The Laryngoscope

Automatic Segmentation of Vestibular Schwannoma From MRI Using Two Cascaded Deep Learning Networks

Sophia Marie Häußler, MD; Christian S. Betz, MD, PhD; Marta Della Seta, MD; Dennis Eggert, Dr. rer. nat; Alexander Schlaefer, MD, PhD; Debayan Bhattacharya

Objective: Automatic segmentation and detection of vestibular schwannoma (VS) in MRI by deep learning is an upcoming topic. However, deep learning faces generalization challenges due to tumor variability even though measurements and segmentation of VS are essential for growth monitoring and treatment planning. Therefore, we introduce a novel model combining two Convolutional Neural Network (CNN) models for the detection of VS by deep learning aiming to improve performance of automatic segmentation.

Methods: Deep learning techniques have been employed for automatic VS tumor segmentation, including 2D, 2.5D, and 3D UNet-like architectures, which is a specific CNN designed to improve automatic segmentation for medical imaging. Specifically, we introduce a sequential connection where the first UNet's predicted segmentation map is passed to a second complementary network for refinement. Additionally, spatial attention mechanisms are utilized to further guide refinement in the second network.

Results: We conducted experiments on both public and private datasets containing contrast-enhanced T1 and high-resolution T2-weighted magnetic resonance imaging (MRI). Across the public dataset, we observed consistent improvements in Dice scores for all variants of 2D, 2.5D, and 3D CNN methods, with a notable enhancement of 8.86% for the 2D UNet variant on T1. In our private dataset, a 3.75% improvement was reported for 2D T1. Moreover, we found that T1 images generally outperformed T2 in VS segmentation.

Conclusion: We demonstrate that sequential connection of UNets combined with spatial attention mechanisms enhances VS segmentation performance across state-of-the-art 2D, 2.5D, and 3D deep learning methods.

Key Words: artificial intelligence, machine learning, MRI, vestibular schwannoma.

Level of Evidence: 3

***Laryngoscope*, 00:1–8, 2025**

INTRODUCTION

Vestibular schwannoma (VS, a.k.a. acoustic neuroma) is a benign tumor originating from the Schwann cells, which are surrounding nerves and therefore supporting the neurons. In particular, VS develop around the eighth cranial nerve (vestibulocochlear nerve) or its branches (predominantly around the inferior or superior vestibular nerve, rarely around the cochlear branch).[1] The highest incidence peak of VS is in the age group of 65 years and above and according to Marinelli et al. in the age group of 70 years and above prevalence ranges between 20.6 and 212.4 per 100 000 persons.[2–4] Therefore, it seems reasonable to enhance imaging workflow for monitoring VS growth, which is also important in patients with bilateral VS.[5–8] Current MRI protocols for VS involve contrast-enhanced (ce) T1-weighted and high-resolution (hr) T2-weighted images, but there are few studies suggesting that there is no significant difference in accurately detecting VS with T1 or T2 images.[9] Therefore, automatic segmentation of MRI recently has become a useful innovation for enhancing imaging workflow and therefore aiding patient management.[10] There are models in the literature with attempts of VS segmentation employing 2D,[11,12] 2.5 D, and 3D[13] convolutional neural networks (CNN). 2D CNNs do not optimally leverage both in-plane and through-plane information, but the advantage is less parameters and therefore lower computational demands.[14] Whereas 3D CNNs are more precise by utilizing both in-plane and through-plane information, albeit by the cost of increased parameters.

We introduce a two-stage Convolutional Neural Network (CNN) model: the first CNN generates an initial tumor region estimate, and the second CNN refines it using the predicted segmentation mask and input image.

From the Department of Otorhinolaryngology (S.M.H., C.S.B., D.E., D. B.), University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Institute of Radiology, Charité-Universitätsmedizin Berlin (M.D.S.), Berlin Humboldt Universität zu Berlin and Berlin Institute of Health, Berlin, Germany; and the Institute of Medical Technology and Intelligent Systems (A.S., D.B.), Hamburg University of Technology, Hamburg, Germany.

Editor's Note: This Manuscript was accepted for publication on December 12, 2024.

Presented at the 95th Annual Meeting of the German Society of ORL-HNS 2024.

The authors declare no conflicts of interests.

This work was partially funded by grant number KK5208101KS0, by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School) from University Medical Center Hamburg-Eppendorf and TUHH i3 initiative.

Send correspondence to Sophia Marie Häußler, Department of Otorhinolaryngology, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany. Email: s.haeussler@uke.de

DOI: 10.1002/lary.31979

Spatial attention (SA) is incorporated by leveraging the first CNN's encoder features to enhance segmentation accuracy in the second CNN. To assess our method, we conducted experiments on 2D, 2.5D, and 3D variations of our architecture, applied to T1- and T2-weighted MRI separately.

MATERIAL AND METHODS

This study focuses on enhancing the segmentation performance of existing CNNs used in VS segmentation in hr MRI. This involves duplicating a CNN with a similar architecture as a first CNN prediction and connecting them in the second step. The second CNN receives input from both the original image and the first CNN's prediction. The initial prediction of the VS tumor location is provided by the first network, and the second network refines this prediction by considering the original input and the first CNN's output. To further enhance performance, spatial attention is applied to the encoder features of the second CNN using the encoder features of the first CNN.

Dataset and Implementation Detail

Our method underwent evaluation using two distinct datasets: a public VS segmentation dataset[15] and an in-house dataset. Therefore, reliability of the proposed CNN could be tested on a larger number of datasets. The in-house dataset consisted of 2D MRI of 102 patient datasets with neuroradiologically diagnosed VS, and we selected the MRI datasets of 96 patients with cMRI including T1- and T2-weighted sequences. The dataset of 6 patients was excluded resulting from poorer resolution not suitable for further processing. MRI was acquired with a 1.5 or 3 Tesla scanner using intravenous gadolinium contrast. The MRI protocol included ceT1- and hrT2-weighted sequences, T1 VIBE (volumetric interpolated breath-hold examination), fat-saturated post-gadolinium sequences, and T2w SPACE sequences and/or 3D CISS (constructive interference in steady-state) sequence. The slice thickness was 0.6–1 mm. MRI identified VS in all cases. Specifically, all slices with visible tumor with intrameatal extension and most prominently visible gross tumor volume were extracted and annotated by an otolaryngologist subspecialized in otology as well as a radiologist subspecialized in otolaryngology. Altogether, 497 MRI slices (251 T1, 246 T2) were annotated with the open annotation tool labelme (http://labelme.csail.mit.edu/Release3.0) as shown in Figure 1, which demonstrates the annotations of case 27. Intrameatal and extrameatal parts of the VS were marked separately to enable the most precise prediction of the tumor volume. Subsequently, pairs of T1 and T2 images from each patient were retained which belong to the same slicing and were used for the calculations. Overall, each patient in the dataset had an average of 2.59 ± 0.88 T1 and T2 images, each accompanied by a mask for corresponding T1 and T2 images. To isolate the relevant VS area from MRI, crops were obtained based on the dimensions of the rectangle encapsulating the largest mask. The in-house dataset was divided into training (72), validation (16), and test (8) sets.

The public dataset contained 3D MRI with 484 image sets from 242 VS patients, each subject to both T1 and T2 scans. These patients were randomly distributed into training (155), validation (39), and test (48) sets. In the preprocessing step, we extracted cubic boxes measuring 100 mm $\times$ 100 mm $\times$ 50 mm, centered on the tumor mass. Both f_1 and f_2 were implemented similarly to the existing 3D CNN for VS segmentation. Our 3D architecture is the same as used by Wang et al.[16] Each network comprised 5 encoding levels followed by 5 decoding levels, with channel numbers in each encoder level set at 16, 32, 64, 80, and 96. The 2D and 2.5D variants shared a comparable architecture, differing only in the use of 2D convolution in 2D

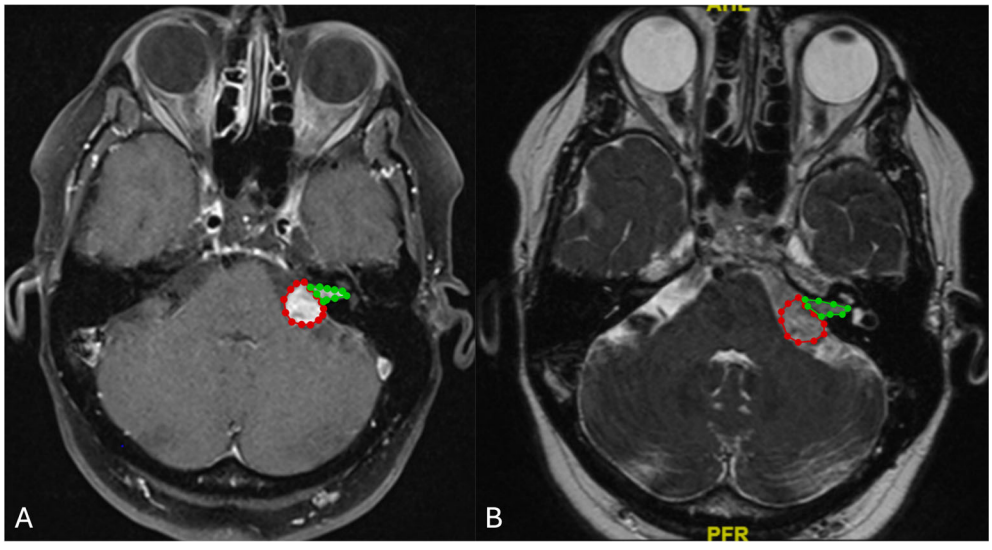

Fig. 1. Annotation of Case 27 with intrameatal (green) and extrameatal (red) VS (Koos III). (A) axial T1-weighted MRI. (B) axial T2-weighted MRI. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

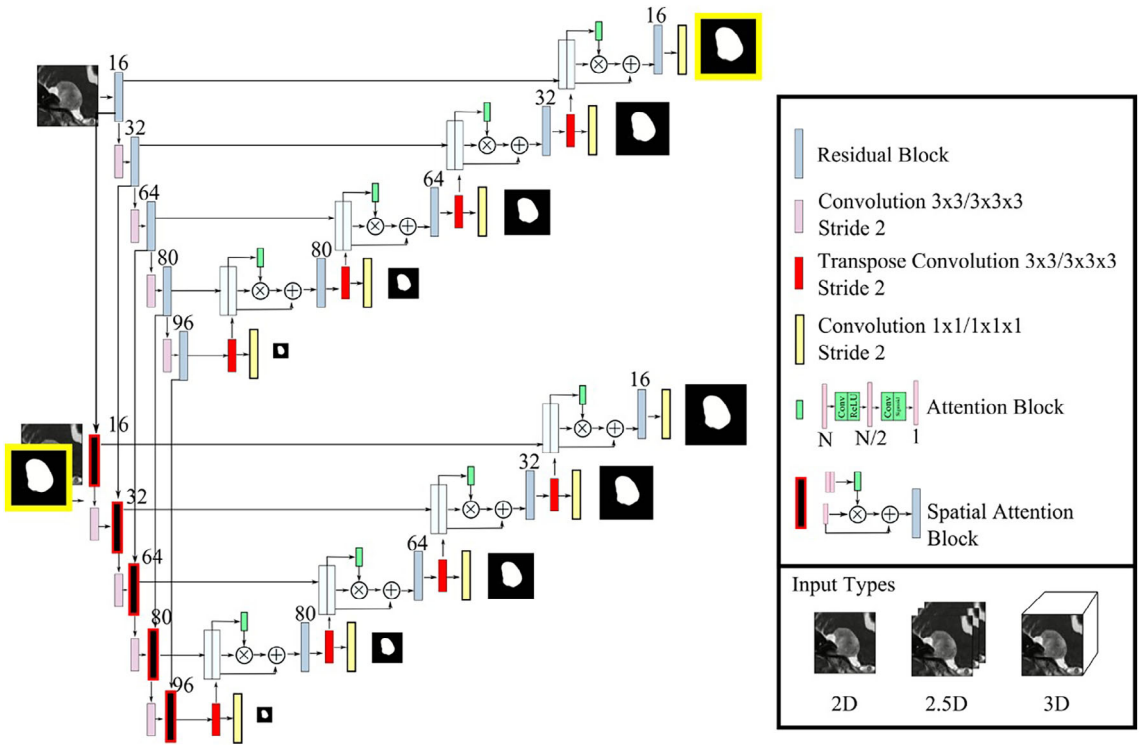

Fig. 2. Proposed architecture: Convolution operations are using 2D or 3D kernels based on the CNN's dimensionality (2D, 2.5D, or 3D). In the 2.5D case, the first two levels use 2D convolutions, while the rest employ 3D convolutions. The smaller resolution segmentation maps denote deep supervision. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

models instead of 3D convolution. In 2.5D models, the first two levels featured 2D convolution, while the last three levels used 3D convolution, following the approach by Wang et al.[16] Augmentation techniques, including random horizontal and vertical flips, random rotation, and Gaussian noise, were applied to all variants. Deep supervision was applied for improved gradient flow. An initial learning rate of 10–3 was utilized with cosine annealing. Optimization was carried out using the Adam optimizer with a momentum of 0.9. Each model was trained for 100 epochs, with 2.5D and 3D models using a batch size of 2, and 2D models using a batch size of 32. Fivefold cross-validation was employed for all our experiments. On the public dataset, segmentation performance was assessed using the Dice score (Dice), 95% Hausdorff distance (HD 95), and relative volume error (RVE). On the in-house dataset, the evaluation was based solely on Dice and HD95 metrics. The Dice score, also known as similarity coefficient, estimates the similarity of two sets of data, introduced by Dice[17]. Specifically, it is a reproducibility validation metric with a range from 0 (no spatial overlap of two sets of data) to 1 (complete overlap). HD 95 specifies the 95th percentile of the distance between surface points of one set (groundtruth) compared with the other set (predicted image). RVE is defined as the ratio of the absolute error of the predicted image (or estimation) to the real one.

Spatial attention mechanisms enhance the performance of CNNs by allowing the model to focus on specific regions of interest in an image while disregarding irrelevant background information. This is particularly crucial in medical imaging tasks, such as the segmentation of vestibular schwannomas from MRI scans, where the target structures are often small and can be obscured by noise.[18]

Our method, illustrated in Figure 2 operates on VS images represented by X. The input image $x \in X$, where for 2D, $x \in \mathbf{R}^{H \times W}$, and for 2.5D and 3D, $x \in \boldsymbol{R}^{H \times W \times D}$, with H, W, and D being the height, width, and depth of the image, respectively.

We define $f^1(x)$ as the first UNet-shaped CNN, with $f^1{}_{\text{enc}}(x)$ denoting its encoder features. The second CNN is fed both the input image x, the predicted segmentation map $f^1(x)$, and the encoder features $f^1{}_{\text{enc}}$, resulting in the refined segmentation mask y:

$$y = f^2\left(x, f^1(x) | f^1{}_{\text{enc}}(x)\right). \tag{1}$$

The encoder features $f^1{}_{\text{enc}}$ are used to introduce spatial attention to the encoder features of f^2. Let li represents the encoder features from the i-th encoder level of f^1, and mi represents the encoder features from the i-th encoder level of f^2. The encoder features passed to the higher levels of the encoder backbone of f^2 are expressed as:

$$m'_i = m_i + m_i \otimes g(m_i \circ l_i), \tag{2}$$

here, $\circ$ signifies concatenation, $\otimes$ represents element-wise multiplication, and g is a function incorporating 2D/3D convolution, Rectified Linear Unit (ReLU),

followed by a 1×1 convolution and sigmoid activation. Equation (2) encapsulates the spatial attention mechanism applied by f^1 to f^2. The Convolutional Block Attention Module (CBAM)[18] is a notable implementation of spatial attention. CBAM operates in two stages: channel attention and spatial attention. The spatial attention component generates a 2D attention map that highlights significant areas of the feature map. This process involves: (1) Global Average Pooling: Aggregating feature information across spatial dimensions to capture global context. (2) Convolutional Layer: Applying a convolution operation to learn spatial features, which helps in identifying which parts of the image are most relevant for the task at hand. (3) Sigmoid Activation: Producing a normalized attention map that indicates the importance of different spatial locations. By applying this attention map to the original feature map, the network can enhance its focus on critical areas, thereby improving segmentation accuracy.

Statistical analysis was performed, and tables were constructed using Microsoft Excel 2016 (Microsoft Corporation, Redmond, WA, USA).

RESULTS

The in-house dataset comprises 96 patient files with radiologically diagnosed VS and hr MRI. We included 45 male patients and 51 female patients in this study with a mean age of 53.52 ± 15.32 years. Of these patients, 16 were diagnosed with NF 2, and of these, 13 had bilateral VS. See Table I for patient demographic data. A total of 497 MRI slices were extracted from the

TABLE I.
Patient Demographic Data of the In-House Dataset.

Patient number (*n*)	Gender male/female	Age in years (mean ± standard deviation)	Unilateral/bilateral VS (*n*)	Neurofibromatosis type 2 (*n*)	Koos grade 1–4 (*n*)
96	45/51	53.52 ± 15.32	83/13	16	1: 22 2: 31 3: 27 4:16

TABLE II.
Experimental Results on the Public Dataset and In-House Dataset Showing the Segmentation Performance Using T1 and T2.

		Dice		HD95 (mm)		RVE (%)	
Dataset	Method	T1	T2	T1	T2	T1	T2
Public	2D	0.79 ± 0.09	0.73 ± 0.04	11.26 ± 7.28	12.2 ± 2.67	26.87 ± 19.09	30.86 ± 9.92
Public	2D own	**0.86 ± 0.01**	**0.75 ± 0.01**	**6.37 ± 1.10**	**10.17 ± 1.62**	**14.83 ± 0.97**	**27.47 ± 3.79**
Public	2.5D	0.70 ± 0.08	0.59 ± 0.08	16.20 ± 6.95	15.87 ± 6.04	37.28 ± 32.05	68.62 ± 6.41
Public	2.5D own	**0.79 ± 0.04**	**0.67 ± 0.03**	**6.49 ± 3.31**	**8.03 ± 1.28**	**17.69 ± 5.40**	**51.50 ± 11.69**
Public	3D	0.85 ± 0.01	0.81 ± 0.02	2.34 ± 0.31	3.52 ± 0.80	15.05 ± 1.94	21.50 ± 5.08
Public	3D own	**0.89 ± 0.01**	**0.82 ± 0.01**	**2.15 ± 0.13**	**3.18 ± 0.52**	**11.14 ± 1.61**	**19.48 ± 4.03**
In-house	2D	0.80 ± 0.03	0.55 ± 0.06	18.25 ± 1.33	32.03 ± 3.92	—	—
In-house	2D own	**0.83 ± 0.01**	0.51 ± 0.08	**17.60 ± 1.91**	32.14 ± 3.86	—	—

Mean and standard deviation values are reported with the first value for T1 and the second value for T2. The best results are highlighted (bold).

TABLE III.
Ablation Study of Our Proposed Method With and Without Spatial Attention.

		Dice		HD95 (mm)		RVE (%)	
Method	SA	T1	T2	T1	T2	T1	T2
2D	–	0.82 ± 0.04	0.74 ± 0.01	7.83 ± 3.82	**9.22 ± 1.18**	19.10 ± 4.81	**22.77 ± 3.81**
2D	+	**0.86 ± 0.01**	**0.75 ± 0.01**	**6.37 ± 1.10**	10.17 ± 1.62	**14.83 ± 0.97**	27.47 ± 3.79
2.5D	–	0.78 ± 0.05	0.66 ± 0.05	**5.58 ± 1.77**	8.79 ± 2.16	19.50 ± 9.37	**49.21 ± 10.33**
2.5 D	+	**0.79 ± 0.04**	**0.67 ± 0.03**	6.49 ± 3.31	**8.03 ± 1.28**	**17.69 ± 5.40**	51.50 ± 11.69
3D	–	0.88 ± 0.01	0.79 ± 0.02	2.16 ± 0.36	3.41 ± 0.79	14.72 ± 4.33	20.66 ± 4.54
3D	+	**0.89 ± 0.01**	**0.82 ± 0.01**	**2.15 ± 0.13**	**3.18 ± 0.52**	**11.14 ± 1.61**	**19.48 ± 4.03**

The best results are highlighted (bold).

dataset (251 T1 images, 246 T2 images) and were subsequently annotated.

Our experiments were performed on the public ($n = 242$) and the in-house dataset ($n = 96$). The mean and standard deviation metric values are presented for both T1 and T2 images in Table II, which shows the calculations with the existing CNN[13,16] and the experiments with our own CNN. We observe improvements across all metrics using our method on the public dataset. Notably, the highest Dice score is achieved with 3D (own/new CNN) using T1 images, reaching 0.89, compared with 0.85 reported for the pre-existing 3D CNN. This represents a 4.75% increase. Additionally, HD95 and RVE show percentage decreases of 8.11% and

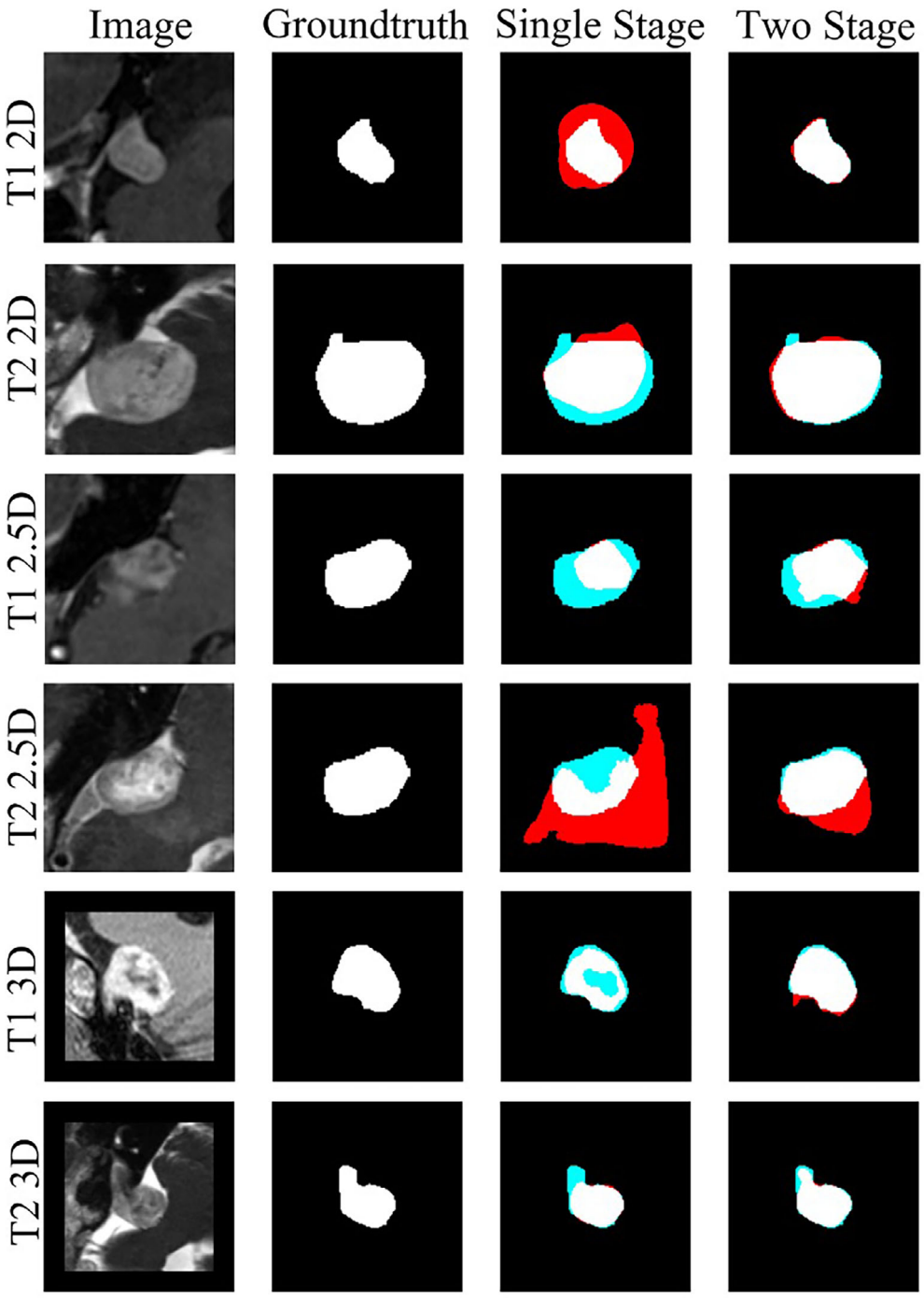

Fig. 3. Qualitative evaluation showing the input, ground truth, single-stage CNN, and two-stage CNN (own) for T1 and T2 for 2D, 2.5D, and 3D CNN variants. The red and cyan pixels indicate false positive (FP) and false negative (FN) predictions, respectively. The white pixels indicate overlap with ground truth. The two-stage network refines the initial hypothesis of the first-stage network to reduce the FP and FN. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

25.98%, respectively, and therefore improvement with the new CNN.

For T2 images, the highest Dice score is 0.82 for 3D (own/new CNN), compared with 0.81 for pre-existing 3D CNN. In general, we observe a similar trend in the 2D and 2.5D (own) variants, with 2D variants exhibiting relatively better performance than 2.5D variants on the public dataset. When using our in-house dataset, we again find that T1 is the superior modality for VS segmentation compared with T2. Specifically, the Dice score for 2D (own) using T1 images reaches 0.83, while the 2D CNN achieves 0.80. It is worth noting that the 2D (own)

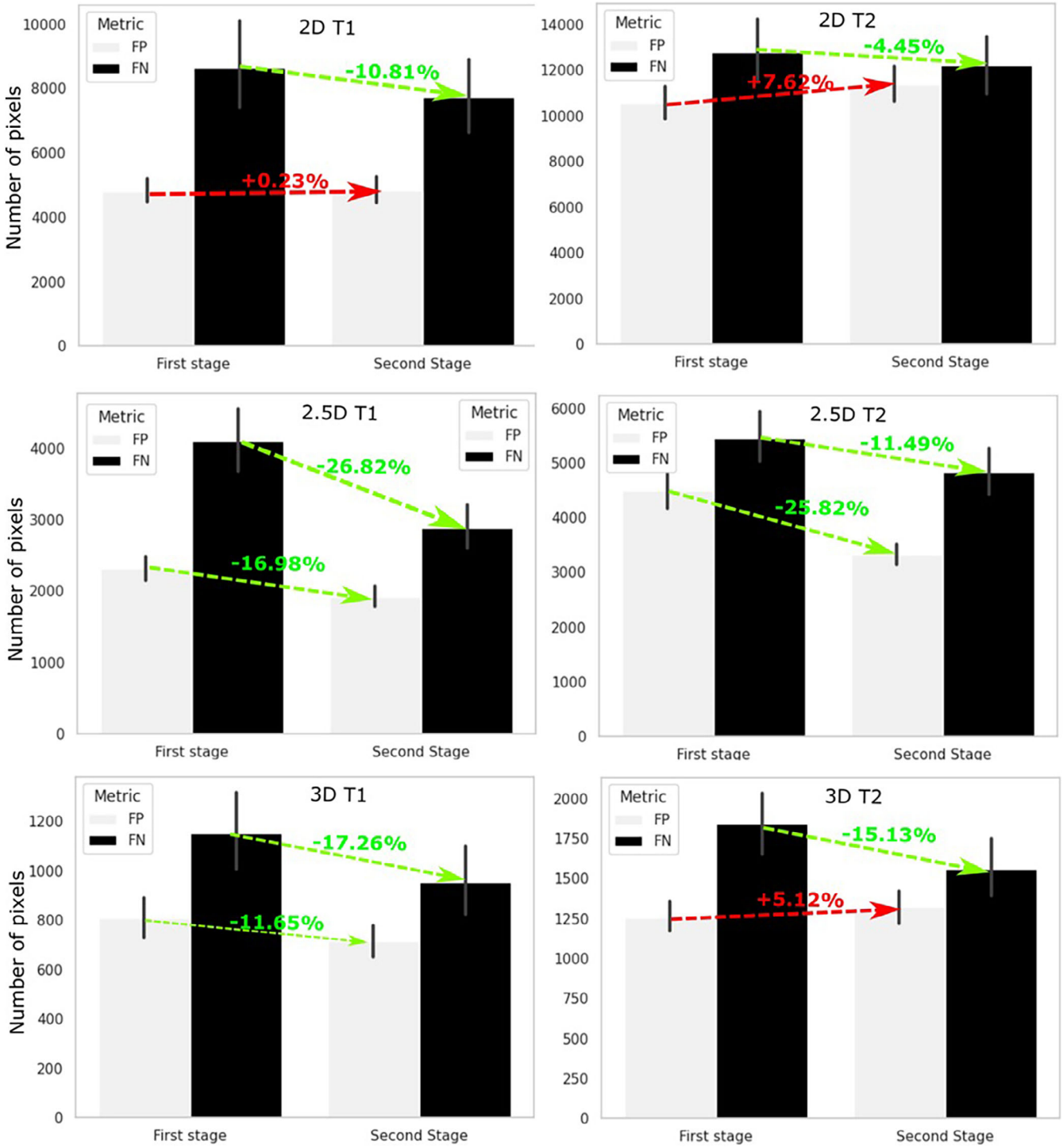

Fig. 4. Qualitative results showing that the second CNN effectively reduces false positive (FP) areas identified by the first CNN, demonstrating sequential segmentation refinement. FN, false negative. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

performs poorer for T2 images of the in-house dataset, with a decrease in Dice score of 7.26%.

In our ablation study shown in Table III, experiments with and without spatial attention were conducted on the public dataset. The primary performance enhancement is attributed to the cascading networks, with spatial attention making additional contribution. Specifically, the 3D (own/new CNN) demonstrates a 24.23% improvement in RVE with spatial attention, and a marginal 1.13% increase in Dice for T1. Similar consistent improvements are observed in 2D variants, with a 4.87% increase in Dice for T1 images. The impact of spatial attention on T2 images remains inconclusive.

Qualitative results (Figs. 3 and 4) show that the second CNN effectively reduces false positive areas identified by the first CNN, demonstrating sequential segmentation refinement. Our code is available at https://github.com/mtec-tuhh/VSSegCode.

DISCUSSION

Current methods have primarily utilized T1 and T2 MRIs for segmentation tasks, typically employing 2D, 2.5D, or 3D CNNs.[13,16] In our study, we establish that regardless of the CNN's dimension or modality, existing methods can enhance segmentation performance by employing our two-stage approach. This entails taking the output of an initial CNN and the input image and feeding them into another CNN with a similar architecture. We note consistent improvements in segmentation metrics across 2D, 2.5D, and 3D CNN variants when connected in a sequential manner for public dataset. This performance boost appears to result from the initial CNN generating an initial hypothesis outlining the probable tumor region, which the second CNN can refine. Additionally, the false positives and false negatives are reduced based on the initial hypothesis of the first network. Qualitative results (Fig. 4) show that the second CNN effectively reduces false positive areas identified by the first CNN, demonstrating sequential segmentation refinement.

Experiments conducted on a public dataset reveal improved Dice Score, HD95, and RVE, with the highest performance achieved in the 3D setting. Although most gains stem from the sequential configuration, the performance can be further boosted through spatial attention using first and second CNN's encoder features. In our study, we enhanced our CNN framework by integrating spatial attention to improve the segmentation of VS. Instead of employing global averaging, we utilized features from the first CNN to refine the feature representations in the second CNN. Specifically, this was achieved by first multiplying the features from the first CNN with those of the second CNN, followed by an additive operation with the original second CNN features. This multiplication step introduces spatial attention, enabling the network to focus on relevant regions, thereby improving localization accuracy and enhancing the Dice score.

Notably, contrast-enhanced T1 proves more effectively than T2 for segmenting VS on both public and in-house datasets, consistent with prior research about automatic segmentation of VS.[15,19] These findings are supported by our in-house dataset, and T2 performance is notably poorer, likely due to challenging tumor delineation and limited training data, causing errors to propagate and worsen performance. Contradictory, previous literature about regular VS imaging suggests that non-contrast hrT2 is sufficient for the monitoring of VS[20] and Strauss et al.[9] postulate that T1 and T2 are equivalent for postoperative surveillance of residual VS. Regarding the investigation of sensorineural hearing loss, T2 seems to be sufficient for the detection of intrameatal or cerebellopontine angle tumors, but non-contrast MRI might miss intralabyrinthine lesions.[21] Therefore, contrast-enhanced T1 is an important sequence in the detection and surveillance of VS and is more accurate than T2 in automatic segmentation as shown in this study.

Our work has limitations. First, it is retrospective and based on a limited number of data from a single center and an open dataset, necessitating prospective trials for clinical validation. Second, our in-house dataset sample includes a limited number of patients and images with planimetric annotations. We were able to show with our experiments on the public dataset that the highest performance in Dice Score HD95 and RVE were found in the 3D setting. To prove these results, a bigger own sample size as well as follow-up MRI scans and volumetric annotations would be helpful in a follow-up study. Third, our proposed architecture increases parameter count by over 2°ø compared with the initial CNN, requiring further experiments to explore smaller networks' efficacy for refinement. Lastly, the sequential configuration may exacerbate errors if the initial estimation is inaccurate.

In summary, the goal of automated VS segmentation is to aid in VS growth monitoring.

We present a method enhancing VS segmentation performance of all current state-of-the-art CNN in 2D, 2.5D, and 3D version for VS, validated on public and in-house datasets containing T1 and T2. The setting with T1 and 3D seems to provide the best performance for automatic segmentation of VS.

CONCLUSION

The contributions of our study are threefold. First, we demonstrate the improvement in the performance of 2D, 2.5D, and 3D CNNs used in VS segmentation by duplicating and sequentially connecting them. Second, to enhance performance further, spatial attention is introduced through the utilization of the first CNN's encoder features to modulate the second CNN's features. Spatial attention mechanisms play a vital role in enhancing the performance of CNNs for medical imaging tasks by allowing models to focus on relevant features while ignoring irrelevant data. This capability is particularly beneficial for accurately segmenting small and complex structures, ultimately improving clinical workflows and patient outcomes. Third, we conduct extensive experiments on both publicly available and in-house datasets on T1- and T2-weighted MRI to illustrate the effectiveness of our approach.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

BIBLIOGRAPHY

1. Roosli C, Linthicum FH Jr, Cureoglu S, Merchant SN. What is the site of origin of cochleovestibular schwannomas? *Audiol Neurootol*. 2012;17:121-125.
2. Goldbrunner R, Weller M, Regis J, et al. EANO guideline on the diagnosis and treatment of vestibular schwannoma. *Neuro Oncol*. 2020;22:31-45.
3. Marinelli JP, Grossardt BR, Lohse CM, Carlson ML. Prevalence of sporadic vestibular schwannoma: reconciling temporal bone, radiologic, and population-based studies. *Otol Neurotol*. 2019;40:384-390.
4. Stangerup SE, Caye-Thomasen P. Epidemiology and natural history of vestibular schwannomas. *Otolaryngol Clin North Am*. 2012;45:257-268.
5. Abeshi A, Ferri GG. A bilateral vestibular schwannoma is not always related to neurofibromatosis type 2. *J Int Adv Otol*. 2023;19:263-265.
6. Lee TJ, Chopra M, Kim RH, Parkin PC, Barnett-Tapia C. Incidence and prevalence of neurofibromatosis type 1 and 2: a systematic review and meta-analysis. *Orphanet J Rare Dis*. 2023;18:292.
7. Asthagiri AR, Parry DM, Butman JA, et al. Neurofibromatosis type 2. *Lancet*. 2009;373:1974-1986.
8. Mautner VF, Lindenau M, Baser ME, et al. The neuroimaging and clinical spectrum of neurofibromatosis 2. *Neurosurgery*. 1996;38:880-885; discussion 885–886.
9. Strauss SB, Stern S, Lantos JE, et al. High-resolution T2-weighted imaging for surveillance in postoperative vestibular Schwannoma: equivalence with contrast-enhanced T1WI for measurement and surveillance of residual tumor. *AJNR Am J Neuroradiol*. 2022;43:1792-1796.
10. Kujawa A, Dorent R, Connor S, et al. Automated Koos classification of vestibular schwannoma. *Front Radiol*. 2022;2:837191.
11. George-Jones NA, Wang K, Wang J, Hunter JB. Automated detection of vestibular schwannoma growth using a two-dimensional U-net convolutional neural network. *Laryngoscope*. 2021;131:E619-e624.
12. Sager P, Näf L, Vu E, et al. Convolutional neural networks for classifying laterality of vestibular schwannomas on single MRI slices-a feasibility study. *Diagnostics*. 2021;11:1676.
13. Wang H, Qu T, Bernstein K, Barbee D, Kondziolka D. Automatic segmentation of vestibular schwannomas from T1-weighted MRI with a deep neural network. *Radiat Oncol*. 2023;18:78.
14. Guotai Wang JS, Li W, Dorent R, et al. Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss. In: Shen, D., et al. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science*. Vol 11765. Cham: Springer; 2019. https://doi.org/10.1007/978-3-030-32245-8_30.
15. Shapey J, Kujawa A, Dorent R, et al. Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm. *Sci Data*. 2021;8:286.
16. Wang G, Shapey J, Li W, et al. Automatic segmentation of vestibular Schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss medical image computing and computer assisted intervention – MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II. Shenzhen, China: Springer-Verlag, 264–272. 2019.
17. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297-302.
18. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*. Vol 11211. Cham: Springer; 2018. https://doi.org/10.1007/978-3-030-01234-2_1.
19. Martinez-Perez R, Ung TH, Youssef AS. The 100 most-cited articles on vestibular schwannoma: historical perspectives, current limitations, and future research directions. *Neurosurg Rev*. 2021;44:2965-2975.
20. Forgues M, Mehta R, Anderson D, et al. Non-contrast magnetic resonance imaging for monitoring patients with acoustic neuroma. *J Laryngol Otol*. 2018;132:780-785.
21. Annesley-Williams DJ, Laitt RD, Jenkins JP, Ramsden RT, Gillespie JE. Magnetic resonance imaging in the investigation of sensorineural hearing loss: is contrast enhancement still necessary? *J Laryngol Otol*. 2001;115:14-21.