clinical resolution using the Discovery CT750 HD Clinical Scanner (GE Healthcare, Chicago, IL) equipped with GE's Gemstone CT detector. Slice thickness was set to 0.625 mm with the scanner operating at an x-ray voltage of 120 kV. The resolution of each sample's scan was 625 μm and the acquisition time for each sample was approximately 37 seconds. Each of these clinical volumes were aligned and registered rigidly with micro-CT references from the same subject using the publicly available 3D Slicer software (https://www.slicer.org/) [50]. Micro-CT images were manually segmented by multiple experts simultaneously including a surgeon (SKA), anatomist (KVO), and medical trainee (MB). Consensus interpretation was achieved using the 3D model on each slice to establish a gold standard for the segmentation.

Manual labeling of eight temporal bone structures, including the sigmoid sinus, facial nerve, inner ear (including the cochlea and semicircular canals), malleus, incus, stapes, internal carotid artery and internal auditory canal, were performed using the region growing approach in the 3D Slicer software using the initial seeds and then manually corrected.

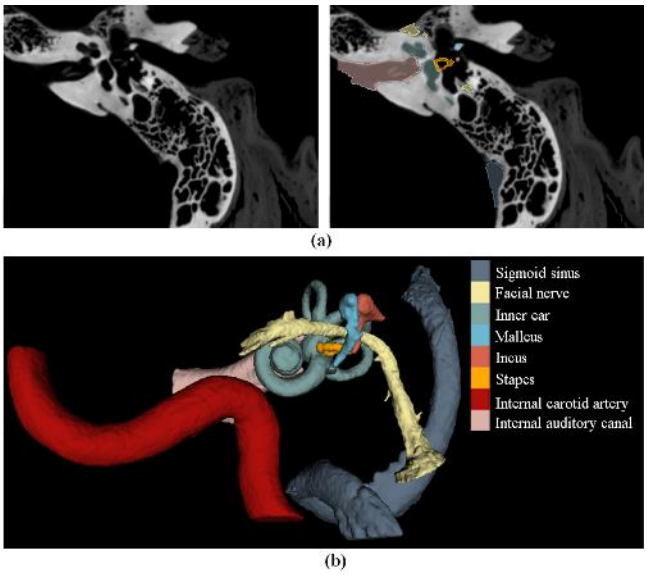


Fig. 1. Temporal bone CT scans and label maps. (a) left: a 2D slice of a micro-CT image volume; right: slice through manual segmentation label volume overlaid on micro-CT. (b) 3D model of label map. Labels are sigmoid sinus (dark gray), facial nerve (yellow), inner ear (light gray), malleus (blue), incus (light red), stapes (orange), internal carotid artery (dark red) and internal auditory canal (pink).

### *B. Cone-beam CT*

Cone-beam CT scans of two specimens were provided by Med-El GmbH (Innsbruck, Austria) (https://www.medel.com/ca/), a leading manufacturer of auditory implants. The scans were created from the XCAT system, a mobile CT scanner from Xoran Technologies LLC (Ann Arbor, USA). The system consists of a cone-beam scanner and creates volumes with a voxel size of 0.1 mm. Ground truth labels were again manually segmented by experts in the same fashion.

## IV. Methods

The building blocks of the proposed patch-wise and densely-connected 3D FCN (PWD-3DNet) are shown in Fig. 2. This work was motivated by the DenseVNet architecture [45], however the proposed algorithm differs in some respects. The DenseVNet segmentation method consists of dense training by utilizing the full-size volumes of the input to the network and then resampling them to a smaller voxel size, whereas the current algorithm adopted a balanced window sampling strategy. Since the regions of interest in the temporal bone training set were extremely small (compared to the abdominal structures in Gibson et al. [45]), we avoided resampling input volumes to a smaller size to preserve the spatial content of small structures. To reduce the computational burden caused by the large size of the input as well as to extract local context from the input volumes, 32 sub-volumes were extracted with balanced distribution among multiple classes of anatomical organs by using discrete ground truth labels sampler weights. The proposed balanced strategy creates equal contributions among regions of interest with various voxel sizes in each training epoch. In this way, each class of label is sampled with the same probability as the other classes to overcome any class imbalance. This method reduces the randomness in the window samples which accordingly decreases the number of redundant computations. The objective function was also modified to a Dice loss, which corrects the class imbalance by a square in the denominator and reduces the risk of bias toward larger labels. Furthermore, in DenseVNet [45] a low-resolution trainable spatial was used prior to the network's output. In our proposed algorithm, we excluded that low-resolution spatial to make the predictions less dependent on prior information and better prepared for inconsistencies in medical images. The activation function and upsampling units were also changed to a parametric function and B-spline method, respectively.

The architecture of the proposed PWD-3DNet, as shown in the schematic of Fig. 2, consists of 3D balanced patch sampling, convolutional layers, dense connections, and decoder-encoder units. Convolutional layers include a 3D convolution kernel, a mini-batch gradient descent, dropout optimization and PReLU nonlinearity. Each dense connection, motivated by [45] and Huang et al. [51], consists of four consecutive 3×3×3 convolutional volumetric kernels of stride 1. The inputs of these dense convolutions were zero-padded by one voxel in each dimension to keep the feature-map size fixed so that they could be concatenated easily [51]. This convolutional structure

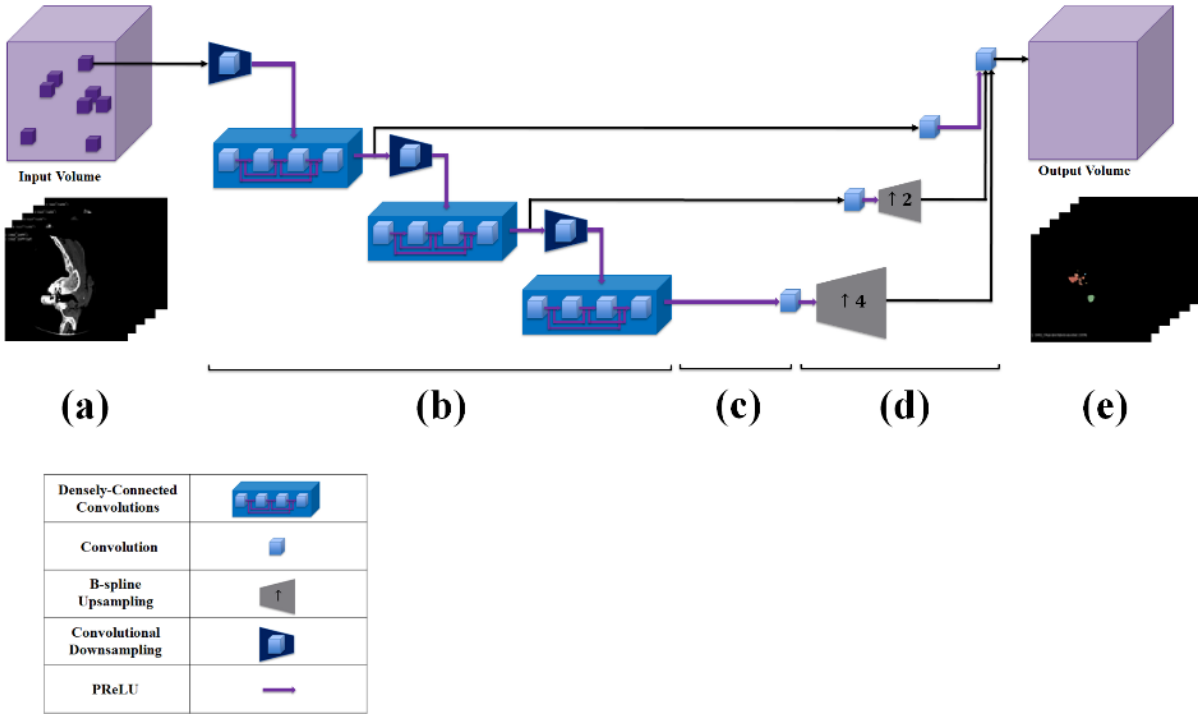


Fig. 2. Architecture of the proposed PWD-3DNet segmentation algorithm: a) extracting sub volumes using a balanced-weighted patch sampling, b) Three encoder units constructed by strided convolutions and densely connected stacks of features, c) skip connections across layers to preserve low-level features, d) decoders constructed by two B-spline upsampling units to decompress data, e) concatenated lapel map volumes with the original segmentation resolution.

creates a dense connectivity of convolution channels by flowing the feature map from each layer as the input to the subsequent layers which makes the network compact and memory efficient, due to a fewer number of channels. This also improves the performance of the network in perceiving global information of the anatomy, preserving spatial context, and providing richer patterns due to diversified features. Finally, by using this approach, the gradient of error can be propagated to earlier layers more directly. Batch-normalization is used to reduce the number of input feature maps and accelerate the training process. The downsampling and upsampling units appear in a V-shape with skip connections. Along the compression stages, the resolution of data is reduced through three downsampling units that are constructed by 3×3×3 convolutions with stride 2 to halve the dimension of feature maps in three steps for computational efficiency and enlarge the receptive field of features. Two B-spline upsampling units gradually decompress data at two scales of 4 and 2, to reach the original segmentation resolution, which limits the artifacts induced by transpose convolution due to uneven overlaps and improves the accuracy of the predictions. Although this could be at the cost of losing delineation power, qualitative segmentation results confirm the good overall delineation power of the proposed method. Furthermore, due to using predefined interpolations instead of learnable parameters, the training speed is improved [45]. The skip connection is a single 3×3×3 convolution to improve the convergence time as well as keeping higher resolution and fine-grained information to the final segmentation predictions and preserving spatial structure; low-level features would be otherwise lost in the downsampling stages of the network [35]. Afterward, the results of the compression and decompression layers are all concatenated and passed through the last convolutional layer as the main classifier. This builds segmentation results with the same size as input volume by calculating the posterior log-probability of the class label for each volume voxel. Final labelmap is aggregated by concatenating the labelmaps obtained from the argmax. A summary of the proposed network parameters, including size of inputs, outputs and convolutional kernel at each layer, has been included in Table I. The optimal kernel size of 3×3×3 was adopted for all convolutions to allow for a relatively low number of parameters and more regularization [34].

TABLE I
PARAMETERS OF PWD-3DNET ALGORITHM.

| Layer | Input size | Output size | Kernel | Stride |
|---|---|---|---|---|
| **Convolutional downsampling** | 144×144×144 | 72×72×72 | 3×3×3 | 2 |
| **Dense convolution** | 72×72×72 | 72×72×72 | 3×3×3 | 1 |
| **Convolutional downsampling** | 72×72×72 | 36×36×36 | 3×3×3 | 2 |
| **Dense convolution** | 36×36×36 | 36×36×36 | 3×3×3 | 1 |
| **Convolutional downsampling** | 36×36×36 | 18×18×18 | 3×3×3 | 2 |
| **Dense convolution** | 18×18×18 | 18×18×18 | 3×3×3 | 1 |
| **Skip convolution** | 72×72×72 | 72×72×72 | 3×3×3 | 1 |
| **Skip convolution** | 36×36×36 | 36×36×36 | 3×3×3 | 1 |
| **B-spline upsampling** | 36×36×36 | 72×72×72 | - | - |
| **Skip convolution** | 18×18×18 | 18×18×18 | 3×3×3 | 1 |
| **B-spline upsampling** | 18×18×18 | 72×72×72 | - | - |
| **Output-layer convolution** | 72×72×72 | 72×72×72 | 1×1×1 | 1 |

### A. Patch-wise analysis method

Window-sampling analysis refers to strategies to generate patches from each image volume. Multiple strategies have been applied in the literature including grid, uniform, and weighted sampling strategies [33]:

1) *Grid window sampling*
   In the grid window sampling method, a sliding window moves on the image and extracts the image patches densely from every adjacent location of the volume to cover the whole size of the image. The size of sub-volumes is smaller or equal to the image volume size. If the size of the image is not divisible by the patch size, the windows are sampled with the minimal possible overlap [48]. Fig. 3a shows the grid patching strategy for a sample volume. In this method, a huge number of calculations are redundant which adds to the algorithm run-time and causes inefficient memory usage.
2) *Uniform window sampling*
   In the uniform window sampling method, the number of image patches is smaller than the previous approach to reduce computational burden and redundant calculations. First, the sampler finds all feasible locations of the image windows on the input image and distributes the targeted number of patches randomly. Figure 3b illustrates uniform sampling.
3) *Background/foreground sampling*
   The background/foreground strategy devotes an equal number of samples to the background and foreground regions [49], as shown in Fig. 3c. However, the probability of patch distribution among different regions of interest is random. Thus, the regions of interest might not contribute equally in the learning iterations. Compared to uniform patching the background/foreground sampling is controlled, however it still makes the training data unbalanced.
4) *Balanced window sampling*
   In balanced window sampling, the sampler distributes the defined number of patches to the regions of interest equally by using the discrete ground truth labels as sampler weights. Therefore, each class of label is sampled with the same probability as the other classes [48], as shown in Fig. 3d. Compared to the uniform and background/foreground techniques, which suffer from class imbalance, this method leads to a relatively balanced ratio of labels and reduces the randomness in subsampling, which accordingly decreases the number of redundant computations.

In this project, we adopted the balanced-weighted sampling method to overcome the class imbalance that occurs in algorithms with uniform and background/foreground patching strategies. This method also reduces the computational redundancy that occurs in grid sampling.

### B. Similarity loss function

The choice of loss function is another important design consideration which can be highly influential in the performance of a segmentation algorithm. The final predictions at the output level of our network are processed through a softmax multi-classification which calculates the probability of each voxel as belonging to the background or to one of the eight temporal bone structures of interest. In order to optimize the learning performance, the accuracy of predicted voxels in comparison with the ground-truth voxels needs to be evaluated by calculating a cost function. Dice similarity score (DSS) is commonly used as the cost function, which calculates the ratio between 1) the intersection of the predicted and ground-truth regions of interest and 2) the sum of the voxels of both regions of interest. The Dice similarity coefficient between the predicted volume $\hat{V}$ and the ground-truth volume $V$ is defined as follows [35],

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

where $M$ is the number of classes in the label map. Each class represents a temporal bone anatomic structure region of interest, with the background label considered as one class with the value of 0. $v_m^j$ and $\hat{v}_m^j$ denote the $j^{th}$ voxels of the predicted volume and the ground-truth volume of class $m$, respectively. The quantity of mean Dice has a range between 0 and 1 [35]. The loss function was calculated by subtracting DSS from 1. The learning process aimed to minimize the loss value.

One of the main issues in medical image segmentation is the classification of an anatomy of interest that occupies a very small region, the size of which is a fraction of the full image size. This results in class imbalance and causes the learning algorithm to be deceived by local minima [35]. In recent years, many techniques have been investigated to overcome this issue. Including the class differences effect in the cost function is one of the most effective techniques in reducing class imbalance. The loss function in this work (adopted from [35]), does not require assigning weights to different classes due to including a square in the denominator. The performance of the segmentation algorithm was improved by preventing the algorithm from being biased towards the predictions of larger regions of interest and mislabeling smaller sized regions.

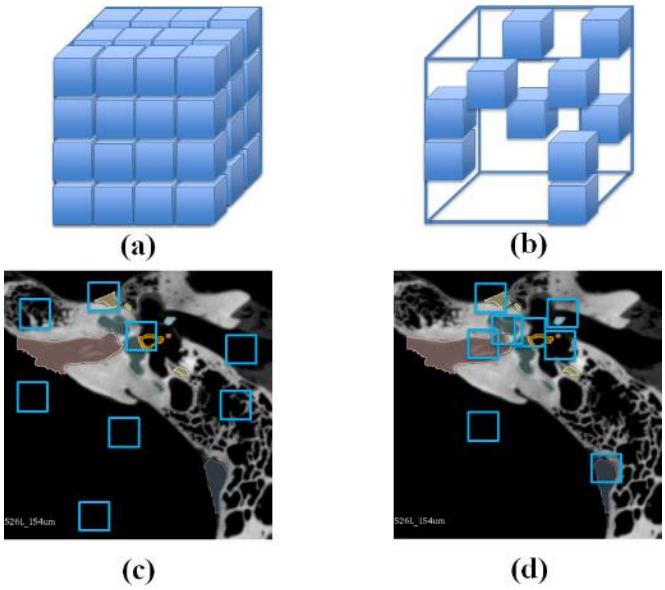


Fig. 3. (a) Grid window-sampling, (b) uniform window-sampling, (c) background/foreground sampling and (d) proposed balanced window sampling.

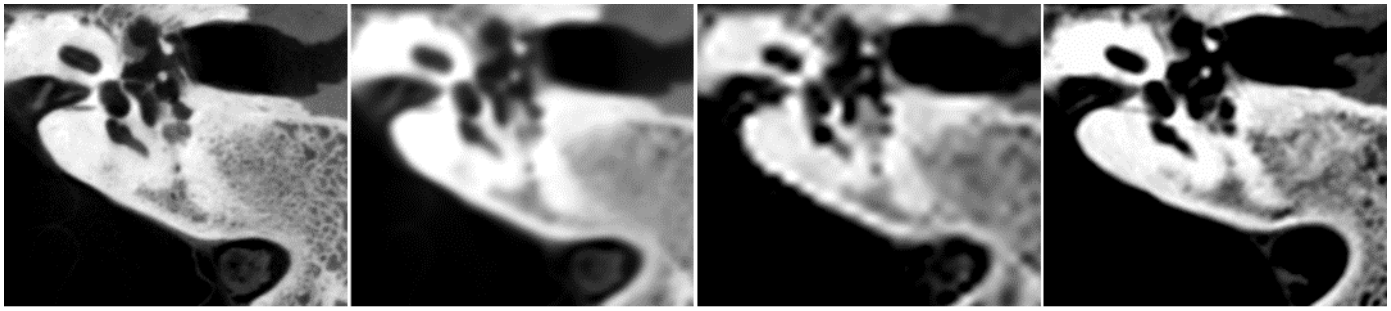


Fig. 4. Lateral view of one slice of the temporal bone CT scans in augmentation layers: (a) micro-CT, (b) blurred micro-CT, (c) resampled micro-CT and (d) clinical CT.

### *C. Augmentation layers*

Novel augmentation layers were introduced to the network training set to enlarge the input dataset, regularize the trained models and enable them to generalize themselves to new unseen test samples with different data acquisition parameters. By augmenting the aforementioned images, the proposed network is more robust against various imaging modalities and scanner settings. Augmentation layers are also more effective and straightforward than regularizing the hyper-parameters iteratively. The augmented layers include the following images (from each of the following datasets, we had equal numbers of images in the training set).

1) *Blurred micro-CT*
   The first group of augmented images consists of blurred versions of high-resolution micro-CTs, as shown in Fig. 4b. To include the blur effect, the images were smoothed with a Gaussian filter with standard deviation $\sigma$=0.45. This data was included because we aimed to enable our algorithm to successfully segment various types of CT scans. For example, certain acquisition protocols are optimized for either bone or soft tissue imaging. CT scans with soft tissue protocols have protocols resulting in less sharpness of the bony boundaries.
2) *Resampled micro-CT*
   The second set of augmented volumes, shown in Fig. 4c, were micro-CTs downsampled to a larger slice thickness (1×1×1 mm) first and then upsampled to the original voxel size (0.154×0.154×0.154 mm) using B-spline interpolation. By using this approach, we intended to purposefully lose some information in the images and emulate artefacts from CT scans from other sites that were acquired at a lower resolution (larger voxel size; up to 1 or 2 mm).
3) *Clinical CT*
   The final set of data consisted of clinical scans of the same temporal bone samples as in the micro-CT dataset, as shown in Fig. 4d. As micro-CT is not performed in human patients, this augmented data prepared the network for the types of CT scans typically acquired in clinical settings. This acquisition of this data has been described in Section III (Materials). These clinical scans were aligned to the micro-CT references and resampled by B-spline interpolation to the same voxel size.

### *D. Inference*

To predict the segmentation labels of an unseen input volume using our trained models, image patches with the same size as the training sub-volumes were extracted from the test scans. The window sampling strategy used in the prediction stage was grid sampling to cover the whole area on the input image. In our experiments in the following section, we evaluated the performance of the proposed network with various amounts of overlap between adjacent patches. Although overlapping requires more calculations and adds to the testing run-time, it was shown to lead to better overall performance. A trade-off between the accuracy of predictions and the required processing time is needed.

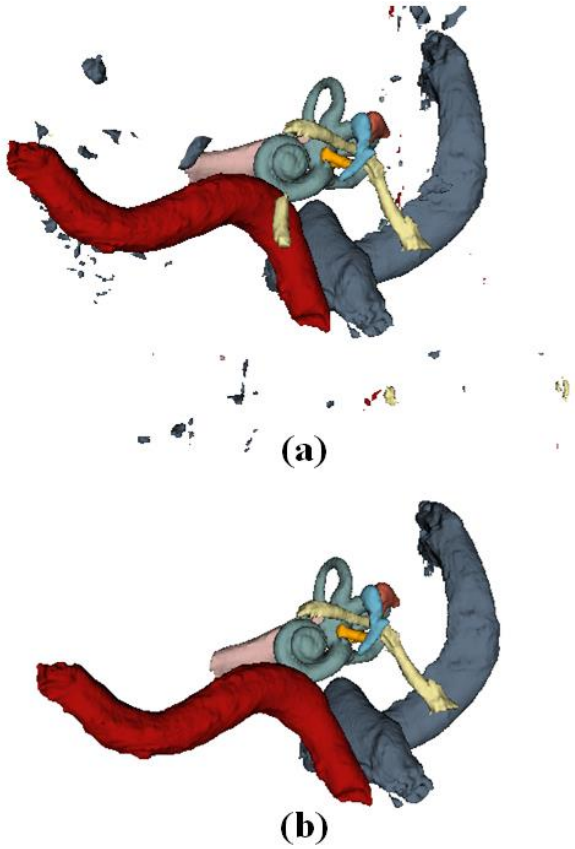


Fig. 5. 3D model of segmentation prediction of one subject in our test set: (a) before post-processing (with disconnected components) and (b) after post-processing. Labels are sigmoid sinus (dark gray), facial nerve (yellow), inner ear (light gray), malleus (blue), incus (light red), stapes (orange), internal carotid artery (dark red) and internal auditory canal (pink).

The final predicted labels contained some expected noise and disconnected components within a single structure's labelmap as shown in Fig. 5a. These types of isolated islands in the segmentation have been previously described [33]. They were automatically removed using the connected-component island removal filter in 3D Slicer as a morphological post-processing operation, as illustrated in Fig. 5b [41].

## V. Experimental Evaluation

The evaluation of PWD-3DNet was four-fold. First, the accuracy of predictions of the algorithm was calculated on all test samples with different resolutions, image artefacts and imaging parameters. Second, the overlapping effect in sub-sampling the input volumes was assessed. Third, the performance of our proposed algorithm was evaluated against other segmentation methods. The impact of dense training compared to the patch-wise strategy was also evaluated. Fourth, the effectiveness of the augmented layers was evaluated by validating the performance of PWD-3DNet against imaging protocols. Details of the implementation and evaluation are discussed in the following sections.

### *A. Implementation setting*

We evaluated the segmentation accuracy of the proposed pipeline on our dataset of CT scans. As explained in section IV-C, before augmentation we had 78 micro-CT and clinical CT scans from 39 specimens from our center in our dataset which was augmented with two more scans per specimen, leading to a total of 156 scans. For each subject, we had four scans including one volumetric micro-CT image with 0.154×0.154×0.154 mm voxel spacing, a smoothed version and a resampled version of micro-CT scan and a clinical scan with 0.650×.0650×0.650 mm voxel spacing, as shown in Fig. 4. The smoothed version of the micro-CT scan was created using a Gaussian filter with standard deviation $\sigma = 0.45$. The other version of micro-CT was utilized by resampling the image to a lower resolution (1×1×1 mm voxel spacing) and then upsampling to the original voxel size of 0.154×0.154×0.154 mm. One-hundred and forty scans, from 35 specimens, were randomly split into a training and validation subset, with 126 and 14 CT scans each (90 and 10 percent, respectively). Inference images consisted of 18 unseen scans; 16 images of four specimens from our center and 2 cone-beam CTs of two specimens from another center, as shown in Fig. 6. The blurred and resampled micro-CTs were used in inference to emulate less sharpness of the bony boundaries acquired from soft tissue acquisition protocols and artefacts from CT scans with lower resolutions (larger voxel size; up to 1 or 2 mm), respectively, and assess the segmentation performance in those scenarios. The cone-beam CT images were resampled to a 0.154×0.154×0.154 mm voxel size. These two samples were adopted to validate the proposed segmentation method against imaging protocols, scanned by a different model of scanner and different settings from the training set acquired at our site. Additionally, the brain is in close proximity to the temporal bone structure in these two images, making the segmentation task more difficult.

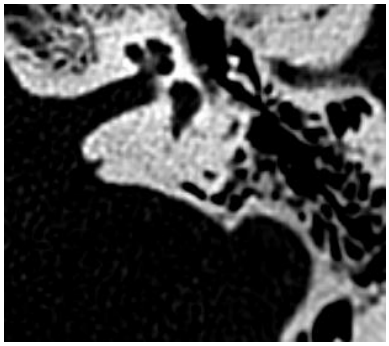

Fig. 6. Lateral view of one slice of a temporal bone cone-beam CT scan in our dataset.

The network was trained using a workstation with Titan RTX GPU for 20800 iterations (52 epochs) with a mini-batch size of 10, Adam optimizer and learning rate of 0.001. The average volume size was 641×594×705 voxels. For training, we used the full-size volumes and sampled each at 32 locations using the balanced-weighted window sampling method with the patch-size of 144×144×144 voxels. Based on the trial experiments using different patch sizes, smaller values led to more noise, disconnected components caused by larger structures and drastically reduced segmentation accuracy. Whereas, with a larger patch size, the trained models did not learn the characteristics of small structures sufficiently and could not predict those regions successfully. Histogram standardization and whitening normalization from the Niftynet open source platform [39] were utilized and training data was augmented with random rotation of each orthogonal plane with an angle in the range of [-10, 10] and spatial rescaling by transforming volumes in a range of [0.9, 1.1] of the original size. Various hyper-parameters were tuned and based on the trial and error on the model performance the optimal values were selected.

### *B. PWD-3DNet predictions of key structures with test data variability*

The accuracy of segmentation predictions of the proposed algorithm was evaluated in terms of Dice similarity score (DSS), symmetric 95% Hausdorff distances (HD) and Jaccard score (JS) compared to the ground truth (manual segmentations) and predicted label maps for each structure of the temporal bone. In the inference stage, the 144×144×144 patches were extracted from every location of the test volume with 36×36×36 voxel overlap between adjacent sub-volumes. Table II shows the average DSS, HD and JS values for each temporal bone structure on all 18 test samples.

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 |

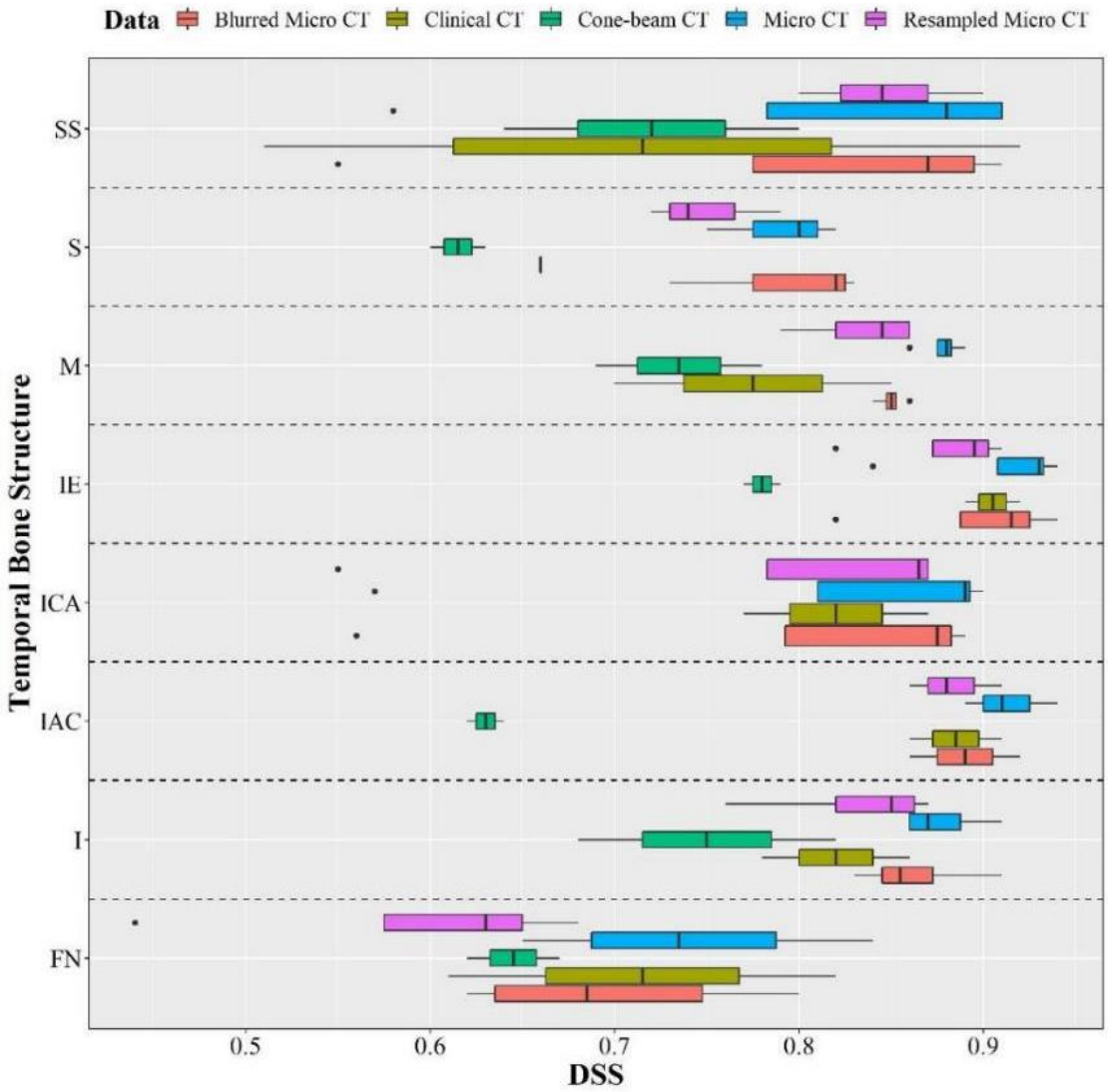


Fig. 7. DSS of the proposed segmentation algorithm for eight temporal bone structures, sigmoid sinus (SS), facial nerve (FN), inner ear (IE), malleus (M), incus (I), stapes (S), internal carotid artery (ICA) and internal auditory canal (IAC),versus various types of inference data including micro-CT, clinical CT and blurred micro-CT, resampled micro-CT and cone-beam CT.

The DSS metrics are on the order of 0.80 or greater for all structures except for facial nerve and stapes which are hardly visible on low resolution scans in Fig. 4.

We also evaluated the performance of our proposed PWD-3DNet with different image types. Figure 7 shows the average DSS of each temporal bone organ for various categories of inference volumes including micro-CT, clinical CT and blurred micro-CT and resampled micro-CT with an average volume size of 641×594×705 voxels and voxel size of 0.154×0.154×0.154 mm and cone-beam CT with an average volume size of 256×235×512 voxels and voxel size of 0.33×0.33×0.31 mm, which had different image acquisition parameters. As shown in this figure, the inner ear and facial nerve had the highest and lowest DSS values for all types of scans, respectively. Cone-beam CT had the lowest DSS metric since the imaging protocols were different from our training set. However, for all structures DSS values were above 0.62, that validates the increased capability of the proposed method to generalize itself against variations imaging parameters in multi-site data, which is one of the main disadvantages of CNN-based segmentations. Fig. 8 shows 3D models of predicted labels for four data types of one sample's CT scans compared to the manual segmentations. As shown, the micro-CT and blurred micro-CT had similar segmentation models. However, the prediction accuracy was decreased for clinical and resampled micro-CTs due to lower resolutions. However, acceptable predictions on segmenting the stapes and facial nerve were observed, which are difficult structures to segment for a human expert (Fig. 4 compares the visibility for each structure). Figure 9 shows one slice from CT scans of four subjects in our test set with their manual and predicted labels superimposed onto the image. As illustrated in this figure, for large structures like the sigmoid sinus, subject 3 had better overlap with the ground truth. Whereas, for small and hardly visible structures like the stapes and facial nerve, subjects 2 and 3 showed better matches, respectively. In general, the qualitative segmentation results in Fig. 8 and 9 confirm the good overall performance of the proposed method.

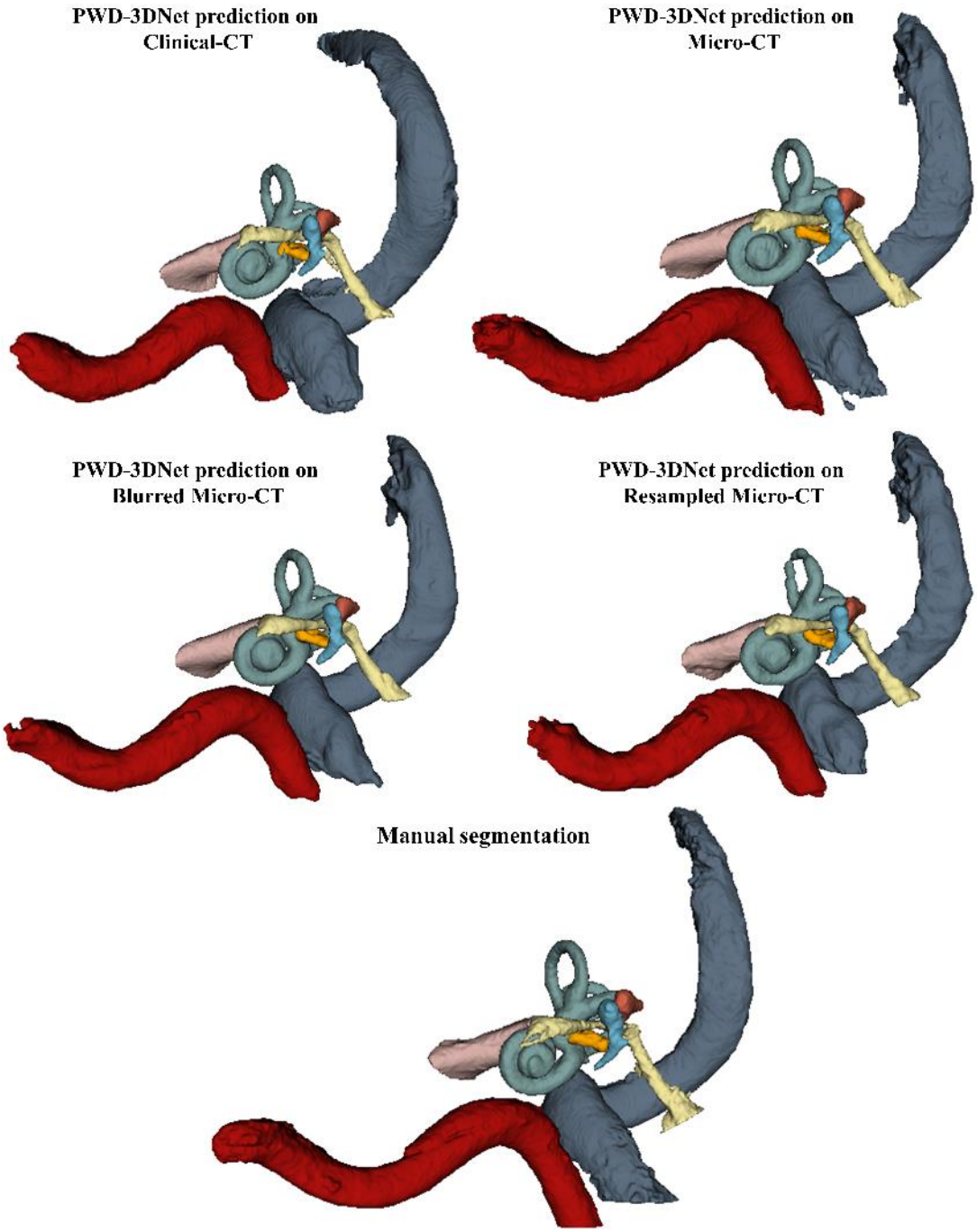


Fig. 8. 3D models of segmentations (predictions and manual) from four data types of one subject in our test set. Labels are sigmoid sinus (dark gray), facial nerve (yellow), inner ear (light gray), malleus (blue), incus (light red), stapes (orange), internal carotid artery (dark red) and internal auditory canal (pink).

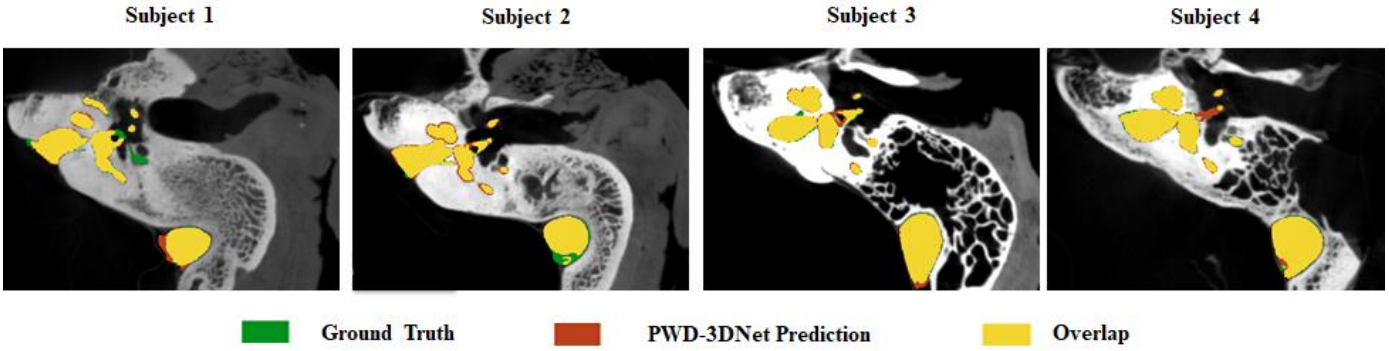


Fig. 9. Lateral view of one slice of a temporal bone CT scan with ground truth and predicted segmentation labels of the sigmoid sinus, facial nerve, inner ear, malleus, incus, stapes, internal carotid artery and internal auditory canal. Green, red and yellow represent the ground truth, predicted segmentation labels and the overlap area between the predicted labels and ground truth, respectively.

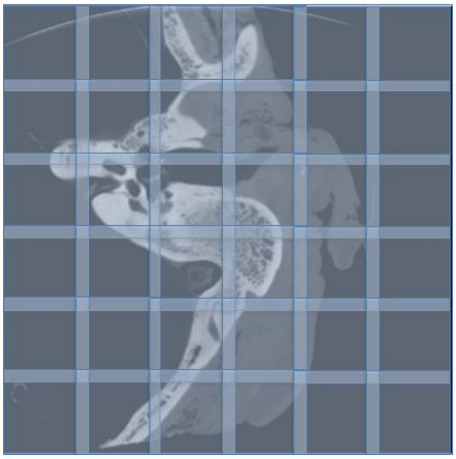

Fig. 10. Overlap of sliding windows in inference stage.

### C. Effect of overlapping in window sampling

In the inference stage, the image windows were extracted from every location on the input image, similar to the grid sampling in section IV-A-1, and their voxels were labeled based on the learned models during training. In this experiment, we intended to explore the influence of overlapping between adjacent windows by including some voxels as the patch borders as shown in Fig. 10. Figure 11 shows a line-plot of DSS values versus overlapping size (voxels) for different classes of labels (temporal bone structures). As illustrated in Fig. 11, non-overlapping blocks or a small border size reduces the accuracy of segmentation of some structures such as the stapes for a border size of 0, 6 and 12 voxels. This is because very small structures become undetectable at the border of neighboring subregions. As the size of the border increases, DSS value rises. However, a very large border size may cause redundant calculations, which increases the computational time substantially and does not improve predictions. IAC could not be predicted with very small overlapping sizes (the first four overlapping sizes of 0, 6, 12 and 18). Figure 12 shows the inference run-time (in logarithmic scale) versus overlap size (in voxels) for different labels. Since computation time is an important factor in terms of application feasibility, it was important to have a trade-off between the accuracy and run-time of the predictions. Moreover, as illustrated in Fig. 11, DSS does not change for overlap sizes larger than 36 voxels. Overlapping the extracted patches by 24 and 36 voxels resulted in the best performance at an acceptable inference speed. Although border size of 24 and 36 voxels showed the exact same accuracies, after validating the network performance in the following sections, the latter was selected as the inference gold standard for our proposed algorithm. The overlap size was the same value for all sub-volumes.

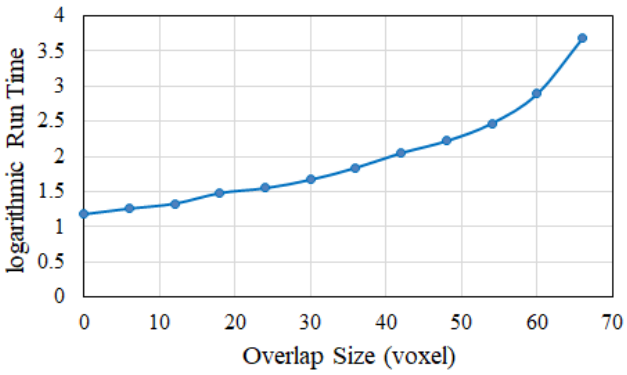


Fig. 12. Inference run-time of the proposed algorithm in predicting eight labels simultaneously on one CT volume versus sub-volume border size.

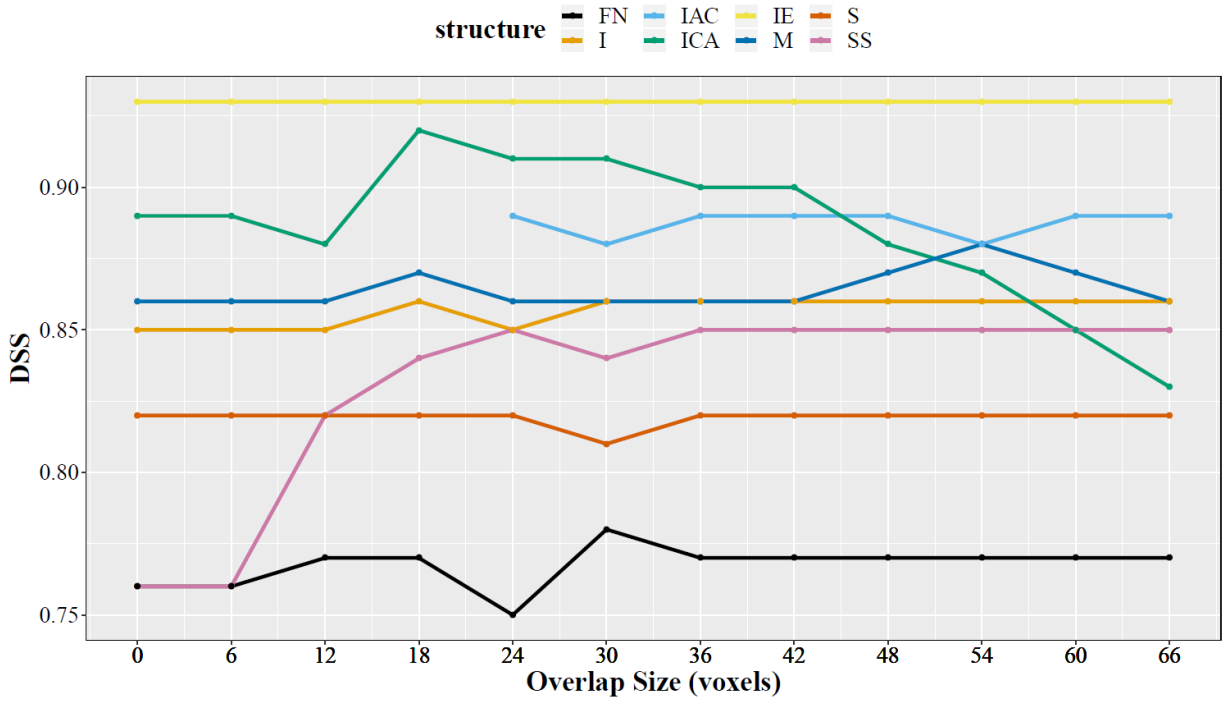


Fig. 11. DSS of the proposed algorithm for eight temporal bone structures versus inference sub-volume border size. Structures are sigmoid sinus (SS), facial nerve (FN), inner ear (IE), malleus (M), incus (I), stapes (S), internal carotid artery (ICA) and internal auditory canal (IAC).

### *D. Comparison to previously published algorithms*

In order to assess the effectiveness of the proposed segmentation algorithm, we compared the DSS of PWD-3DNet to other state-of-the-art techniques. The augmented samples were excluded from the training set to reduce the learning time. In this experiment, we also intended to investigate the effect of the proposed patch-wise strategy. For this reason, we implemented DenseVNet [45] and DeepMedic [33] as two successful segmentation methods using the Niftynet platform and DeepMedic open-source software (https://github.com/Kamnitsask/ deepmedic), respectively. We used the default network settings. All three networks were trained on 35 micro-CT scans with 31 training and four validation volumes (90 and 10 percent, respectively). Four micro-CT samples from four subjects were left out of the training dataset to be used to test the algorithm. DenseVNet adopted the full-size images as the input and resampled them to a smaller size of 144×144×144 voxels. DeepMedic [33] employed patch sampling with sub-volumes of size 25×25×25 voxels and uniform distribution. Table III shows the calculated DSSs and JSs between the predicted and the ground truth labels for each class of labels for our proposed network compared to DenseVNet [45] and DeepMedic [33]. Results showed that the proposed method is significantly better in segmenting multiple structures on temporal bone CTs. DeepMedic showed better performance than DenseVNet on larger structures whereas, for small organs, it did not perform well and the reason may be that uniform sampling in DeepMedic extracts equal number of image windows from random locations on the input foreground and background volumes. Therefore, there is an unbalanced distribution among the selected sub-volumes from regions of interest. This strategy increases the risk of ignoring some labels, specifically small regions of interest, in each training iteration and thus the network does not learn the characteristics of some regions sufficiently. Furthermore, DeepMedic is a multi-scale dual-pathway network where global information is preserved as well as local details from the same input at two resolutions. However, the dual-path training structure increases the learning-time unnecessarily. DenseVNet employs dense-training strategy which does not perform well for our dataset as described in the previous section (section IV.A). As shown in Table III, dense training works better on larger regions of interest but is not successful for segmentation of tiny structures (i.e. it could not predict the malleus, incus, and stapes labels) since the network might lose some important local information through image compression to a smaller size.

TABLE III
COMPARISON OF THE AVERAGE DICE SIMILARITY SCORE (DSS) AND JACCARD SCORE (JS) FOR PREDICTIONS OF EACH TEMPORAL BONE STRUCTURE WITH DIFFERENT ALGORITHMS (NON-AUGMENTED) ON 4 SUBJECTS' MICRO-CT VOLUMES. 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 |
|---|---|---|---|---|---|---|---|---|
| **DenseVNet [45]** | | | | | | | | |
| **DSS** | 0.58 | 0.35 | 0.58 | - | - | - | 0.71 | 0.75 |
| **JS** | 0.41 | 0.21 | 0.41 | - | - | - | 0.55 | 0.60 |
| **DeepMedic [33]** | | | | | | | | |
| **DSS** | 0.48 | 0.35 | 0.87 | 0.80 | 0.84 | 0.30 | 0.84 | 0.65 |
| **JS** | 0.32 | 0.21 | 0.77 | 0.67 | 0.72 | 0.18 | 0.72 | 0.48 |
| **PWD-3DNet** | | | | | | | | |
| **DSS** | **0.84** | **0.80** | **0.95** | **0.89** | **0.89** | **0.82** | **0.91** | **0.90** |
| **JS** | **0.72** | **0.67** | **0.90** | **0.80** | **0.80** | **0.69** | **0.83** | **0.82** |

Dashed line indicates there was no segmentation result.

### *E. Investigating the effect of augmentation layers*

Based on the experimental results in the previous section and the calculated DSS for the proposed PWD-3DNet in Table III, the algorithm is more accurate when using only micro-CT samples in the training set. However, the main reason for augmenting samples with lower resolutions and image artefacts in the training set was to increase the robustness of the models against variations in scan resolution and characteristics utilized at other imaging sites with different scanners. To assess the effectiveness of the proposed augmentations, we tested the models from both training settings (the models learned from 35 micro-CTs only and the models learned from 140 micro-CTs, clinical CTs and blurred and resampled micro-CTs). Table IV compares the segmentation results for the augmented (Aug) and non-augmented (NAug) networks by calculating the average DSS for different types of data in the test set (including 18 scans). As shown in Table IV, the DSS from augmented network was not improved in the networks trained on micro-CT scans.

TABLE IV
AVERAGE DSS AND STANDARD DEVIATION (STD) OF THE SEGMENTATION RESULTS FOR THE AUGMENTED (AUG) AND NON-AUGMENTED (NAUG) NETWORKS ON DIFFERENT DATA TYPES IN THE TEST SET AND THE P-VALUES OF THE STRUCTURES. TEMPORAL BONE 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 |
|---|---|---|---|---|---|---|---|---|
| **Micro CT** (DSS±STD) | | | | | | | | |
| Aug | 0.81±0.16 | 0.74±0.08 | 0.91±0.05 | 0.88±0.01 | 0.88±0.02 | 0.79±0.04 | 0.81±0.16 | 0.91±0.03 |
| NAug | **0.84±0.13** | **0.80±0.06** | **0.95±0.01** | **0.89±0.02** | **0.89±0.03** | **0.82±0.07** | **0.91±0.06** | **0.90±0.02** |
| **Clinical CT** (DSS±STD) | | | | | | | | |
| Aug | **0.72±0.29** | **0.72±0.15** | **0.90±0.02** | **0.78±0.11** | **0.82±0.06** | **0.66±0.01** | **0.82±0.07** | **0.88±0.04** |
| NAug | 0.64±0.23 | 0.62±0.14 | 0.77±0.06 | 0.69±0.1 | 0.68±0.04 | 0.60±0.01 | 0.64±0.08 | 0.75±0.08 |
| **Blurred Micro CT** (DSS±STD) | | | | | | | | |
| Aug | 0.80±0.17 | **0.70±0.08** | **0.90±0.05** | **0.85±0.01** | **0.86±0.03** | **0.79±0.06** | **0.80±0.16** | **0.89±0.03** |
| NAug | **0.81±0.16** | 0.38±0.09 | 0.88±0.09 | 0.81±0.02 | 0.85±0.05 | 0.47±0.12 | 0.60±0.14 | 0.81±0.03 |
| **Resampled Micro CT** (DSS±STD) | | | | | | | | |
| Aug | **0.84±0.04** | 0.60±0.11 | 0.88±0.04 | **0.84±0.03** | **0.83±0.05** | **0.75±0.04** | **0.78±0.16** | **0.88±0.03** |
| NAug | 0.80±0.04 | **0.67±0.07** | **0.89±0.02** | 0.78±0.04 | 0.82±0.07 | 0.63±0.06 | 0.62±0.17 | 0.79±0.07 |
| **P-Values** | | | | | | | | |
| | 0.041 | 0.119 | 0.016 | 0.0002 | 0.0735 | 0.007 | 1.2e-7 | 1.3e-4 |

However, the average DSS values were numerically increased after augmentation in clinical CT, blurred micro-CT and resampled micro-CT for majority of anatomical features. To confirm the increase statistically, for each anatomical feature, a paired *t*-test was used to test the null hypothesis that the DSS was increased of in samples from augmented network at 5% significance level. Paired *t*-test showed statistically significant increase of DSS in augmented network for 6 out of 8 anatomical features including, FN, IAC, ICA, IE, M and S.

## VI. Discussion and Summary

The current work presents the first fully-automated algorithm for multi- structure segmentation of temporal bone CT scans with accuracy and speed that surpasses current manual and automated segmentation techniques.

We proposed a densely-connected fully convolutional neural network with patch-wise balanced training, PWD-3DNet. The PWD-3DNet segmentation algorithm was evaluated in several assessments, including on CT scans with different voxel sizes, scanner settings and image artefacts, and compared to other state-of-the-art multi-organ segmentation techniques. PWD-3DNet outperformed the previously published methods likely because the proposed subsampling method increased the learning balance among different classes by using the discrete ground truth labels as sampler weights and therefore, including samples from all labels in each training iteration. The balanced sampling method was chosen for the present study as this method overcomes the class imbalance that is introduced in grid sampling, uniform and background/foreground methods, because the sampling is not random and the sampler weights are the discrete ground truth labels. In particular, for temporal bone segmentation with a combination of both small and large-scale key structures, this approach performed significantly better. Based on the experimental results, grid sampling, uniform or background/foreground methods led to poor segmentation results for very small regions in the temporal bone anatomy such as the stapes, malleus and incus because their size is a fraction of the full image size. By using balanced sampling, compared to uniform, background/foreground and grid sampling strategies, the learning algorithm is not deceived by local minima which prevents the algorithm from being biased towards the predictions of larger regions of interest and mislabeling smaller sized regions. Also, compared to grid sampling, the balanced strategy reduces the computational redundancy that occurs in grid sampling. For the above reasons, we believe balanced sampling is the most appropriate method for use in this study. In this work, we proposed augmentation layers in the training set by including clinical CT scans from the same subjects at a clinical resolution as well as resampled and blurred (smoothed) versions of micro-CT volumes. This was done to increase the robustness and accuracy of our trained models against various scanning resolutions and imaging artefacts that may be introduced by scanners used in other imaging centers, and to prepare the network for more difficult segmentation tasks. For example, some scanners are designed for scanning both soft tissue and bony structures. The scans produced by these devices are therefore blurry with a lower level of sharpness. Also, some image acquisition protocols acquire images with large voxel spacing to reduce the level of harmful radiations. The augmented layers used in the current study therefore increase the accuracy of segmenting temporal bone organs on volumes created from various resolutions and data acquisition parameters. Furthermore, this validates the efficacy of our proposed augmentation to increase the capability of our trained models to deal with challenges that may be encountered in cases of multi-site imaging acquisition protocols.

Experimental results revealed that PWD-3DNet had superior performance in predicting hardly visible structures on low resolution images which is useful in clinical applications. Also, it had an average DSS of 0.64 for predicting eight temporal bone labels on cone-beam CT scans another site with various parameters from the training set. This suggest that our algorithm could potentially segment images from other sites obtained using different scanners and acquisition parameters.

Furthermore, we showed that including borders in the sub-volumes extracted from the image in the inference stage increased the accuracy of segmentation. However, the particular number of overlapping voxels should be selected based on a trade-off between the accuracy and speed of predictions. The automated segmentation run-time of the proposed algorithm for an unseen micro-CT test sample with a volume size of 641×594×705 and border size of 36×36×36 is 65.12 secs, which is significantly faster than any manual or semi-automated temporal bone segmentation pipeline. The average accuracy (DSS value) for segmentation of eight structures of a temporal bone anatomy was 0.86.

In conclusion, the Dice similarity scores of the proposed algorithm for eight critical structures of the temporal bone, (sigmoid sinus, facial nerve, inner ear, malleus, incus, stapes, internal carotid artery and internal auditory canal), is higher than previously used segmentation methods. The superiority of speed and accuracy creates the opportunity of adopting the proposed PWD-3DNet for automated segmentations in patient-specific surgical planning and rehearsal. The adopted dataset in the current study had some limitations and the pathological data and diseased temporal bones have not been studied. Future work will focused on segmentation of additional structures on temporal CTs and validating the algorithm by augmenting pediatric CT scans and samples from patients with cochlear implants and acoustic neuroma diseases. In addition, deep learning-based segmentation techniques have some limitations in generalization. The required data to study the performance of the proposed algorithm on a variety of imaging protocols was not available. multi-site data will be added to the training set which may lead to enhanced segmentation predictions when the imaging protocol varies. More importantly, we will propose a loss function by considering some initial information regarding the size and distances of multiple labels to improve the segmentation balance among small and large structures. Furthermore, by resampling the ground truth from smaller voxel sizes and validating the learning performance against them, we can improve the segmentation accuracy and create more detailed predictions from multiple structures.

## ACKNOWLEDGEMENT

Funding for this work was provided jointly by the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council's Collaborative Health Research Project grant program. We gratefully acknowledge Lauren H. Siegel, with the Department of Otolaryngology-Head and Neck Surgery, Western University, for help with the proofreading.

## REFERENCES

[1] J. D. Durrant, and J. H. Lovrinic, "Anatomy of the ear," in *Bases of Hearing Science*, 3rd ed. Maryland: Williams & Wilkins, 1995, ch. 4, sec. 2, pp. 102–137.

[2] K. Van Osch et al., "Morphological analysis of sigmoid sinus anatomy: clinical applications to neurotological surgery," *J Otolaryngol Head Neck Surg*., vol. 48, no. 1, pp. 2-8, Jan. 2019.

[3] T. Liang, "Atlas-based segmentation of temporal bone anatomy," M.Sc. thesis, Dept. Electrical and Computer Engineering., The Ohio State Univ., Columbus, Ohio, 2017.

[4] S. Chan et al., "High-fidelity haptic and visual rendering for patient-specific simulation of temporal bone surgery," *Comput. Assist. Surg*., vol. 21, no. 1, pp. 85-101, 2016.

[5] P. Lu et al., "Facial nerve image enhancement from CBCT using supervised learning technique," in *Proc IEEE Eng. Med. Biol. Soc*., 2015, pp. 2964-2967.

[6] P. Lu et al., "Highly accurate facial nerve segmentation refinement from CBCT/CT imaging using a super-resolution classification approach," *IEEE Trans. Biomed. Eng*., vol. 65, no. 1, pp. 178-188, 2018.

[7] K. Hassan et al., "Evaluation of software tools for segmentation of temporal bone anatomy," *Stud. Health Technol. Inform*., vol. 220, pp. 103-133, 2016.

[8] J. Minnema et al., "CT image segmentation of bone for medical additive manufacturing using a convolutional neural network," *Comput. Biol. Med.*, vol. 103, pp. 130-139, Dec. 2018.

[9] K. Souadih et al., "Automatic forensic identification using 3D sphenoid sinus segmentation and deep characterization", *Med. & Biol. Eng. & Comput*., vol. 58, no. 2, pp. 291-306, Dec. 2019.

[10] G. Luo et al., "Multi-views fusion CNN for left ventricular volumes estimation on cardiac MR images," *IEEE Trans. Biomed. Eng*., vol. 65, no. 9, pp. 1924-1934, Sep. 2018.

[11] G. Yang et al., "Simultaneous left atrium anatomy and scar segmentations via deep learning in multiview information with attention," *Future Generation Computer Systems*, vol. 107, pp. 215-228, May. 2020.

[12] L. Li et al., "Atrial scar quantification via multi-scale CNN in the graph-cuts framework," *Medical Image Analysis*, vol. 60, pp. 101595, Feb. 2020.

[13] G. Yang et al., "Fully automatic segmentation and objective assessment of atrial scars for long-standing persistent atrial fibrillation patients using late gadolinium-enhanced MRI," *Med. Phys.*, vol. 45, pp. 1562-1576, Apr. 2018.

[14] G. Yang et al., "Multiview sequential learning and dilated residual learning for a fully automatic delineation of the left atrium and pulmonary veins from late gadolinium-enhanced cardiac MRI images," in *Proc. EMBC*, 2018, pp. 1123-1127.

[15] Y. Mo et al., "The deep poincaré map: a novel approach for left ventricle segmentation," in *Proc. MICCAI*, 2018, pp. 561-568.

[16] L. Zhang et al., "Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons," *Journal of Medical Imaging*, vol. 6, pp. 024001-1-12, Apr. 2019.

[17] A. Ali et al., "A deep learning based approach to skin lesion border extraction with a novel edge detector in dermoscopy images," in *Proc. IJCNN*, 2019.

[18] J. Liu et al., "A cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas," *IEEE Trans. Biomed. Eng*., vol. 65, no. 9, pp. 1943-1952, Sep. 2018.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist Interv.*, 2015, pp. 234-241.

[20] J. Fauser et al., "Toward an automatic preoperative pipeline for image-guided temporal bone surgery," *Int. J. Comput. Assist Radiol. Surg*., vol. 14, pp. 967-976, Mar. 2019.

[21] H. Chen et al., "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446-455, Apr. 2018.

[22] N. Gerber et al., "A multiscale imaging and modelling dataset of the human inner ear," *Sci. Data*, vol. 4:170132, 2017.

[23] X. Li et al., "A 3D deep supervised densely network for small organs of human temporal bone segmentation in CT images," *Neural Networks*, vol. 124, pp. 75–85, Apr. 2020.

[24] M. Seemann et al, "Evaluation of the middle and inner ear structures: comparison of hybrid rendering, virtual endoscopy and axial 2D source images," *Eur. Radiol*., vol. 9, pp. 1851-1858, 1999.

[25] K. A. Powell et al., "Atlas-based segmentation of temporal bone surface structures," *Int. J. Comput. Assist Radiol. Surg*., vol. 14, no. 8, pp. 1267-1273, 2019.

[26] T. Okada et al., "Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors," *Med. Image Anal*., vol. 26, no. 1, pp. 1-18, 2015.

[27] T. Tong et al., "Discriminative dictionary learning for abdominal multi organ segmentation," *Med. Image Anal*., vol. 23, no. 1, pp. 92-104, 2015.

[28] Z. Xu et al., "Evaluation of six registration methods for the human abdomen on clinically acquired CT," *IEEE Trans. Biomed. Eng*., vol. 63, no. 8, pp. 1563–1572, Aug. 2016.

[29] J. H. Noble et al., "Automatic segmentation of the facial nerve and chorda tympani in CT images using spatially dependent feature values," *Med. Phys*., vol. 35, no. 12, pp. 5375-5384, 2008.

[30] J. H. Noble et al., "Automatic identification and 3D rendering of temporal bone anatomy," *Otol. Neurotol*., vol. 30, no. 4, pp. 436-442, 2009.

[31] F. A. Reda et al., "Model-based segmentation of the facial nerve and chorda tympani in pediatric CT scans," in *Proc. Medical Imaging 2011: Image Processing*, vol. 7962, 2011.

[32] A. Prasoon et al., "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," Lecture Notes in Computer Science, pp. 246-253, 2013.

[33] K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal*., vol. 36, pp. 61-78, Oct. 2017.

[34] W. Li et al., "On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task," in *Proc. Inf. Process. Med. Imaging*, 2017, pp. 348-360.

[35] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis*., 2016, pp. 565-571.

[36] Y. Liu et al., "Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention," *IEEE Access*, vol. 7, pp. 163626 - 163632, Nov. 2019.

[37] X. Zhuang et al., "Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge," *Medical Image Analysis*, vol. 58, pp. 101537, Dec. 2019.

[38] Z. Shi et al., "Bayesian VoxDRN: a probabilistic deep Voxelwise dilated residual network for whole heart segmentation from 3D MR images," in *Proc. MICCAI*, 2018, pp. 569-577.

[39] M. Li et al., "MV-RAN: Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis," *Comput. Biol. Med.*, vol. 120, pp. 103728, May. 2020.

[40] A. Boutillon et al., "Multi-structure bone segmentation in pediatric MR images with combined regularization from shape priors and adversarial network," *Computer Science, Engineering*, Sep. 2020.

[41] O. Charron et al., "Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network," *Comput. Biol. Med*., vol. 95, pp. 43-54, Apr. 2018.

[42] H. Dong et al., "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Proc. MIUA*, 2017, pp. 506-517.

[43] S. R. Hashemi et al., "Asymmetric loss functions and deep densely connected networks for highly imbalanced medical image segmentation: application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721-1735, Dec. 2018.

[44] J. J. Cerrolaza et al., "Computational Anatomy for Multi-Organ Analysis in Medical Imaging: A Review," *Medical Image Analysis*, v. 56, pp. 44-67, Aug. 2019.

[45] E. Gibson et al., "Automatic multi-organ segmentation on abdominal CT with dense V-Networks," *IEEE Trans. Med. Imaging*, vol. 37, no. 8, pp. 1822-1834, Aug. 2018.

[46] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993-2024, Oct. 2015.
[47] K. Kamnitsas et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. International Conference on IPMI*, 2017, pp. 597–609.
[48] E. Gibson et al., "NiftyNet: a deep-learning platform for medical imaging," *Comput. Methods Programs Biomed*., vol. 158, pp. 113-122, May. 2018.
[49] C. B. Rossello, "Brain lesion segmentation using convolutional neuronal networks," B.Sc. thesis, Telecom BCN, Polytechnic University of Catalonia, Catalonia, Spain, 2018.
[50] A. Fedorov et al., "3D Slicer as an Image Computing Platform for the Quantitative Imaging Network," *Magn. Reson. Imaging*, vol. 30, no. 9, pp. 1323-1341, Nov. 2012.
[51] G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700-4708.