공학석사 학위논문

MP3DU: Multi-Projection 3D U-Net for Automatic Segmentation of Temporal Bone Structures in CT images

MP3DU: CT 영상 내 자동적 측두골 구조물
영상분할을 위한 다중 투영 3차원 U-Net

2022년 08월

서울대학교 대학원
협동과정 바이오엔지니어링 전공
전 보 성

MP3DU: Multi-Projection 3D U-Net for Automatic Segmentation of Temporal Bone Structures in CT images

MP3DU: CT 영상 내 자동적 측두골 구조물 영상분할을 위한 다중 투영 3차원 U-Net

지도교수 이 원 진

이 논문을 공학석사 학위논문으로 제출함

2022년 08월

서울대학교 대학원
협동과정 바이오엔지니어링 전공
전 보 성

전보성의 공학석사 학위논문을 인준함

2022년 07월

위 원 장 허 민 석 (인)

부 위 원 장 이 원 진 (인)

위 원 허 경 회 (인)

Master Dissertation

MP3DU: Multi-Projection 3D U-Net for Automatic Segmentation of Temporal Bone Structures in CT images

August 2022

Interdisciplinary Program in Bioengineering
College of Engineering
Seoul National University

Bo Soung Jeoun

MP3DU: Multi-Projection 3D U-Net for Automatic Segmentation of Temporal Bone Structures in CT images

Academic advisor **Won-Jin Yi**

Submitting a Master Dissertation

August 2022

Interdisciplinary Program in Bioengineering

College of Engineering

Seoul National University

Bo Soung Jeoun

Confirming the Master Dissertation written by Bo Soung Jeoun

July 2022

Chair Min Suk, Heo, D.D.S., Ph.D. (Seal)

Vice Chair Won Jin, Yi, Ph.D. (Seal)

Examiner Kyung Hoe, Huh, Ph.D. (Seal)

Abstract

MP3DU: Multi-Projection 3D U-Net for Automatic Segmentation of Temporal Bone Structures in CT images

Bo Soung Jeoun
Interdisciplinary Program in Bioengineering
College of Engineering
Seoul National University

Background: The inner ear surgery such as cochlea implantation and tumor removal requires accurate identification and comprehension of temporal bone structures to make appropriate preoperative planning. However, it is considered to be challenging locate and understand the critical temporal bone structures, facial nerve, cochlea, and ossicle, due to their small sizes and anatomical variations. In addition, the low contrast of temporal bone computed tomography (CT) causes blurry boundaries of anatomical structures so it causes confusion to distinguish anatomical structures.

Though, it is required to the otologists to acquire segmentation of temporal bone structures manually. Therefore, a multi-projection 3-dimensional (3D) U-Net (MP3DU) was proposed for automatic segmentation of temporal bone structures in CT images.

Materials and Methods: In this study, 381 temporal bone CT of normal condition were collected from the 418 patients who were diagnosed inner ear diseases. The MP3DU was designed based on 3D U-Net that has 3D encoder-decoder architecture with multi-projection maps generated from 3D volume input. The 3D contextual information and structural shape information simultaneously complement and optimize the segmentation performance during training in end-to-end manner. The multi-projection maps of MP3DU minimizes the feature loss while passing through 3D encoder-decoder architecture.

Result: The MP3DU achieved 0.81 dice similarity coefficient score (DSC), 0.71 jaccard index (JI), 0.81 precision (PR), and 0.84 recall (RC) in 2-dimensional (2D) performance metrics, and 0.34 relative volume difference (RVD), and 0.43 volume of error (VOE) in 3D performance metrics for the whole temporal bone structures which outperformed than other popular deep learning networks. Also, fewer false positives and negatives were observed from segmentation results than in other networks. In particular, a tubular structure, facial nerve, had improved segmentation results maintaining its anatomical shape well and achieving the highest evaluation metric of all others.

Conclusion: The proposed network, MP3DU, could provide the automatic segmentation of temporal bone structures by improving the structural shape and 3D contextual information through multi-projection maps with 3D encoder-decoder architecture.

Keyword: 3D Segmentation of the temporal bone structures, CT image, Deep Learning Network, Multi-projection

Student Number: 2020-24482

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CT	Computed tomography
3D	3-dimensional
2D	2-dimensional
MP3DU	Multi-projection 3D U-Net
DSC	Dice similarity coefficient score
JI	Jaccard index
PR	Precision
RC	Recall
VOE	Volume of error
RVD	Relative volume difference
CNN	Convolutional neural network
3DU	3D U-Net
2DU	2D U-Net

Introduction

Inner ear surgery is an otologic procedure that requires extensive knowledge of radiology and surgical anatomy for patients' safety [1]. Thus, the accurate identification of critical structures and appropriate comprehension of their complexity are essential for pre- and intra-operative planning of inner ear surgery [2]. Specifically, cochlea implantation, the most commonly practiced otological procedure, is highly influenced by anatomical variability so it is important to understand the orientation and geometry of structures [1, 3]. The temporal bone computed tomography (CT) is largely used for diagnosis and surgical plan for inner ear surgery to discern temporal bone structures since it provides otologists crucial insights into inherent anatomical information [1, 4, 5]. However, it is challenging to precisely distinguish the interested temporal bone structures such as facial nerve, cochlea, and ossicle due to their small sizes and pathologic variations as well as the

difficulties derived from multiple unrelated structures like air cells [1, 4]. Especially, facial nerve, which is a tubular structure that travels from stylomastoid foramen to internal auditory meatus, is more critical to locate accurately due to the risk of temporal or permanent facial paralysis during procedures [6-8]. Yet, it is considered to be demanding since it has high topological differences and low visibility with unclear boundaries caused by the lack of contrast from CT [6, 9] (**Figure 1**).

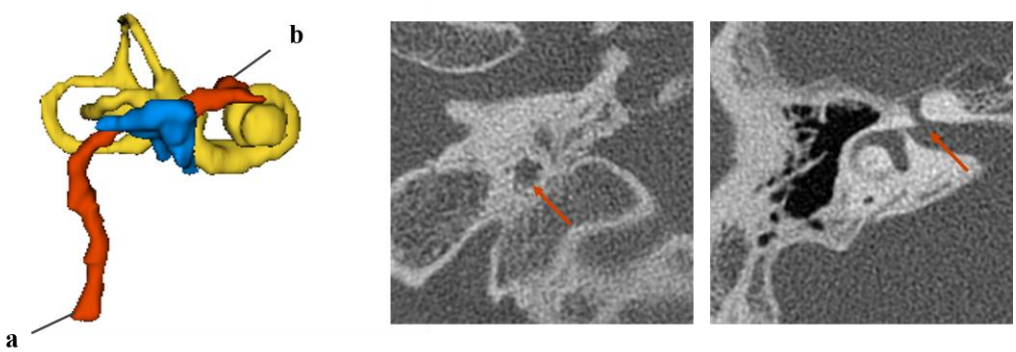

Figure 1. The example of critical temporal bone structures. The facial nerve, cochl ea, and ossicle are visualized in red, yellow, and blue, respectively. The (a), and (b) denotes stylomastoid foramen and internal auditory meatus for facial nerve. The co rresponding temporal bone computed tomographic image slices for (a) and (b) are displayed with red arrows indicating facial nerve, from left to right.

Furthermore, CT needs to be reviewed routinely by the experts through 2-dimensional (2D) image slices even if it is composed of 3-dimensional (3D) volume [1, 4, 10]. This process requires mental compilation to translate and assemble the information acquired from 2D image slices into 3D information [1, 4]. Thus, commercially available volume rendering techniques have been used for 3D reconstruction [4, 10, 11]. However, unlike large anatomical structures such as lungs and large vessels, these techniques are not likely to render accurate results for small structures for otological applications. In addition, it may generate artifacts around the inner ear boundaries due to the large variations of intensity values in CT [12]. In the end, the segmentation of temporal bone structures necessitates the interactions and manual efforts from the experts to obtain reliable segmentation results, which is labor-intensive, tedious, and time-consuming [4, 10, 11, 13, 14].

In early works, atlas-based approaches and other customized solutions were studied for automatic segmentation of temporal bone structures [9, 15-18]. However, these methods not only require a large amount of data but are also highly limited to averaged shape model from the collected dataset so the segmentation performance was likely to fail if the input image diverged from the atlas [4, 19]. Recently, deep learning has been widely applied for automatic medical image segmentation particularly using convolutional neural networks (CNNs) and achieved superior performance to traditional approaches [20-26]. Yet, the existing deep learning networks are mostly used to segment solid organs with relatively clear boundaries, such as kidneys and livers [27-29]. Especially, 3D CNNs are

rarely used for the segmentation of small, complicated, and tubular anatomical structures such as temporal bone structures due to the possibility of the feature disappearance during training [14]. Still, 3D CNNs are desirable in the medical imaging field since these take spatial context information from the volume into account for volumetric segmentation [30].

In this study, a multi-projection 3D U-Net (MP3DU) was proposed for automatic segmentation of temporal bone in CT images. It was hypothesized that the 3D encoder-decoder architecture learns volumetric contextual information and the multi-projection maps of each anatomical structure compensate for the feature loss that may occur through 3D encoder-decoder architecture while training by providing structural information. In particular, the MP3DU was designed to overcome the low visibility and high topological variations, and yield more accurate segmentation results with a tubular structure such as facial nerve.

Materials and Methods

Participants and Data acquisition

The patients of 381 (221 females and 170 males; mean age 50.93 ± 15.24 years) who were diagnosed sudden sensory neural hearing loss or otitis media at the Gachon University Gil Hospital (2012-2018). The obtained CT was separated into left and right and the sides that showed normal condition were collected and the rest which had diseases and inflammations were excluded, therefore, total 418 CT were acquired. The patient data were obtained at 120 kVp and 180 mAs using CT (SOMATOM Definition; Siemens Healthcare, Munich, Germany). The CT images had dimensions of 512×512×z pixels, which z were varied from 60 to 96, voxel sizes ranging from 0.13×0.13×0.6 to 0.16×0.16×0.6 mm^3, and 16-bit depth. This study was performed with approval from the institutional review board of the Gachon University Gil Hospital (GCIRB2020-339) and in accordance with the Declaration of Helsinki.

Data preparation

The temporal bone structures, facial nerve, cochlea, ossicle, were manually annotated by four otologists using a software (AVIEW KOREA for Windows 10; Coreline, Seoul, Korea). We used the cropped images consisting of 48 slices of 256×256 pixels that were centered at whole regions containing temporal bone structures in order to reduce the memory requirement. Zero-padding was performed to maintain the input volume of the same length for all patients showing different lengths of anatomical structures (**Figure 2**).

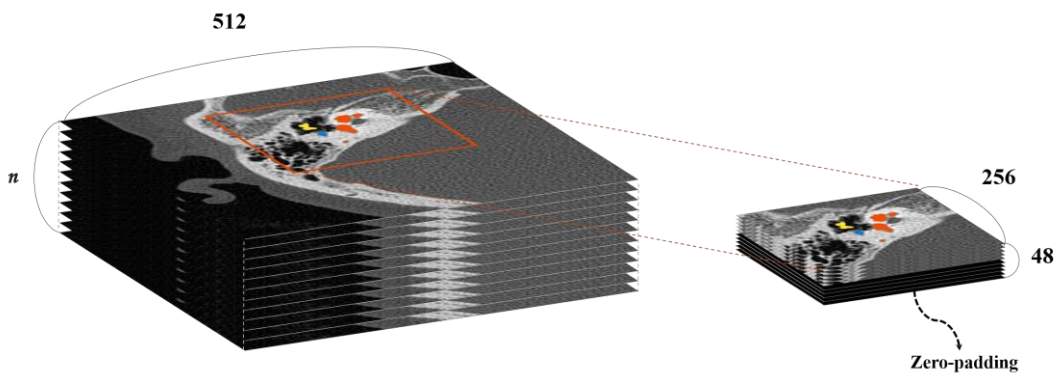

Figure 2. The example of data preprocess for multi-projection 3D U-Net (MP3DU). The original images of 512×512×z were cropped into 256×256×48 for memory requirement. Zero-padding was done to provide the same length of volume to the network. *n* denotes the number of z which varies from patient to patients.

We estimated the minimally required sample size to detect significant differences in the accuracy between the MP3DU and the other networks, when both assessed the same subjects in CT images. We designed to capture a mean accuracy-difference and a standard deviation of 0.05 and 0.10 between the MP3DU and the other networks. Based on an effect size of 0.5, a significance level of 0.05, and a statistical power of 0.80, we obtained a sample size of N = 128 (G* Power for Windows 10, Version 3.1.9.7; Universität Düsseldorf, Düsseldorf, Germany) [31, 32]. Eventually, we split the CT dataset with ratio of 0.8, and 0.2, thus 331 volumes (13086 slices) and 87 volumes (3373 slices) were selected randomly, and used for training, and test dataset.

MP3DU

The MP3DU was designed based on 3D encoder-decoder architecture with multi-projection maps from each anatomical structure (**Figure 3**). While training, the MP3DU generated the 2D multi-projection images of each facial nerve, cochlea, and ossicle from the input volume of temporal bone structures directly, which the network was able to learn the overall volume and structural shapes of each structure simultaneously (Multi-projection outputs and the output volume in **Figure 3**). It was optimized in an end-to-end where the segmentation outputs of temporal bone structures were generated directly from the input volumes of the CT images.

Through the encoder, MP3DU learned to capture contextual information from input volume and enable precise location. The input volume underwent the convolutional blocks of each stage, which were comprised of two repeated modules of two 3×3×3 3D convolutions, batch normalization, ReLU, and 2×2×2 max-pooling sequentially. The first initial convolutional block produced 256 feature maps as output, and the number of feature maps gradually decreased from 256 to 128, 64, and 32. The features from at the end of the encoder passed the corresponding up-sampling layer and fed to the convolutional blocks in decoder path, which were composed of two repeated modules of two 3×3×3 3D convolutions, batch normalization, ReLU, and 2×2×2 up-sampling. The number of feature maps gradually increased from 32 to 64, 128, and 256 after each stage of decoder path. The encoder and decoder were connected by skip connections to maintain the features.

The 3D volume loss and multi-projection map losses from the 2D projections simultaneously encouraged the network to learn the structural information of the temporal bone structures, especially the tubular features of facial nerve. The multi-projection map losses were calculated by the 2D projection map generated from axial plane of each anatomical structure (**Figure 4**). The dice similarity coefficient score (DSC) was used for the two loss functions. The loss function ($L_{total} = DL_{vol} + DL_{mp}$) of the MP3DU consisted of 3D volume loss (DL_{vol}) for the entire canal volume, and the multi-projection map losses as sum of the 2D projection losses from each facial nerve (DL_{fn}), cochlea (DL_{cc}), and ossicle (DL_{os}), respectively (L_{total} in **Figure 3**).

The proposed networks were trained using an Adam optimizer, and the learning rate of 0.00025 was reduced on plateau by a factor of 0.5 every 25 epochs in 300 epochs with the batch size of 1. They were implemented with Python3 based on Keras with a Tensorflow backend using a single NVIDIA Titan RTX GPU 24G.

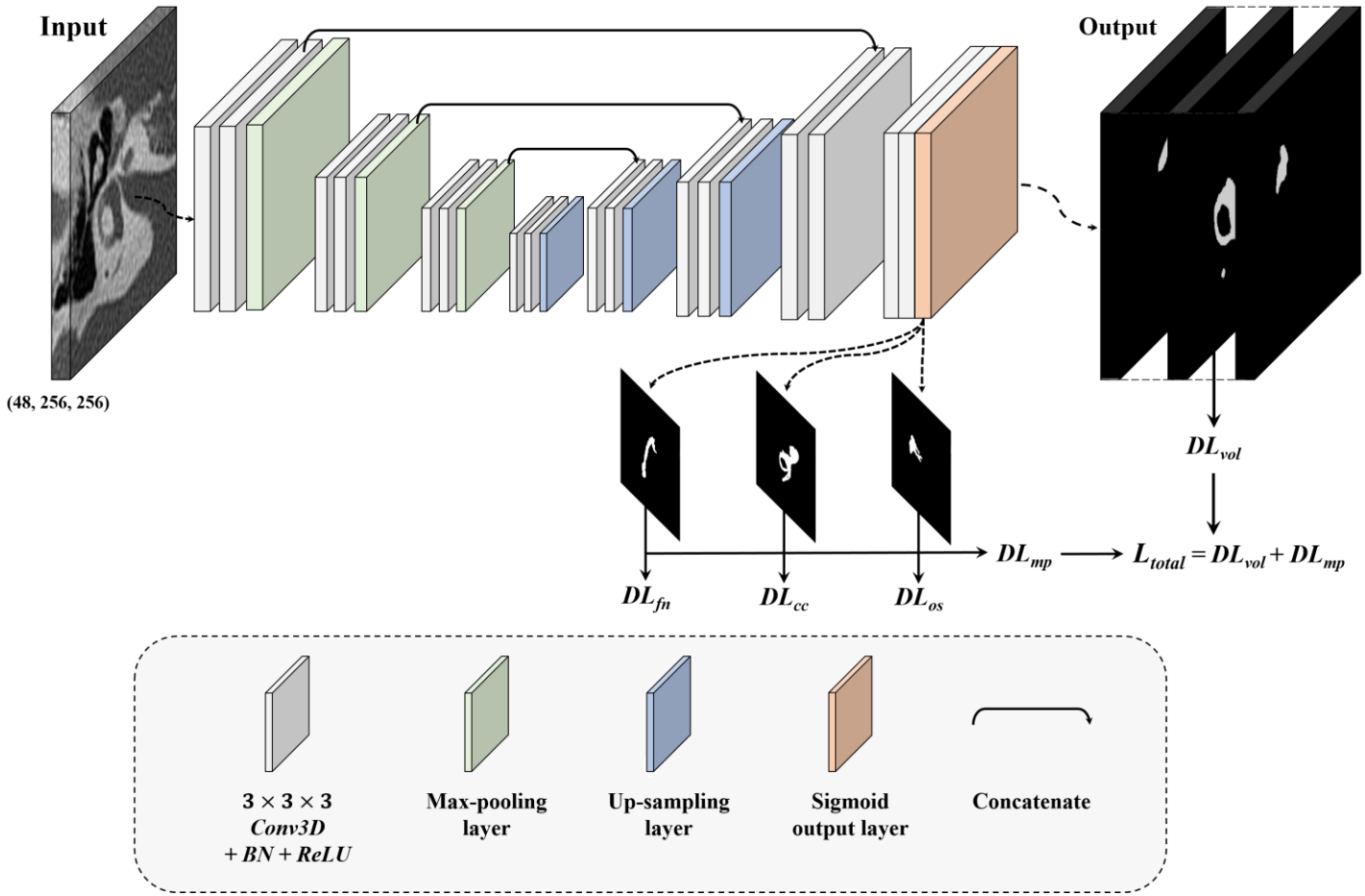

Figure 3. The multi-projection 3D U-Net (MP3DU) architecture with a 3D encoder-decoder with multi-projection maps. The 3D encoder-decoder was utilized to capture spatial information of anatomical structures, and multi-projection maps were used to learn overall shape of each anatomical structure.

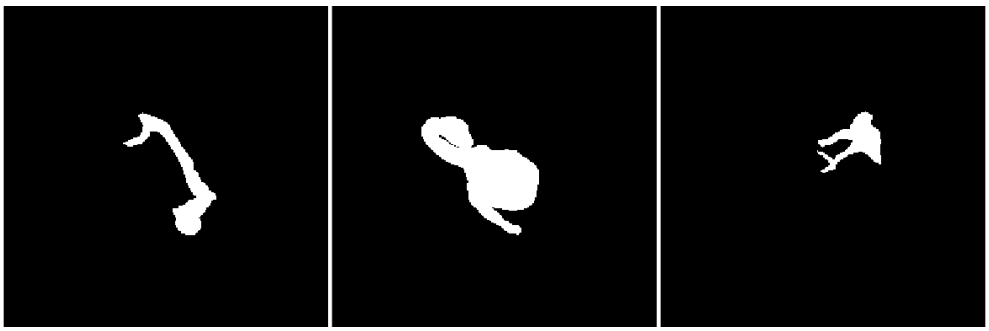

Figure 4. The example of multi-projection maps of each anatomical structures generated from 3D input volume during training of multi-projection 3D U-Net (MP3DU). The example of projection map from facial nerve, cochlea, and ossicle are visualized from left to right.

Performance evaluation

The segmentation performance of temporal bone structures by MP3DU was compared with those by other networks of 2D U-Net [33], EfficientNet [34], 3D U-Net [35]. To evaluate the performances quantitatively, the networks were compared with the 2D segmentation performance metrics of DSC ($DSC = \frac{2TP}{2TP+FN+FP}$), Jaccard index ($JI = \frac{TP}{TP+FN+FP}$), precision ($PR = \frac{TP}{TP+FP}$), recall ($RC = \frac{TP}{TP+FN}$) among networks, where TP, FP, and FN denoted true positives, false positives, and false negatives, and also 3D volumetric performance metrics of volume of error ($VOE = 1 - \frac{V_{gt} \cap V_{pred}}{V_{gt} \cup V_{pred}}$), and relative volume difference ($RVD = \frac{|V_{gt} - V_{pred}|}{V_{gt}}$), where V_{gt} and V_{pred} represented the number of voxels for the ground truth and for the predicted volume, respectively. The higher values of DSC, JI, PR, and RC, and the lower values of VOE, and RVD indicated better segmentation performance. The paired two-tailed t-tests (SPSS Statistics for Windows 10, Version 26.0; IBM, Armonk, New York, USA) was used to compare performances between MP3DU and others. The statistical significance level was set at 0.05.

Results

The performances of MP3DU, 3D U-Net, EfficientNet, and 2D U-Net were evaluated for a total of 87 temporal bone structures not used for training. **Table 1** shows the quantitative results of the segmentation performance for the whole temporal bone structures by each network. The MP3DU achieved the highest values of 0.81 DSC, 0.71 JI, 0.81 PR, and 0.84 RC in 2D performance metrics, and also the lowest values of 0.34 RVD, and 0.43 VOE in 3D performance metrics for the whole temporal bone structures (**Table 1**). In addition, the quantitative evaluation for each anatomical structure, facial nerve, cochlea, and ossicle, were performed respectively as well as the whole structures and the results showed better values for both 2D and 3D evaluation metrics (**Table 1**). Especially, the results for facial nerve segmentation of MP3DU outperformed other networks achieving the highest values of 0.75 DSC, 0.63 JI, 0.73 PR, and 0.79 RC, and the lowest values of 0.60 RVD, and 0.40 VOE

(**Table 1**). Therefore, MP3DU showed better results than all the other networks in DSC, JI, PR, RC, RVD, and VOE (**Table 1**).

The performance of the networks for whole temporal bone structures is also plotted in boxplots in **Figure 5**. Though the data dispersion and whisker length of MP3DU seemed similar to other networks, it had a relatively small number of outliers and better median scores (**Figure 5**). Specifically, the boxplots of facial nerve segmentation results were plotted in **Figure 6**. As the **Figure 5**, even though MP3DU represented similar data dispersion and whisker length to other networks, it had the best median score among others (**Figure 6**).

		DSC	JI	PR	RC	RVD	VOE
Whole structures	MP3DU	0.81 ± 0.13	0.71 ± 0.12	0.81 ± 0.13	0.84 ± 0.13	0.34 ± 0.06	0.43 ± 0.03
	3D U-Net	0.79 ± 0.05*	0.68 ± 0.05*	0.81 ± 0.05	0.81 ± 0.08*	0.36 ± 0.07*	0.44 ± 0.03*
	EfficientNet	0.80 ± 0.04†	0.69 ± 0.05†	0.83 ± 0.04†	0.81 ± 0.06†	0.37 ± 0.09†	0.44 ± 0.03†
	2D U-Net	0.79 ± 0.13‡	0.69 ± 0.12‡	0.81 ± 0.13‡	0.82 ± 0.13‡	0.36 ± 0.05‡	0.44 ± 0.03‡
Facial nerve	MP3DU	0.75 ± 0.14	0.63 ± 0.15	0.73 ± 0.18	0.79 ± 0.17	0.60 ± 0.10	0.40 ± 0.15
	3D U-Net	0.73 ± 0.10*	0.61 ± 0.10*	0.76 ± 0.11*	0.76 ± 0.13*	0.61 ± 0.08*	0.45 ± 0.18*
	EfficientNet	0.73 ± 0.09†	0.60 ± 0.10†	0.76 ± 0.12†	0.78 ± 0.11†	0.61 ± 0.08†	0.44 ± 0.24†
	2D U-Net	0.71 ± 0.14‡	0.60 ± 0.14‡	0.72 ± 0.18‡	0.75 ± 0.17‡	0.61 ± 0.10‡	0.44 ± 0.14‡
Cochlea	MP3DU	0.84 ± 0.15	0.74 ± 0.14	0.84 ± 0.15	0.85 ± 0.15	0.46 ± 0.06	0.40 ± 0.07
	3D U-Net	0.82 ± 0.04*	0.71 ± 0.05*	0.84 ± 0.07	0.84 ± 0.09*	0.47 ± 0.04*	0.40 ± 0.07
	EfficientNet	0.83 ± 0.03†	0.72 ± 0.04†	0.84 ± 0.05	0.80 ± 0.06†	0.47 ± 0.05†	0.43 ± 0.06†
	2D U-Net	0.83 ± 0.15‡	0.73 ± 0.14‡	0.84 ± 0.16	0.83 ± 0.15‡	0.47 ± 0.07‡	0.41 ± 0.07‡
Ossicle	MP3DU	0.83 ± 0.16	0.73 ± 0.15	0.83 ± 0.15	0.84 ± 0.16	0.22 ± 0.03	0.22 ± 0.03
	3D U-Net	0.82 ± 0.10*	0.72 ± 0.10*	0.85 ± 0.06*	0.83 ± 0.11*	0.23 ± 0.02*	0.22 ± 0.02
	EfficientNet	0.82 ± 0.08†	0.72 ± 0.09†	0.85 ± 0.07†	0.83 ± 0.10†	0.23 ± 0.03†	0.23 ± 0.02†
	2D U-Net	0.81 ± 0.18‡	0.72 ± 0.16‡	0.82 ± 0.18‡	0.82 ± 0.18‡	0.23 ± 0.05‡	0.22 ± 0.04

Table 1. The evaluation metric calculation results for whole temporal bone structures and each anatomical structure from multi-projection 3D U-Net (MP3DU), 3D U-Net, EfficientNet, and 2D U-Net. Mean (SD) Dice simila rity coefficient score (DSC), Jaccard index (JI), precision (PR), recall (R C), volume of error (VOE), and relative volume difference (RVD) are de monstrated in each column (*: significant difference between MP3DU and 3D U-Net ($p < 0.05$), †: between MP3DU and EfficientNet ($p < 0.05$), ‡: between MP3DU and 2D U-Net ($p < 0.05$)).

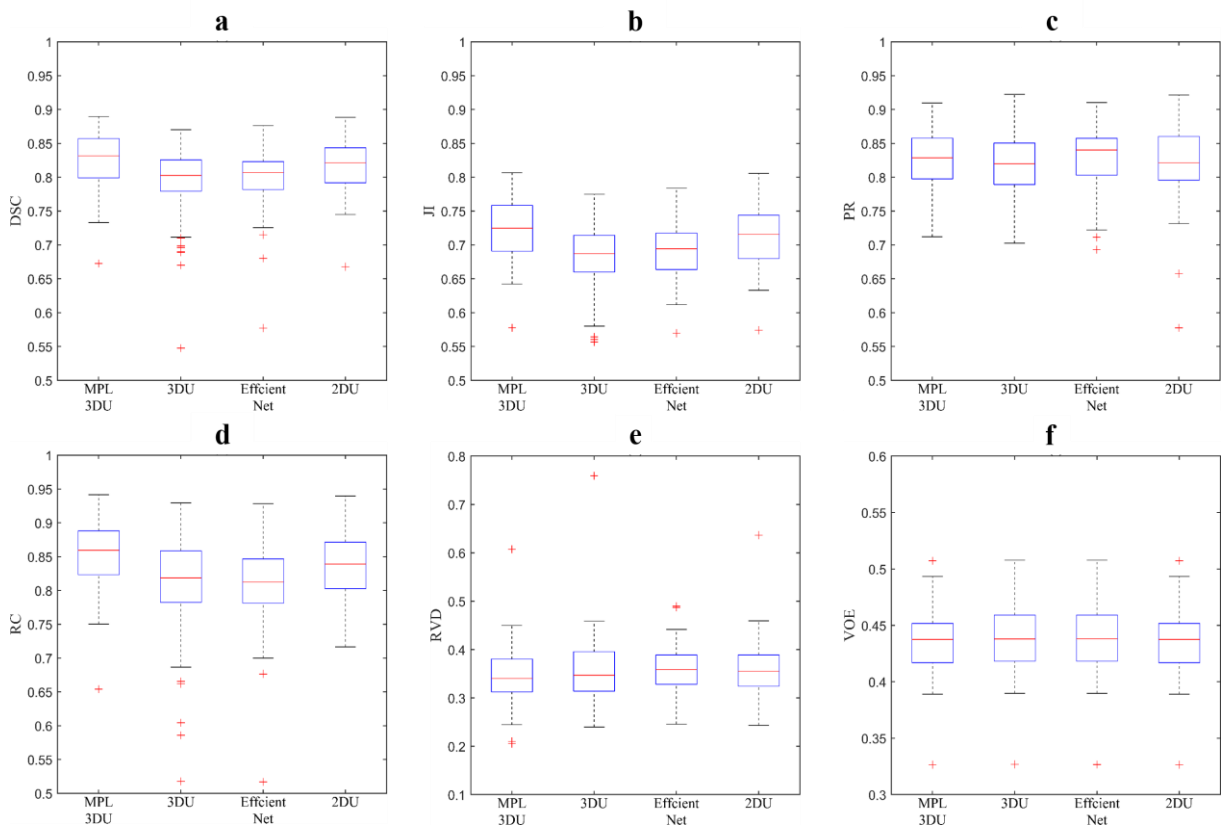

Figure 5. The boxplots of whole temporal bone structure segmentation performance results of the (a) Dice similarity coefficient score (DSC), (b) Jaccard index (JI), (c) precision (PR), (d) recall (RC), (e) relative volume difference (RVD), (f), and volume of error (VOE) for the deep learning networks, multi-projection 3D U-Net (MP3DU), 3D U-Net (3DU), EfficientNet, and 2D U-Net (2DU). Each box contains the first and third quartile of data. The medians are located inside of the boxes, visualized as red lines. The whiskers are extended above and below each box in ±1.5 times the interquartile range (IQR), and the outliers are visualized as red + marks defining values 1.5 IQR away from the box.

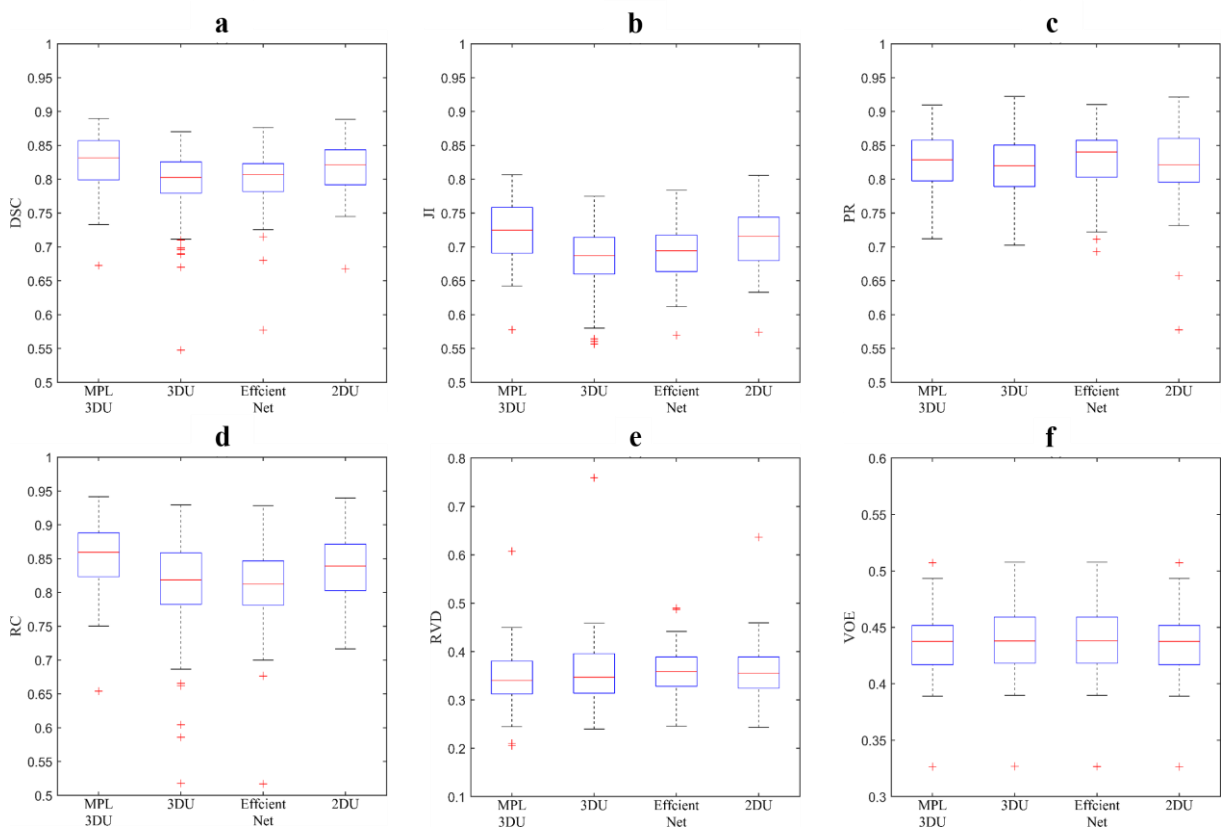

Figure 6. The boxplots of facial nerve segmentation performance results of the (a) Dice similarity coefficient score (DSC), (b) Jaccard index (JI), (c) precision (PR), (d) recall (RC), (e) relative volume difference (RVD), (f), and volume of error (VOE) for the deep learning networks, multi-projection 3D U-Net (MP3DU), 3D U-Net (3DU), EfficientNet, and 2D U-Net (2DU). Each box contains the first and third quartile of data. The medians are located inside of the boxes, visualized as red lines. The whiskers are extended above and below each box in ±1.5 times the interquartile range (IQR), and the outliers are visualized as red + marks defining values 1.5 IQR away from the box.

In **Figure 7**, the MP3DU exhibited more accurate predictions with more true positives and less false positives and false negatives compared to other networks in each different temporal bone structure. In the 3D segmentation results, the MP3DU exhibited better prediction results with less false positives and false negatives compared to other networks (**Figure 8**). Especially, the MP3DU predicted more accurately the entire facial nerve volume and exhibited improved structural continuity to other networks (**Figure 8**). The DSC for the entire test dataset were plotted from the stylomastoid foramen to the internal auditory meatus, and the MP3DU generally exhibited less variations of the performances compared to other networks (**Figure 9**). The MP3DU demonstrated the most consistent performances with the less fluctuations of true segmentation compared to the others throughout the entire facial nerve volume (**Figure 9**). As a result, the MP3DU represented the best segmentation performances overall and also for tubular structure such as facial nerve.

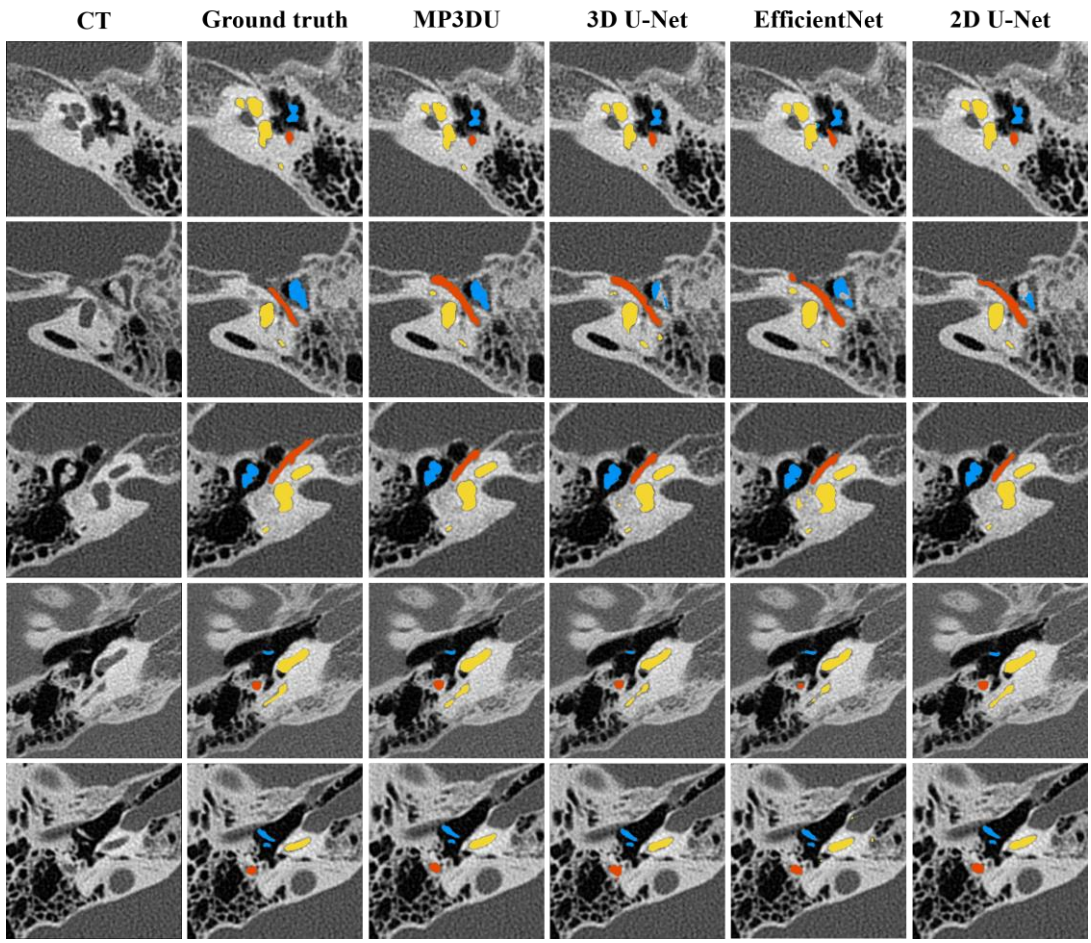

Figure 7. Segmentation images for the deep learning networks of multi-projection 3D U-Net (MP3DU), 3D U-Net, EfficientNet, and 2D U-Net. Each temporal bone structure is shown in CT images. The facial nerve, cochlea, and ossicle are visualized as red, yellow, and blue respectively.

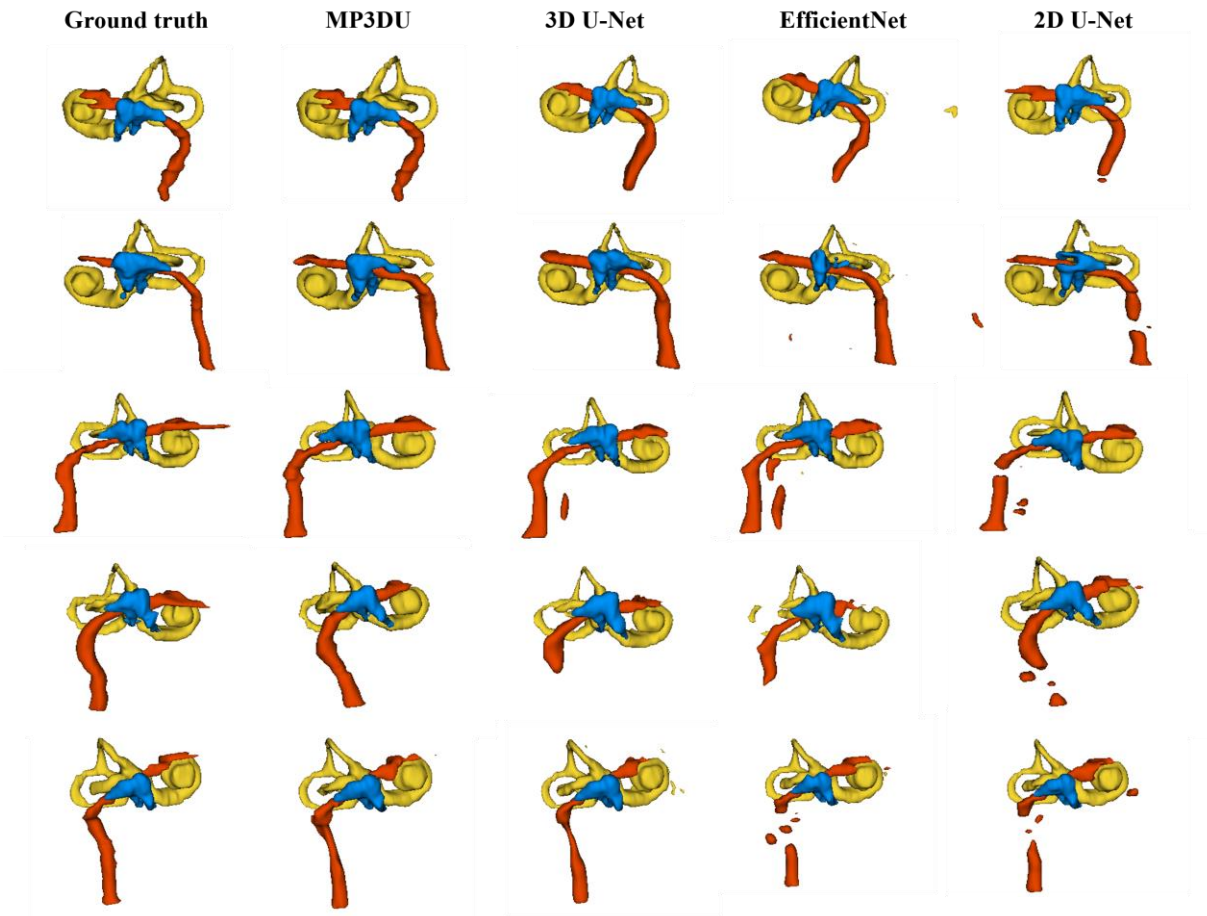

Figure 8. The 3D reconstructed temporal bone structures for the ground truth and segmentation results of the multi-projection 3D U-Net (MP3DU), 3D U-Net, EfficientNet, and 2D U-Net displayed from the left to right. The facial nerve, cochlea, and ossicle are visualized as red, yellow, and blue respectively.

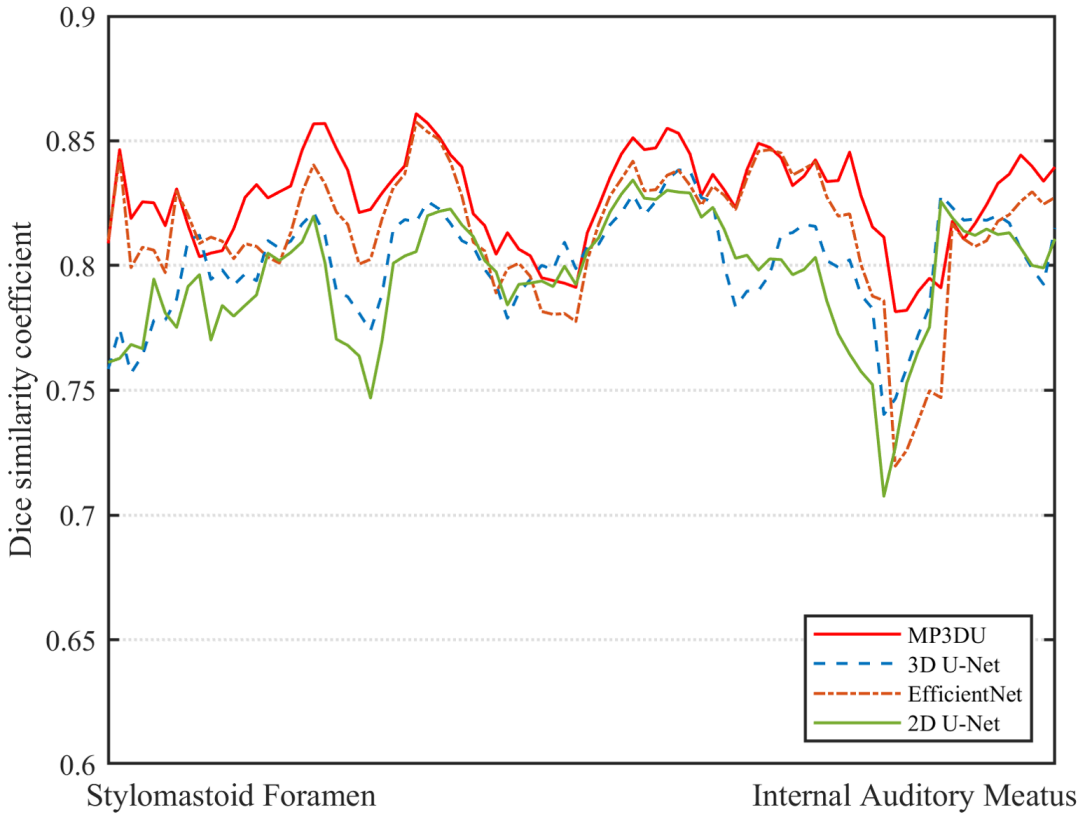

Figure 9. The Dice similarity coefficient score (DSC) line plots of facial nerve from the stylomastoid foramen to the internal auditory meatus for multi-proejction 3D U-Net (MP3DU), 3D U-Net, EfficientNet, and 2D U-Net.

Discussion

In this study, the MP3DU was proposed which learned 3D anatomical contextual information of the structures through 3D encoder-decoder architecture and the structural shapes from multi-projection maps by multi-projection losses in order to segment the temporal bone structures. The MP3DU was able to learn the volumetric information with the structural information of the temporal bone structures simultaneously. During training, The MP3DU obtained complementarily optimized features from 3D volume loss and multi-projection losses. Therefore, the MP3DU showed improved performance of automatic segmentation of the temporal bone structures by combining anatomical context information and structural shape information, resulting in higher accuracy throughout the entire volume in the CT images.

The accurate identification of the critical temporal bone structures in inner ear is an essential prerequisite for the preoperative planning of otological procedures such as cochlea implantation and tumor removal [1-3]. However, it is considered to be difficult to understand the complications and geometric information of structures for several reasons. First, the size of temporal bone structures is small and have large pathological variations. Second, CT, the most commonly used 3D imaging technique for inner ear diagnosis [1, 4, 5], does not provide enough information to distinguish temporal bone structures, especially facial nerve due to its low contrast compared to other areas [6, 9]. This may affect facial nerve to be seen with blurry boundaries, and the otologists to experience difficulties distinguishing it from unrelated structures [6, 9]. Though it is challenging to comprehend the temporal bone structures through CT images, it is crucial to acquire accurate segmentation of temporal bone structures for successful preoperative planning eventually.

The proposed network, MP3DU, was compared with other popular segmentation networks such as 2D U-Net, EfficientNet, and 3D U-Net. In segmentation performances of anatomical structures in CT images, 2D U-Net and EfficientNet exhibited slightly lower accuracies compared to the 3D networks in general. Especially, for the facial nerve, false negatives and positives were observed at a higher rate around the stylomastoid foramen and internal auditory meatus area due to the unclear boundaries of soft tissues affected by low contrast and confusion caused by unrelated structures such as air cells. Since the 2D networks were not able to learn the 3D contextual features of the temporal bone structures in CT images, the

2D networks exhibited coarser 3D segmentation volumes with more fluctuations of 3D performance accuracy from the stylomastoid foramen to the internal auditory meatus regions. In terms of learning 3D spatial contextual information between image slices of the 3D anatomical structures, 3D U-Net was generally expected to generate more accurate segmentation results compared to 2D networks [35]. In the present study, the 3D U-Net predicted the more accurate segmentation of the temporal bone structures with fewer false positives and negatives compared to the 2D U-Net and EfficientNet. However, the 3D U-Net had still limitations in segmenting the facial nerve regions with unclear boundaries accurately by only learning 3D spatial information between image slices. Moreover, it exhibited inaccurate segmentation results with disconnections around the stylomastoid foramen and internal auditory meatus area, since other bigger structures such as cochlea and ossicle affected learning of smaller features of facial nerve while training.

The network in this study showed an improvement for segmentation of temporal bone structures. The false positives and negatives were barley observed than other networks. Specifically, MP3DU demonstrated the better overall structural shape compared to 3D U-Net because its spatial information was complemented with the structural information by learning through multi- projection maps. Therefore, the MP3DU represented the most accurate segmentation of the entire volume of temporal bone structure compared to the other networks by simultaneously learning structural shape through multi-projection maps.

In the MP3DU, the multi-projection maps complementarily provided

structural shape information from each anatomical projection map to spatial contextual information of 3D encoder-decoder architecture. Unlike other temporal bone structures such as cochlea and ossicle, in the facial nerve areas with low visibility of unclear boundaries in CT images, the MP3DU exhibited the best outcomes with continuous and consistent facial nerve volumes from the stylomastoid foramen to the internal auditory meatus. The MP3DU especially surpassed other networks by showing continuous facial nerve volumes around the stylomastoid foramen and internal auditory meatus area where it is considered to be challenging due to unrelated similar structures like air cells, and in areas that had large topological variations in CT images [1, 4]. Compared with previous study using 3D U-Net [14, 30], our MP3DU achieved 0.75 of DSC while two previous studies reported 0.74, and 0.70 of DSC [14, 30]. Compared with two previous studies [14, 30], the MP3DU showed substantially enhanced performance of facial nerve segmentation in CT image.

The primary reason for improved segmentation performance by MP3DU was that its network architecture was constructed to complimentary learn the 3D anatomical context information of the temporal bone structure through 3D encoder-decoder architecture and the structural shape information by multi-projection maps. In the MP3DU, the complementary context information was successfully learned in the proposed network architecture, leading to maintain accurate anatomical structure volumes overall, particularly for facial nerve from the stylomastoid foramen to the internal auditory meatus areas. As a result, the most beneficial point of the proposed

network is that the simultaneous learning process from 3D volume and multi-projection maps minimizes the feature loss that may be caused by the 3D encoder-decoder and optimizes the segmentation results during training. This could improve the segmentation performance accuracy of temporal bone structures, especially for facial nerve in difficult areas with low contrast and unclear boundaries in CT images.

Based on the segmentation of the temporal bone structures in inner ear on CT images, The MP3DU could be used for clinical application. The advantage of automatic segmentation of the temporal bone structures using the MP3DU was that it could provide accurate identification of the temporal bone structures in the inner ear, maintaining accuracy and consistency when conducting large amounts of analysis. The MP3DU achieved automatic 3D segmentation of temporal bone structures. Specifically, the proposed networks showed improvement for the tubular structure such as facial nerve with large topological variations and unclear boundaries in low contrast area of CT images, which helps operators to reduce the workload and the time required for the segmentation of those. Moreover, it would be more useful for a preoperative planning procedure for cochlea implantation and tumor removal by automatic and accurate identification of the temporal bone structures.

Though, there are several limitations to overcome for our study. First, it is necessary to optimize memory use for efficient network learning and GPU usage. As there was a problem with reducing the memory requirements for dealing with large amounts of volumetric data when using 3D networks on GPU, we ended up using

the cropped images with smaller dimensions than the original. This data preprocessing task required additional time and labor. Second, there is a possibility of limitation of the generalization ability of our study due to using internal data from a single organization. Thus, the proposed network needs to be trained and evaluated for datasets from multiple organizations. In the future, we will improve the generalization ability and clinical efficacy of the MP3DU by using datasets from multiple organizations or devices with various acquisition settings.

Conclusions

In this study, the network for automatic segmentation of critical temporal bone structures to plan safe and efficient inner ear surgeries was proposed based on 3D encoder-decoder architecture with multi-projection maps. The MP3DU was designed based on a 3D U-Net with multi-projection maps which are generated from 3D volume input in order to complementally learn anatomical contexts and structural shape information. As a result, the MP3DU achieved substantially enhanced performances compared to other networks such as 3D U-Net, EfficientNet, and 2D U-Net in 2D and 3D performances. Furthermore, MP3DU demonstrated improved segmentation performance of temporal bone structures in volume. In particular, it was observed that facial nerve was maintained with more accurate structural shape than other networks due to the complementary learning with multi-projection maps. The MP3DU could be contributed to accurate and automatic identification of the temporal

bone structures for the preoperative planning such as cochlea implantation and tumor extraction to avoid any surgical complications.

References

1. Neves CA, Tran ED, Kessler IM, Blevins NH. *Fully automated preoperative segmentation of temporal bone structures from clinical CT scans.* Sci Rep 2021; **11**: 116.
2. Meng J, Li S, Zhang F, Li Q, Qin Z., *Cochlear size and shape variability and implications in cochlear implantation surgery.* Otol Neurotol 2016; **37**: 1307-1313.
3. Meng J, Li S, Zhang F, Li Q, Qin Z. *Multi-atlas segmentation of the facial nerve from clinical CT for virtual reality simulators.* Int J Comput Assis Radiol Surg 2020; **15**: 259-267.
4. Hussain R, Lalande A, Girum KB, Guigou C, Bozorg Grayeli A. *Automatic segmentation of inner ear on CT-scan using auto-context convolutional neural network.* Sci Rep 2021; **11**: 4406.

5. Vlastarakos PV, Kiprouli C, Pappas S, Xenelis J, Maragoudakis P, Troupis G, et al. *CT scan versus surgery: how reliable is the preoperative radiological assessment in patients with chronic otitis media?* Eur Arch Otorhinolaryngol 2012; **269**: 81-86.

6. Dong B, Lu C, Hu X, Zhao Y, He H, Wang J. *Towards accurate facial nerve segmentation with decoupling optimization.* Phys Med Biol 2022; **67:** 065007.

7. Chhabda S, Leger DS, Lingam RK. *Imaging the facial nerve: A contemporary review of anatomy and pathology.* Eur J Radiol 2020; **126**: 108920.

8. Gupta S, Mends F, Hagiwara M, Fatterpekar G, Roehm PC. *Imaging the facial nerve: a contemporary review.* Radiol Res Pract 2013; **2013**: 248039.

9. Noble JH, Warren FM, Labadie RF, Dawant BM. *Automatic segmentation of the facial nerve and chorda tympani in CT images using spatially dependent feature values.* Med Phys 2008; **35**: 5375-5384.

10. Wei Q, Hu Y, Gelfand G, Macgregor JH. *Segmentation of lung lobes in high-resolution isotropic CT images.* IEEE Trans Biomed Eng 2009; **56**: 1383-1393.

11. Oliveira DA, Feitosa RQ, Correia MM. *Segmentation of liver, its vessels and lesions from CT images for surgical planning.* Biomed Eng OnLine 2011; **10**: 30.

12. Ferreira A, Gentil F, Tavares JM. *Segmentation algorithms for ear image*

data towards biomechanical studies. Comput Methods Biomech Biomed Eng 2014; **17**: 888-904.

13. Bonne NX, Dubrulle F, Risoud M. *How to perform 3D reconstruction of skull base tumours.* European Annals of Otorhinolaryngology, Head and Neck Diseases, 2017; **134**: 117-120.

14. Lv Y, Ke J, Xu Y, Shen Y, Wang J, Wang J. *Automatic segmentation of temporal bone structures from clinical conventional CT using a CNN approach.* Int J Med Robot 2021: **17**: e2229.

15. Nakashima S, Sando I, Takahashi H, Fujita S. *Computer-aided 3-D reconstruction and measurement of the facial canal and facial nerve. I. Cross-sectional area and diameter: preliminary report.* Laryngoscope 1993; **103**: 1150-1156.

16. Noble JH, Dawant BM, Warren FM, Labadie RF. *Automatic identification and 3D rendering of temporal bone anatomy.* Otol Neurotol 2009; **30**: 436-442.

17. Powell KA, Kashikar T, Hittle B, Stredney D, Kerwin T, Wiet GJ. *Atlas-based segmentation of temporal bone surface structures.* Int J Comput Assis Radiol Surg 2019; **14**: 1267-1273.

18. Noble JH, Labadie RF, Majdani O, Dawant BM. *Automatic segmentation of intracochlear anatomy in conventional CT.* IEEE Trans Biomed Eng 2011; **58**: 2625-2632.

19. Heimann T, Meinzer HP. *Statistical shape models for 3D medical image*

segmentation: a review. Med Image Anal 2009; **13**: 543-563.

20. Jiang F, Grigorev A, Rho S, Tian Z, Fu Y, Jifara W, Adil K, Liu S. *Medical image semantic segmentation based on deep learning.* Neural Computing and Applications 2018; **29**: 1257-1265.

21. Du C, Gao S. *Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network.* IEEE Access 2017; **5**: 15750-15761.

22. Micheli-Tzanakou E. *Artificial neural networks: an overview.* Network: Comput Neural Syst 2011; **22**: 208-230.

23. Couteaux V, Si-Mohamed S, Renard-Penna R, Nempont O, Lefevre T, Popoff A, et al. *Kidney cortex segmentation in 2D CT with U-Nets ensemble aggregation.* Diagn Interv Imaging 2019; **100**: 211-217.

24. Yang Y, Jiang H, Sun Q. *A Multiorgan Segmentation Model for CT Volumes via Full Convolution-Deconvolution Network.* Biomed Res Int 2017; **2017**: 6941306.

25. Zhang T, Wang X, Xu X, Chen CP. *GCB-Net: Graph convolutional broad network and its application in emotion recognition.* IEEE Trans Affective Comput 2019; **13**: 379-388.

26. Zhao C, Carass A, Lee J, He Y, Prince JL. *Whole brain segmentation and labeling from CT using synthetic MR images*. *International Workshop on Machine Learning in Medical Imaging* 2017; Springer.

27. Man Y, Huang Y, Feng J, Li X, Wu F. *Deep Q learning driven CT pancreas*

segmentation with geometry-aware U-Net. IEEE Trans Med Imaging 2019; **38**: 1971-1980.

28. Li X, Hong Y, Kong D, Zhang X. *Automatic segmentation of levator hiatus from ultrasound images using U-net with dense connections.* Phys Med Biol 2019; **64**: 075015.

29. Liu Z, Song YQ, Sheng VS, Wang L, Jiang R, Zhang X, et al. *Liver CT sequence segmentation based with improved U-Net and graph cut.* Expert Syst Appl 2019; **126**: 54-63.

30. Nikan S, Van Osch K, Bartling M, Allen DG, Rohani SA, Connors B, et al. *PWD-3DNet: A deep learning-based fully-automated segmentation of multiple structures on temporal bone CT scans.* IEEE Trans Image Process 2020; **30**: 739-753.

31. Faul F, Erdfelder E, Buchner A, Lang AG. *Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses.* Behav Res Methods 2009; **41**: 1149-1160.

32. Faul F, Erdfelder E, Lang AG, Buchner A. *G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences.* Behav Res Methods 2007; **39**: 175-191.

33. Ronneberger O, Fischer P, Brox T. *U-net: Convolutional networks for biomedical image segmentation*. *International Conference on Medical image computing and computer-assisted intervention* 2015; Springer.

34. Tan M, Le Q. *Efficientnet: Rethinking model scaling for convolutional*

neural networks. *International conference on machine learning* 2019; PMLR.

35. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. *3D U-Net: learning dense volumetric segmentation from sparse annotation*. *International conference on medical image computing and computer-assisted intervention* 2016; Springer.

국문초록

MP3DU: CT영상 내 자동적 측두골 구조물 영상분할을 위한 다중 투영 3차원 U-Net

연구 배경: 인공와우 이식수술과 같은 내이 수술은 술중 발생할 수 있는 여러 크고 작은 부작용을 피하기 위해 정확한 술전 계획이 필요하다. 이를 위해 내이 속 측두골 구조물에 대한 해부적 정보, 위치 등에 대한 정보의 정확한 이해가 필수적이다. 여러 측두골 구조물 중에서도 중요하게 여겨지는 대표적인 것들로 얼굴신경, 달팽이관, 이소골이 있으며 해당 구조물들은 구조적 변형이 크고 크기가 작아 실질적인 이해가 어려운 것으로 여겨진다. 이러한 구조물들을 비교적 쉽게 파악하기 위해 내이 수술 전 측두골 CT영상을 취득하게 되지만 CT영상의 낮은 대비로 인해 구조물들 간의 경계가 모호해져 이비인후과 전문가도 구분이 어려운 문제가 있다. 그럼에도 불구하고 해당 구조물들에 대한 정확한 정보는 반드시 획득되어야 하기에 그 과정에서 이비인후과 전문가들의 수동적 영상분할은 무조건적으로 발생한다. 따라서 본 연구는 이러한 불편함을 줄이고자 측두골 구조물에 대한 자동 영상분할을 달성하고자 CT영상 다중 투영 3차원 U-Net을 제안하였다.

연구 방법: 연구를 위해 418명의 환자로부터 381개의 CT영상을 수집하였다. 해당 환자들은 내이 관련 질병으로 내원한 것으로 진단 과정에서 촬영된 CT영상 중 병변이 없는 정상 내이 영상만 사용하였다. 다중 투영 3차원 U-Net은 의료영상 영상분할에 많이 사용되는 3차원 U자형 신경망의 구조를 바탕으로 각 구조물에 대한 2차원의 다중 투영 영상을 접목하여 자동 영상분할을 달성하고자 하였다. 본 연구에서 제안하는 네트워크의 3차원 인코더-디코더 구조는 3차원 맥락 정보를 제공하며 그와 동시에 3차원 정보로부터 얻어진 2차원 다중 투영 영상이 전체적인 구조적 형태 정보를 딥러닝 학습 중 동시에 제공하며 상호보완적 결과를 얻고자 하였다.

연구 결과: 딥러닝을 이용한 측두골 영상분할 결과를 비교하기 위해 의료영상 영상분할에 많이 사용되는 2D U-Net, EfficientNet, 그리고 해당 네트워크의 기본 구조인 3D U-Net을 사용하였다. 본 연구가 제안한 네트워크인 다중 투영 3차원 U-Net이 전체 측두골 구조물 영상분할 결과로 2차원 성능 지표로 0.81의 DSC, 0.71의 JI, 0.81의 PR, 0.84의 RC를, 3차원 성능 지표로 0.34의 RVD와 0.43의 VOE를 달성하였으며 이는 다른 비교군 대비 높은 결과를 보였음을 관찰하였다. 또한 영상분할 결과의 3차원 모델링 비교 결과, 제안한 네트워크가 다른 비교 네트워크 대비 위양성과 위음성이 적게 관찰되었다.

결론: 본 연구는 CT영상에서 측두골 구조물 자동 영상분할을 수행하는 다중 투영 3차원 U-Net 제안하였으며 해당 네트워크는 3차원 맥락 정보와 각 구조물의 2차원 다중 투영 영상이 상호보완적으로 최적의 결과를 학습한 것을 확인할 수 있었으며 결과적으로 측두골 구조물 영상분할에 있어서 개선된 성능을 보였다.

주요어: 측두골 구조물, 3차원 영상분할, 컴퓨터단층영상, 딥러닝, 다중 투영

학 번: 2020-24482