MRI-based Quantitative AI approaches for Vestibular Schwannoma Care

Yunjie Chen

Colophon

MRI-based Quantitative AI approaches for Vestibular Schwannoma Care
Yunjie Chen

ISBN: 000-00-0000-000-0
Thesis layout & cover designed by Yunjie Chen
Printed by xxxxxxxx, the Netherlands

MRI-based Quantitative AI approaches for Vestibular Schwannoma Care

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. ,

volgens besluit van het college voor promoties

te verdedigen op xxxxxxxx x x xxxx

klokke xx:xx uur

door

Yunjie Chen

geboren te Shanghai, China

in 1993

Promotor:	Prof. dr. ir. M. Staring
Co-promotor:	Dr. Qian Tao
Leden promotiecommissie:	Prof. dr. Boudewijn Lelieveldt *Leiden University Medical Center, Leiden* Prof. dr. Clarisa Sánchez *University of Amsterdam, Amsterdam* Dr. Hugo Kuijf *TNO, Den Haag* Dr. Jonathan Shapey *King's College London, London*

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands. This work was carried out in the ASCI graduate school.

Financial support for the publication of this thesis was kindly provided by:
LKEB, Department of Radiology of Leiden University Medical Center, The Netherlands.
Library of Leiden University.

Contents

List of abbreviations

MRI	magnetic resonance imaging
CT	computed tomography
PET	positron emission tomography
RF	radio frequency
GBCA	gadolinium-based contrast agent
VS	vestibular schwannoma
CPA	cerebellopontine angle
CN	cranial nerve
LINAC	Linear accelerator
AI	artificial intelligence
DL	deep learning
SGD	stochastic gradient descent
CNN	convolutional neural network
INR	implicit neural representation
MLP	Multilayer Perceptron
3D	three-dimensional
T1	T1-weighted MRI
T2	T2-weighted MRI
T1ce	contrast-enhanced T1-weighted MRI
DWI	diffusion-weighted imaging
FLAIR	T2-fluid-attenuated inversion recovery
DCE-MRI	dynamic contrast-enhanced MRI

TE	echo time
TR	repetition time
SE	Spin echo
GR	Gradient echo
GAN	generative adversarial network
VAE	variational auto-encoder
FiLM	feature-wise linear modulation
TTUR	Two Time-scale Update Rule
PSNR	peak signal-to-noise ratio
SSIM	structural similarity index
S2S	surface-to-surface distance
HD	Hausdorff distance
RVD	relative volume difference
RVE	relative volume error
ET	enhanced tumor
WT	whole tumor
TC	tumor core
FOV	field of view
IQR	interquartile range
SDF	signed distance function
PE	positional encoding
MF	Magnetic field
ICC	interclass correlation coefficient
ROC	Receiver operating characteristic
CI	confidence interval

1

Introduction

1.1 Head and neck MRI

With recent advances in imaging technology, Magnetic Resonance Imaging (MRI) has proven to be sensitive and reliable in head and neck radiology [1, 2]. Unlike computed tomography (CT) and positron emission tomography (PET), MRI is a non-invasive medical imaging technology that does not involve the use of ionizing radiation [3]. The MRI scanning starts from a strong external magnetic field that aligns the directions of the spin angular momentum of protons (hydrogen nucleus) in the tissue. The protons then absorb energy from external Radio Frequency (RF) pulses, temporarily perturbing this alignment. After a short period, the protons release the absorbed energy, which is called relaxation, and in so doing emit RF energy. The images are obtained by applying the Fourier transformation to convert the signals from the frequency domain into the spatial domain. By manipulating the sequence of RF pulses applied and collected, different MRI contrasts are acquired, providing a comprehensive evaluation of anatomical structures, facilitating the identification of a wide range of traumatic, inflammatory, and neoplastic lesions [4].

One distinguishing characteristic of MRI is its wide variety of imaging sequences. By varying the pulse sequence parameters, specific tissue characteristics can be highlighted. For instance, T1- and T2-weighted imaging are standard techniques for structural assessment. Diffusion-weighted imaging and dynamic contrast-enhanced imaging are promising tools that provide biomarkers to guide treatment in head and neck cancer [5, 6]. Besides the structural MRI sequences, functional MRI detects and studies brain activities by measuring changes in local oxygenation of blood [7]. Among all sequences, contrast-enhanced T1-weighted magnetic resonance imaging with a gadolinium-based contrast agent (GBCA) is exceptionally valuable in the non-invasive detection and characterization of brain tumors. By accumulating in tissues with rich vascularity or interstitial space, GBCAs offer enhanced visibility of tumor lesions on MR imaging.

With appropriate selection and interpretation of imaging protocols, head and neck

MRI serves as an indispensable tool, particularly valuable for tumor assessment. Its superior contrast resolution enables precise delineation of the lesion margins and adjacent anatomical structures. The absence of ionizing radiation also makes MRI suitable for serial imaging during long-term surveillance or treatment monitoring.

1.2 Vestibular schwannoma

Vestibular schwannomas (VS) are benign tumors that arise from vestibulocochlear nerves, and are the most common neoplasm of the cerebellopontine angle (CPA) in adults [8]. Recent epidemiological data estimate the prevalence of VS at approximately 1 in 2,000 adults, and this increases to about 1 in 500 among individuals aged 70 years and older [9].

Disease presentation

Vestibular schwannomas are generally slow-growing tumors, with approximately 60% showing no or minimal progression over time [8]. The most common symptoms of VS include hearing loss (90%) [10], balance disorders (61%) [10, 11], and asymmetric tinnitus (55%) [12]. Hearing loss is the most common symptom that is mainly caused by dysfunction of the cochlear nerve [13]. Affected patients usually experience progressive, unilateral hearing impairment in the ear ipsilateral to the tumor [14, 13], and, in some cases, sudden hearing loss may occur in the contralateral ear, leading to loss of binaural hearing [15, 16]. Patients with hearing loss are more likely to develop tinnitus. Tinnitus can occur before and after intervention, and the severity varies across patients [17, 18]. Of all the symptoms, balance disorders, including vertigo and dizziness, have the most profound effect on quality of life [19, 20, 21]. Patients with large tumors exerting pressure on the brain stem and cerebellum may experience trigeminal hypoesthesia, secondary trigeminal neuralgia, cerebellar dysmetria and ataxia, or slowly progressive hydrocephalus without altered consciousness [22]. Notably, there is only a weak correlation between tumor size and the severity of the symptoms [23]. Similarly, symptom progression does not reliably correlate with tumor growth [23]. Figure 1.1a shows the microanatomy affected by vestibular schwannoma.

Treatment strategy

Treatment options for vestibular schwannoma include wait-and-scan surveillance, radiosurgery, microsurgery, or a combination of these approaches. The treatment plan is chiefly guided by the tumor size and changes in size over time, but also by the symptoms and other patient-specific factors [24].

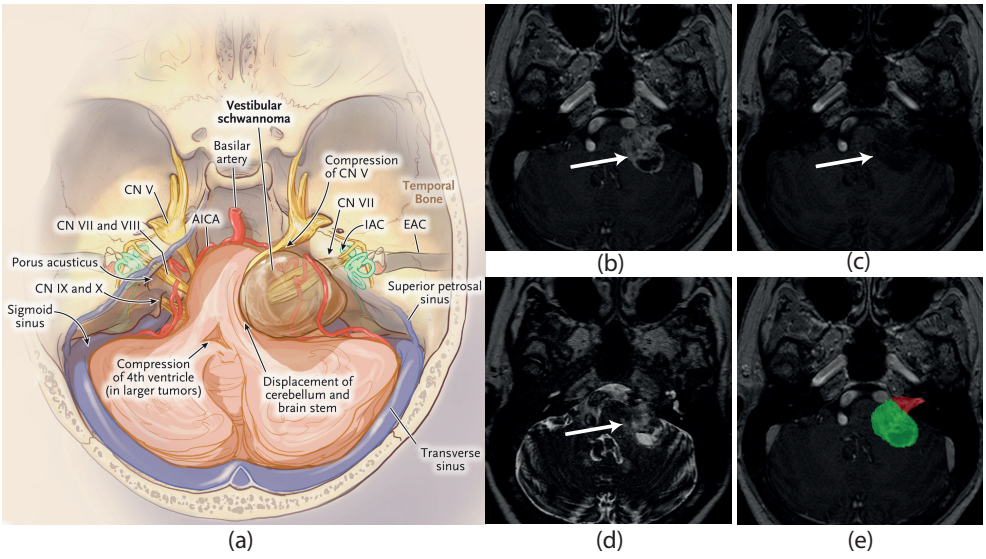


Figure 1.1: Disease presentation of vestibular schwannoma (a) Microanatomy and structures affected by vestibular schwannomas [8]. Large vestibular schwannomas may compress cranial nerves (CN), the brain stem, and cerebellum. (b) Vestibular schwannoma on post-contrast T1. The lesion is indicated by the white arrow. (c) Vestibular schwannoma on pre-contrast T1. (d) Vestibular schwannoma on high-resolution T2. (e) The segmentation mask of vestibular schwannoma overlaid on post-contrast T1. The green mask shows the extrameatal portion, and the red mask shows the intrameatal portion.

Wait-and-scan policy The wait-and-scan approach (Observation) is a preferred strategy, particularly for the incidental, asymptomatic VS [25]. After diagnosis, MRI exams and audiologic evaluation are commonly performed every 6 months or annually. Panels b – e of Figure 1.1 shows vestibular schwannoma in different MRI sequences and corresponding tumor mask. The wait-and-scan approach aims to identify a fast-growing tumor or a more aggressive process mimicking a vestibular schwannoma. Tumor progression on two consecutive MRIs is commonly defined as an increase in diameter of more than 2 mm. In case of tumor progression, more active treatment is advised.

Microsurgery Microsurgical resection can be performed on tumors of all sizes and is the primary approach for large tumors associated with symptomatic brain-stem compression, hydrocephalus, trigeminal neuralgia or neuropathy, or a combination of these complications [26]. The middle fossa, translabyrinthine, and retrosigmoid approaches are the three primary microsurgical approaches. Postoperatively, patients are at risk of hearing loss and facial nerve dysfunction, and the risk is directly

proportional to tumor size [27, 28]. To reduce the risk of permanent facial nerve paralysis, a small remnant of the tumor may be intentionally left adherent to the facial nerve or brainstem. However, approximately 30% of tumors continue to grow after incomplete tumor resection [29, 30].

Radiosurgery During stereotactic radiosurgery, highly conformal radiation is delivered to the lesion of interest targeted by non-contrast-enhanced computed tomography or contrast-enhanced MRI. By selectively treating the tumor and maximally sparing the surrounding tissue, radiosurgery helps to prevent tumor growth [8]. Popular techniques include gamma-knife radiosurgery [31] and Linear accelerator (LINAC)-based systems [32]. Patients with tumors that have a diameter of less than 2.5 cm in the cerebellopontine angle are usually considered to be candidates for radiosurgery. However, for individuals with contraindications to microsurgery, radiosurgery offers a viable alternative [33].

Unfortunately, none of the treatments can eradicate the tumor without the risk of damage to the auditory-vestibular organs or nerves. Therefore, active treatment is generally not aimed at alleviating the audiovestibular symptoms most patients present with, but maintaining the quality of life is an important aim of vestibular schwannoma management, in addition to tumor control [34, 35, 36, 37].

Data-driven vestibular schwannoma care

The variability in vestibular schwannoma presentation encourages personalized patient management, and the use of data science and information technology in vestibular schwannoma care is rapidly increasing. However, utilizing large-scale healthcare data remains challenging. First, although large-scale datasets have the potential to provide unbiased insights and evidence-based recommendations, it is difficult to merge and summarize heterogeneous information from multiple sources due to the various imaging protocols and guidelines. Second, data inconsistency, including low-quality scans, missing data, and limited longitudinal data, restricts the robustness of developed models. Last but not least, traditional data-driven approaches heavily rely on hand-crafted features, while there is no reliable imaging biomarker for vestibular schwannoma care. Approaches that can more efficiently and intelligently utilize large-scale medical data are urgently needed.

Recent advances in artificial intelligence (AI), particularly deep learning, have demonstrated strong potential in radiology. AI supports radiologists in providing more precise and patient-centered care through applications such as automated tumor segmentation, image reconstruction, image registration, and survival analysis. These developments highlight the promise of AI-driven approaches in facilitating

more consistent, objective, and personalized management strategies for patients with vestibular schwannoma.

1.3 Artificial intelligence in medical image computing

AI has demonstrated remarkable progress in healthcare research, particularly in MRI analysis [38]. While radiologists require years of training before they can reliably interpret and assess medical images, an AI model can typically be trained within hours to days for a specific task. Besides assisting image interpretation, AI also holds great potential to support the training of junior physicians by providing annotation guidance [39, 40].

Medical image computing

Medical image computing aims to use computational and mathematical methods for solving problems pertaining to medical imaging data to support a wide range of healthcare activities, including diagnosis assistance, therapy planning, follow-up management, and biomedical research. Classical methods have been successfully applied to tasks such as image reconstruction, image registration, organ segmentation, and lesion classification. For instance, level-set methods have been widely explored for medical image segmentation [41, 42]. Mutual information has become one of the most popular similarity measures for multi-modal image registration [43, 44]. In the field of image reconstruction, filtered back-projection and compressed sensing have been developed to address the inverse problem, offering efficient and accurate solutions for CT reconstruction [45, 46].

Deep learning in computer vision

As mentioned above, traditional machine learning algorithms heavily rely on hand-crafted features and strong mathematical assumptions, which limit their adaptability to complex and diverse imaging tasks. Thanks to the advancement of computing capabilities and the availability of large-scale annotated datasets, deep learning has rapidly emerged as an alternative approach in medical image computing [38, 47]. Unlike conventional methods, deep learning models consist of stacked computational kernels that extract feature representations directly from raw data. These models are typically trained using gradient-based optimization methods and their extensions, such as stochastic gradient descent (SGD) and Adam [48]. Numerous studies have demonstrated that deep learning can effectively extract and leverage information from vast amounts of data, thereby enabling robust and reliable predictions. Among various deep learning architectures, convolutional neural networks (CNNs) are particularly

well-suited for medical image analysis due to their ability to capture hierarchical image features. For instance, nnUnet [49] has established a strong benchmark of medical image segmentation by providing a self-adapting framework across diverse datasets. Moreover, CNNs have also been applied to either the raw data domain or image data to accelerate image reconstruction with low noise and potentially low radiation dose [50]. In the area of image registration, CNN-based models also have demonstrated strong performance in predicting displacement vector fields [51]. Although CNNs dominate the field of computer vision and image processing, limitations such as high memory consumption and inherent difficulties in handling data with variable resolution, which is common for MRI, impede their application in radiology.

Implicit neural representations

To address the limitations of CNN, neural fields, also known as implicit neural representations (INRs) or coordinate-based networks, are increasingly popular in medical image analysis [52, 53]. Different from CNNs, the core idea of INRs is to represent a complex signal on a continuous spatial or spatiotemporal domain. A typical INR algorithm usually contains latent variables for conditioning and a Multilayer Perceptron (MLP) that takes 3D/2D coordinates as input and outputs a quantity such as an RGB intensity or a signed distance function. INRs have been used to solve a wide range of problems, including 3D reconstruction [54, 55], image super-resolution [56, 57], and deformable image registration [58]. Due to their resolution-agnostic nature and computational efficiency, INRs offer significant advantages in medical image computing [53]. Leveraging the MLP decoder, the implementation of INRs is usually not constrained by a fixed resolution [56], which makes them well-suited to medical datasets acquired using various image protocols. This architecture also eliminates the need for costly discretization and significantly reduces the memory requirements and computation consumption [59], which is particularly beneficial for real-time image reconstruction in image-guided interventions.

1.4 Challenges

Deep learning-based model outperforms traditional methods in a wide range of tasks, including lesion classification, tumor segmentation, and multi-model registration. However, the application of deep learning to head and neck MRI, particularly in the context of vestibular schwannoma management, still faces several significant challenges. First, although tumor measurements chiefly drive the management strategy of VS, there is, to my knowledge, no automated vestibular schwannoma segmentation tool that has been validated using a multicenter, multivendor dataset yet. Due to the rarity of vestibular schwannoma, it is difficult to collect a large-scale dataset. The

variability of imaging protocols across different medical centers also poses a particular challenge for model deployment. In addition, although contrast-enhanced images using GBCA remain the gold standard for the diagnosis of VS, increasing concerns, including long-term toxicity [60] and negative environmental impact [61], raise growing interest in reducing the use of GBCA. Moreover, while tumor progression is the key to vestibular schwannoma care, data-driven tumor growth prediction remains poorly studied. Previous studies on tumor growth predictions typically rely on strong mathematical assumptions, such as the reaction-diffusion model [62, 63, 64], or on advanced imaging techniques [65], which are usually not available during normal clinical routine. Existing models have been evaluated on small-scale datasets and have not been assessed in the context of longitudinal VS care. Consequently, the clinical impact of deep learning in VS management still remains unclear.

1.5 Thesis outline

This thesis aims to develop and evaluate the applicability of AI approaches to intracranial tumor diagnosis and management, particularly vestibular schwannomas, in head and neck MRI. The thesis covers three major challenges in vestibular schwannoma care: precise tumor measurement, incomplete multi-sequence MRI, and tumor growth prediction. The research topics of each chapter are summarized in Figure 1.2.

Chapter 2 presents a deep learning model to automatically segment vestibular schwannoma from gadolinium-enhanced T1-weighted and T2-weighted MRI. A retrospective study is conducted using the data acquired from multiple centers, which employ different MRI scanners and scan protocols. The applicability of the deep learning model in automatic vestibular schwannoma measurement was demonstrated by quantitative analysis and a clinical reader study.

Chapter 3 and 4 present a novel image translation model that synthesizes missing MRI sequences from the existing ones. We introduced implicit neural representations to this task, and the experiments have demonstrated the superiority of the proposed method over traditional image translation models. The model was evaluated qualitatively and quantitatively in downstream tumor segmentation tasks with missing MRI sequences (**Chapter 3**). By applying the proposed model to restore contrast-enhanced images from the low-dose images, we have shown that deep-learning image translation can also help reduce the need for contrast dose (**Chapter 4**).

Although predicting the potential growth of vestibular schwannoma is the key factor for clinical decision-making and treatment planning, it has not been well-studied. **Chapter 5 and 6** present a data-driven vestibular schwannoma growth prediction approach and its extension that predicts the volumetric tumor growth

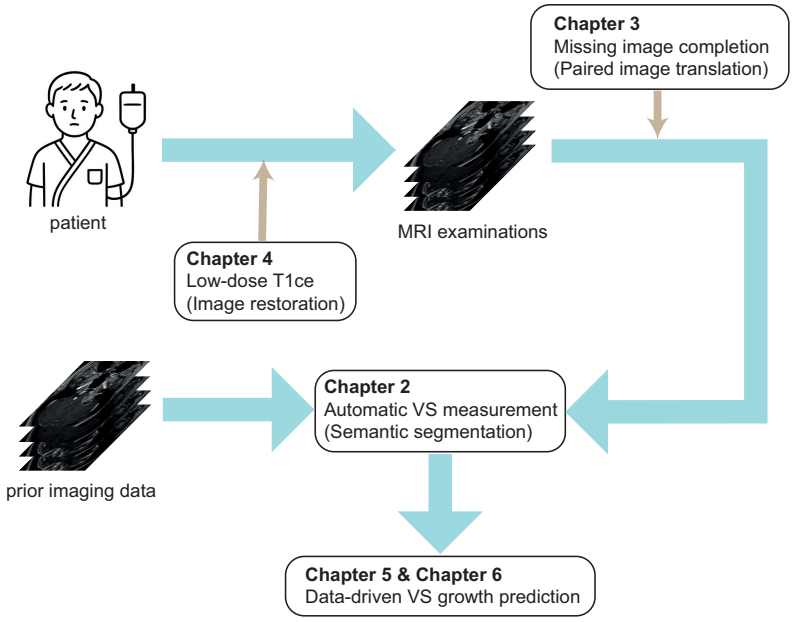


Figure 1.2: The overview of the research topics of this thesis. Patients either presenting with previously diagnosed vestibular schwannoma (follow-up scans) or exhibiting hearing loss symptoms (baseline scans) underwent multi-sequence MRI. To deal with the data acquired with various imaging protocol, we proposed an image translation model for missing image completion (**Chapter 3**). This approach also enables dose reduction in contrast-enhanced MRI by restoring low-dose T1ce (**Chapter 4**). We then developed a semantic segmentation model that automatically delineates and measures vestibular schwannoma from MRI examinations (**Chapter 2**). With the volumetric measurements obtained from the current and prior MRI examinations, we developed a deep learning model for data-driven tumor growth prediction and performed evaluation in clinical setting (**Chapter 5** & **Chapter 6**).

solely based on prior MR images. The model was developed based on implicit neural representations and recurrent neural networks. A retrospective study was conducted using a multi-center longitudinal dataset to verify the clinical impact of the proposed model.

In **Chapter 7**, we summarize the work and implications of the previous chapters, and discuss the future of AI approaches for intracranial imaging.

References

[1] F. J. Wippold. "Head and neck imaging: the role of CT and MRI". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 25.3 (2007), pages 453–465.

[2] J. C. Junn, K. A. Soderlund, and C. M. Glastonbury. "Imaging of head and neck cancer with CT, MRI, and US". In: *Seminars in nuclear medicine*. Volume 51. 1. 2021, pages 3–12.

[3] R. C. Semelka, D. M. Armao, J. Elias, and W. Huda. "Imaging strategies to reduce the risk of radiation in CT studies, including selective substitution with MRI". In: *Journal of Magnetic Resonance Imaging: an Official Journal of the International Society for Magnetic Resonance in Medicine* 25.5 (2007), pages 900–909.

[4] G. Widmann, B. Henninger, C. Kremser, and W. Jaschke. "MRI sequences in head & neck radiology–state of the art". In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Volume 189. 05. 2017, pages 413–422.

[5] F. K.-h. Lee, A. D. King, B. B.-Y. Ma, and D. K.-w. Yeung. "Dynamic contrast enhancement magnetic resonance imaging (DCE-MRI) for differential diagnosis in head and neck cancers". In: *European journal of radiology* 81.4 (2012), pages 784–788.

[6] M. Connolly and A. Srinivasan. "Diffusion-weighted imaging in head and neck cancer: technique, limitations, and applications". In: *Magnetic Resonance Imaging Clinics* 26.1 (2018), pages 121–133.

[7] R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2024.

[8] M. L. Carlson and M. J. Link. "Vestibular schwannomas". In: *New England Journal of Medicine* 384.14 (2021), pages 1335–1348.

[9] J. P. Marinelli, B. R. Grossardt, C. M. Lohse, and M. L. Carlson. "Prevalence of sporadic vestibular schwannoma: reconciling temporal bone, radiologic, and population-based studies". In: *Otology & Neurotology* 40.3 (2019), pages 384–390.

[10] C. Matthies and M. Samii. "Management of 1000 vestibular schwannomas (acoustic neuromas): clinical presentation". In: *Neurosurgery* 40.1 (1997), pages 1–10.

[11] M. L. Carlson, Ø. V. Tveiten, C. L. Driscoll, et al. "Long-term dizziness handicap in patients with vestibular schwannoma: a multicenter cross-sectional study". In: *Otolaryngology–Head and Neck Surgery* 151.6 (2014), pages 1028–1037.

[12] A. D. Sweeney, M. L. Carlson, N. T. Shepard, et al. "Congress of neurological surgeons systematic review and evidence-based guidelines on otologic and audiologic screening for patients with vestibular schwannomas". In: *Neurosurgery* 82.2 (2018), E29–E31.

[13] J. Gan, Y. Zhang, J. Wu, et al. "Current understanding of hearing loss in sporadic vestibular schwannomas: a systematic review". In: *Frontiers in Oncology* 11 (2021), page 687201.

[14] S. G. Harner, D. A. Fabry, and C. W. Beatty. “Audiometric findings in patients with acoustic neuroma”. In: *Otology & Neurotology* 21.3 (2000), pages 405–411.

[15] S. Early, C. E. Rinnooy Kan, M. Eggink, et al. “Progression of contralateral hearing loss in patients with sporadic vestibular schwannoma”. In: *Frontiers in Neurology* 11 (2020), page 796.

[16] J. E. Saunders, W. M. Luxford, K. K. Devgan, and B. L. Fetterman. “Sudden hearing loss in acoustic neuroma patients”. In: *Otolaryngology—Head and Neck Surgery* 113.1 (1995), pages 23–31.

[17] K. Kameda, T. Shono, K. Hashiguchi, et al. “Effect of tumor removal on tinnitus in patients with vestibular schwannoma”. In: *Journal of neurosurgery* 112.1 (2010), pages 152–157.

[18] G. Andersson, A. Kinnefors, L. Ekvall, and H. Rask-Andersen. “Tinnitus and translabyrinthine acoustic neuroma surgery”. In: *Audiology and Neurotology* 2.6 (1997), pages 403–409.

[19] E. Myrseth, P. Møller, T. Wentzel-Larsen, et al. “Untreated vestibular schwannoma: vertigo is a powerful predictor forhealth-related quality of life”. In: *Neurosurgery* 59.1 (2006), pages 67–76.

[20] S. K. Lloyd, A. V. Kasbekar, D. M. Baguley, and D. A. Moffat. “Audiovestibular factors influencing quality of life in patients with conservatively managed sporadic vestibular schwannoma”. In: *Otology & Neurotology* 31.6 (2010), pages 968–976.

[21] M. L. Carlson, O. V. Tveiten, C. L. Driscoll, et al. “Long-term quality of life in patients with vestibular schwannoma: an international multicenter cross-sectional study comparing microsurgery, stereotactic radiosurgery, observation, and nontumor controls”. In: *Journal of neurosurgery* 122.4 (2015), pages 833–842.

[22] G. Grinblat, M. Dandinarasaiah, I. Braverman, et al. “Large and giant vestibular schwannomas: overall outcomes and the factors influencing facial nerve function”. In: *Neurosurgical Review* 44.4 (2021), pages 2119–2131.

[23] N. S. Patel, A. E. Huang, E. M. Dowling, et al. “The influence of vestibular schwannoma tumor volume and growth on hearing loss”. In: *Otolaryngology–Head and Neck Surgery* 162.4 (2020), pages 530–537.

[24] M. L. Carlson, Ø. V. Tveiten, M. Lund-Johansen, et al. “Patient motivation and long-term satisfaction with treatment choice in vestibular schwannoma”. In: *World neurosurgery* 114 (2018), e1245–e1252.

[25] R. Goldbrunner, M. Weller, J. Regis, et al. “EANO guideline on the diagnosis and treatment of vestibular schwannoma”. In: *Neuro-oncology* 22.1 (2020), pages 31–45.

[26] M. Bailo, N. Boari, A. Franzin, et al. “Gamma Knife radiosurgery as primary treatment for large vestibular schwannomas: clinical results at long-term follow-up in a series of 59 patients”. In: *World neurosurgery* 95 (2016), pages 487–501.

[27] M. L. Carlson, E. X. Vivas, D. J. McCracken, et al. “Congress of neurological surgeons systematic review and evidence-based guidelines on hearing preservation outcomes in patients with sporadic vestibular schwannomas”. In: *Neurosurgery* 82.2 (2018), E35–E39.

[28] C. G. Hadjipanayis, M. L. Carlson, M. J. Link, et al. “Congress of neurological surgeons systematic review and evidence-based guidelines on surgical resection for the treatment of patients with vestibular schwannomas”. In: *Neurosurgery* 82.2 (2018), E40–E43.

[29] M. L. Carlson, M. J. Link, C. L. Driscoll, et al. “Working toward consensus on sporadic vestibular schwannoma care: a modified Delphi study”. In: *Otology & Neurotology* 41.10 (2020), e1360–e1371.

[30] P. Romiyo, E. Ng, D. Dejam, et al. “Radiosurgery treatment is associated with improved facial nerve preservation versus repeat resection in recurrent vestibular schwannomas”. In: *Acta Neurochirurgica* 161.7 (2019), pages 1449–1456.

[31] N. Boari, M. Bailo, F. Gagliardi, et al. “Gamma Knife radiosurgery for vestibular schwannoma: clinical results at long-term follow-up in a series of 379 patients”. In: *Journal of neurosurgery* 121.Suppl_2 (2014), pages 123–142.

[32] W. A. Friedman, P. Bradshaw, A. Myers, and F. J. Bova. “Linear accelerator radiosurgery for vestibular schwannomas”. In: *Journal of neurosurgery* 105.5 (2006), pages 657–661.

[33] B. D. Milligan, B. E. Pollock, R. L. Foote, and M. J. Link. “Long-term tumor control and cranial nerve outcomes following gamma knife surgery for larger-volume vestibular schwannomas”. In: *Journal of neurosurgery* 116.3 (2012), pages 598–604.

[34] C. Fuentealba-Bassaletti, O. M. Neve, B. F. van Esch, et al. “Vestibular complaints impact on the long-term quality of life of vestibular schwannoma patients”. In: *Otology & Neurotology* 44.2 (2023), pages 161–167.

[35] M. Peris-Celda, C. S. Graffeo, A. Perry, et al. “Beyond the ABCs: hearing loss and quality of life in vestibular schwannoma”. In: *Mayo Clinic Proceedings*. Volume 95. 11. 2020, pages 2420–2428.

[36] M. L. Carlson, Ø. V. Tveiten, C. L. Driscoll, et al. “What drives quality of life in patients with sporadic vestibular schwannoma?” In: *The Laryngoscope* 125.7 (2015), pages 1697–1702.

[37] P. L. Rigby, S. B. Shah, R. K. Jackler, et al. “Acoustic neuroma surgery: outcome analysis of patient-perceived disability”. In: *Otology & Neurotology* 18.4 (1997), pages 427–435.

[38] A. Hosny, C. Parmar, J. Quackenbush, et al. “Artificial intelligence in radiology”. In: *Nature Reviews Cancer* 18.8 (2018), pages 500–510.

[39] C. Preiksaitis and C. Rose. “Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review”. In: *JMIR medical education* 9 (2023), e48785.

[40] C. K. Boscardin, B. Gin, P. B. Golde, and K. E. Hauer. “ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity”. In: *Academic Medicine* 99.1 (2024), pages 22–27.

[41] C. Li, R. Huang, Z. Ding, et al. “A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI”. In: *IEEE transactions on image processing* 20.7 (2011), pages 2007–2016.

[42] D. Cremers, M. Rousson, and R. Deriche. “A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape”. In: *International journal of computer vision* 72.2 (2007), pages 195–215.

[43] J. P. Pluim, J. A. Maintz, and M. A. Viergever. “Mutual-information-based registration of medical images: a survey”. In: *IEEE transactions on medical imaging* 22.8 (2003), pages 986–1004.

[44] S. Klein, M. Staring, K. Murphy, et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pages 196–205.

[45] M. J. Willemink and P. B. Noël. “The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence”. In: *European radiology* 29.5 (2019), pages 2185–2195.

[46] G.-H. Chen, J. Tang, and S. Leng. “Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets”. In: *Medical physics* 35.2 (2008), pages 660–663.

[47] G. Litjens, T. Kooi, B. E. Bejnordi, et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pages 60–88.

[48] D. P. Kingma. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[49] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pages 203–211.

[50] L. R. Koetzier, D. Mastrodicasa, T. P. Szczykutowicz, et al. “Deep learning image reconstruction for CT: technical principles and clinical prospects”. In: *Radiology* 306.3 (2023), e221257.

[51] B. D. De Vos, F. F. Berendsen, M. A. Viergever, et al. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical image analysis* 52 (2019), pages 128–143.

[52] Y. Xie, T. Takikawa, S. Saito, et al. “Neural fields in visual computing and beyond”. In: *Computer Graphics Forum*. Volume 41. 2. 2022, pages 641–676.

[53] A. Molaei, A. Aminimehr, A. Tavakoli, et al. “Implicit neural representation in medical imaging: A comparative survey”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pages 2381–2391.

[54] J. J. Park, P. Florence, J. Straub, et al. “DeepSDF: Learning continuous signed distance functions for shape representation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pages 165–174.

[55] T. Amiranashvili, D. Lüdke, H. B. Li, et al. “Learning shape reconstruction from sparse measurements with neural implicit functions”. In: *International Conference on Medical Imaging with Deep Learning*. 2022, pages 22–34.

[56] Y. Chen, S. Liu, and X. Wang. “Learning continuous image representation with local implicit image function”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pages 8628–8638.

[57] J. McGinnis, S. Shit, H. B. Li, et al. “Single-subject Multi-contrast MRI Super-resolution via Implicit Neural Representations”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pages 173–183.

[58] J. M. Wolterink, J. C. Zwienenberg, and C. Brune. “Implicit neural representations for deformable image registration”. In: *International Conference on Medical Imaging with Deep Learning*. 2022, pages 1349–1359.

[59] A. Essakine, Y. Cheng, C.-W. Cheng, et al. “Where do we stand with implicit neural representations? a technical and performance survey”. In: *arXiv preprint arXiv:2411.03688* (2024).

[60] S. Coimbra, S. Rocha, N. R. Sousa, et al. “Toxicity mechanisms of gadolinium and gadolinium-based contrast agents—a review”. In: *International Journal of Molecular Sciences* 25.7 (2024), page 4071.

[61] H. M. Dekker, G. J. Stroomberg, A. J. Van der Molen, and M. Prokop. “Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents”. In: *Insights into Imaging* 15.1 (2024), page 62.

[62] T. Roque, L. Risser, V. Kersemans, et al. “A DCE-MRI driven 3-D reaction-diffusion model of solid tumor growth”. In: *IEEE transactions on medical imaging* 37.3 (2017), pages 724–732.

[63] K. C. Wong, R. M. Summers, E. Kebebew, and J. Yao. “Tumor growth prediction with reaction-diffusion and hyperelastic biomechanical model by physiological data fusion”. In: *Medical Image Analysis* 25.1 (2015), pages 72–85.

[64] K. C. Wong, R. M. Summers, E. Kebebew, and J. Yao. “Pancreatic tumor growth prediction with elastic-growth decomposition, image-derived motion, and FDM-FEM coupling”. In: *IEEE transactions on medical imaging* 36.1 (2016), pages 111–123.

[65] S. M. Schouten, D. Lewis, S. Cornelissen, et al. “Dynamic contrast-enhanced and diffusion-weighted MR imaging for predicting tumor growth of sporadic vestibular schwannomas: a prospective study”. In: *Neuro-oncology* 27.4 (2025), pages 1116–1127.

2

Fully automated 3D vestibular schwannoma segmentation with and without gadolinium-based contrast material: a multicenter, multivendor study

This chapter was adapted from:

Neve, O.M.*, Chen, Y.*, Tao, Q., Romeijn, S.R., de Boer, N.P., Grootjans, W., Kruit, M.C., Lelieveldt, B.P., Jansen, J.C., Hensen, E.F., Verbist, B.M., Staring, M., 2022. Fully automated 3D vestibular schwannoma segmentation with and without gadolinium-based contrast material: a multicenter, multivendor study. Radiology: Artificial Intelligence, 4(4), p.e210300

Abstract

Purpose

To develop automated vestibular schwannoma measurements on contrast-enhanced T1- and T2-weighted MRI.

Material and methods

MRI data from 214 patients in 37 different centers were retrospectively analyzed between 2020 and 2021. Patients with hearing loss (134 positive for vestibular schwannoma [mean age ± SD, 54 ± 12 years; 64 men] and 80 negative for vestibular schwannoma) were randomly assigned to a training and validation set and to an independent test set. A convolutional neural network (CNN) was trained using five-fold cross-validation for two models (T1 and T2). Quantitative analysis, including Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error, was used to compare the computer and the human delineations. An observer study was performed in which two experienced physicians evaluated both delineations.

Results

The T1-weighted model showed state-of-the-art performance with a mean S2S distance of less than 0.6 mm for the whole tumor and the intrameatal and extrameatal tumor parts. The whole tumor Dice index and Hausdorff distance were 0.92 and 2.1 mm in the independent test set. T2-weighted images had a mean S2S distance less than 0.6 mm for the whole tumor and the intrameatal and extrameatal tumor parts. THe whole tumor Dice index and Hausdorff distance were 0.87 and 1.5 mm in the independent test set. The observer study indicated that the tool was similar to human delineations in 85-92% of cases.

Conclusion

The CNN model detected and delineated vestibular schwannomas accurately on contrast-enhanced T1 and T2-weighted MRI and distinguished the clinically relevant difference between intrameatal and extrameatal tumor parts.

2.1 Introduction

Vestibular schwannomas are rare, benign intracranial tumors arising from the neurilemma of the vestibular nerve. Initial symptoms usually comprise hearing loss, tinnitus, and balance disturbance. Approximately 60% of tumors show no or minimal progression over time, and 40% either are very large at presentation or show progression during follow-up [1]. Small- to medium-sized tumors are not life-threatening and are generally conservatively managed, at least initially, using surveillance with repeated MRI examinations. Conversely, patients with large tumors at presentation or with tumors that progress during follow-up may need intervention through radiation therapy or surgery. There are no reliable predictors for tumor progression.

Tumor progression is determined according to the extrameatal manual diameter measurements at subsequent MRI examinations [2]. However, these two-dimensional (2D) measurements have considerable error, resulting in inter- and intraannotator differences of 10-40 % [3, 4, 5]. The more accurate three-dimensional (3D) volume measurements have not been widely applied in clinical practice since these measurements are time-consuming [3, 4, 5, 6].

To address this problem, several automated segmentation tools have been developed in recent years [7, 8, 9]. The reported tools were trained for volume measurement of vestibular schwannoma on gadolinium-enhanced T1-weighted MRI and sometimes additional T2-weighted MRI. These tools are increasingly based on deep learning methods, which yield state-of-the-art performance in many vision tasks, including medical image segmentation. Deep convolutional neural networks (CNNs), particularly the U-Net architecture, can reach expert-level performance in various organ segmentation tasks from clinical MRI [8]. Although many variants of the UNet have been proposed and demonstrated task-specific improvements, recent insights suggest that rather than the architecture, careful selection of the hyperparameters and training strategy can have an important effect on performance [9]. The no-new-UNet framework, abbreviated nnU-Net, indeed demonstrated this for several organs and imaging modalities [10, 11]. As such, we propose the application of nnU-Net to address vestibular schwannoma segmentation in our clinical setting.

This study aimed to develop a deep learning CNN model to automatically detect and segment vestibular schwannoma in 3D from T2-weighted and gadolinium-enhanced T1-weighted MRI, acquired from multiple centers using different MRI scanners and scan protocols. We additionally carried out a carefully designed observer study, based on the concept that the radiologists' visual observation of the segmentation results can be a direct, important evaluation of segmentation quality. In addition to conventional

Table 2.1: Characteristics of Patients with Vestibular Schwannoma

Characteristic	value
No. of Patients	134
Mean age(y)	54 ± 12
Men	64 (48%)
Cystic component	63 (47%)
Tumor size	
Intrameatal only	28 (21%)
Small (0-10 mm)	19 (14%)
Medium (11-20 mm)	26 (19%)
Moderately large (21-30 mm)	24 (18%)
Large (31-40 mm)	24 (18%)
Giant (>40 mm)	13 (10%)

Note: Data presented with a plus/minus sign are the means ± SDs. Other data are presented as numbers of patients, with percentages in parentheses.

metrics, the observer study highlights the applicability of our model in a clinical setting.

2.2 Materials and methods

This retrospective study was performed at the Leiden University Medical Center, a tertiary referral center for vestibular schwannoma, in 2020-2021. The institutional review board approved the study protocol (G19.115) and waived the obligation to obtain informed consent.

Patients and Data

In total, 214 patients who underwent MRI because of hearing loss were included in the study, with 134 patients who were vestibular schwannoma positive (mean age, 54 ± 12 years; 64 men) and 80 who were vestibular schwannoma negative. The selection of patients with vestibular schwannoma included a wide spectrum of patient and tumor characteristics, such as patient age, sex, tumor size, and tumor consistency. All positive patients were adults with a unilateral vestibular schwannoma and at least one gadolinium-enhanced T1-weighted MRI examination. High-resolution T2-weighted images were available in 112 patients. MRI scans obtained after surgery or irradiation were excluded. Available MRI examinations were originally performed in 37 different hospitals with 12 different MRI scanners from three major MRI vendors. The MRI scans of negative cases, included to optimize detection performance, were solely acquired at the Leiden University Medical Center in adult patients with hearing

Table 2.2: Technical Information of Patients with Vestibular Schwannoma

Technical MRI features	Contrast-enhanced T1-weighted MRI	T2-weighted MRI
No. of patients	134	112
In-plane resolution (mm)	0.35×0.35 $(0.27 \times 0.27 - 1.0 \times 1.0)$	0.29×0.29 $(0.23 \times 0.23 - 0.70 \times 0.70)$
In-plane matrix	400×400 $(256 \times 208 - 560 \times 560)$	512×512 $(256 \times 192 - 768 \times 652)$
TE (ms)	9 $(2.38 - 20)$	200 $(1.53 - 297)$
TR (ms)	602.10 $(8.76 - 2200)$	2400 $(4.47 - 5000)$
Section thickness (mm)	1.0 $(0.9 - 5.0)$	0.6 $(0.5 - 1.8)$

Note: Unless otherwise noted, data are presented as medians, with ranges in parentheses. TE = echo time, TR = repetition time.

loss before cochlear implantation, and provided no demographic data because of previous anonymization. Patients' characteristics and technical information are shown in Table 2.1 and Table 2.2, respectively.

In positive cases, the intra- and extrameatal components [2] and the whole tumor were manually delineated by two annotators independently (O.M.N., a physician with 3 years of experience, and S.R.R., a technical physician with 2 years of experience) with gadolinium-enhanced T1-weighted MRI, supervised and, when necessary, corrected by a senior head-and-neck radiologist (B.M.V.). Two senior radiologists (M.C.K. and B.M.V., with 18 and 21 years of experience) trained both annotators. Delineation was performed using Vitrea software, v7.14.2.227 (Vital Images Inc., Minnetonka, MN, USA). The delineation was automatically propagated to T2-weighted MRI after rigid image registration using Elastix [12, 13]. The complete data set was split into a training and validation set (80% from 26 centers), and an independent test set (20% from 11 different centers) on which the model was not trained, see Figure 2.1 for details. This was done to mimic clinical deployment, in which new cases may be slightly different from the data seen in the training phase and possibly bear an unknown distribution shift [14].

Furthermore, the publicly available data set by Shapey et al. [15] was used as an additional external test of the contrast-enhanced T1-weighted model (n=242). This dataset contained 47 post-surgery scans, which were omitted from the analysis.

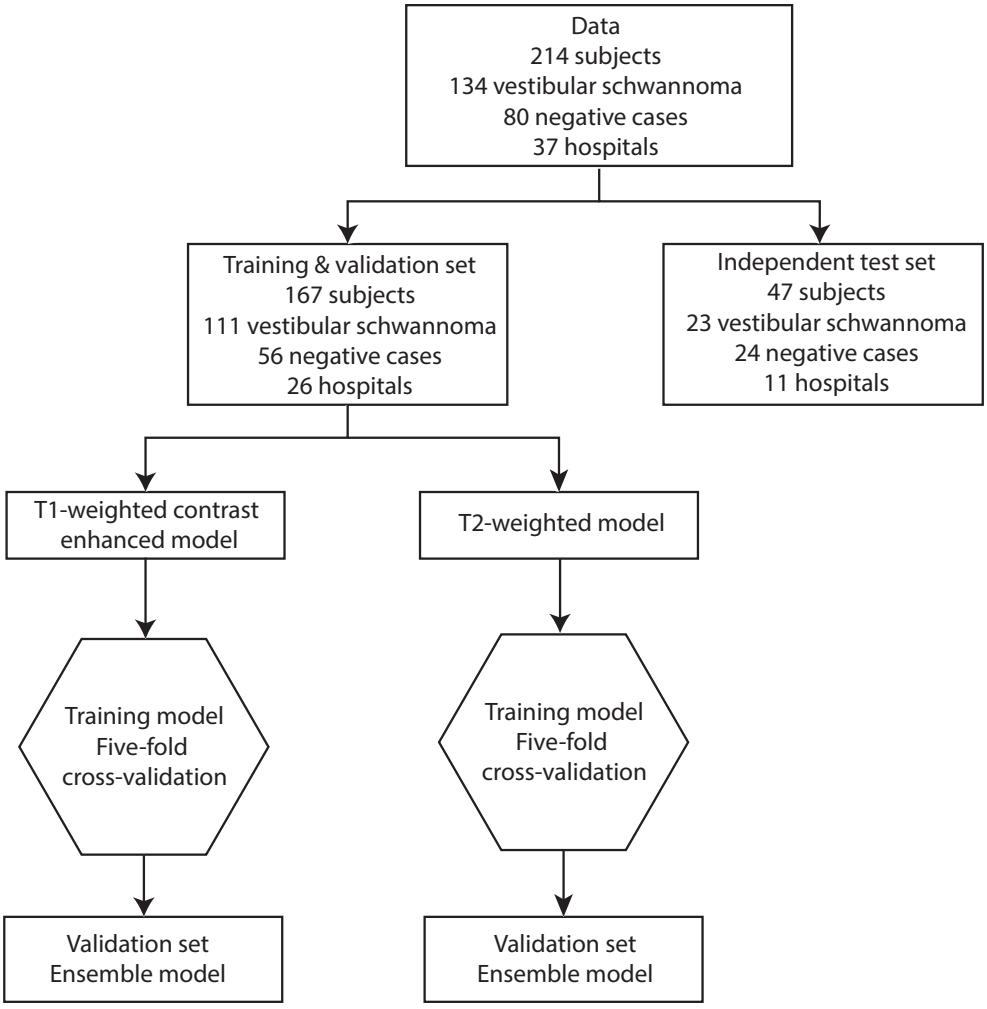


Figure 2.1: Flowchart of data. Patients were randomly assigned to the training and validation set (80%) and the independent test set (20%). Positive cases were randomly assigned on the basis of the hospital where the scan was acquired, so the independent test set contained data from 11 hospitals that were not used to train the algorithm. For training and validation, five-fold cross-validation was used. The mean of the five models is the ensemble model. This ensemble model was evaluated in the independent test set.

CNN Architecture and Training

NnUNet is a deep learning-based segmentation method that automatically selects one of three network architectures, includes preprocessing and postprocessing methods, and performs automatic tuning of hyperparameters [10]. In this study, a 3D U-Net

with five encoder and decoder layers was selected, using randomly cropped 3D image patches of size $320 \times 320 \times 20$ voxels as network input during training. The network was trained as a multi-class segmentation task to automatically segment both the intra- and extrameatal components of the tumor. Two 3D nnU-Nets were trained (one for contrast-enhanced T1 and one for T2-weighted MRI) from scratch with He initialization. Five-fold cross-validation was used, generating five models that were merged by averaging the softmax scores. To deal with multi-center settings, z-scoring normalization was performed on each image independently. All the training images were then resampled to the median spacing of the training dataset using third-order spline interpolation. Training was performed on an NVIDIA Tesla V100 graphics processing unit with 16 GB memory using the PyTorch (v1.7.1) library.

Observer Study

An observer study was performed to test whether the CNN could perform as well as human delineation on contrast-enhanced T1-weighted images. The T1-weighted annotations were propagated to T2-weighted MRI; therefore, the observer study was conducted only for the T1-weighted images. A user interface was created (Figure. 2.2), showing a gadolinium-enhanced T1-weighted image and the registered T2-weighted image in the top row and the human and automatic delineation in random order on the bottom row, projected on the gadolinium-enhanced T1-weighted scan. Observers could scroll through the MRI, manually adjust its brightness and contrast, and toggle the segmentations on and off for optimal assessment. The observers were a head-and-neck radiologist (B.M.V.) and a skull-base otorhinolaryngologist (E.F.H., with 18 years of experience), blinded to case information and delineation type (human or automated). The observers were asked to rate and compare the two delineations by answering two separate questions about the intra- and extrameatal part and the whole tumor: (a) Which delineation is better (annotation 1, annotation 2, or comparable), and (b) Is the annotation quality satisfactory (yes or no). In a consensus meeting, cases in which observers did not agree were discussed. The consensus results are presented in the section on outcomes of the observer study.

Testing and Statistical Analysis

All test images were resampled in the same way as the training data, and a sliding window approach was used to predict images with a window size of $320 \times 320 \times 10$ voxels, which is the same as the network's input size. The step size is half of the window size, and a Gaussian weighted function was applied in aggregating the predictions. To eliminate false detection, connected component-based postprocessing was performed. Only the largest connected component in the predictions was kept. Tumor detection

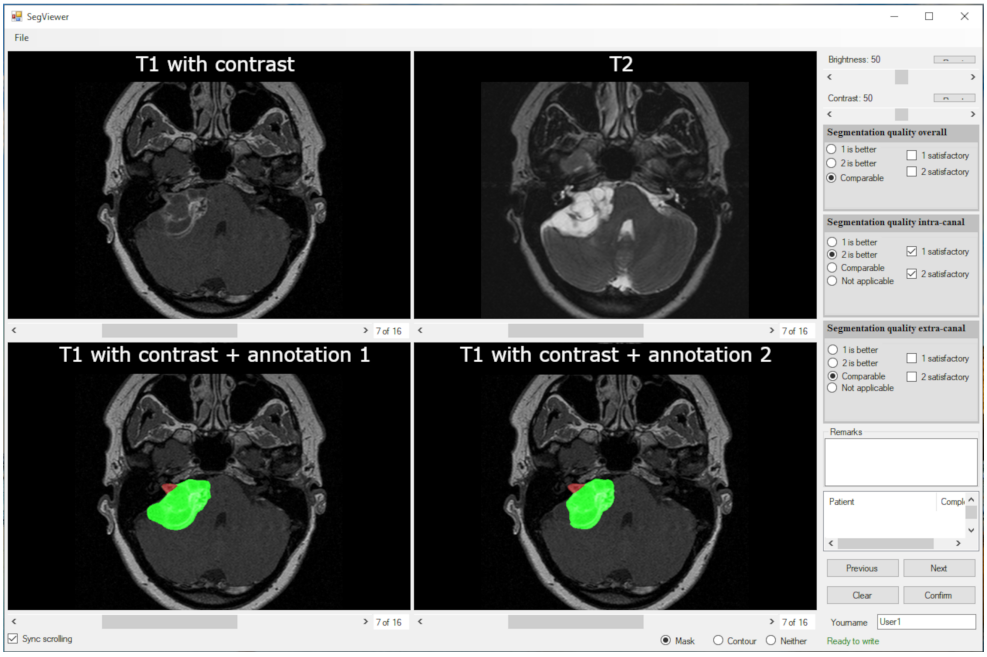


Figure 2.2: Observer study interface. The top row shows the clean, gadolinium-enhanced T1-weighted MRI and T2-weighted MRI. The bottom row shows the convolutional neural network and human annotations, randomized to the left and the right pane, respectively. The multiple-choice questions for each observer are shown on the right side of the interface. The observers could additionally add free-text comments.

by the CNN was defined as at least one voxel being detected. The performance was evaluated using the Dice index (measuring overlap of the delineations), 95th percentile Hausdorff distance (indicating the maximum distance between delineations), surface-to-surface (S2S) distance (indicating the mean distance between delineations), and the relative volume error (RVE) (indicating the difference in volume in percentage). One of the annotator' s (O.M.N., annotator 1) delineations were used for training and quantitative evaluation. The results were plotted in box-and-whisker plots. Furthermore, interannotator variability was investigated. Differences between the prediction performance of each annotator and the interannotator variability were tested using the Wilcoxon signed-rank test. In addition, a post hoc analysis of T1-model performance was conducted with respect to tumor size, according to the classification by Kanzaki et al. [2]. To avoid group sizes that were too small per category, the validation and test sets were pooled, and a Kruskal-Wallis test was performed. P values less than .05 were considered to indicate statistically significant differences. Observer agreement before the consensus meeting on the satisfactory degree for segmentation

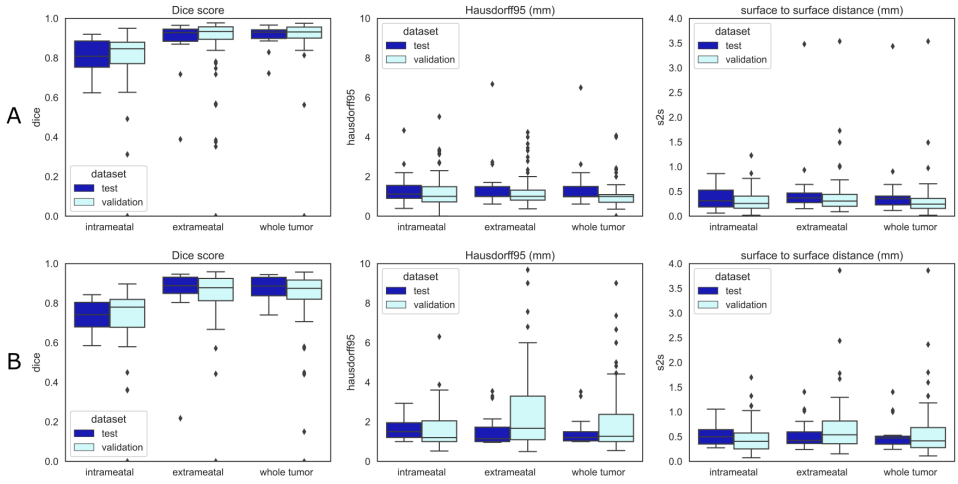


Figure 2.3: Quantitative boxplots of convolutional neural network tumor segmentation performance. The Dice, 95 % Hausdorff (Hausdorff95) distance, and surface-to-surface distance (S2S) measures are shown from left to right. (A) Boxplots of the contrast-enhanced T1 model. (B) Results of the T2-weighted model. Validation set results are shown in sky blue and independent test set results in dark blue.

and human delineation was expressed as percentage agreement. All analyses were performed in Python (v3.8.2) with NumPy (v1.20.2), SciPy (v1.3.3), and the sklearn (v0.23.2) library.

2.3 Results

The CNN detected tumors with 100 % sensitivity and 99.1 % specificity for the validation set and 100 % sensitivity and 100 % specificity for the test set. The algorithm calculated the segmentation with a median runtime of 78 seconds per patient.

Performance with Contrast-enhanced T1-weighted MRI

The results of the CNN on contrast-enhanced T1-weighted MRI are shown in Table 2.3 and Figure 2.3A. S2S distance of the whole tumor was 0.31 ± 0.36 mm in the validation set and 0.47 ± 0.67 mm in the independent test set. These S2S distances are around the in-plane voxel size and lower than the slice thickness. The whole tumor Hausdorff distance in the independent test set was 2.10 ± 3.34 mm; it was 1.34 ± 0.84 mm and 2.18 ± 3.43 mm in the intra- and extrameatal parts, respectively. All the median Hausdorff distances were below the 2 mm threshold, which is often used in clinical practice to define 2D growth [1]. T1 model performance on the independent test set

Table 2.3: Quantitative Results of Contrast-enhanced T1-weighted Model

Variable	Dice		95% Hausdorff (mm)		S2S (mm)		RVE (%)	
	Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
Validation set								
Whole tumor	0.91 ± 0.10	0.93	1.13 ± 1.45	1.00	0.31 ± 0.36	0.24	7.59 ± 8.10	4.88
Intrameatal	0.78 ± 0.21	0.85	1.26 ± 0.78	1.00	0.31 ± 0.20	0.26	19.7 ± 43.5	11.5
Extrameatal	0.83 ± 0.26	0.93	1.43 ± 1.67	1.00	0.41 ± 0.43	0.31	12.0 ± 21.6	4.94
Independent test set								
Whole tumor	0.92 ± 0.05	0.93	2.10 ± 3.34	1.00	0.47 ± 0.67	0.36	10.2 ± 9.1	7.1
Intrameatal	0.81 ± 0.08	0.81	1.34 ± 0.84	1.12	0.37 ± 0.23	0.32	14.7 ± 14.8	6.8
Extrameatal	0.89 ± 0.12	0.93	2.18 ± 3.43	1.00	0.52 ± 0.68	0.37	12.1 ± 16.9	6.5
Publicly available dataset by Shapey et al. [15]								
Whole tumor	0.88 ± 0.04	0.88	1.31 ± 0.22	1.30	0.39 ± 0.12	0.37	27.6 ± 11.9	26.1

Note: Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error (RVE) of the model compared with annotator 1 in the validation set, independent test set, and publicly available dataset by Shapey et al. [15]. The publicly available dataset seems to have structurally smaller ground truths, as can be seen in Figure 2.11 (Appendix).

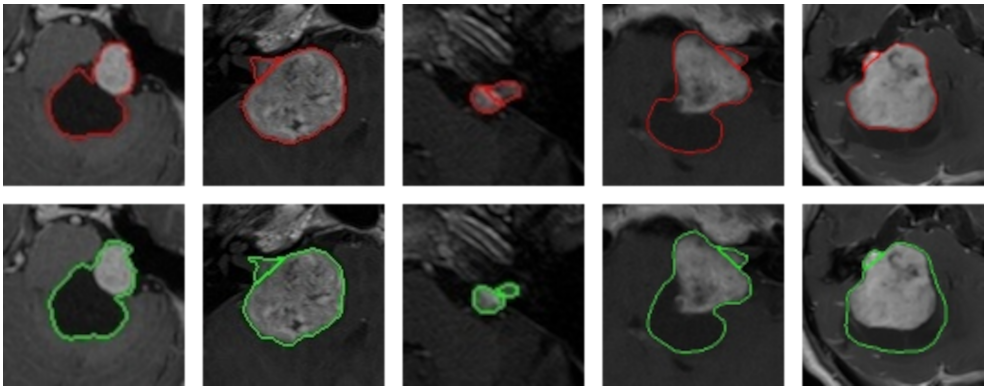


Figure 2.4: Examples of cystic, large, small vestibular schwannoma whole tumor annotations, including the separation between the intra- and extrameatal tumor parts, of contrast-enhanced T1-weighted MRIs. The top row shows the convolutional neural network (CNN) predictions in red, and the bottom row shows the delineation of annotator 1 in green. The first, fourth, and fifth tumors are potentially hard to delineate for the CNN due to the large peripheral cystic tumor parts. The Dice scores of these patients were 0.96, 0.96, 0.91, 0.93, and 0.72, respectively, and the surface-to-surface distances (mm) were 0.39, 0.21, 0.24, 0.35, and 3.44, respectively.

was similar to the results in the validation set, indicating robust external validity. Remarkably, the independent test set had higher mean Hausdorff properties compared to the median due to two outliers (cystic tumor) in the test set that influenced the Hausdorff distance and its standard deviation. Dice indices for the whole tumor were above 0.91 ± 0.10 and 0.92 ± 0.05 in both sets, and RVE was $7.6 \pm 4.9\,\%$ and $10.2 \pm 9.1\,\%$, with lower values for the intra- and extrameatal parts of the tumor due to the sensitivity of Dice and RVE to small volumes. Figure 2.4 shows some examples of the T1 model compared with annotator 1.

The CNN model, when applied to the publicly available dataset of Shapey et al. [15], performed at the same level as with the independent test set, with a mean Dice index of 0.88 ± 0.04, a mean Hausdorff distance of $1.31 \pm 0.22\,\text{mm}$, a mean S2S distance of $0.39 \pm 0.12\,\text{mm}$, and an RVE of $26.0 \pm 11.9\,\%$.

Performance with T2-weighted MRI

The results of the whole tumor and the intra- and extrameatal parts are summarized in Table 2.4 and Figure 2.3B. S2S distances ranged between 0.46 ± 0.28 and 1.00 ± 3.75 for all tumor parts in both data sets. Hausdorff distance of the whole tumor in the validation set was $3.12 \pm 9.28\,\text{mm}$, with a smaller value in the independent test set ($1.52 \pm 0.76\,\text{mm}$). Whole tumor Dice indexes were 0.82 ± 0.19 and 0.87 ± 0.06, and

Table 2.4: Quantitative Results of T2-weighted Model

Variable	Dice		95% Hausdorff (mm)		S2S (mm)		RVE (%)	
	Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
Validation set								
Whole tumor	0.82 ± 0.19	0.87	3.12 ± 9.28	1.27	1.00 ± 3.75	0.42	24.5 ± 98.9	7.60
Intrameatal	0.69 ± 0.23	0.78	1.60 ± 0.95	1.20	0.46 ± 0.28	0.40	14.5 ± 18.7	8.39
Extrameatal	0.77 ± 0.28	0.88	2.70 ± 3.19	1.67	0.82 ± 1.01	0.54	30.9 ± 73.3	18.5
Independent test set								
Whole tumor	0.87 ± 0.06	0.89	1.52 ± 0.76	1.21	0.54 ± 0.31	0.47	12.1 ± 10.8	9.01
Intrameatal	0.74 ± 0.08	0.74	1.64 ± 0.59	1.50	0.52 ± 0.20	0.50	12.6 ± 21.2	5.27
Extrameatal	0.85 ± 0.17	0.89	1.60 ± 0.92	1.14	0.56 ± 0.33	0.42	22.3 ± 14.9	20.0

Note: Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error (RVE) of the model compared with annotator 1 in the validation set and independent test set.

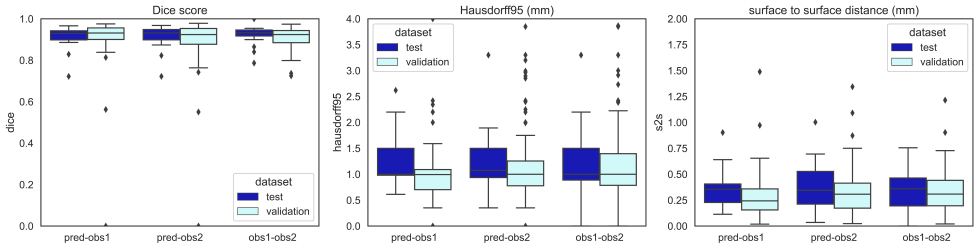


Figure 2.5: Quantitative measures of whole tumor convolutional neural network performance compared with the two annotators on contrast-enhanced T1-weighted MRI scans. Interannotator variability is also shown (obs 1- obs 2). The Dice indices, 95% Hausdorff distance (Hausdorff95), and surface-to-surface (S2S) distance boxplots are shown. The validation set results are shown in sky blue, and the independent test set in dark blue. The box extends from the first to the third quartile, with a line at the median. Whiskers extend from the box 1.5 times the interquartile range. Data points outside the whiskers were plotted individually. pred = CNN prediction, obs = observer.

RVE values ranged from $12.1 \pm 10.8\,\%$ and $24.5 \pm 98.8\,\%$ in both data sets. Intrameatal tumors had worse Dice indexes (0.69 ± 0.23 and 0.74 ± 0.08), and RVE ($14.5 \pm 18.7\,\%$ and $12.6 \pm 21.2\,\%$), likely due to the low contrast between the tumor and adjacent petrous bone in T2-weighted images. Overall T2 performance was slightly degraded compared to postcontrast T1. However, S2S distances below $1\,\text{mm}$ indicate acceptable performance.

Inter-annotator Variability

Comparisons between the T1-weighted model and the two annotators and between the two annotators are shown in Table 2.5 and Figure 2.5. The comparison between both annotators shows the whole tumor interannotator variability, resulting in a Dice index around 0.91 and RVE of 7-9 %. When the model was compared to each annotator in both datasets, S2S distances were similar and below $0.5\,\text{mm}$. The model was trained on annotator 1, but the results compared with annotator 2 are similar for all quantitative measures.

Performance by Tumor Size

In the supplemental material (Figure 2.10), the results of the performance per size category are shown. Whole tumor results show a pattern of higher Dice indexes for larger tumors, which was expected because the Dice index is sensitive to size. S2S was very similar in all size groups ($<0.5\,\text{mm}$), although S2S was slightly greater in

Table 2.5: Comparison of the Model with Annotators and Interannotator Variability

Variable	Dice			95% Hausdorff (mm)			S2S (mm)			RVE (%)		
	Mean ± SD	*P*	Median	Mean ± SD	*P*	Median	Mean ± SD	*P*	Median	Mean ± SD	*P*	Median
Validation set												
CNN – ann 1	0.91 ± 0.10	< .001	0.93	1.13 ± 1.45	< .001	1.00	0.31 ± 0.36	< .001	0.24	7.59 ± 8.10	.21	4.88
CNN – ann 2	0.90 ± 0.11	.40	0.92	1.33 ± 1.52	.18	1.00	0.36 ± 0.36	.58	0.31	10.1 ± 9.8	.35	7.1
ann 1 – ann 2	0.91 ± 0.05	N/A	0.92	1.27 ± 0.82	N/A	1.00	0.34 ± 0.20	N/A	0.31	9.01 ± 9.14	N/A	6.40
Independent test set												
CNN – ann 1	0.92 ± 0.05	.56	0.93	2.10 ± 3.34	.83	1.00	0.48 ± 0.67	.67	0.35	10.2 ± 9.1	.28	7.1
CNN – ann 2	0.91 ± 0.05	.69	0.93	2.08 ± 3.41	.94	1.07	0.50 ± 0.68	.96	0.35	9.69 ± 9.19	.57	7.72
ann 1 – ann 2	0.92 ± 0.04	N/A	0.93	1.20 ± 0.65	N/A	1.00	0.34 ± 0.19	N/A	0.36	6.93 ± 5.32	N/A	4.53

Note: Dice index, Hausdorff distance, surface-to-surface distance (S2S), and relative volume error (RVE) of the model compared with annotator (ann) 1, annotator 2, and both annotators of the contrast-enhanced T1-weighted model. Results of the validation set and the independent test set are shown. CNN = convolutional neural network. *P* values denote the Wilcoxon signed rank test between this quantitative score and the corresponding score of annotator 1-annotator 2 (the third row).

larger tumors ($P < 0.001$). Results of intra- and extrameatal tumor parts show stable performance, except for four outliers in the small tumors (inaccurate extrameatal segmentation) and three outliers in giant tumors (false intrameatal tumor detection). In these tumors, there were some differences between model and human delineation for a completely intrameatal tumor with or without a tiny extrameatal part (small) or an extrameatal tumor with or without an intrameatal part (giant).

Outcomes of Observer Study

Agreement between the two observers before the consensus meeting on whole tumor segmentation quality was 131 of 134 (98%) for the human annotators and 127 of 134 (95%) for the CNN.

CNN segmentations of the whole tumor were considered comparable to the human segmentations in 103 of 111 (93%) of cases in the validation set and 20 of 23 (87%) in the test set. The CNN segmentations were rated better than the human segmentations in 2 of 111 (2%) and 2 of 23 (9%) of cases in the two datasets, respectively. Intrameatal segmentations were rated as similar to or better than human segmentations in 100 of 106 (94%) and 22 of 23 (96%) in the validation and test sets, respectively. For extrameatal segmentations, these percentages were 83 of 97 (86%) and 18 of 22 (82%).

In addition, the observers considered 104 of 111 (94%, validation set) and 20 of 23 (87%, test set) of whole tumor CNN segmentations satisfactory. Intrameatal tumor parts were considered satisfactory in 100 of 104 (94%, validation set) and 22 of 23 (96%, test set) of segmentations. Extrameatal tumor parts were considered satisfactory in 90 of 97 (93%, validation set) and 18 of 22 (82%) (test set) of segmentations. For human segmentations of the intrameatal tumor, 98 of 104 (94%) in the validation and 23 of 23 (100%) in the test set were rated satisfactory. Other satisfaction levels of the human segmentations were 110 of 111 (99%, validation set) and 22 of 23 (96%) (test set) for the whole tumor and 89 of 97 (92%, validation set) and 21 of 22 (95%, test set) for the extrameatal tumor part.

2.4 Discussion

To our knowledge, this is the first study of a multicenter, multivendor automated segmentation tool for vestibular schwannoma. The developed 3D CNN tool measured tumor volume with high accuracy on contrast-enhanced T1-weighted MRI scans and T2-weighted MRI scans. The S2S distances were between 0.4 mm and 0.9 mm, which was lower than the median section thickness of 1.0 mm. The observer study suggests that the tool performs similarly to human delineation in 87-93% of the cases.

The contrast-enhanced T1-weighted MRI model provided excellent S2S distances and Dice indexes. However, the standard deviations of the Hausdorff distances were remarkably large in the test set because of two outliers, which contained peripheral cysts in the extrameatal part. The model did have difficulties with tumors containing large peripheral cysts (see supplemental material Figure 2.6 and Figure 2.7 for examples), which were sometimes partially included by the model.

Evaluation of the model on the publicly available dataset of Shapey et al. [15] showed robust performance on contrast-enhanced T1-weighted images. The ground-truth delineations of Shapey et al. are smaller than those used in the current study, as shown in supplemental Figure 2.11, reducing Dice index from 0.93 to 0.88 [7]. When erosion (3×3 kernel) was performed on model delineation, the Dice index improved again to 0.93 ± 0.03, supporting this observation. The delineations by Shapey et al. were used for radiotherapy purposes, where preventing damage to the surrounding tissue is important, warranting conservative delineation. We did not compare the T2-weighted images of the publicly available dataset to those in our dataset, given differences in the imaging characteristics (echo time and repetition time) and region of interest (whole brain vs. cerebellopontine angle region).

In our study, CNN performance on T2-weighted MRI was slightly less accurate, with more uncertainty, compared with the contrast-enhanced T1-weighted images. This was particularly the case in polycystic tumors, where the tumor border was hard to distinguish from the cerebrospinal fluid solely on T2 (Figure 2.8 and Figure 2.9). In one case, the model could not distinguish a small tumor obliterating the internal meatus. In another single case, the model detected the contralateral eye as a false-positive volume outside the region of interest.

The RVE values of the whole tumor were 8-12 %, compared with 9-10 % inter-annotator volume differences. Only the T2 model in the validation set had a larger RVE of 25 %. The performance of our CNN compared with human volume measurement is below previously reported interannotator variabilities ranging from 15-20 % [3, 4, 5], and also below the generally accepted threshold of 20 % before volume increase is considered growth. Use of 2D measurements is advised in the consensus guidelines, but these measurements have high intra-observer variabilities ranging from 10-40 % [2, 3, 4, 5]. Volume measurement is more accurate; in addition, the proposed tool can reduce the workload, which has been a barrier for clinical adoption, enabling the shift from 2D measurement. Since documented detection and evaluation of tumor growth are main factors that indicate the need for treatment, be it surgical removal or irradiation, this is of notable clinical relevance.

A distinct attribute in vestibular schwannoma research is the integration of an

observer study. Determining a ground truth is necessary in artificial intelligence imaging studies. The reliability of the ground truth is uncertain when human observer performance is suboptimal, as described above. Our observer study allowed evaluation of the comparability between CNN segmentation and human segmentation, the reference standard. Our results showed that the CNN tool performs similarly to human observers in most cases, supporting the quantitative results that the tool is feasible and robust for use in clinical practice. Whole tumor delineations performed slightly better than the extrameatal delineations, which should be considered when the tool is used in clinical practice because extrameatal tumor progression is of particular interest for treatment decisions.

Previously proposed artificial intelligence tools for vestibular schwannoma segmentation were performed on data from a single center [5, 6, 7]. In clinical practice, however, diagnostic and follow-up scans are often obtained in different centers using a variety of scanners and MRI protocols. In addition to its documented performance in a multicenter, multivendor setting, our method contains three features that make the tool more suitable for clinical practice compared with previous automated vestibular schwannoma delineation methods. First, the tool can distinguish between the intra- and extrameatal parts of the tumor. This distinction is important for clinical decision-making, as extension and progression of the extrameatal part usually determine the need for intervention. For this reason, current tumor staging systems are based mainly on the extrameatal dimensions of the tumor, while the intrameatal part is not measured [2, 16]. Second, the proposed tool can also delineate on solely T2-weighted MRI. Given the ongoing debate on the use of gadolinium-based contrast material, this is a valuable feature [17]. Third, unlike previous models, our network is a fully 3D network that enables complete use of intersection information.

This study had some inherent limitations. First, the study was performed using retrospective MRI data. Although this is an accepted method for the development of a new tool, some bias may be introduced by using older MRI examinations with suboptimal image quality and resolution. Therefore, accuracy and efficacy should also be investigated in prospective studies before clinical implementation and use. Second, for training of the T2 model, the registered human T1 delineations were used. This might have resulted in a suboptimal ground truth for the T2 model, although the reported tumor size correlations between T1 and high-resolution T2 were high [18, 19]. Third, the model is only trained on data before treatment and cannot be used for follow-up after surgery or radiation therapy without retraining.

Implementation of the CNN tool in clinical practice could lead to more accurate volume measurements of vestibular schwannoma at diagnosis and during follow-up, while reducing the workload of radiologists. Tumor volume change over time is a

decisive factor in clinical decision-making, and future research should focus on the performance of the tool in a prospective study and its effect on clinical practice. The tool might be improved using postprocessing to reduce the false-positive volumes outside the region of interest. In addition, the algorithm used for the development of the tool could be adapted to analyze other slow-growing skull-base pathologies, such as meningiomas, that are typically approached by a wait-and-scan policy [20].

The proposed CNN model delineated vestibular schwannoma from MRI with excellent accuracy, similar to human performance in most cases. The CNN tool made the clinically relevant distinction between intra- and extrameatal tumor parts. The study shows the feasibility of automatically detecting and evaluating vestibular schwannoma with or without contrast material administration in large datasets acquired from multiple medical centers and MRI vendors.

2.5 Acknowledgements

This work was supported by a strategic fund of the Leiden University Medical Center and the China Scholarship Council (grant 202008130140).

Appendix

To better understand how our method performs in different modalities and different kinds of tumors, we demonstrate ten selected test samples for each model and their corresponding prediction and human annotations, including outliers and well-performing cases. In each sample, we selected five slices for demonstration. Predictions are shown in red, and corresponding human annotations are shown in green.

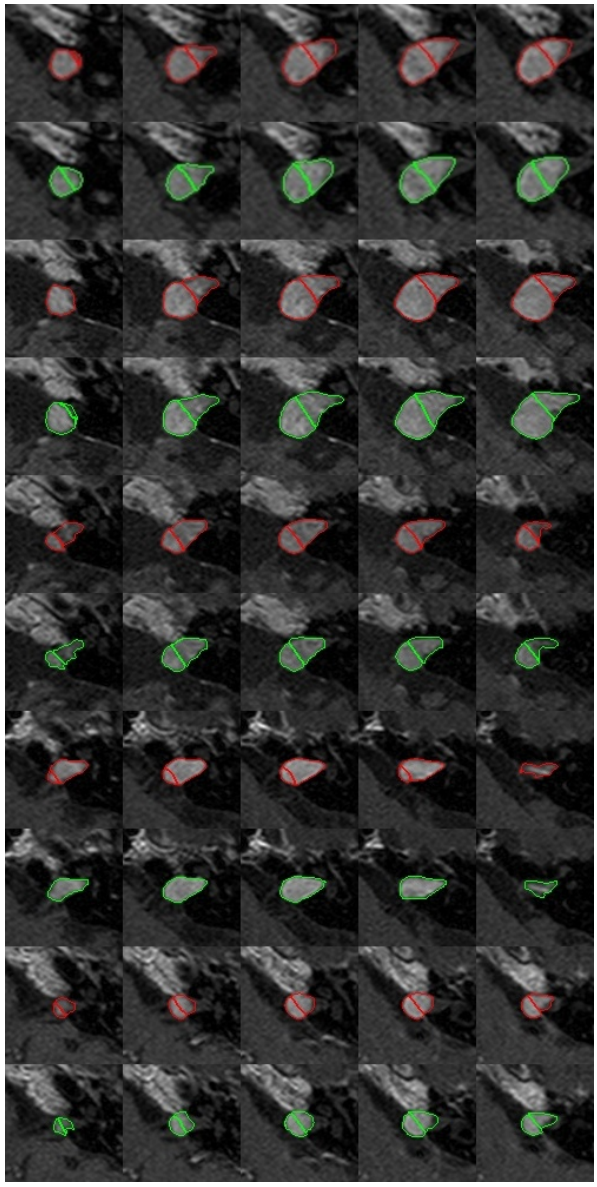


Figure 2.6: Examples of tumor segmentation of contrast T1-weighted images. Odd rows show the CNN prediction of several sections from each patient in red, and even rows show the corresponding delineation of annotator 1 in green. The Dice scores of these patients are 0.95,0.96, 0.94, 0.93, and 0.92 respectively, and the surface-to-surface distances (mm) are 0.11, 0.09, 0.10, 0.10, and 0.11 respectively.

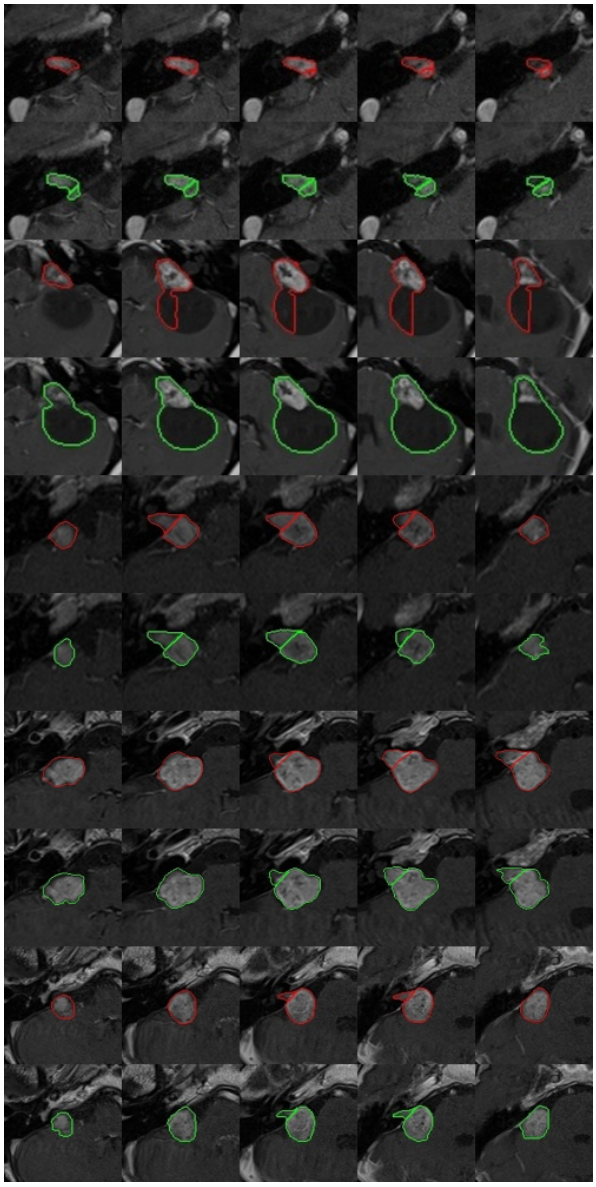


Figure 2.7: Examples of tumor segmentation of contrast T1-weighted images. Odd rows show the CNN prediction of several sections from each patient in red, and even rows show the corresponding delineation of annotator 1 in green. The dice scores of these patients are **0.86**, **0.56**, **0.93**, **0.95**, and **0.96** respectively, and the surface-to-surface distances (**mm**) are **0.34**, **3.54**, **0.11**, **0.10**, and **0.10** respectively.

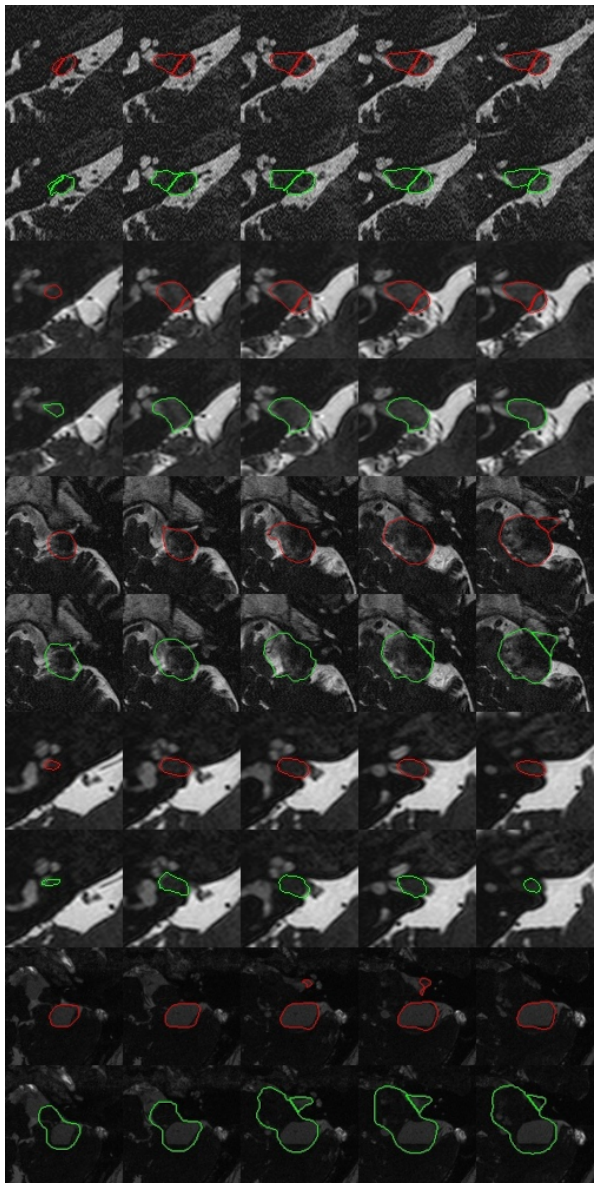


Figure 2.8: Examples of excellent tumor segmentation of T2-weighted images. Odd rows show the CNN prediction of several sections from each patient in red, and even rows show the corresponding delineation of annotator 1 in green. The human T2 annotations were transposed from the contrast-enhanced T1 delineation. The dice scores of these patients are 0.89, 0.92, 0.89, 0.86, 0.44 respectively, and the surface-to-surface distances (mm) are 0.17, 0.11, 1.00, 0.15, 4.75 respectively.

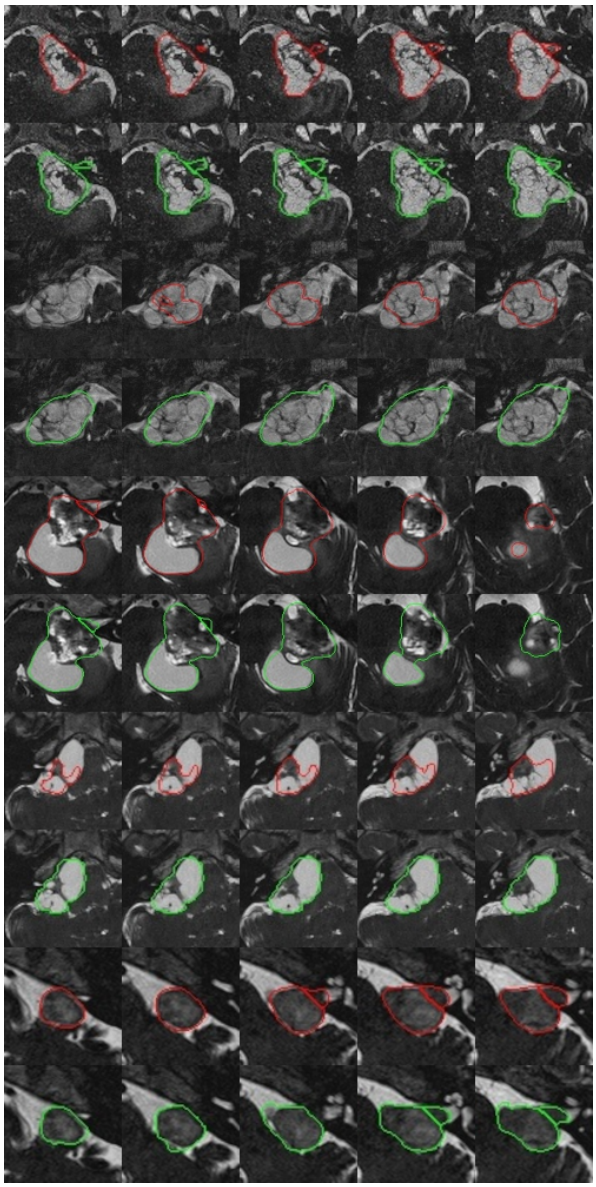


Figure 2.9: Examples of suboptimal tumor segmentation of T2-weighted images. Odd rows show the CNN prediction of several sections from each patient in red, and even rows show the corresponding delineation of annotator 1 in green. The human T2 annotations were transposed from the contrast-enhanced T1 delineation. The dice scores of these patients are 0.86, 0.57, 0.85, 0.72, and 0.85 respectively, and the surface-to-surface distances (mm) are 1.40, 3.86, 1.03, 2.36, and 0.81 respectively.

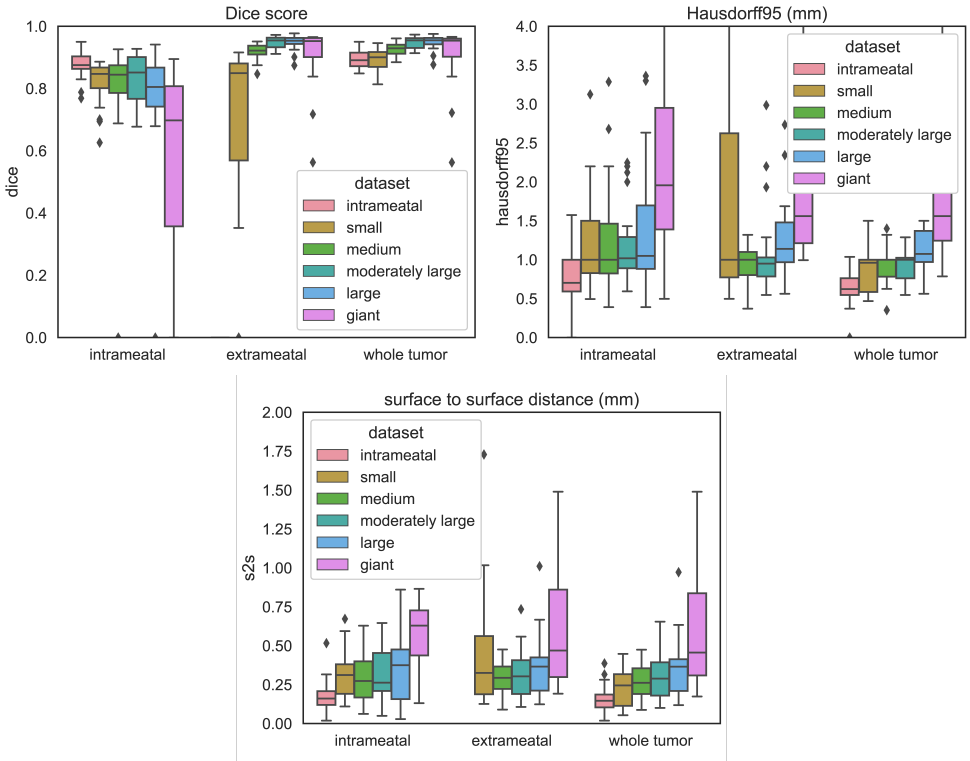


Figure 2.10: Performance on contrast-enhanced T1 by size. Performance of the CNN model by size on the complete dataset. Tumor size is classified using the consensus guideline criteria as proposed by Kanzaki et al. Differences in the determination of the intra- and extrameatal tumor parts in small and giant tumors cause the very wide range of Dice Index and Hausdorff distance, since sometimes a tumor is classified as fully intra or extrameatal by the model, but the human delineation also contains a small intra/extrameatal part or vice versa.

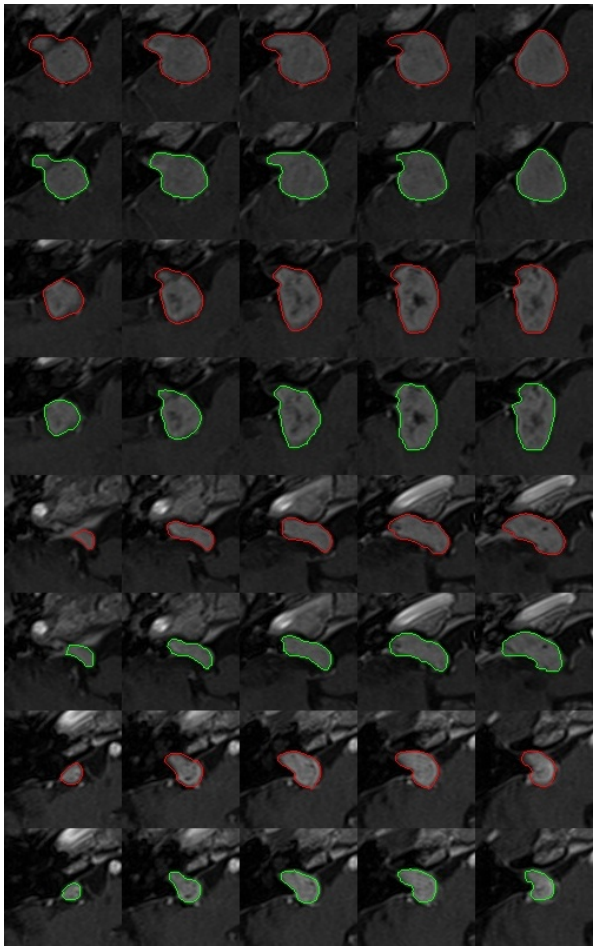


Figure 2.11: Example results on the VS-SEG dataset from Shapey et al. [15]. Odd rows show the CNN prediction of several slices from each patient in red, and even rows show the corresponding manual annotation in green. Predictions are usually larger than the manual annotations.

References

[1] M. L. Carlson and M. J. Link. "Vestibular schwannomas". In: *New England Journal of Medicine* 384.14 (2021), pages 1335–1348.

[2] J. Kanzaki, M. Tos, M. Sanna, and D. A. Moffat. "New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma". In: *Otology & neurotology* 24.4 (2003), pages 642–649.

[3] J. Varughese, T. Wentzel-Larsen, F. Vassbotn, et al. "Analysis of vestibular schwannoma size in multiple dimensions: a comparative cohort study of different measurement techniques". In: *Clinical Otolaryngology* 35.2 (2010), pages 97–103.

[4] S. MacKeith, T. Das, M. Graves, et al. "A comparison of semi-automated volumetric vs linear measurement of small vestibular schwannomas". In: *European Archives of Oto-Rhino-Laryngology* 275 (2018), pages 867–874.

[5] R. van de Langenberg, B. J. de Bondt, P. J. Nelemans, et al. "Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements". In: *Neuroradiology* 51 (2009), pages 517–524.

[6] K. A. Lees, N. M. Tombers, M. J. Link, et al. "Natural history of sporadic vestibular schwannoma: a volumetric study of tumor growth". In: *Otolaryngology–Head and Neck Surgery* 159.3 (2018), pages 535–542.

[7] J. Shapey, G. Wang, R. Dorent, et al. "An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI". In: *Journal of neurosurgery* 134.1 (2019), pages 171–179.

[8] C.-c. Lee, W.-K. Lee, C.-C. Wu, et al. "Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery". In: *Scientific reports* 11.1 (2021), page 3106.

[9] N. A. George-Jones, K. Wang, J. Wang, and J. B. Hunter. "Automated detection of vestibular schwannoma growth using a two-dimensional U-Net convolutional neural network". In: *The Laryngoscope* 131.2 (2021), E619–E624.

[10] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pages 203–211.

[11] F. Isensee, J. Petersen, A. Klein, et al. "Self-adapting framework for u-net-based medical image segmentation". In: *Preprint]. arXiv* 10 (2018).

[12] S. Klein, M. Staring, K. Murphy, et al. "Elastix: a toolbox for intensity-based medical image registration". In: *IEEE transactions on medical imaging* 29.1 (2009), pages 196–205.

[13] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, et al. "Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease". In: *Frontiers in neuroinformatics* 7 (2014), page 50.

[14] R. Rai, L. C. Holloway, C. Brink, et al. “Multicenter evaluation of MRI-based radiomic features: A phantom study”. In: *Medical physics* 47.7 (2020), pages 3054–3063.

[15] J. Shapey, A. Kujawa, R. Dorent, et al. “Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm”. In: *Scientific Data* 8.1 (2021), page 286.

[16] W. T. Koos, J. D. Day, C. Matula, and D. I. Levy. “Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas”. In: *Journal of neurosurgery* 88.3 (1998), pages 506–512.

[17] K. Buch, A. Juliano, K. M. Stankovic, et al. “Noncontrast vestibular schwannoma surveillance imaging including an MR cisternographic sequence: is there a need for postcontrast imaging?” In: *Journal of neurosurgery* 131.2 (2018), pages 549–554.

[18] A. M. Tolisano, C. C. Wick, and J. B. Hunter. “Comparing linear and volumetric vestibular schwannoma measurements between T1 and T2 magnetic resonance imaging sequences”. In: *Otology & Neurotology* 40.5S (2019), S67–S71.

[19] F. B. Pizzini, A. Sarno, I. B. Galazzo, et al. “Usefulness of high resolution T2-weighted images in the evaluation and surveillance of vestibular schwannomas? Is gadolinium needed?” In: *Otology & Neurotology* 41.1 (2020), e103–e110.

[20] I. R. Whittle, C. Smith, P. Navoo, and D. Collie. “Meningiomas”. In: *The Lancet* 363.9420 (2004), pages 1535–1543.

3

Conditional neural fields with shift modulation for multi-sequence MRI translation

This chapter was adapted from:

Chen, Y., Staring, M., Wolterink, J.M. and Tao, Q., 2023. Local implicit neural representations for multi-sequence MRI translation. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5)

Chen, Y., Staring, M., Neve, O.M., Romeijn, S.R., Hensen, E.F., Verbist, B.M., Wolterink, J.M., Tao, Q., 2024. CoNeS: Conditional neural fields with shift modulation for multi-sequence MRI translation. The journal Machine Learning for Biomedical Imaging 2, 657 – 685

Abstract

Multi-sequence magnetic resonance imaging (MRI) has found wide applications in both modern clinical studies and deep learning research. However, in clinical practice, it frequently occurs that one or more of the MRI sequences are missing due to different image acquisition protocols or contrast agent contraindications of patients, limiting the utilization of deep learning models trained on multi-sequence data. One promising approach is to leverage generative models to synthesize the missing sequences, which can serve as a surrogate acquisition. State-of-the-art methods tackling this problem are based on convolutional neural networks (CNNs), which usually suffer from spectral biases, resulting in poor reconstruction of high-frequency fine details. In this paper, we propose Conditional Neural Fields with Shift Modulation (CoNeS), a model that takes voxel coordinates as input and learns a representation of the target images for multi-sequence MRI translation. The proposed model uses a multi-layer perceptron (MLP) instead of a CNN as the decoder for pixel-to-pixel mapping. Hence, each target image is represented as a neural field that is conditioned on the source image via shift modulation with a learned latent code. Experiments on BraTS 2018 and an in-house clinical dataset of vestibular schwannoma patients showed that the proposed method outperformed state-of-the-art methods for multi-sequence MRI translation both visually and quantitatively. Moreover, we conducted spectral analysis, showing that CoNeS was able to overcome the spectral bias issue common in conventional CNN models. To further evaluate the usage of synthesized images in clinical downstream tasks, we tested a segmentation network using the synthesized images at inference. The results showed that CoNeS improved the segmentation performance when some MRI sequences were missing and outperformed other synthesis models. We concluded that neural fields are a promising technique for multi-sequence MRI translation. Our code is available at `https://github.com/cyjdswx/CoNeS.git`.

3.1 Introduction

Multi-sequence magnetic resonance imaging (MRI) plays a key role in radiology and medical image computing. One advantage of MRI is the availability of various pulse sequences, such as T1-weighted MRI (T1), T2-weighted MRI (T2), T1-weighted with contrast (T1ce), and T2-fluid-attenuated inversion recovery MRI (FLAIR), which can provide complementary information to clinicians [1]. The importance of the availability of multi-sequence MRI was also indicated by recent deep learning research [2], which shows that the more sequences were used for segmentation, the better results could be obtained. However, due to clinical restrictions on the use of contrast agents and the diversity in imaging protocols in different medical centers, it is difficult and time-consuming to always obtain exactly the same MRI sequences for training and inference, which may damage the generalization and performance of deep learning segmentation models.

One way to tackle this problem is to generate missing sequences from existing images based on the information learned from a set of paired images, known as image-to-image translation. Like in other computer vision tasks, convolutional neural networks (CNNs) with an encoder and decoder architecture are normally used for this specific task [3, 4, 5]. Despite the significant improvement over traditional non-deep-learning methods, these methods still suffer from the limitation of a pixel-wise loss function, such as the ℓ_1 or MSE loss, which tends to result in blurry results with undesirable loss of details in image structures [6, 7]. To overcome this limitation, generative adversarial networks (GANs) were introduced for image-to-image translation and rapidly became a training protocol benchmark for medical image translation [8, 9, 10]. GANs improve translation results both visually and quantitatively, owing to the adversarial learning loss, which penalizes the images that are correctly classified by the discriminator.

However, research showed that generative models that use a CNN as a backbone network consisting of ReLU activation functions and transposed or up-convolutional layers usually suffer from spectral biases [11, 12]. Therefore, these generative models fit low-frequency signals first and may again fail to capture details in image structures during training. Transformers, which instead use multi-head self-attention blocks and multi-layer perceptrons (MLPs), have gained tremendous attention in computer vision research [13, 14]. Due to the absence of convolutional layers, transformers show great potential for preserving fine details and long-range dependencies and have recently been applied to medical image translation [15, 7]. However, despite the numerous efforts made by these studies, such as hybrid architectures and image patch-based processing, the training of transformers is still considered heavy and data-demanding [16, 17]. The inherently high computational complexity of the transformer block and

expensive memory cost of low-level tasks, such as denoising and super-resolution, further complicate the application of transformers in medical image translation [18].

To address these limitations, we propose image-to-image translation using neural fields [19]. In contrast to CNN or transformer-based methods, a neural field represents the target images on a continuous domain using a coordinate-based network, which can be conditioned on the information extracted from the source images. We previously proposed an image-to-image translation approach based on neural fields [20]. Here, we substantially extend this model by proposing **Co**nditional **Ne**ural fields with **S**hift modulation (CoNeS). In contrast to traditional deep learning computer vision techniques, CoNeS parameterizes the target images as neural fields that can be queried on a grid to provide pixel-wise predictions. Specifically, we use an MLP as the decoder to map the voxel coordinates to the intensities on the target images. To capture instance-specific information, we condition the neural fields on the latent codes extracted from the source images. By applying shift modulation, the neural fields can be further varied across the coordinates to enhance their ability to preserve high-frequency signals.

Although plenty of work has shown great progress in medical image translation, most previous works have been evaluated based on image similarity metrics, and only a few papers have evaluated the benefits of using synthesized images for downstream analysis. [21] fine-tuned a segmentation model with synthesized cardiac images to improve the performance of different modalities; [22] introduced a variational auto-encoder (VAE) based network for image translation-based data augmentation to improve the generalization capabilities of a segmentation model. In practice, however, it would be more straightforward and beneficial to use the synthesized images directly without fine-tuning or training a new network. In this study, we perform downstream experiments using a pre-trained segmentation model to further evaluate different image translation models.

The main contributions of our work are:

- We developed a novel generative adversarial network for medical image translation based on conditional neural fields. In the proposed model, we build neural fields on the coordinates across the image to fit the target image. To improve the performance and the stability of the model, we introduce shift modulation, which conditions the neural fields on the output of a hypernetwork.
- We evaluated the proposed model by synthesizing various MRI sequences on two paired multi-sequence brain MRI datasets. The results show that the proposed model outperforms state-of-the-art methods both visually and quantitatively.

We additionally performed spectral analysis, which indicates that our method is not affected by spectral biases in the way that traditional CNN-based generative models are.

- We compare different medical image translation models in downstream tasks by testing a segmentation model with the synthesized images. Our experiments indicate that by applying image translation, we can improve segmentation performance for incomplete MRI acquisition, and our synthesized images outperform the state-of-the-art methods.

3.2 Related work

Missing MRI sequences Several studies have dealt with the missing MRI sequences problem in medical image analysis [23]. One early idea was to translate all available sequences into a shared latent space for downstream analysis. Following this idea, [24] developed the Hetero-Modal Image Segmentation (HeMIS) method, where sequence-specific convolutional layers are applied to each image sequence for establishing a common representation which enables robust segmentation when some images are missing. Later, [25] and [26] introduced knowledge distillation to improve the segmentation performance in the same situation. In such a model, a network using all modalities as input (teacher network) and another network using a subset of them (student network) are optimized synchronously. During training, information from all modalities is distilled from the teacher to the student network to improve the performance of the student network. Recently, [13] developed a transformer-based model for Alzheimer's classification that can handle missing data scenarios. All these models managed to build a robust model for the situation that only a part of the modalities are available. However, since the missing MRIs are not explicitly constructed, it is still difficult for medical doctors to interpret and make decisions with these methods in clinical practice.

Image-to-image translation Image-to-image translation, on the contrary, focuses on synthesizing missing images from existing ones based on prior knowledge learned from the dataset. By predicting the missing images, clinicians can offer comprehensive diagnoses and also find an explanation of the results in downstream analysis. Recent progress in generative modeling, such as generative adversarial networks (GANs), variational auto-encoders (VAE), and diffusion models, has shown extraordinary capabilities in image generation [6, 27]. In the domain of medical image translation, [28] proposed pGAN based on a conditional GAN combined with a pixel-wise loss and a perceptual loss. [29] proposed a multi-modal GAN (MM-GAN) that extends the idea by using multi-modality imputation for arbitrary input and output modalities. Recently,

[30] proposed mustGAN that enhanced the synthesis performance by aggregating multiple translation streams. Inspired by the recent progress of the transformer model, [7] proposed ResViT based on a hybrid architecture that consists of convolutional operators and transformer blocks. Although promising, most studies focus on the image quality of the output images, and only a few have extended their work to the use of synthesized images in downstream tasks [31, 32, 21].

Neural fields Neural fields, also known as implicit neural representations (INRs) or coordinate-based networks, are increasingly popular in computer vision and medical image analysis [19, 33]. The core idea behind neural fields is that neural networks are not used to learn an operator between signals, as in CNNs or vision transformers, but to *represent* a complex signal on a continuous spatial or spatiotemporal domain. Neural fields can be used to solve a wide range of problems, including 3D scene reconstruction and generative shape modeling. [34] proposed DeepSDF which learns a continuous signed distance function to represent 3D surfaces. One distinguished benefit of using neural fields is the capability to handle data with variable resolution because of the absence of up-sampling architectures. Inspired by this, [35] proposed a Local Implicit Image Function (LIIF) for image super-resolution, which also shows the potential of handling image generation. Similarly in the field of medical imaging, [36] performed multi-contrast MRI super-resolution via neural fields without any high-resolution training data. [37] proposed to use INRs to represent a transformation function for deformable image registration. [38] proposed to reconstruct anatomical shapes from sparse measurements via neural fields. Recently, [39] developed Spatially-Adaptive Pixelwise Networks (ASAP-Net), which is most relevant to our work, to speed up image-to-image translation by locally conditioned MLPs. Different from prior work, the neural fields in CoNeS are conditioned on a latent code varying across the coordinates through shift modulation, inspired by [40]. Specifically, CoNeS consists of a global MLP shared by the whole image and a varying latent code, which determines pixel-wise affine transformations to modulate the neural fields.

3.3 Methods

Model overview

To formulate the problem, let $\boldsymbol{I}_t = \{I_t^i\}_{i=0}^{N_t}$ be the set of missing MRI sequences and $\boldsymbol{I}_s = \{I_s^i\}_{i=0}^{N_s}$ be the set of available MRI sequences, where N_t is the number of target sequences and N_s the number of source sequences. We assume that all images from an instance are co-registered so that there is no extra deformation between the images. As a result, our problem is identical to learning a mapping function $\Phi : \boldsymbol{I}_s \rightarrow \boldsymbol{I}_t$ using

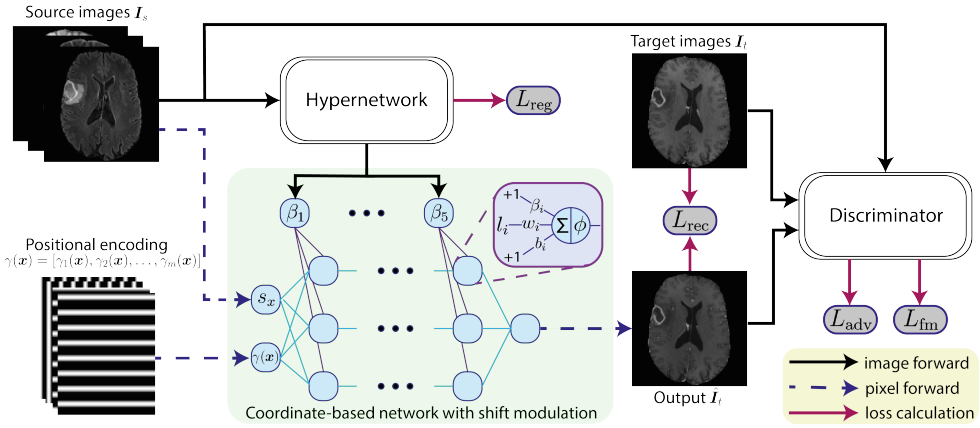


Figure 3.1: The overall architecture of CoNeS. The generator in the proposed models consists of a hypernetwork and a coordinate-based network. We condition the coordinate-based network on a varying latent code, which is generated by the hypernetwork, across coordinates via shift modulation. The conditional discriminator, which takes both the source images and real/fake images as input, further improves the performance of the generator. The proposed model is optimized using a reconstruction loss L_{rec}, an adversarial loss L_{adv}, a feature matching loss L_{fm} and latent code regularization L_{reg}.

the training dataset, which can be applied to all patients in the test dataset and generate the corresponding missing image $\boldsymbol{I}_t$. Similar to traditional GAN models, the proposed model consists of a generator that performs the mapping and a discriminator that aims to tell the real target image and the synthesized one apart. As introduced in pix2pix [6], we apply a conditional discriminator that takes both the source and predicted image as its input. The overall architecture of our approach is shown in Figure 3.1. In the following section, we introduce how to use a coordinate-based network to model conditional neural fields for image-to-image translation.

Coordinate-based network

In a typical neural field algorithm, a quantity defined over spacetime, such as an RGB intensity or a signed distance function, is represented as a neural network that maps coordinates to the quantity. Specifically, in our problem, we train an MLP that takes coordinates as input and outputs intensities of the target MRI sequences. Given a normalized d-dimensional coordinate $\boldsymbol{x} \in \mathbb{R}^d$, where each component lies in $[-1, 1]$, we use $\boldsymbol{t}_x = \{t^i_{\boldsymbol{x}}\}$ and $\boldsymbol{s}_x = \{s^i_{\boldsymbol{x}}\}$ to denote the intensities at position $\boldsymbol{x}$, where t^i_x refers to the intensity value in I^i_t and s^i_x refers to the intensity value in I^i_s, respectively. Hence, the function Φ can be formulated as a pixel-wise mapping that generates intensities over

a d-dimensional space:

$$\boldsymbol{t}_{\boldsymbol{x}} = \Phi(\boldsymbol{x}; \boldsymbol{z}), \tag{3.1}$$

where $\boldsymbol{z}$ is a latent code that contains instance-specific information. During inference, the target images $\hat{\boldsymbol{I}}_t = \{\hat{I}_t^i\}$ are obtained by intensity prediction via sampling from the entire grid via Φ.

A network directly operating on the Cartesian coordinates tends to fit the low-frequency signals first and, as a result, fails to reconstruct the high-frequency image details [41, 11]. One popular approach to overcome this problem is to map the Cartesian coordinates to a higher dimensional space via positional encoding $\gamma : \mathbb{R}^d \to \mathbb{R}^m$. In the proposed model, we use sinusoidal functions to perform positional encoding as follows [42]:

$$\gamma(\boldsymbol{x}) = [\gamma_1(\boldsymbol{x}), \gamma_2(\boldsymbol{x}), \ldots, \gamma_m(\boldsymbol{x})], \tag{3.2}$$

$$\gamma_{2i}(\boldsymbol{x}) = \sin(2^{i-1}\pi\boldsymbol{x}), \tag{3.3}$$

$$\gamma_{2i+1}(\boldsymbol{x}) = \cos(2^{i-1}\pi\boldsymbol{x}), \tag{3.4}$$

where m is a frequency parameter. Positional encoding can also be seen as Fourier feature mapping of the Cartesian coordinates. By using positional encoding as the input of the MLP, we enable the network to fit the neural field containing high-frequency variation.

Conditional neural fields

To let the neural field adapt to different input images, we condition it on a set of latent codes $\boldsymbol{z}$, which contain instance-specific information. In the proposed model, we introduce a hypernetwork H that generates the latent code from the source images: $\boldsymbol{z} = H(\boldsymbol{I}_s)$. By extracting $\boldsymbol{z}$, we can then vary and adapt the neural fields to different instances. Below, we explain how we obtain the latent code $\boldsymbol{z}$ and how the proposed method parameterizes the neural fields with the conditioning via $\boldsymbol{z}$.

A hypernetwork refers to an extra neural network that generates parameters for the main network [43]. The main network behaves like a typical neural network, while the hypernetwork encodes information from the inputs and transfers the information to the main network via the generated parameters. For clarity, we use $\boldsymbol{z}_i = [\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i]$ to denote the latent code used by the i-th layer of the MLP, where $\boldsymbol{\alpha}_i$ are the weights and $\boldsymbol{\beta}_i$ are the biases, both generated by H. Hence, for each layer of the MLP, we have:

$$\boldsymbol{l}_{i+1} = \phi(\boldsymbol{\alpha}_i \boldsymbol{l}_i + \boldsymbol{\beta}_i), \tag{3.5}$$

where $\boldsymbol{l}_i$ is the input feature of the i-th layer, and ϕ is the activation function. Both $\boldsymbol{\beta}_i$ and $\boldsymbol{l}_i$ are column vectors of size $n_{i+1} \times 1$ and $\boldsymbol{\alpha}_i$ is a matrix of size $n_{i+1} \times n_i$, where n_i is the number of neurons of the i-th layer. Inspired by ASAP-Net [39], we vary the neural field of each pixel by varying the latent code across the coordinates, which can be denoted as $\boldsymbol{z}_i(\boldsymbol{x}) = [\boldsymbol{\alpha}_i(\boldsymbol{x}), \boldsymbol{\beta}_i(\boldsymbol{x})]$, to improve the representation capability. We use $H_{\boldsymbol{x}}$ to represent the latent code mapping for each pixel, and thus, Φ can be denoted as:

$$t_{\boldsymbol{x}} = \Phi(\gamma(\boldsymbol{x}); \boldsymbol{z}(\boldsymbol{x})) = \Phi(\gamma(\boldsymbol{x}); H_{\boldsymbol{x}}(\boldsymbol{I}_s)), \tag{3.6}$$

and each layer of the MLP can be denoted as:

$$\boldsymbol{l}_{i+1}(\boldsymbol{x}) = \phi(\boldsymbol{\alpha}_i(\boldsymbol{x})\boldsymbol{l}_i(\boldsymbol{x}) + \boldsymbol{\beta}_i(\boldsymbol{x})), \tag{3.7}$$

where $\boldsymbol{l}_i(\boldsymbol{x})$ refers to the i-th input feature at position $\boldsymbol{x}$. Different from ASAP-Net, we adapt the bottom-up pathway from a feature pyramid network [44] as the hypernetwork H, which outputs the latent code $\boldsymbol{z}(\boldsymbol{x})$ for each pixel with a feasible memory cost (detailed in Section 3.4).

By conditioning neural fields on varying latent codes across the coordinates, we can improve the representation capability of the network and better model the structure details [19, 45]. However, the number of parameters also increases with the spatial expansion of the neural fields, which may induce high computational costs and damage the performance due to over-fitting. This problem may become worse with larger input images. To compact the model while maintaining spatially varying neural fields, we propose to condition the neural network through feature-wise linear modulation (FiLM) [46]. Instead of generating all parameters of the MLP per pixel, an affine transformation (scale and shift) is applied to every neuron of a single, global MLP. Thus, each layer of the one MLP can be denoted as:

$$\boldsymbol{l}_{i+1}(\boldsymbol{x}) = \overline{\boldsymbol{\alpha}_i}(\boldsymbol{x})\phi(\boldsymbol{w}_i\boldsymbol{l}_i + \boldsymbol{b}_i) + \boldsymbol{\beta}_i(\boldsymbol{x}), \tag{3.8}$$

where the weights and biases of the MLP are now replaced by trainable parameters $\boldsymbol{w}_i$ and $\boldsymbol{b}_i$ that are shared by all coordinates, and $\overline{\boldsymbol{\alpha}_i}(\boldsymbol{x})$ is an $n_{i+1} \times n_{i+1}$ matrix that performs scaling to the neurons of the i-th layer by left matrix multiplication. Thus, we can obtain a modified neural field for each coordinate with fewer parameters. Furthermore, research shows that by using shifts only, which is so-called shift modulation, we can achieve comparable results with half of the parameters [40]. In this case, $\overline{\boldsymbol{\alpha}_i}$ is an identity matrix and all latent codes are used as shift parameters: $\boldsymbol{z}_i(\boldsymbol{x}) = \boldsymbol{\beta}_i(\boldsymbol{x})$. In practice, we split the biases of the MLP into two parts: trainable biases $\boldsymbol{b}_i$ and biases $\boldsymbol{\beta}_i(\boldsymbol{x})$ generated by H:

$$\boldsymbol{l}_{i+1}(\boldsymbol{x}) = \phi(\boldsymbol{w}_i\boldsymbol{l}_i + \boldsymbol{b}_i + \boldsymbol{\beta}_i(\boldsymbol{x})). \tag{3.9}$$

The hypernetwork H is optimized together with the MLP during training. In the experimental section, we will show that by using shift modulation, our model can achieve better performance at reduced complexity.

In addition to shift modulation, we also condition the neural fields on the source images directly. Different from the latent codes, the pixel intensities provide first-hand, uncoded local information. We concatenate the image intensities from all the source images as an additional input of the MLP. The mapping function of the neural fields, therefore, becomes:

$$t_{\boldsymbol{x}} = \Phi(\gamma(\boldsymbol{x}), \boldsymbol{s}_{\boldsymbol{x}}; H_{\boldsymbol{x}}(\boldsymbol{I}_s)). \tag{3.10}$$

Loss function

Like the standard GAN model, the discriminator and the generator in the proposed model are optimized alternately. In each iteration, we train the discriminator using the hinge loss [47]:

$$L_D = \mathbb{E}_{\boldsymbol{I}_t, \boldsymbol{I}_s}[\max(0, 1 - D(\boldsymbol{I}_t, \boldsymbol{I}_s))] + \mathbb{E}_{\boldsymbol{I}_s}[\max(0, 1 + D(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_s))], \tag{3.11}$$

where D is the discriminator and $\mathbb{E}$ is the expectation over the whole dataset.

The generator is trained by a loss function L that contains a reconstruction loss, an adversarial loss, a feature matching loss, and latent code regularization.

Reconstruction loss To ensure the synthesized images are as close to the real images as possible, we apply a reconstruction loss that maximizes the similarity between ground truth $\boldsymbol{I}_t$ and output images $\hat{\boldsymbol{I}}_t$, which are obtained by intensity prediction via sampling from the entire grid. We use the ℓ_1 loss function as suggested in [6]:

$$L_{\text{rec}} = \mathbb{E}_{\boldsymbol{I}_t, \boldsymbol{I}_s}[\|\hat{\boldsymbol{I}}_t - \boldsymbol{I}_t\|_1]. \tag{3.12}$$

Adversarial loss Adversarial loss is applied to enforce that the generated images are good enough to fool the discriminator. Like the discriminator loss, we use the hinge function, which is defined as:

$$L_{\text{adv}} = -\mathbb{E}_{\boldsymbol{I}_s}[\log D(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_s)]. \tag{3.13}$$

Feature matching loss To stabilize the training, we apply a feature matching loss introduced by [48]. Specifically, we feed both the real and generated images to the discriminator and extract the intermediate features from each forward pass. The two groups of intermediate features are matched using the ℓ_1 loss function. Hence, the

feature matching loss is defined as:

$$L_{\text{fm}} = \mathbb{E}_{\boldsymbol{I}_t, \boldsymbol{I}_s} \sum_{i=1}^{T} \frac{1}{N_i} [\| D_i(\boldsymbol{I}_t, \boldsymbol{I}_s) - D_i(\hat{\boldsymbol{I}}_t, \boldsymbol{I}_s) \|_1]. \tag{3.14}$$

where D_i denotes the i-th feature layer of the discriminator and N_i denotes the number of elements in each layer. T is the total number of layers of the discriminator.

Latent code regularization Last, we apply the ℓ_2 norm to $\boldsymbol{z}$ as a latent code regularization to stabilize the training:

$$L_{\text{reg}} = \mathbb{E}_{\boldsymbol{I}_s} \| H_{\boldsymbol{x}}(\boldsymbol{I}_s) \|_2. \tag{3.15}$$

Overall loss The overall loss function then becomes

$$L = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{fm}} L_{\text{fm}} + \lambda_{\text{reg}} L_{\text{reg}}, \tag{3.16}$$

where λ_{rec}, λ_{adv}, λ_{fm}, and λ_{reg} are the weights of the loss functions.

3.4 Experiments and results

Dataset

To evaluate the proposed translation model, we conducted experiments on two datasets: (1) BraTS 2018 [49] and (2) an in-house Vestibular Schwannoma MRI (VS) dataset [50].

BraTS 2018 BraTS 2018 is a multi-sequence brain MRI dataset for tumor segmentation. The dataset consists of 285 patients for training and 66 patients for validation. Each patient has four co-registered MRI sequences: T1 (1-6 mm slice thickness), T1ce 1-6 mm slice thickness), T2 (2-6 mm slice thickness) and FLAIR (2-6 mm slice thickness). The tumor mask that includes the non-enhanced tumor, the enhanced tumor, and the edema was delineated by experts from multiple centers as a segmentation ground truth. All the scans in BraTS 2018 are resampled to 1 mm isotropic resolution.

Vestibular schwannoma MRI dataset The VS dataset is MRI scans of patients with vestibular schwannoma, which is a benign tumor arising from the neurilemma of the vestibular nerve. 191 patients were collected from 37 different hospitals using 12 different MRI scanners. In our study, 147 patients were selected for training, and the remaining 44 patients were the validation set. All patients have a gadolinium-enhanced T1-weighted MRI (shortened to T1ce) and a high-resolution T2 (shortened to T2).

The spatial resolution of the T1ce ranges from $0.27 \times 0.27 \times 0.9$ to $1.0 \times 1.0 \times 5.0$ mm, and the spatial resolution of T2 scans ranges from $0.23 \times 0.23 \times 0.5$ to $0.7 \times 0.7 \times 1.8$ mm. The intra- and extrameatal tumor was manually delineated by four radiologists. Different from BraTS 2018, there are only two sequences available in the VS dataset, and the high resolution of the T2 offers better visibility of the lesion, but may also degrade the image quality, which makes the image translation more challenging on this dataset.

Experimental setup

Network architecture The hypernetwork H of the proposed model is adapted from feature pyramid network [44]. H consists of four convolutional modules containing [2, 4, 23, 3] residual blocks in sequence. Each convolutional module is followed by a 3×3 convolutional smoothing layer and up-sampling layer, computing a feature map that is downscaled by a factor of 2. We take the output of the last module, which has the same size as the input resolution, as the latent code $\boldsymbol{z}$. Adapted from [39], the MLP in the proposed model contains five 64-channel layers. The Leaky ReLU function with a negative slope of 0.2 is applied as the activation function after all intermediate layers. The output layer is followed by a Tanh function which can constrain the range of the intensities to $[-1, 1]$. The discriminator is a 2D convolutional neural network that takes both the source image and prediction as input, both as a whole. The network contains five 4×4 convolutional blocks followed by a Leaky ReLU function except for the last layer. The strides and number of filters of the blocks are $[2, 2, 2, 1, 1]$ and $[64, 128, 256, 512, 1]$ respectively. Like pix2pix [6], the discriminator down-samples the inputs by 8 and penalizes structures at the scale of patches.

Pre-processing Registration was applied to the VS dataset before training. We considered the T1ce as the fixed image and performed rigid registration with the T2, for which we used Elastix software [51]. All images from both datasets were then normalized to the range of $[-1, 1]$, and the background was cropped based on the bounding box of the foreground before training to reduce the image size. Both sequences in the VS dataset were resampled to 0.29×0.29 mm in-plane resolution, which is the median value of the T1ce domain. During training, random cropping was conducted on the images, with a cropping size of 160×128 for BraTS 2018 and a cropping size of 320×320 for the VS dataset, respectively.

Implementation details All experiments were conducted using Python 3.10 and PyTorch 1.12.1 on a mixed computation server equipped with Nvidia Quadro RTX 6000 and Nvidia Tesla V100 GPUs. The models were trained by the Adam optimizer using the Two Time-scale Update Rule (TTUR) training scheme, in which the generator

Table 3.1: Quantitative comparison of different image translation models on BraTS 2018. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to the proposed method.

model	T1ce translation		T1 translation		T2 translation		FLAIR translation	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
pix2pix	30.1 ±2.65†	0.941 ±0.014†	27.0 ±3.69	0.945 ±0.013†	28.0 ±2.63†	0.926 ±0.062†	27.6 ±3.03†	0.910 ±0.097†
pGAN	30.7 ±3.18†	0.943 ±0.015†	**27.5 ±3.72**	0.945 ±0.015†	29.2 ±2.80†	0.943 ±0.020†	28.5 ±3.26†	0.916 ±0.095†
ResViT	29.2 ±2.37†	0.935 ±0.014†	25.0 ±2.60†	0.918 ±0.014†	26.6 ±2.30†	0.923 ±0.020†	24.7 ±2.08†	0.876 ±0.092†
ASAP-Net	30.8 ±2.97†	0.948 ±0.017†	27.3 ±3.79	0.948 ±0.015†	28.6 ±2.74†	0.940 ±0.019†	28.4 ±3.10†	0.916 ±0.098†
CoNeS (proposed)	**31.2 ±3.11**	**0.951 ±0.017**	27.3 ±4.03	**0.953 ±0.014**	**29.6 ±3.03**	**0.950 ±0.021**	**29.1 ±2.99**	**0.926 ±0.097**

and discriminator have a different initial learning rate [52]. We found that an initial learning rate of 1×10^{-4} for the generator and an initial learning rate of 4×10^{-4} for the discriminator worked best for our experiments. The learning rates were further decayed using a linear learning rate scheduler. Adapted from the choices of hyperparameters in [39], we set the frequency parameter $m = 6$ for positional encoding. We set $\lambda_{\text{adv}} = 1.0$ and $\lambda_{\text{rec}} = 100.0$, which gives us the best balance between sharp results and fewer artifacts as suggested in [6]. Both L_{fm} and L_{reg} help to stabilize the training, while large λ_{reg} and λ_{fm} may lead to poor reconstruction performance. We set $\lambda_{\text{fm}} = \lambda_{\text{reg}} = 10.0$, which ensures a stable training while maintaining reconstruction performance [48, 39]. Lastly, we focus on 2D image translation in this paper and hence use 2D coordinates ($d = 2$).

Benchmark overview We compared our model with the following state-of-the-art methods: (1) pix2pix: pix2pix is a GAN-based image translation model, which consists of a UNet-based generator and a patch-based discriminator [6]; (2) pGAN: pGAN is a GAN-based image translation model using ResNet which follows an encoder-bottleneck-decoder architecture as backbone [28]. Perceptual loss is introduced to improve the results; (3) ResViT: ResViT is an image translation model that combines pGAN with a transformer-based information bottleneck; (4) ASAP-Net: ASAP-Net is a neural field-based image translation model [39]. Different from the proposed model, ASAP-Net parameterizes patch-wise neural fields, which are conditioned through a UNet-shape hypernetwork without a shared MLP. For all implementations, we used the official GitHub repositories provided by the authors. We used the ℓ_1 loss as a reconstruction loss for all the benchmark methods. We used the least square loss

Table 3.2: Quantitative comparison of different image translation models after cropping on BraTS 2018. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T1 translation**		**T2 translation**		**FLAIR translation**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
pix2pix	19.9 ±3.30†	0.610 ±0.092†	15.7 ±4.41	0.636 ±0.089†	19.9 ±2.85†	0.658 ±0.082†	18.1 ±3.31†	0.585 ±0.081†
pGAN	20.6 ±3.73†	0.646 ±0.096†	**16.1** **±4.61**	0.663 ±0.099	20.8 ±3.23†	0.721 ±0.093†	18.9 ±3.70†	0.636 ±0.086†
ResViT	20.2 ±3.56†	0.612 ±0.098†	15.1 ±4.32†	0.599 ±0.090†	19.5 ±3.50†	0.672 ±0.093†	17.0 ±3.26†	0.545 ±0.111†
ASAP-Net	20.4 ±3.67†	0.634 ±0.115†	15.7 ±4.48	0.626 ±0.103†	20.3 ±3.05†	0.669 ±0.089†	18.5 ±3.60†	0.593 ±0.086†
CoNeS (proposed)	**20.9** **±3.66**	**0.667** **±0.099**	15.8 ±4.44	**0.666** ±0.094	**21.5** **±3.35**	**0.739** **±0.095**	**19.6** **±3.49**	**0.663** **±0.084**

function [53] as an adversarial loss for pix2pix, pGAN, and ResViT. Like the proposed method, we used the hinge loss function [47] as an adversarial loss for ASAP-net. All the benchmark methods were trained using hyperparameters that were optimized by the original authors on the same dataset (BraTS). We trained ResViT with the pre-trained network as suggested in [7], while all other models were trained from scratch.

Multi-sequence MRI translation

We first examined the quality of the images generated from the proposed model. Theoretically, CoNeS can be applied to any number of missing or present sequences by adapting input and output channels to N_s and N_t. For simplicity, we assumed one sequence was missing for all the patients during inference ($N_t = 1$), and thus, we trained models that generate one MRI sequence from the other sequences in the dataset for evaluation. Specifically, four image translation experiments were performed on BraTS 2018: (1) T1, T2, FLAIR → T1ce (shortened to T1ce translation); (2) T1ce, T2, FLAIR → T1 (shortened to T1 translation); (3) T1ce, T1, FLAIR → T2 (shortened to T2 translation); and (4) T1ce, T1, T2 → FLAIR (shortened to FLAIR translation). We used two different metrics for quantitative analysis in our study: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Both the synthesized images and real images were normalized to $[0, 1]$ before evaluation. Wilcoxon signed-rank test between each benchmark and the proposed model was performed on all image translation experiments.

The quantitative results are listed in Table 3.1. As shown in the table, the proposed

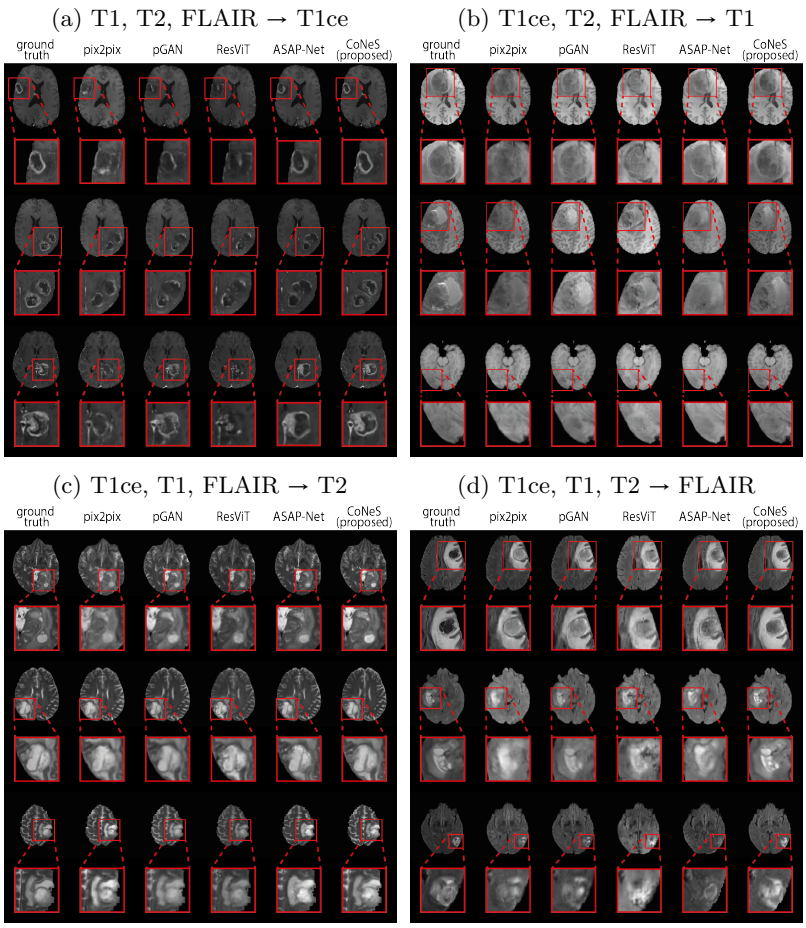


Figure 3.2: Comparison results of different image translation models on BraTS 2018: (a) T1, T2, FLAIR → T1ce; (b) T1ce, T2, FLAIR → T1; (c) T1ce, T1, FLAIR → T2; (d) T1ce, T1, T2 → FLAIR. For each translation experiment, three examples are selected for display. Each column shows the ground truth and translation results of the different models. Zoomed-in results indicated in red rectangles are shown below the whole images.

model performs significantly better ($p < .05$) than other state-of-the-art methods in most metrics, except that pGAN obtains higher PSNR on T1 translation. Using T1ce translation on BraTS 2018 as an example, the PSNR and SSIM of the proposed model on BraTS 2018 are **31.2 dB** and **0.951**, which outperforms pix2pix by **1.1 dB** PSNR and **1.0 %** SSIM, pGAN by **0.5 dB** PSNR and **0.8 %** SSIM, ResViT by **2.0 dB** PSNR and **1.6 %** SSIM, and ASAP-Net by **0.4 dB** PSNR and **0.3 %** SSIM. Translation examples

Table 3.3: Quantitative comparison of different image translation models on VS dataset. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; All metrics of the benchmarks in this table show significant differences ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T2 translation**	
	PSNR	SSIM	PSNR	SSIM
pix2pix	21.1 ± 1.39	0.602 ± 0.068	21.4 ± 1.78	0.506 ± 0.121
pGAN	21.6 ± 1.55	0.635 ± 0.077	22.2 ± 2.04	$\mathbf{0.575 \pm 0.131}$
ResViT	21.0 ± 1.58	0.575 ± 0.090	21.5 ± 1.80	0.489 ± 0.110
ASAP-Net	20.4 ± 1.24	0.552 ± 0.061	20.9 ± 1.97	0.500 ± 0.117
CoNeS (proposed)	$\mathbf{21.9 \pm 1.69}$	$\mathbf{0.638 \pm 0.077}$	$\mathbf{22.6 \pm 2.03}$	0.560 ± 0.126

Table 3.4: Quantitative comparison of different image translation models after cropping on VS dataset. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to the proposed method.

model	**T1ce translation**		**T2 translation**	
	PSNR	SSIM	PSNR	SSIM
pix2pix	14.8 ± 3.28	$0.415 \pm 0.122^{\dagger}$	$16.6 \pm 1.72^{\dagger}$	$0.321 \pm 0.084^{\dagger}$
pGAN	$14.2 \pm 3.31^{\dagger}$	$0.417 \pm 0.133^{\dagger}$	$16.8 \pm 1.98^{\dagger}$	0.372 ± 0.134
ResViT	14.5 ± 2.75	$0.400 \pm 0.106^{\dagger}$	$16.7 \pm 1.66^{\dagger}$	$0.342 \pm 0.099^{\dagger}$
ASAP-Net	$13.0 \pm 2.93^{\dagger}$	$0.340 \pm 0.132^{\dagger}$	$15.3 \pm 1.64^{\dagger}$	$0.300 \pm 0.101^{\dagger}$
CoNeS (proposed)	$\mathbf{15.0 \pm 3.17}$	$\mathbf{0.451 \pm 0.118}$	$\mathbf{17.3 \pm 1.58}$	$\mathbf{0.379 \pm 0.101}$

are shown in Figure 3.2 in which we can see that the proposed model can recover more detailed structures, such as the contrast-enhanced tumor in T1ce, which is clinically highly relevant.

Both the PSNR and SSIM show global similarity, while the quality of the region around the tumor is more clinically interesting. To further evaluate the proposed model, we cropped the images using the bounding box of the tumor region and then evaluated the similarity of these sub-images using the aforementioned metrics. The bounding box was generated from the segmentation results of nnUNet [54] for the reason that the segmentation ground truths of BraTS 2018 validation set are not available. The results are listed in Table 3.2. As we can see, the proposed model also performs significantly better ($p < .05$) in most tasks within this sub-region, which is consistent with our observation from zoomed-in results in Figure 3.2. We observed that the performance of the proposed model decreased after cropping due to the lack

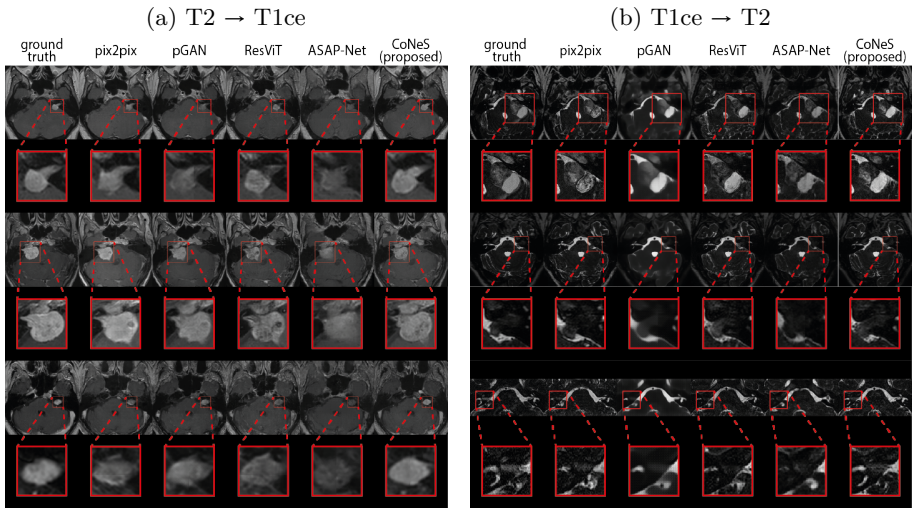
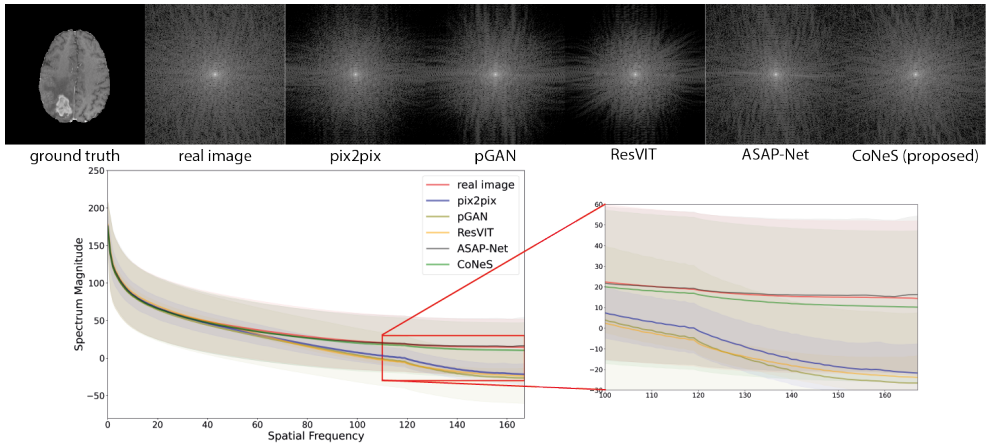


Figure 3.3: Comparison results of different image translation models on the VS dataset: (a) T2 → T1ce; (b) T1ce → T2. For each translation experiment, three examples are selected for display. Each column shows the ground truth and translation results of the different models. Zoomed-in results indicated in red rectangles are shown below the images.

of background. Again using T1ce translation as an example, the PSNR and SSIM of the proposed model are **20.9 dB** and **0.667**, which outperforms pix2pix by **1.0 dB** PSNR and **5.7 %** SSIM, pGAN by **0.3 dB** and **2.1 %** SSIM, ResViT by **0.7 dB** and **5.5 %** SSIM, and ASAP-Net by **0.5 dB** and **3.3 %** SSIM.

Next, we performed two image translation experiments on the VS dataset: (1) T2 → T1ce (shortened to T1ce translation) and (2) T1ce → T2 (shortened to T2 translation). We again evaluated the entire image as well as the cropped region around the tumor, similar to BraTS 2018. Both quantitative results are listed in Table 3.3 and Table 3.4. All models struggle with the VS dataset and show decreased performance compared to BraTS 2018, and CoNeS still performs significantly better ($p < .05$) in most of the metrics. Taking T1ce translation as an example, CoNeS obtains a PSNR of **21.9 dB** and a SSIM score of **0.638**, which outperforms pix2pix by **0.8 dB** PSNR and **3.6 %** SSIM, pGAN by **0.3 dB** PSNR and **0.3 %** SSIM, ResViT by **0.9 dB** PSNR and **6.3 %** SSIM, and ASAP-Net by **1.5 dB** PSNR and **8.6 %** SSIM. Qualitatively, we can observe improved synthesized images using the proposed model as shown in Figure 3.3. It is worth pointing out that although pGAN obtained better SSIM scores (**0.575**) in T2 translation, the visualization suggests that our results contain more informative details, while pGAN's results are blurry.

(a) Spectral analysis on BraTS 2018



(b) Spectral analysis on the VS dataset

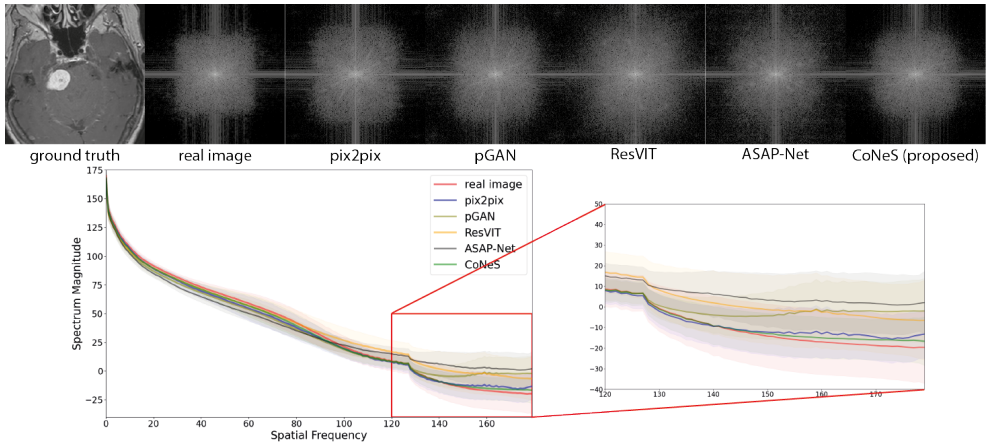


Figure 3.4: Spectral analysis of different image translation models. (a) and (b) show the analysis results on BraTS 2018 and the VS dataset, respectively. For each analysis, the Fourier transform of different synthesized images and the real image are shown in the top row. The bottom row shows the spectral distribution, in which the high-frequency range is zoomed in by the red rectangle.

Spectral analysis

Research has shown that CNN-based generative models with up-sampling layers usually struggle with reproducing the spectral distribution correctly [12, 55]. On the contrary, coordinate-based networks like CoNeS build a direct pixel-to-pixel mapping without any up-sampling layer. In this section, we further evaluated the synthesized images in

the frequency domain to demonstrate the improvement we obtained by performing spectral analysis on the T1ce translation model of both datasets. Specifically, we applied a 2D Fourier transform to all synthesized results as well as the real images, and then calculated a 1D representation of the 2D spectrum using Azimuthal integration [12]. Azimuthal integration is defined as an integration over the radial frequencies:

$$\mathrm{AI}(\omega_k) = \int_0^{2\pi} \|\mathscr{F}(\omega_k \cos\theta, \omega_k \sin\theta)\omega_k\|_2 d\theta, \tag{3.17}$$

for $k = 0, \ldots, M/2-1$, and where $\mathscr{F}(m,n)$ is the Fourier Transform of a 2D image, θ is the radian, ω_k is the spatial frequency and M is the length of a square image.

A log transformation was performed to the 2D spectrum for better visualization, and we calculated the average 1D representation over the dataset to avoid biased sampling. As shown in Figure 3.4, both ASAP-Net and CoNeS, which are coordinate-based networks, can reproduce the spectrum over the entire frequency range on BraTS 2018. Specifically, all the spectral curves are very close in the low-frequency range (spatial frequency < 50), which enables the generative models to reconstruct the general structure of the images. However, the spectral curves of GAN-based models dramatically drop in the high-frequency range (spatial frequency > 75), while the curves of ASAP-Net and CoNeS remain close to the real distribution. This shows that neural fields are able to overcome the spectral biases issue of convolutional neural networks. On the VS dataset, all the models yield higher spectrum magnitudes in the high-frequency range compared to the real images, which suggests that these translation models might add high-frequency noise to the synthesized images. Consistent with the similarity measurement results, ASAP-Net is not robust enough to reproduce the spectrum on the VS dataset and may induce more artifacts. On the contrary, CoNeS still outputs images whose spectrum is closest to the real images among all the translation models. The results indicate that by using neural fields conditioned via shift modulation, CoNeS is able to keep the representation capability and reproduce the spectrum distribution.

Synthesized images for tumor segmentation

To further examine the impact of synthesized images in downstream analysis, we performed tumor segmentation using the synthesized images at inference. To do this, we first adopted the architecture from nnUnet [54] and trained a segmentation network that uses all the sequences in the dataset as input. Note that all the images were normalized to a range of $[-1,1]$ during training to make the input channels consistent with the synthesized images. During inference, we tested the segmentation model with synthesized images and compared the results with the performance of the model when filling the missing channel with zeros, called zero imputation, in our experiments.

Table 3.5: Results of using different images for segmentation inference on BraTS 2018. The real sequences used are indicated by ✓, the missing ones by ✗, and the ones replaced by synthesized images by ○. The mean of Dice scores and 95% HD (mm) of the enhanced tumor (ET), the whole tumor (WT), and the tumor core (TC) are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates a significant difference ($p < .05$) compared to inference using synthesized images from CoNeS.

input sequences				method	Dice			95 % HD (mm)		
T1ce	T1	T2	FLAIR		ET	WT	TC	ET	WT	TC
✓	✓	✓	✓	N/A	0.770	0.888	0.822	4.49	6.27	8.92
✗	✓	✓	✓	zero imputation	0.068†	0.845†	0.362†	27.9†	8.80†	18.1†
○	✓	✓	✓	pix2pix	0.191†	0.850†	0.537†	15.0†	8.10†	15.0†
○	✓	✓	✓	pGAN	0.317†	0.858†	0.598†	14.9†	8.01†	14.0†
○	✓	✓	✓	ResViT	0.223†	0.858†	0.555†	15.0†	7.87†	14.1†
○	✓	✓	✓	ASAP-Net	0.332†	0.866†	0.597†	13.3†	**6.95**	13.2†
○	✓	✓	✓	CoNeS	**0.386**	**0.870**	**0.662**	**13.1**	7.23	**13.0**
✓	✗	✓	✓	zero imputation	0.717†	0.865†	0.753†	6.53†	7.86†	11.3†
✓	○	✓	✓	pix2pix	0.747	0.869†	0.780†	5.07†	7.70†	9.84†
✓	○	✓	✓	pGAN	0.747†	0.868†	0.779†	5.60†	7.61†	10.2†
✓	○	✓	✓	ResViT	0.751†	0.869†	0.784†	**4.62**	7.39†	9.75†
✓	○	✓	✓	ASAP-Net	0.753†	0.881	0.806	5.43†	**6.73**	9.39
✓	○	✓	✓	CoNeS	**0.764**	**0.885**	**0.808**	5.30	7.05	**8.94**
✓	✓	✗	✓	zero imputation	0.748†	0.835†	0.752†	5.64†	8.67†	11.6†
✓	✓	○	✓	pix2pix	0.761	0.862†	0.784†	3.90†	7.53†	9.82†
✓	✓	○	✓	pGAN	0.767	0.872†	0.797†	3.83†	7.34†	9.27
✓	✓	○	✓	ResViT	0.759	0.855†	0.788†	4.19†	8.18†	9.74
✓	✓	○	✓	ASAP-Net	0.764	0.880†	0.817	3.84†	6.50†	9.05
✓	✓	○	✓	CoNeS	**0.778**	**0.886**	**0.829**	**3.15**	**6.01**	**8.34**
✓	✓	✓	✗	zero imputation	0.679†	0.403†	0.690†	27.8†	30.3†	23.2†
✓	✓	✓	○	pix2pix	0.760	0.805†	0.771†	**3.74**	9.75†	11.1†
✓	✓	✓	○	pGAN	0.766	0.833†	0.777†	4.60	8.59†	10.3†
✓	✓	✓	○	ResViT	0.783	0.768†	0.752†	5.16†	11.7†	11.5†
✓	✓	✓	○	ASAP-Net	**0.785**	0.823†	0.808†	3.80†	9.36†	**9.36**†
✓	✓	✓	○	CoNeS	0.768	**0.853**	**0.809**	4.30	**7.56**	9.38

For simplicity, we again assumed one specific sequence was missing and replaced this sequence while keeping the rest unchanged. Similar to the image translation experiments, we compared the segmentation performance using synthesized images generated from the proposed model to the other images via the Wilcoxon signed-rank test.

The tests were performed for each MRI sequence (T1ce, T1, T2, and FLAIR) on BraTS 2018. The performance was evaluated using three specific categories: 1) enhanced tumor (ET); 2) tumor core (TC, non-enhanced tumor, and edema); and 3) the whole tumor (WT, enhanced tumor, non-enhanced tumor, and edema). Dice score and 95 % Hausdorff distance (95 % HD) of all the three categories are reported for quantitative evaluation in Table 3.5. We can see that the presence of sequences dramatically influences the performance of the segmentation model. For instance, when the T1ce is missing, the Dice score of the enhanced tumor is **0.068** because the enhanced information is only visible in the T1ce. As expected, most of the metrics

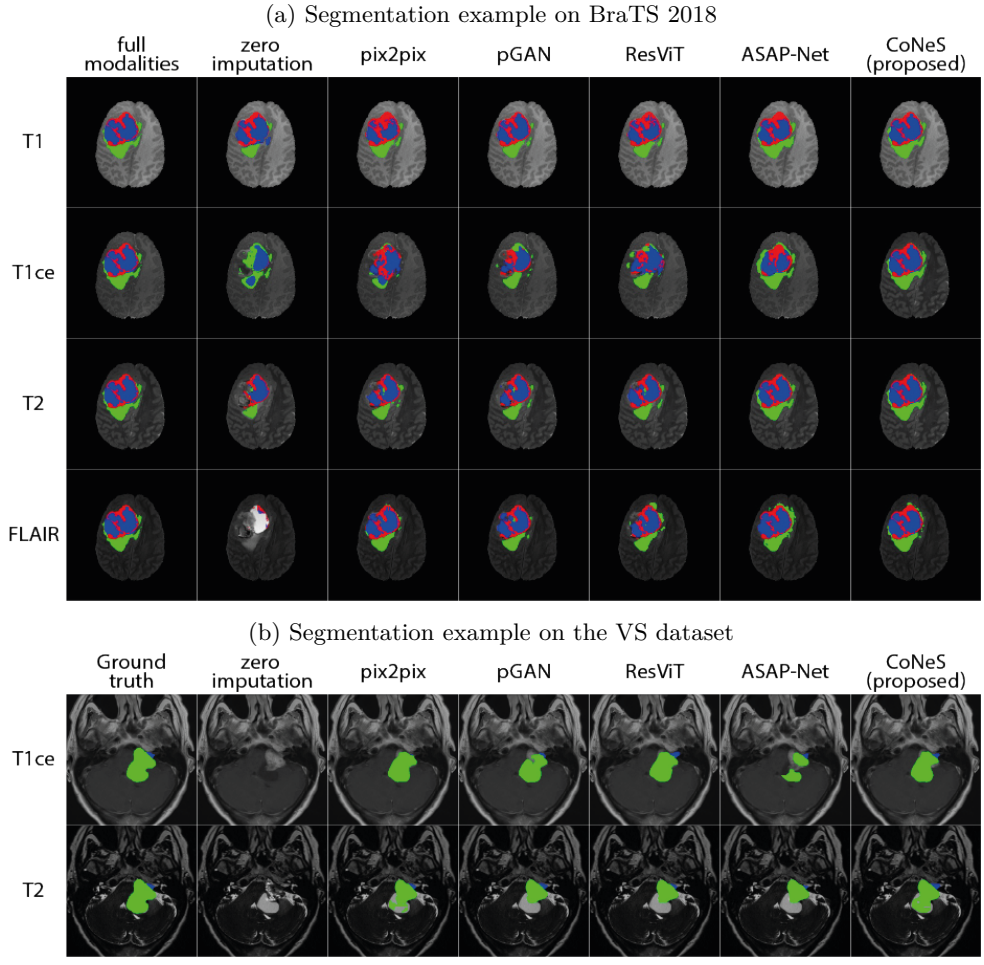


Figure 3.5: The results of segmentation experiments: (a) A segmentation example on BraTS 2018 and (b) an example on the VS dataset. The rows show the segmentation results with different MRI sequences replaced. The columns show ground truth (for BraTS 2018, segmentation results with full sequences) and segmentation results using different synthesized images.

show that inference with synthesized images performs worse than inference with full sequences. However, we also noticed that when the real T2 or FLAIR were replaced with synthesized ones, we obtained a lower mean 95 % HD. This occurs due to the influence of certain outliers. For example, sometimes the model can identify the enhanced tumor at the wrong position using real images, leading to a large 95 % HD, while the other inferences using synthesized images completely miss the tumor. When we removed three outliers, the mean of 95 % HD of the enhanced tumor became

2.99 mm, which is better than the others.

The best results among all the inferences using synthesized images (including zero imputation) for each sequence were highlighted in Table 3.5. The results indicate that using synthesized images for inference can significantly improve the segmentation performance and that the synthesized images of our model yield the best segmentation performance with a significant difference ($p < .05$) among all the translation models. Using the inferences without T1ce as examples, the Dice scores of the proposed model are 0.386, 0.870, and 0.662 in the enhanced tumor, the whole tumor, and the tumor core respectively. In comparison, the proposed model outperforms zero imputation by 31.8 %, 2.5 %, and 30.0 %, pix2pix by 19.5 %, 2.0 %, and 12.5 %, pGAN by 6.9 %, 1.2 %, and 6.4 %, ResViT by 16.3 %, 1.2 %, and 10.7 %, and ASAP-Net by 5.4 %, 0.4 %, and 6.5 %. The 95 % HDs of the proposed model are 13.1 mm, 7.23 mm, and 13.0 mm in the enhanced tumor, the whole tumor, and the tumor core respectively. In comparison, the proposed model outperforms zero imputation by 14.8 mm, 1.57 mm, and 5.1 mm, pix2pix by 1.9 mm, 0.87 mm, and 2.0 mm, pGAN by 1.8 mm, 0.78 mm, 1.0 mm, ResViT by 1.9 mm, 0.64 mm, 1.1 mm. Although ASAP-Net obtained a higher 95 % HD (6.95 mm) in the whole tumor, we did not observe significant differences between it and the proposed model. Some example segmentation results are presented in Figure 3.5. It is worth noting that the synthesized T1 of CoNeS performs better in segmentation than the ones from pGAN, although we got higher PSNR for pGAN in the former experiment.

We also performed the same segmentation experiments on the VS dataset. We evaluated the performance using three specific categories: 1) intrameatal tumor; 2) extrameatal tumor); and 3) the whole tumor (including intra- and extrameatal tumor). Dice score and 95 % HD of all three categories are reported in Table 3.6. Similarly to BraTS 2018, all the synthesized images compensate for the performance loss due to the drop of sequences, and the proposed model performs significantly better ($p < .05$) than the other models. For instance, the synthesized T1ce generated by the proposed model obtained Dice scores of 0.567, 0.714, and 0.749 in the intrameatal tumor, the extrameatal tumor, and the whole tumor respectively. In comparison, the proposed model outperforms zero imputation by 56.6 %, 68.4 %, and 72.1 %, pix2pix by 9.6 %, 2.8 %, and 3.6 %, pGAN by 12.6 %, 3.8 %, and 8.8 %, ResViT by 2.7 %, 0.1 %, and 3.0 %, and ASAP-Net by 25.9 %, 22.2 %, and 24.1 %. The 95 % HDs of the proposed model are 2.33 mm, 3.54 mm, and 4.05 mm in the intrameatal tumor, the extrameatal tumor, and the whole tumor respectively. These results outperform zero imputation by 5.71 mm, 25.37 mm, and 30.05 MM, pix2pix by 0.21 mm, 2.13 mm, and 2.45 mm, pGAN by 0.15 mm, 3.03 mm, and 3.82 mm, ASAP-Net by 0.77 mm, 3.72 mm, and 8.35 mm. We observed that ResViT obtained lower 95 % HD (3.32 mm) in the extrameatal

Table 3.6: Results of using different images for segmentation inference on the VS dataset. The real sequences used are indicated by ✓, the missing ones by ✗, and the ones replaced by synthesized images by ∘. The mean of Dice scores and 95% HD (mm) of the intrameatal tumor (IT), the extrameatal tumor (ET), and the whole tumor (WT) are reported. The highest values per column are indicated in boldface; The † after each metric of the benchmarks indicates significant differences ($p < .05$) compared to inference using synthesized images from CoNeS.

input sequences		method	**Dice**			95 % **HD** (mm)		
T1ce	T2		IT	ET	WT	IT	ET	WT
✓	✓	N/A	0.761	0.853	0.896	1.34	1.71	1.45
✗	✓	zero imputation	0.001†	0.030†	0.028†	8.04†	28.91†	34.1†
∘	✓	pix2pix	0.471†	0.686†	0.713†	2.54	5.67†	6.50
∘	✓	pGAN	0.441†	0.676†	0.661†	2.48†	6.57†	7.87†
∘	✓	ResViT	0.540†	0.713†	0.719	2.36	**3.32**†	5.73
∘	✓	ASAP-Net	0.308†	0.492†	0.508†	3.10†	7.26†	12.4†
∘	✓	CoNeS	**0.567**	**0.714**	**0.749**	**2.33**	3.54	**4.05**
✓	✗	zero imputation	0.184†	0.397†	0.400†	4.06†	18.0†	22.2†
✓	∘	pix2pix	0.713†	0.856†	0.874†	1.54	2.19†	2.09
✓	∘	pGAN	0.701†	0.839†	0.844†	1.82†	2.60†	2.58†
✓	∘	ResViT	0.716†	0.831†	0.862	1.61	2.48†	2.32
✓	∘	ASAP-Net	0.677†	0.834†	0.854†	1.95†	2.63†	2.45†
✓	∘	CoNeS	**0.746**	**0.858**	**0.878**	**1.40**	**2.09**	**1.96**

tumor, however, the proposed model still performs better than ResViT in most of the experiments. Example segmentation results are displayed in Figure 3.5.

Ablation study

Ablation studies were performed to verify the benefits of individual components in the proposed method. For simplicity, we trained a baseline CoNeS model that translates T1ce from T2 on BraTS 2018. We first examined the added value of the source image as input to the MLP by removing the intensity value from the input channel. In this case, the neural fields are conditioned on the latent code only. Next, we compared shift modulation against a full hypernetwork where all the parameters of the MLP are generated. Last, we trained the proposed method without the adversarial loss to show the contribution of the discriminator in our model. Wilcoxon signed-rank tests were performed between the baseline model and ablated models. Quantitative and qualitative results are shown in Table 3.7 and Figure 3.6. We noticed that although the model without adversarial loss achieves marginally better SSIM and PSNR, the results, especially the tumor region, are visually blurry, which shows that the adversarial loss helps the model to reconstruct more details and outputs more realistic images. Apart from this, the proposed model obtained the best results among the ablated models that

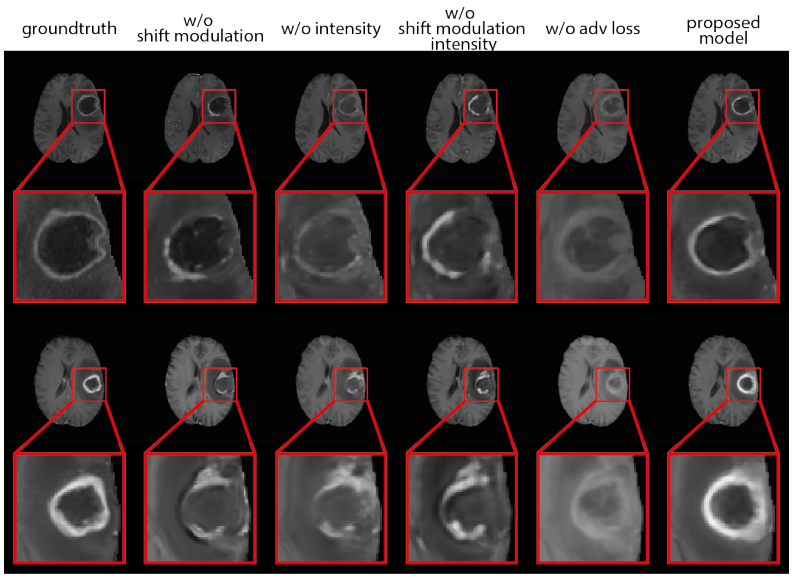


Figure 3.6: Example results of the ablated models. Zoomed-in results indicated with red rectangles are shown below the full images.

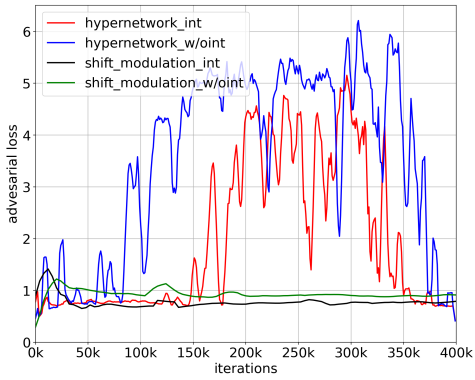
Table 3.7: Quantitative comparison of ablated models on BraTS 2018. The mean value and standard deviation of PSNR and SSIM are reported. The highest values per column are indicated in boldface; The † after each metric indicates a significant difference ($p < .01$) compared to the proposed model (the bottom row).

shift modulation	intensity	adversarial loss	#param generated	PSNR	SSIM
no	no	yes	$14.5k$	$29.6 \pm 2.13^{\dagger}$	$0.933 \pm 0.013^{\dagger}$
no	yes	yes	$14.5k$	29.9 ± 2.32	$0.938 \pm 0.013^{\dagger}$
yes	no	yes	$0.26k$	30.0 ± 2.13	$0.938 \pm 0.013^{\dagger}$
yes	yes	no	$0.26k$	$\mathbf{30.2 \pm 2.33}$	$\mathbf{0.943 \pm 0.013}^{\dagger}$
yes	yes	yes	$0.26k$	30.0 ± 2.22	0.941 ± 0.014

include the adversarial loss and showed significant differences ($p < .01$) in SSIM. These results also show that shift modulation helps to reduce the parameters from $14.5k$ to $0.26k$, which is the number of neurons in the MLP, without loss of representation capability. Moreover, although the instance-specific information is already encoded in the latent code, conditioning the network on intensity directly can still add extra information and improve performance.

We next demonstrated the stability of the models by comparing the loss curves of the ablated models. Both the adversarial loss L_{adv} and the total loss L are shown in

(a) Adversarial loss curves



(b) Total generator loss curves

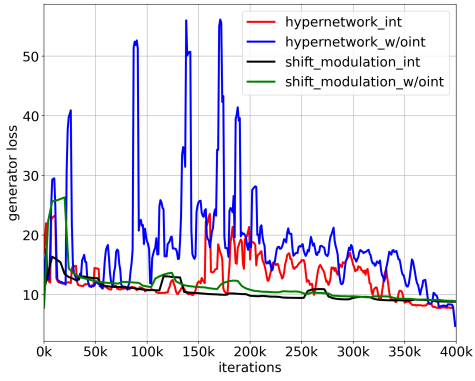


Figure 3.7: Training loss curves of the ablated models. (a) adversarial loss (b) total generator loss including reconstruction loss, adversarial loss, feature matching loss, and latent code regularization. The models using shift modulation show more stable training loss against the models using a full hypernetwork.

Figure 3.7. We observe that both losses of the models using the full hypernetwork fluctuated substantially and L_{adv} increased midway through training. On the contrary, both loss curves of the models using shift modulation remained stable throughout the learning. The experiments suggested that by reducing the number of parameters generated, shift modulation is able to improve the stability of the image translation model.

3.5 Discussion and conclusion

In this work, we proposed CoNeS, a novel conditional neural fields-based model for MRI translation. We modeled the image translation problem using neural fields,

which are conditioned on the source images, and learned latent codes through a coordinate-based network. The proposed model adapts the predicted neural fields by varying the latent codes across coordinates to ensure better local representations. We then introduced a shift modulation strategy for the conditioning to reduce the model complexity and stabilize the training. We compared the proposed model with state-of-the-art image translation models and our experiments showed that CoNeS performs better in the entire image scope as well as the tumor region, which is clinically relevant. Through visualization results, we also showed that the proposed method can reproduce more structural details, while the other methods' results are visually more blurry. The transformer-based model (ResViT) performed on par with the other methods in our experiments, where on natural images they have been reported to outperform those [56, 57]. Our datasets, however, are considerably smaller than what is used in the domain of natural images, while transformer-based models are considered data-demanding.

We performed a spectral analysis to demonstrate improvements in image translation when using neural fields. As expected, all the CNN-based models and ResViT, which is a hybrid transformer model containing transposed layers during decoding, were unable to reproduce high-frequency signals due to their spectral bias [11, 12]. In contrast, the proposed model was able to preserve the high-frequency information and reconstruct the spectrum in the entire frequency range on both datasets. We also observed that ASAP-Net, a neural field-based benchmark, did not show consistent performance across the two datasets and could not reproduce the spectral distribution on the VS dataset either. These results are consistent with prior studies demonstrating that the full hypernetwork, in which all the parameters of the main network are generated, is sensitive to its initialization and difficult to optimize [58]. The ablation studies further indicated that compared to a full hypernetwork, the conditioning via shift modulation can make the training of neural fields more stable and maintain the representation capability. Furthermore, the results also showed that by introducing the adversarial loss, the predicted images are more realistic and contain more textural detail, although the quantitative metrics (SSIM and PSNR) are slightly lower than the model without the adversarial loss. The reason may be that SSIM and PSNR are not able to measure the benefits of the adversarial loss, which is in line with the conclusion in previous research [15, 7].

To evaluate the value of synthesized MRI in downstream analysis, we performed tumor segmentation experiments. We first demonstrated that dropping sequences during inference of a segmentation model can significantly damage the performance, which shows the complementary importance of multiple MRI sequences in segmentation. We next tested the segmentation model using different synthesized images and

compared the results with the inference using incomplete input images. The experiments demonstrated that image translation models can significantly improve segmentation accuracy by replacing the missing input channel with synthesized images. Furthermore, the images generated by our proposed CoNeS model performed best among the state-of-the-art methods in most of the experiments, which is consistent with the visual improvement observed in the translation experiments. Nevertheless, we found that synthesized images cannot fully replace real images, and a baseline model trained on all real images performed best.

One limitation of our work is that in the clinic, the availability of MRI sequences may vary from patient to patient [59]. The proposed model, however, cannot handle arbitrarily missing sequences, unless separate models are trained for each case. Further work would be adapting the proposed model to random incomplete MRI scans by incorporating techniques like learning disentangled representations [60] or latent representation fusion [61]. Moreover, the choice of the positional encoding frequency m may bias the network to fit the signal of a certain bandwidth [62]. To ease the optimization and improve the generalization, it may be worthwhile to integrate periodic activation functions [63] in our design instead of positional encoding for better representation capability.

In summary, we presented a neural fields-based model that synthesizes missing MRI from other sequences with excellent performance, which can be further integrated into the downstream analysis. All experiments showed improved performance compared to state-of-the-art translation models, while the spectrum analysis and ablation studies demonstrated the strengths of the proposed model over traditional CNN and neural fields models. Neural fields hold great promise in MRI translation to solve the missing MRI sequence problem in the clinic.

3.6 Acknowledgements

This work was supported by the China Scholarship Council (grant 202008130140) and by an unrestricted grant of Stichting Hanarth Fonds, The Netherlands (project MLSCHWAN).

References

[1] A. Cherubini, M. E. Caligiuri, P. Péran, et al. “Importance of multimodal MRI in characterizing brain tissue and its potential application for individual age prediction”. In: *IEEE journal of biomedical and health informatics* 20.5 (2016), pages 1232–1239.

[2] M. Cercignani and S. Bouyagoub. “Brain microstructure by multi-modal MRI: Is the whole greater than the sum of its parts?” In: *Neuroimage* 182 (2018), pages 117–127.

[3] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris. “Whole image synthesis using a deep encoder-decoder network”. In: *Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*. 2016, pages 127–137.

[4] T. Joyce, A. Chartsias, and S. A. Tsaftaris. “Robust multi-modal MR image synthesis”. In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. 2017, pages 347–355.

[5] W. Wei, E. Poirion, B. Bodini, et al. “Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis”. In: *Journal of Medical Imaging* 6.1 (2019), pages 014005–014005.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 1125–1134.

[7] O. Dalmaz, M. Yurt, and T. Çukur. “ResViT: residual vision transformers for multimodal medical image synthesis”. In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pages 2598–2614.

[8] H. Li, J. C. Paetzold, A. Sekuboyina, et al. “DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. 2019, pages 795–803.

[9] D. Nie, R. Trullo, J. Lian, et al. “Medical image synthesis with deep convolutional adversarial networks”. In: *IEEE Transactions on Biomedical Engineering* 65.12 (2018), pages 2720–2730.

[10] K. Armanious, C. Jiang, M. Fischer, et al. “MedGAN: Medical image translation using GANs”. In: *Computerized Medical Imaging and Graphics* 79 (2020), page 101684.

[11] N. Rahaman, A. Baratin, D. Arpit, et al. “On the spectral bias of neural networks”. In: *International Conference on Machine Learning*. 2019, pages 5301–5310.

[12] R. Durall, M. Keuper, and J. Keuper. “Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pages 7890–7899.

[13] L. Liu, S. Liu, L. Zhang, et al. "Cascaded multi-modal mixing transformers for Alzheimer' s disease classification with incomplete data". In: *NeuroImage* (2023), page 120267.

[14] Y. Jiang, S. Chang, and Z. Wang. "TransGAN: Two pure transformers can make one strong GAN, and that can scale up". In: *Advances in Neural Information Processing Systems*. Volume 34. 2021, pages 14745–14758.

[15] J. Liu, S. Pasumarthi, B. Duffy, et al. "One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation". In: *IEEE Transactions on Medical Imaging* (2023).

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[17] P. Esser, R. Rombach, and B. Ommer. "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pages 12873–12883.

[18] H. Chen, Y. Wang, T. Guo, et al. "Pre-trained image processing transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pages 12299–12310.

[19] Y. Xie, T. Takikawa, S. Saito, et al. "Neural fields in visual computing and beyond". In: *Computer Graphics Forum*. Volume 41. 2. 2022, pages 641–676.

[20] Y. Chen, M. Staring, J. M. Wolterink, and Q. Tao. "Local implicit neural representations for multi-sequence MRI translation". In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. 2023, pages 1–5.

[21] S. Amirrajab, C. Lorenz, J. Weese, et al. "Pathology synthesis of 3d consistent cardiac mr images using 2d vaes and gans". In: *International Workshop on Simulation and Synthesis in Medical Imaging*. 2022, pages 34–42.

[22] Y. Skandarani, N. Painchaud, P.-M. Jodoin, and A. Lalande. "On the effectiveness of GAN generated cardiac MRIs for segmentation". In: *Medical Imaging with Deep Learning*. 2020.

[23] R. Azad, N. Khosravi, M. Dehghanmanshadi, et al. "Medical image segmentation on mri images with missing modalities: A review". In: *arXiv preprint arXiv:2203.06217* (2022).

[24] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. "Hemis: Hetero-modal image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. 2016, pages 469–477.

[25] M. Hu, M. Maillard, Y. Zhang, et al. "Knowledge distillation from multi-modal to mono-modal segmentation networks". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. 2020, pages 772–781.

[26] R. Azad, N. Khosravi, and D. Merhof. “SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities”. In: *International Conference on Medical Imaging with Deep Learning*. 2022, pages 48–62.

[27] B. Kawar, M. Elad, S. Ermon, and J. Song. “Denoising diffusion restoration models”. In: *Advances in Neural Information Processing Systems* 35 (2022), pages 23593–23606.

[28] S. U. Dar, M. Yurt, L. Karacan, et al. “Image synthesis in multi-contrast MRI with conditional generative adversarial networks”. In: *IEEE transactions on medical imaging* 38.10 (2019), pages 2375–2388.

[29] A. Sharma and G. Hamarneh. “Missing MRI pulse sequence synthesis using multi-modal generative adversarial network”. In: *IEEE Transactions on Medical Imaging* 39.4 (2019), pages 1170–1183.

[30] M. Yurt, S. U. Dar, A. Erdem, et al. “mustGAN: multi-stream Generative Adversarial Networks for MR Image Synthesis”. In: *Medical image analysis* 70 (2021), page 101944.

[31] J. E. Iglesias, E. Konukoglu, D. Zikic, et al. “Is synthesizing MRI contrast useful for inter-modality analysis?” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2013, pages 631–638.

[32] G. Van Tulder and M. de Bruijne. “Why does synthesized data improve multi-sequence classification?” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015, pages 531–538.

[33] A. Molaei, A. Aminimehr, A. Tavakoli, et al. “Implicit neural representation in medical imaging: A comparative survey”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pages 2381–2391.

[34] J. J. Park, P. Florence, J. Straub, et al. “DeepSDF: Learning continuous signed distance functions for shape representation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pages 165–174.

[35] Y. Chen, S. Liu, and X. Wang. “Learning continuous image representation with local implicit image function”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pages 8628–8638.

[36] J. McGinnis, S. Shit, H. B. Li, et al. “Single-subject Multi-contrast MRI Super-resolution via Implicit Neural Representations”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pages 173–183.

[37] J. M. Wolterink, J. C. Zwienenberg, and C. Brune. “Implicit neural representations for deformable image registration”. In: *International Conference on Medical Imaging with Deep Learning*. 2022, pages 1349–1359.

[38] T. Amiranashvili, D. Lüdke, H. B. Li, et al. “Learning shape reconstruction from sparse measurements with neural implicit functions”. In: *International Conference on Medical Imaging with Deep Learning*. 2022, pages 22–34.

[39] T. R. Shaham, M. Gharbi, R. Zhang, et al. “Spatially-adaptive pixelwise networks for fast image translation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pages 14882–14891.

[40] E. Dupont, H. Kim, S. A. Eslami, et al. “From data to functa: Your data point is a function and you can treat it like one”. In: *International Conference on Machine Learning*. 2022, pages 5694–5725.

[41] B. Mildenhall, P. P. Srinivasan, M. Tancik, et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pages 99–106.

[42] E. D. Zhong, T. Bepler, J. H. Davis, and B. Berger. “Reconstructing continuous distributions of 3D protein structure from cryo-EM images”. In: *International Conference on Learning Representations*. 2020.

[43] D. Ha, A. M. Dai, and Q. V. Le. “HyperNetworks”. In: *International Conference on Learning Representations*. 2017.

[44] T.-Y. Lin, P. Dollár, R. Girshick, et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pages 2117–2125.

[45] S. Peng, M. Niemeyer, L. Mescheder, et al. “Convolutional occupancy networks”. In: *European Conference on Computer Vision*. 2020, pages 523–540.

[46] E. Perez, F. Strub, H. De Vries, et al. “FiLM: Visual reasoning with a general conditioning layer”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 32. 1. 2018.

[47] J. H. Lim and J. C. Ye. “Geometric GAN”. In: *arXiv preprint arXiv:1705.02894* (2017).

[48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, et al. “High-resolution image synthesis and semantic manipulation with conditional GANs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pages 8798–8807.

[49] B. H. Menze, A. Jakab, S. Bauer, et al. “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *IEEE Transactions on Medical Imaging* 34.10 (2014), pages 1993–2024.

[50] O. M. Neve, Y. Chen, Q. Tao, et al. “Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study”. In: *Radiology: Artificial Intelligence* 4.4 (2022), e210300.

[51] S. Klein, M. Staring, K. Murphy, et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pages 196–205.

[52] M. Heusel, H. Ramsauer, T. Unterthiner, et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).

[53] X. Mao, Q. Li, H. Xie, et al. “Least squares generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pages 2794–2802.

[54] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pages 203–211.

[55] I. Anokhin, K. Demochkin, T. Khakhulin, et al. “Image generators with conditionally-independent pixel synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pages 14278–14287.

[56] S. Kim, J. Baek, J. Park, et al. “InstaFormer: Instance-Aware Image-to-Image Translation With Transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pages 18321–18331.

[57] K. Shibasaki, S. Fukuzaki, and M. Ikehara. “4K Real Time Image to Image Translation Network With Transformers”. In: *IEEE Access* 10 (2022), pages 73057–73067.

[58] O. Chang, L. Flokas, and H. Lipson. “Principled weight initialization for hypernetworks”. In: *International Conference on Learning Representations*. 2019.

[59] H. B. Li, G. M. Conte, S. M. Anwar, et al. “The Brain Tumor Segmentation (BraTS) Challenge 2023: Brain MR Image Synthesis for Tumor Segmentation (BraSyn)”. In: *arXiv preprint arXiv:2305.09011* (2023).

[60] L. Shen, W. Zhu, X. Wang, et al. “Multi-domain image completion for random missing input data”. In: *IEEE Transactions on Medical Imaging* 40.4 (2020), pages 1113–1122.

[61] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris. “Multimodal MR synthesis via modality-invariant latent representation”. In: *IEEE Transactions on Medical Imaging* 37.3 (2017), pages 803–814.

[62] P.-S. Wang, Y. Liu, Y.-Q. Yang, and X. Tong. “Spline Positional Encoding for Learning 3D Implicit Signed Distance Fields”. In: *International Joint Conference on Artificial Intelligence*. 2021.

[63] V. Sitzmann, J. Martel, A. Bergman, et al. “Implicit neural representations with periodic activation functions”. In: *Advances in neural information processing systems*. Volume 33. 2020, pages 7462–7473.

4

A deep learning model to reduce agent dose for contrast-enhanced MRI of the cerebellopontine angle cistern

This chapter was adapted from:

Chen, Y.*, Weber, R.*, Neve, O.M., Romeijn, S.R., Hensen, E.F., Wolterink, J.M., Tao, Q., Staring, M., Verbist, B.M., 2025. A deep learning model to reduce agent dose for contrast-enhanced MRI of the cerebellopontine angle cistern (under review)

Abstract

Objectives

To evaluate a deep learning (DL) model for reducing the agent dose of contrast-enhanced T1-weighted MRI (T1ce) of the cerebellopontine angle (CPA) cistern.

Material and methods

In this multi-center retrospective study, MRIs of vestibular schwannoma (VS) patients were used to simulate low-dose T1ce with varying reductions of contrast agent dose. DL models were trained to restore standard-dose T1ce from the low-dose simulation. The image quality and segmentation performance of the DL-restored T1ce were evaluated. A head and neck radiologist was asked to rate DL-restored images in multiple aspects, including image quality and diagnostic characterization.

Results

203 MRI studies from 72 VS patients (mean age, 58.51 ± 14.73, 39 men) were evaluated. As the input dose increased, the structural similarity index measure of the restored T1ce ranged from 0.639 ± 0.113 to 0.993 ± 0.009, and the peak signal-to-noise ratio ranged from 21.60 ± 3.73 dB to 41.40 ± 4.84 dB. At 10 % input dose, using DL-restored T1ce for segmentation improved the Dice from 0.673 to 0.734, the 95 % Hausdorff distance from 2.38 mm to 2.07 mm, and the average surface distance from 1.00 mm to 0.59 mm. Both DL-restored T1ce from 10 % and 30 % input doses showed excellent image quality (3, interquartile range(IQR) [Q3-Q1] = 3-3 and 3, IQR [Q3-Q1] = 4-3), with the latter being considered more informative (4, IQR [Q3-Q1] = 4 – 3).

Conclusion

The DL model improved the image quality of low-dose MRI of the CPA cistern, which makes lesion detection and diagnostic characterization possible with 10-30 % of the standard dose.

4.1 Introduction

Vestibular schwannomas (VSs) are benign tumors arising from the vestibulocochlear nerve. They account for 80% of cerebellopontine angle (CPA) tumors [1]. Magnetic Resonance Imaging (MRI), especially contrast-enhanced T1-weighted (T1ce) MRI with a gadolinium-based contrast agent (GBCA), plays a key role in the noninvasive detection and characterization of CPA lesions [1, 2, 3]. By accumulating in tissues with rich vascularity or interstitial space, GBCAs offer enhanced visibility of tumor lesions on MR imaging [4]. While T1ce is still considered the gold standard for VS diagnosis and postoperative assessment [5, 6, 7, 8], the necessity of a contrast agent in both the diagnostic setting as well as during longitudinal VS management is being questioned [6, 8, 9]. After initial diagnosis, the surveillance of VS is sometimes carried out with T1 without contrast and/or T2-weighted MRI [3]. Given the growing concerns about short- and long-term toxicity [10, 11], the negative environmental impact of GBCA [12], and cost-effectiveness [9, 13], there is an increased interest in reducing the use of GBCA.

The possibilities of GBCA dose reduction have been explored mainly in the field of oncology [14, 15, 16, 17]. To acquire and evaluate low-dose images, the patients typically underwent two separate examinations with an interval of more than 24 hours [14] or a two-step agent injection within 10 minutes [15, 16]. Preliminary research suggested that MRIs with 50-75 % of the standard dose are non-inferior to the standard-dose MRI [14, 17], and the scans with an even lower dose (15 % of the standard dose) may still be equally effective in lesion detection and conspicuity [15, 16] . However, the low signal-noise-ratio of low-dose MRI may potentially add difficulties for accurate lesion delineation and require more experienced clinical experts to interpret the images [18].

Recently, deep learning (DL) models have been increasingly introduced to restore standard-dose T1ce from low-dose MRI scans [18, 19, 20, 21, 22, 23]. Although promising results have been demonstrated, most validations were carried out using a small-scale dataset, and only a limited number of vestibular schwannoma patients were included [19]. In addition, the values of dose reduction in previous studies were usually determined empirically, and the impact of different dose reductions on deep learning models has not been well-examined. In this study, we aimed to develop a deep learning model to restore the image quality to standard-dose MRI of the CPA for VS diagnosis and longitudinal management. We conducted a comprehensive quantitative and qualitative validation using a low-dose T1ce dataset simulated from multi-center data with various dose reductions.

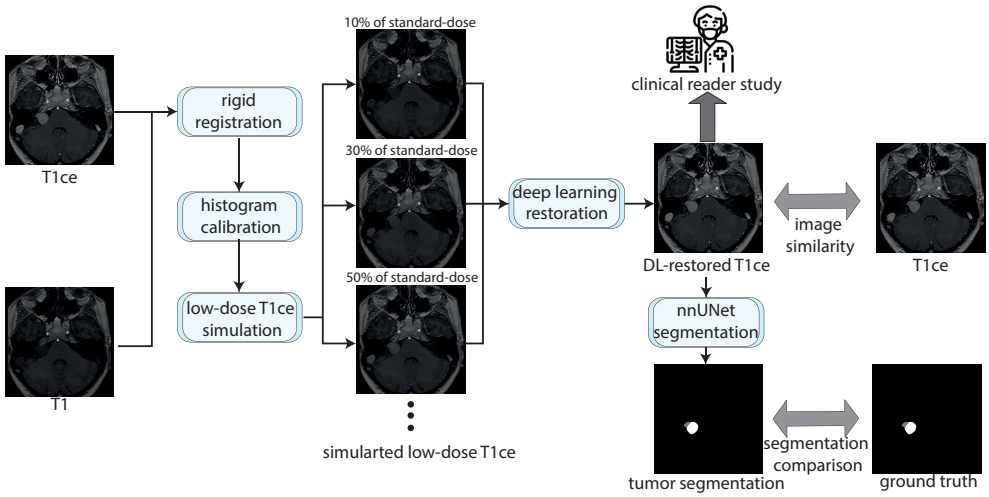


Figure 4.1: The overall pipeline of the study. Nine different low-dose T1ce, ranging from 10 % to 90 % with 10 % increments, were simulated from T1 and calibrated T1ce. The deep learning model was applied to restore T1ce from them. The DL-restored T1ce was evaluated using image similarity metrics, downstream segmentation performance, and a clinical reader study.

4.2 Materials and methods

This retrospective study was performed at Leiden University Medical Center, a tertiary referral center for skull base pathology in the Netherlands. The institutional review board approved the study protocol (G19.115), which granted an exemption for informed consent. The overall pipeline of this study is shown in Figure 4.1.

Patient Population and Data Preparation

MRI studies of Patients who underwent MRI of the CPA, showing a unilateral vestibular schwannoma, were collected from 33 different hospitals. Patients with other CPA pathologies, as well as multiple CPA tumors, were excluded. Every MRI study has both T1 and T1ce examinations in the transverse plane. According to the tumor classification by Kanzaki et al [24], the proportion of intrameatal (only), small, medium, moderately large, large, and giant tumor are 43.8 %, 17.2 %, 26.1 %, 8.4 %, 4.4 %, and 0.0 %, respectively. The detailed characteristics and technical information of the dataset are shown in Table 4.1 and Table 4.2. The intra- and extra-meatal portions of VS were manually delineated by O.M.N., a physician with 3 years of clinical experience, who was trained in performing segmentation and, when necessary, corrected by a senior head-and-neck radiologist (B.M.V.).

Table 4.1: Patient characteristics

Characteristics	
Num of patients	72
Age (year)	58 ± 14.73
Gender (male/female)	39/33
Num of scan	203
Intrameatal only	89 (43.8%)
Small (0-10 mm)	35 (17.2%)
Medium (11-20 mm)	53 (26.1%)
Moderately large (21-30 mm)	17 (8.4%)
Large (31-40 mm)	9 (4.4%)
Giant (>40 mm)	0 (0.0%)

Note: The average age of the first scan of each patient was presented as mean ± standard deviation

Before simulating low-dose T1ce, a rigid registration was performed using Elastix [25] to ensure the spatial alignment between T1 and standard-dose T1ce. Inspired by Müller-Franzes et al. [18], we approximated the signal intensity of the low-dose images via a linear transformation. For contrast-enhanced T1, the relaxation time can be expressed as:

$$\frac{1}{T_1^{ce}} - \frac{1}{T_1} = r \times c \tag{4.1}$$

where T_1^{ce} and T_1 are the post- and pre-contrast T1 relaxation times, respectively, and r and c are the relaxivity and concentration of the contrast agent, respectively. When TR is short, the signal intensity can be approximated as proportional to the inverse of the T1 relaxation time [26], and thus to the concentration of the contrast agent c. Therefore, the low-dose images can be simulated as follows:

$$I_{low-dose} = (1 - \beta\%) \times I_{T1} + \beta\% \times I_{T1ce}\,, \tag{4.2}$$

where I_{T1} and I_{T1CE} refer to the intensity on T1 and T1ce, respectively, and $I_{(}low-dose)$ refers to the intensities on contrast-enhanced T1 with $\beta\%$ of the standard-dose. Intensity calibration was performed on T1ce before simulation to eliminate the influence of different intensity windows used by T1 and T1ce. Specifically, we assumed the intensity histogram of T1 and T1ce had a similar distribution, differing only in translation and scaling. We then applied translation and scaling that was determined

Table 4.2: MRI acquisition parameters

Modality	Pulse sequence	Magnetic field (T)	TE(ms)	TR(ms)	in-plane resolution (mm)	Slice thickness (mm)
T1	SE	1.0	20.0	546.9	0.45×0.45	3.0
		1.5	10.0 (6.0 – 24.0)	550.0 (370.8 – 698.9)	0.37×0.37 ($0.32 \times 0.32 - 0.86 \times 0.86$)	3.0 (1.0 – 6.0)
		3.0	9.0 (7.0 – 16.0)	750.0 (400.0 – 958.4)	0.35×0.35 ($0.27 \times 0.27 - 0.66 \times 0.66$)	1.0 (1.0 – 3.0)
	GR	1.0	6.9 (6.7 – 6.9)	26.0 (23.0 – 30.0)	0.47×0.47 ($0.35 \times 0.35 - 0.90 \times 0.90$)	1.4 (1.2 – 3.0)
		1.5	4.6 (2.4 – 5.6)	25.0 (8.6 – 1690.0)	0.59×0.59 ($0.39 \times 0.39 - 1.00 \times 1.00$)	1.4 (1.0 – 6.0)
		3.0	3.7 (2.2 – 4.7)	20.0 (8.1 – 25.0)	0.63×0.63 ($0.33 \times 0.33 - 0.94 \times 0.94$)	2.0 (0.6 – 3.0)
T1ce	SE	1.0	20.0	546.9	0.45×0.45	3.0
		1.5	11.0 (6.0 – 24.0)	550.0 (370.8 – 698.9)	0.37×0.37 ($0.32 \times 0.32 - 0.86 \times 0.86$)	2.0 (1.0 – 5.0)
		3.0	9.0 (7.0 – 15.0)	786.0 (500.1 – 958.4)	0.35×0.35 ($0.27 \times 0.27 - 0.66 \times 0.66$)	1.0 (1.0 – 3.0)
	GR	1.0	6.9 (6.7 – 6.9)	26.0 (23.0 – 30.0)	0.47×0.47 ($0.35 \times 0.35 - 0.90 \times 0.90$)	1.4 (1.2 – 3.0)
		1.5	4.6 (2.4 – 6.3)	25.0 (8.6 – 2200.0)	0.59×0.59 ($0.39 \times 0.39 - 1.00 \times 1.00$)	1.4 (0.9 – 3.0)
		3.0	3.7 (2.2 – 4.7)	20.0 (8.1 – 25.0)	0.63×0.63 ($0.33 \times 0.33 - 0.94 \times 0.94$)	2.0 (0.6 – 3.0)

Note: The technical features are presented as the median value with the range in parentheses. Note that there is only one SE scan was acquired using a magnetic field of 1.0 T. T1ce contrast-enhanced T1-weighted MRI, TE echo time, TR repetition time, SE Spin echo, GR Gradient echo

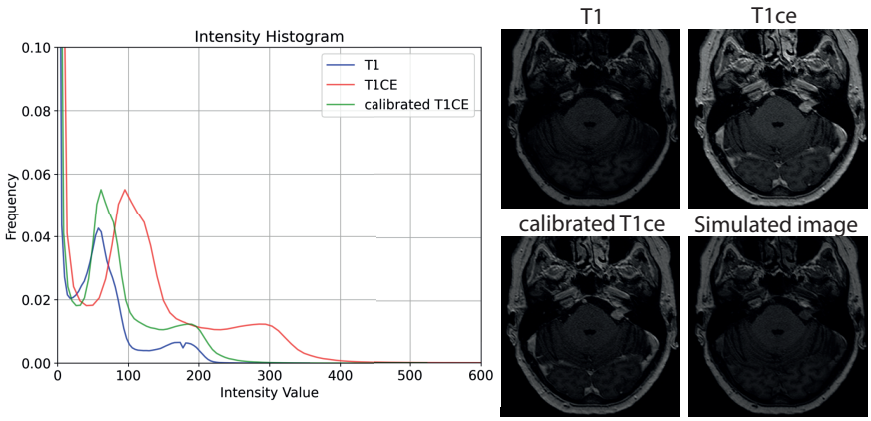


Figure 4.2: An example of low-dose T1ce simulation (20 % of the standard-dose). The histograms of T1, T1ce, and calibrated T1ce are shown on the left, where we can observe an intensity shift between T1ce and T1. T1, T1ce, calibrated T1CE, and a simulated low-dose image are visualized on the right.

by the second peak of each histogram (the first peak is the zero background) to the T1ce. One simulation example is shown in Figure 4.2 with the histograms before and after calibration. All images were normalized between -1 to 1 after simulation. Instead of using one specific proportion of dose reduction, we constructed scans of different simulated doses by varying the parameter β in Eq. (1). As a result, nine different low-dose T1 scans ranging from 10 % to 90 % of the standard dose with a 10 % increment were simulated.

Deep learning model for MRI restoration

We used the deep learning model previously developed by our group [27], which employs implicit neural representations, to restore standard-dose T1ce from simulated low-dose T1ce. Different from traditional convolutional neural network models, implicit neural representations use fully connected layers as the decoder and output 2D images as continuous fields. We trained separate models for each group of low-dose T1ce with different dose reductions, resulting in ten different models (nine models for low-dose T1ce and one model for T1 as input). All images were cropped based on the bounding box of the skull to reduce the workload of the model. Data augmentation, including random flipping, shifting, and cropping, was performed during training. The models were trained by a combination of ℓ_1 reconstruction loss, adversarial loss, and ℓ_2 regularization. We used the Adam optimizer with an initial learning rate of $1e^{-4}$, and followed the Two Time-scale Update Rule [28] to train the model. We refer interested readers to more details of the model architecture and training in reference [27].

The model inference was performed with a sliding window strategy. Specifically, image patches with a size of 128×160 and a step size of half that (64 in height and 80 in width) were used as the input of the model. The results were aggregated using Gaussian filter smoothing. Since the enhanced regions account for only a small proportion of T1ce, we cropped the images based on the bounding box of the tumor and evaluated the image quality of these sub-images. We subsequently tested the DL-restored T1ce in a downstream delineation task using a deep learning segmentation model previously developed by Neve et al. [29] based on nnUNet [30]. That tumor segmentation model was trained on standard-dose T1ce scans to delineate the intra- and extrameatal portions of VS. Both the training and inference were performed on a mixed computation server equipped with NVIDIA Quadro RTX6000 and NVIDIA Quadro RTXA6000 GPUs using Python v3.6.8 and Pytorch v1.10.2. The full codes and experimental settings are available at https://github.com/RianneAr/CPAMRIrestoration.

Clinical reader study

A senior head and neck neuroradiologist with 24 years of experience rated synthetic images based on multiple criteria, as shown in Table 4.3. To prevent bias towards patients with more scans, we randomly select one or two scans per patient from the test set for the assessment. For each study, two DL-restored T1ce images restored from 10 % and 30 % input dose were displayed together with T1 and standard-dose T1ce. The reader study is not blinded, as we aim to evaluate the clinical quality of the restored images, rather than making a visual comparison between the restored images and the ground truth. The radiologist was asked to compare DL-restored MRIs with standard-dose T1ce in terms of image quality of the entire field of view (FOV) and detectability of the lesion. The results were rated on a Likert scale from 1 (DL-restored image is better than the standard-dose T1ce) to 5 (DL-restored image is worse than the standard-dose T1ce). Moreover, the radiologist was asked to identify differences in enhancement patterns, including texture, brightness, and enhancement boundary, between the real and DL-restored T1ce, and to judge if the DL-restored MRI would still be sufficient for diagnostic characterization given observed differences. Lastly, a visual comparison rating between the two DL-restored images on a Likert scale from 1 (image restored from 10 % dose input is much better) to 5 (image restored from 30 % input dose is much better) was performed.

Evaluation and statistical analysis

We evaluated the DL-restored T1ce based on the image quality within the region of interest, defined by the bounding box of VS, as well as the performance on the VS delineation task. A comparison was made between the low-dose T1ce and DL-

Table 4.3: Questions used in clinical reader study

Qualitative evaluation of the DL-restored images					
	1	2	3	4	5
Overall image quality	Better	Slightly better	Similar	Slightly worse	Worse
Detectability of the lesion	Better	Slightly better	Similar	Slightly worse	Worse
Qualitative evaluation of the enhanced pattern of the DL-restored images					
Enhancement pattern	Similar	Different			
Sufficient for diagnostic characterization	Yes	No			
Qualitative comparison between images restored from 10% and 30% dose input on Likert scales					
	1	2	3	4	5
Comparison of diagnostic values	10 % dose is much better	10 % dose is better	10 % dose is equivalent to 30 % dose	30 % dose is better	30 % dose is much better

Note: The overall image quality, detectability of the lesion, and comparison of diagnostic values between images restored from different input doses were rated on a Likert scale. The enhancement pattern and diagnostic characterization were assessed in a binary manner. DL deep learning, T1ce contrast-enhanced T1-weighted MRI.

restored T1ce. The image quality was quantified by measuring the image similarity between the region of interest and the corresponding region in standard-dose T1ce via structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR). The performance on the VS delineation task was quantified by the Dice coefficient, the 95 % Hausdorff distance (95HD), and the average surface distance (ASD). The Wilcoxon signed rank test was performed between the image quality of low-dose images and corresponding DL-restored images. P values less than 0.001 were considered to indicate statistically significant differences. All analyses were performed in Python v3.6.8 with Numpy v1.21.5, SciPy v1.7.3, and scikit-learn v1.0.2.

4.3 Results

A total of 203 MRI studies from 72 patients (mean age, 58.51 ± 14.73, median age, 61 ($29-83$), 39 men) were involved in this study. The data were divided into three groups at the patient level, resulting in a training set of 130 studies, a validation set of 25 studies, and a test set of 48 studies.

Quantitative performance on image similarity

The image quality of low-dose T1ce (before restoration) and DL-restored images is shown in Table 4.4. The DL-restoration demonstrated the most pronounced enhancement at the lowest input dose (0 %) and gradually diminished as the input dose increased. Specifically, the SSIM for low-dose T1ce ranged from 0.562 ± 0.128 to 0.997 ± 0.001, and for DL-restored T1ce ranged from 0.639 ± 0.113 to 0.993 ± 0.009. The PSNR for low-dose T1ce ranged from 19.9 ± 4.08 to $39.9 \pm 4.08 dB$, and for DL-restored T1ce ranged from 21.6 ± 3.73 to 41.4 ± 4.84 dB. The enhancement obtained from AI restoration was statistically significant ($P < .001$) at lower dose levels ($\leq 70\%$) but became marginal for small dose reductions (input dose $\geq 80\%$). Figure 4.3 shows examples of DL-restoration based on different input doses.

Downstream task analysis

Figure 4.4 shows the quantitative results of deep learning segmentation using low-dose MRI and DL-restored MRI. Like the image quality, the enhancement from DL-restored images decreased with the increasing dose. At 10 % input dose, the average Dice across the whole tumor, the intra- and extrameatal portion was 0.734 for DL-restored MRI and 0.674 for low-dose MRI. The average 95HD was 2.07 mm for DL-restored MRI and 2.38 mm for low-dose MRI. The average ASD was 0.59 mm for DL-restored MRI and 1.00 mm for low-dose MRI. At 30 % input dose, the DL-restoration still improved the Dice of both the whole tumor and intrameatal portion by 0.02, but

Table 4.4: Quantitative image quality of DL-restored images compared to low-dose T1ce (before restoration)

Input dose(%)		0	10	20	30	40	50	60	70	80	90
SSIM	before restoration	0.562 ±0.128	0.641 ±0.107	0.717 ±0.086	0.785 ±0.066	0.845 ±0.047	0.896 ±0.032	0.936 ±0.019	0.965 ±0.010	0.985 ±0.004	0.997 ±0.001
	DL-restored	0.639 ±0.113	0.726 ±0.104	0.812 ±0.077	0.881 ±0.053	0.921 ±0.034	0.947 ±0.024	0.966 ±0.018	0.979 ±0.011	0.987 ±0.008	0.993 ±0.009
PSNR(dB)	before restoration	19.9 ±4.08	20.8 ±4.08	21.8 ±4.08	23.0 ±4.08	24.3 ±4.08	25.9 ±4.08	27.9 ±4.08	30.4 ±4.08	33.9 ±4.08	39.9 ±4.08
	DL-restored	21.6 ±3.73	23.0 ±3.82	25.6 ±2.82	28.4 ±3.10	30.1 ±3.21	31.8 ±3.47	33.3 ±3.89	35.1 ±4.29	37.3 ±4.92	41.4 ±4.84

Note: The mean and standard deviation of SSIM and PSNR between the images and the standard-dose T1ce calculated on the tumor area are reported. * indicates significant difference ($P < .001$) between the low-dose MRI and corresponding DL-restored MRI. SSIM structural similarity index measure, PSNR peak signal-to-noise ratio, DL deep learning, T1ce contrast-enhanced T1-weighted MRI.

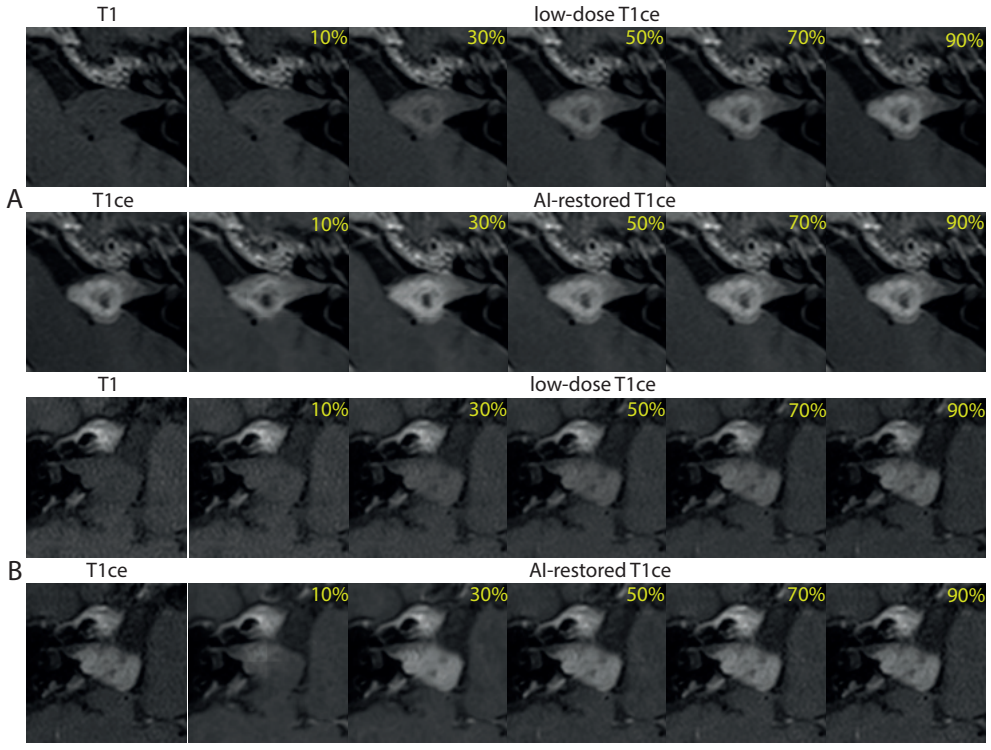


Figure 4.3: Two examples (A and B) of MRI restored from different simulated low-dose T1ce. For each example, the leftmost column shows T1 (upper) and standard-dose T1ce (lower). The second to sixth columns show low-dose T1ce (upper) and their corresponding DL-restored MRI (lower). The input doses in percentage are indicated in yellow on each image. The DL-restored T1ce demonstrates a lesion enhancement, most pronounced at low input dose. However, as illustrated in B (10 %), enhancement and texture restoration may be incomplete at very low input doses.

no significant differences were observed in HD95 and ASD between low-dose MRI and DL-restored MRI. From 30 % upwards, the segmentation performance difference between DL-restored images and low-dose images is less pronounced. Two examples comparing segmentation using DL-restored MRI and low-dose MRI directly are shown in Figure 4.5.

Clinical reader study

Twenty-two scans were randomly selected from the test set for the clinical reader study. The results are shown in Table 4.5. Evaluation of the entire FOV revealed no significant differences in image quality (3, interquartile range(IQR) [Q3-Q1] = 3 – 3 vs. 3, IQR [Q3-Q1] = 4 – 3) and lesion detectability (3, IQR [Q3-Q1] =

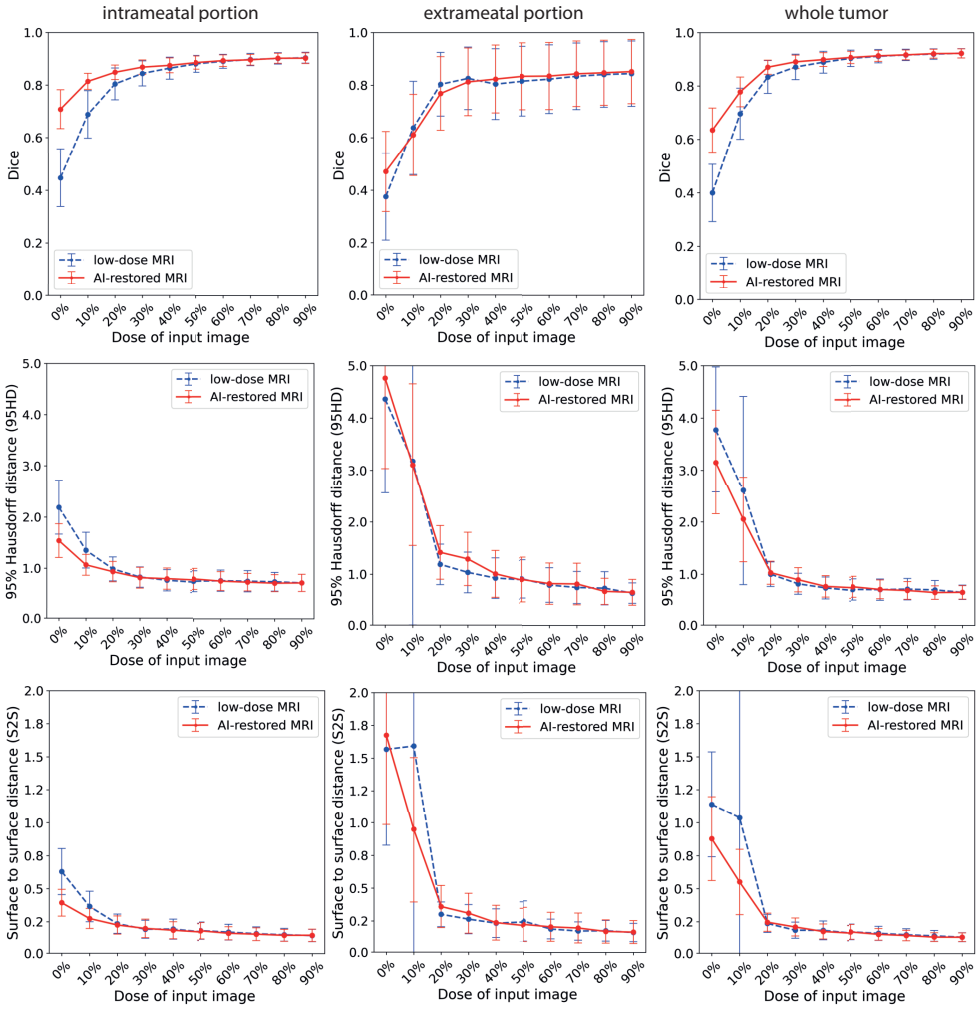


Figure 4.4: Quantitative segmentation results using different dose inputs, with 95 % confidence intervals shown as error bars. In each plot, the segmentation results using DL-restored MRI are indicated as a red solid line, and results using low-dose MRI are indicated as a blue dashed line. The left, middle, and right columns show plots of the intrameatal portion, extrameatal portion, and the whole tumor, respectively.

4 – 3 vs. 3, IQR [Q3-Q1] = 3 – 3) between the DL-restored images of 10 % and 30 % low-dose input compared to standard-dose T1ce. However, the images restored from a 30 % input dose were judged to be more informative than those from a 10 % input dose, with a median rating of 4, IQR [Q3-Q1] = 4 – 3, due to better texture restoration. Specifically, 32% (7 out of 22) of the images restored from 10 % input dose demonstrated similar enhancement patterns to the standard-dose T1ce, while

Table 4.5: The results of the clinical reader study

	Restored from 10 % dose	Restored from 30 % dose
Overall image quality	3 (IQR [Q3-Q1] = 3 – 3)	3 (IQR [Q3-Q1] = 4 – 3)
Detectability of lesion	3 (IQR [Q3-Q1] = 4 – 3)	3 (IQR [Q3-Q1] = 3 – 3)
Enhancement pattern	7/15 (32%)	16/8 (73%)
Sufficiency for diagnostic	10/12 (45%)	19/3 (86%)
comparison	4 (IQR [Q3-Q1] = 4 – 3)	

Note: The median with IQR [Q3-Q1] in parentheses of the Likert scale were reported. Results of enhancement pattern and sufficiency of diagnostic characterization were presented as the number of patients with percentage in parentheses. T1ce contrast-enhanced T1-weighted MRI, IQR interquartile range

this was the case for 73% (16 out of 22) of the images restored from 30 % input dose. The differences in enhancement pattern on scans restored from 10 % input dose included enhancement outside the tumor (8), (partial) lack of enhancement (6), decreased brightness (5), and cystic components being imperceptible (1). Among the six dissimilar scans that were restored from 30 % input dose, four scans showed a slight overestimation of the intrameatal portion, and two scans showed different brightness in the lesion. Given that under- or overestimation of the intrameatal portion has no major impact on clinical decision-making [3], 45% of the images restored from 10 % dose input were deemed sufficient for diagnosis, compared to 86% of the images restored with a 30 % dose input.

4.4 Discussion

In this study, we retrospectively evaluated a deep learning model to restore standard-dose contrast-enhanced MRI of the cerebellopontine angle from simulated low-dose MRI. We evaluated the restored images using multiple metrics to determine the clinical applicability of the model. Using simulated data enabled us to evaluate the proposed model with retrospective data at a large scale and explore the impact of different dose reductions on the deep learning model's performance. Our experiments showed that deep learning models can significantly improve image quality, especially when the input dose is substantially reduced. With DL-restored images, it is possible for lesion detection and further diagnostic characterization with reduced contrast doses.

In previous GBCA dose reduction studies of brain MRI, only a limited number of MRIs of the CPA were included [19, 21, 22]. For instance, of the 83 patients described by Luo et al. [19], who underwent zero-dose, 10 % dose, and standard-dose brain MRI, only five had a VS. Our study, by contrast, included both the diagnostic and follow-up MRI scans of 72 vestibular schwannoma patients from 33 different medical

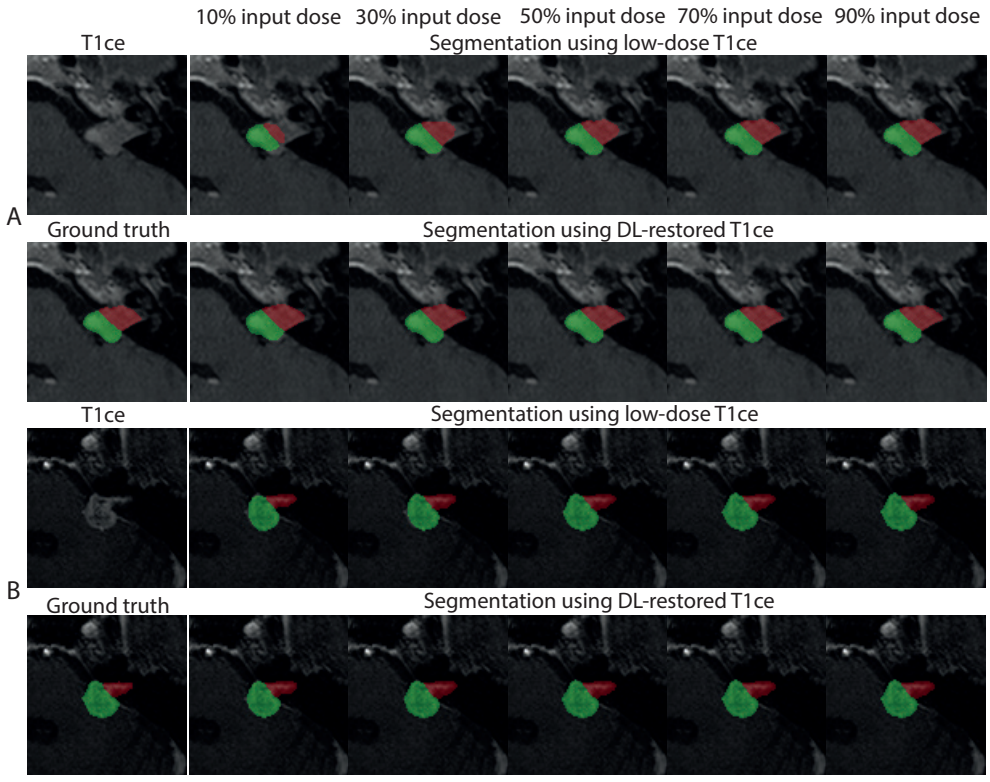


Figure 4.5: Two examples of downstream segmentation experiments. The standard-dose T1ce and corresponding tumor masks are displayed in the leftmost column. Columns 2 to 6 compare segmentation results between low-dose T1ce and DL-restored T1ce. For each patient, the top row shows segmentation using low-dose MRI, and the bottom row shows segmentation using DL-restored MRI. In patient A, segmentation at low input doses underestimates the intrameatal portion (10 %, 30 %, and 50 % input dose) and extrameatal portion (10 % input dose). In contrast, the lesion was correctly segmented on restored T1ce.

centers, which enabled more detailed and robust enhancement evaluations related to CPA tumors. Moreover, due to the restrictions of low-dose imaging protocols, the values of dose reduction were usually determined empirically, and the scope of datasets of previous studies was generally limited. In this retrospective study, we simulated low-dose T1ce with various dose reductions and trained separate models for each group of low-dose T1ce. A comprehensive evaluation on the impact of different dose reductions on deep learning models was conducted.

Different from previous studies, we evaluated the image quality of the DL-restored T1ce based on local image similarity to the standard-dose T1ce via SSIM and

PSNR. In the experiment, DL-restoration demonstrated significant improvement both qualitatively and quantitatively compared to the original low-dose T1ce. We observed that DL-restoration was most useful when the input dose was substantially reduced, and the quantitative improvements decreased with increasing input dose. This is in line with the previous study [23], in which MRI with doses ranging from 5-25 % of the standard dose were used. As a further step, our study validated the effectiveness of the deep learning model in restoring the enhanced regions across a broader range of dose reductions.

In order to show the clinical applicability of the DL-restored images, we investigated the performance of the restored images in a downstream segmentation task. A previous study by Pasumarthi et al. [21] conducted a segmentation evaluation with a test set of limited size ($N = 17$), and the results lack a quantitative statistical comparison. In our study, we used the state-of-the-art nnUNet model for evaluation and conducted a comprehensive study. The results suggest that the AI restoration enhanced the segmentation performance, especially when the GBCA input dose was substantially reduced.

In addition to deep learning segmentation, a clinical reader study was performed to evaluate the clinical impact of DL-restored MRI. According to the radiologist, both images restored from 10 % input dose and 30 % input dose showed equally excellent performance in terms of overall image quality and lesion detectability. However, more different enhancement patterns were observed in images restored from a 10 % input dose. This indicated that images restored from a 30 % input dose demonstrated better texture restoration of the tumor, offering more information on cystic changes, which are associated with a faster growth rate, and post-therapeutic changes. It is also worth noting that there was no pronounced overestimation in the extrameatal portions of all scans, which is important for precise diameter measurements of the follow-up scans. We noticed that the radiologist sometimes rated DL-restored images higher than the real T1ce because of the high signal-to-noise ratio. This is partly due to the natural smoothing effect of the pixel-wise loss, which encourages models to produce pixel-wise averages [31], used in the DL-restoration method.

While using simulation data has the advantage of scalable dose reduction, the inherent disadvantage is that we did not account for the noise from dose reduction and different TR during the simulation, and did not evaluate the model on real low-dose images either. Future studies using real low-dose T1ce scans should be performed to confirm the performance of DL-restoration with the actual reduction of contrast agent. Moreover, the assumption of a linear relationship between contrast dose and enhancement does not account for other factors that may influence lesion conspicuity, such as the type of GBCA [32] and wait time after injection [33]. Although the low-

dose T1ce was simulated based on standard-dose T1ce, from which we know exactly what intensity difference the contrast agent provides, follow-up studies that involve relaxivity information for model development would be promising. Additionally, in this study, the T2 scan, which is commonly also part of VS care, was not used by the DL-restoration model. We expect that the involvement of additional modalities can further improve the performance of DL-based restoration. Lastly, the proposed model has demonstrated a limited contribution to images with high input doses. We attribute this to the fact that images with higher input doses are already sufficient for tumor delineation, meaning there are no significant improvements for the model to be made. Lastly, this study was limited to a retrospective dataset. A prospective study for feasibility evaluation would be valuable before the model deployment.

In conclusion, our study has shown that a deep learning restoration model may improve the image quality of MR imaging of the CPA with reduced contrast agent doses. By using AI restoration, VS lesion detection can be done with only 10 % of the standard dose, and further diagnostic characterization is possible with 30 % of the standard dose. In addition, this deep learning technique could also be applied to other pathologies that are visualized with contrast-enhanced MR imaging, in order to reduce the need for contrast agents on a broader scale.

4.5 Acknowledgements

This work was supported by the China Scholarship Council (grant 202008130140) and by an unrestricted grant of Stichting Hanarth Fonds, The Netherlands (project MLSCHWAN).

References

[1] E. Lin and B. Crane. “The management and imaging of vestibular schwannomas”. In: *American Journal of Neuroradiology* 38.11 (2017), pages 2034–2043.

[2] K. Singh, M. P. Singh, C. Thukral, et al. “Role of magnetic resonance imaging in evaluation of cerebellopontine angle schwannomas”. In: *Indian Journal of Otolaryngology and Head & Neck Surgery* 67.1 (2015), pages 21–27.

[3] L. Lassaletta, L. A. Cervera, X. Altuna, et al. “Clinical practice guideline on the management of vestibular schwannoma”. In: *Acta Otorrinolaringologica (English Edition)* 75.2 (2024), pages 108–128.

[4] J. G. Smirniotopoulos, F. M. Murphy, E. J. Rushing, et al. “Patterns of contrast enhancement in the brain and meninges”. In: *Radiographics* 27.2 (2007), pages 525–551.

[5] D. Annesley-Williams, R. Laitt, J. Jenkins, et al. “Magnetic resonance imaging in the investigation of sensorineural hearing loss: is contrast enhancement still necessary?” In: *The Journal of Laryngology & Otology* 115.1 (2001), pages 14–21.

[6] C. S. Graffeo, W. Sivakumar, S. A. Tavakol, et al. “Congress of Neurological Surgeons Systematic Review and Evidence-Based Guidelines Update for the Role of Imaging in the Management of Patients With Vestibular Schwannomas”. In: *Neurosurgery* (2025).

[7] R. Goldbrunner, M. Weller, J. Regis, et al. “EANO guideline on the diagnosis and treatment of vestibular schwannoma”. In: *Neuro-oncology* 22.1 (2020), pages 31–45.

[8] V. A. R. Silva, J. Lavinsky, H. F. Pauna, et al. “Brazilian Society of Otology task force–Vestibular Schwannoma–evaluation and treatment”. In: *Brazilian journal of otorhinolaryngology* 89.6 (2023), page 101313.

[9] D. H. Coelho, Y. nnU-NetTang, B. Suddarth, and M. Mamdani. “MRI surveillance of vestibular schwannomas without contrast enhancement: clinical and economic evaluation”. In: *The Laryngoscope* 128.1 (2018), pages 202–209.

[10] S. Coimbra, S. Rocha, N. R. Sousa, et al. “Toxicity mechanisms of gadolinium and gadolinium-based contrast agents—a review”. In: *International Journal of Molecular Sciences* 25.7 (2024), page 4071.

[11] J. W. Choi and M. Won-Jin. “Gadolinium deposition in the brain: current updates”. In: *Korean journal of radiology* 20.1 (2019), page 134.

[12] H. M. Dekker, G. J. Stroomberg, A. J. Van der Molen, and M. Prokop. “Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents”. In: *Insights into Imaging* 15.1 (2024), page 62.

[13] M. G. Crowson, D. J. Rocke, J. K. Hoang, et al. “Cost-effectiveness analysis of a non-contrast screening MRI protocol for vestibular schwannoma in patients with asymmetric sensorineural hearing loss”. In: *Neuroradiology* 59.8 (2017), pages 727–736.

[14] B. P. Liu, M. Rosenberg, P. Saverio, et al. “Clinical efficacy of reduced-dose gadobutrol versus standard-dose gadoterate for contrast-enhanced MRI of the CNS: an international multicenter prospective crossover trial (LEADER-75)”. In: *American Journal of Roentgenology* 217.5 (2021), pages 1195–1205.

[15] D. He, A. Chatterjee, X. Fan, et al. “Feasibility of dynamic contrast-enhanced magnetic resonance imaging using low-dose gadolinium: comparative performance with standard dose in prostate cancer diagnosis”. In: *Investigative Radiology* 53.10 (2018), pages 609–615.

[16] F. Pineda, D. Sheth, H. Abe, et al. “Low-dose imaging technique (LITE) MRI: initial experience in breast imaging”. In: *The British Journal of Radiology* 92.1103 (2019), page 20190302.

[17] A. N. Melsaether, E. Kim, E. Mema, et al. “Preliminary study: breast cancers can be well seen on 3T breast MRI with a half-dose of gadobutrol”. In: *Clinical imaging* 58 (2019), pages 84–89.

[18] G. Müller-Franzes, L. Huck, S. Tayebi Arasteh, et al. “Using machine learning to reduce the need for contrast agents in breast MRI through synthetic images”. In: *Radiology* 307.3 (2023), e222211.

[19] H. Luo, T. Zhang, N.-J. Gong, et al. “Deep learning-based methods may minimize GBCA dosage in brain MRI”. In: *European Radiology* 31.9 (2021), pages 6419–6428.

[20] E. Gong, J. M. Pauly, M. Wintermark, and G. Zaharchuk. “Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI”. In: *Journal of magnetic resonance imaging* 48.2 (2018), pages 330–340.

[21] S. Pasumarthi, J. I. Tamir, S. Christensen, et al. “A generic deep learning model for reduced gadolinium dose in contrast-enhanced brain MRI”. In: *Magnetic Resonance in Medicine* 86.3 (2021), pages 1687–1700.

[22] S. Ammari, A. Bône, C. Balleyguier, et al. “Can deep learning replace gadolinium in neuro-oncology?: a reader study”. In: *Investigative Radiology* 57.2 (2022), pages 99–107.

[23] G. Müller-Franzes, L. Huck, M. Bode, et al. “Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast MRI”. In: *European Radiology Experimental* 8.1 (2024), page 53.

[24] J. Kanzaki, M. Tos, M. Sanna, and D. A. Moffat. “New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma”. In: *Otology & neurotology* 24.4 (2003), pages 642–649.

[25] S. Klein, M. Staring, K. Murphy, et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pages 196–205.

[26] R. E. Hendrick. *Breast MRI: fundamentals and technical aspects*. 2008.

[27] Y. Chen, M. Staring, O. M. Neve, et al. “CoNeS: Conditional neural fields with shift modulation for multi-sequence MRI translation”. In: *The journal Machine Learning for Biomedical Imaging* 2.special issue for Generative Models (2024).

[28] M. Heusel, H. Ramsauer, T. Unterthiner, et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).

[29] O. M. Neve, Y. Chen, Q. Tao, et al. “Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study”. In: *Radiology: Artificial Intelligence* 4.4 (2022), e210300.

[30] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pages 203–211.

[31] C. Ledig, L. Theis, F. Huszár, et al. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 4681–4690.

[32] J. Hao, C. Pitrou, and P. Bourrinet. “A comprehensive overview of the efficacy and safety of gadopiclenol: a new contrast agent for MRI of the CNS and body”. In: *Investigative Radiology* 59.2 (2024), pages 124–130.

[33] P. Robert, V. Vives, A.-L. Grindel, et al. “Contrast-to-dose relationship of gadopiclenol, an MRI macrocyclic gadolinium-based contrast agent, compared with gadoterate, gadobenate, and gadobutrol in a rat brain tumor model”. In: *Radiology* 294.1 (2020), pages 117–126.

5

Vestibular schwannoma growth prediction from longitudinal MRI by time-conditioned neural fields

This chapter was adapted from:

Chen, Y., Wolterink, J.M., Neve, O.M., Romeijn, S.R., Verbist, B.M., Hensen, E.F., Tao, Q. and Staring, M., 2024. Vestibular schwannoma growth prediction from longitudinal MRI by time-conditioned neural fields. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 508-518)

Abstract

Vestibular schwannomas (VS) are benign tumors that are generally managed by active surveillance with MRI examination. To further assist clinical decision-making and avoid overtreatment, an accurate prediction of tumor growth based on longitudinal imaging is highly desirable. In this paper, we introduce DeepGrowth, a deep learning method that incorporates neural fields and recurrent neural networks for prospective tumor growth prediction. In the proposed model, each tumor is represented as a signed distance function (SDF) conditioned on a low-dimensional latent code. Unlike previous studies, we predict the latent codes of the future tumor and generate the tumor shapes from it using a multilayer perceptron (MLP). To deal with irregular time intervals, we introduce a time-conditioned recurrent module based on a ConvLSTM and a novel temporal encoding strategy, which enables the proposed model to output varying tumor shapes over time. The experiments on an in-house longitudinal VS dataset showed that the proposed model significantly improved the performance ($\geq 1.6\%$ Dice score and ≥ 0.20 mm 95 % Hausdorff distance), in particular for top 20% tumors that grow or shrink the most ($\geq 4.6\%$ Dice score and ≥ 0.73 mm 95 % Hausdorff distance). Our code is available at `https://github.com/cyjdswx/DeepGrowth`.

5.1 Introduction

Vestibular schwannomas (VS) are intracranial tumors arising from the balance and hearing nerves, of which approximately 40% are progressive and ultimately become life-threatening [1]. In current clinical practice, VS are generally managed by active surveillance with MRI examination and manual tumor diameter measurements [2, 3]. Once significant growth (>2mm difference between two consecutive MRI scans) is detected, the tumors are treated with either radiotherapy or surgery [2, 4]. However, research shows that although 80% of VS show certain growth during observation, only half of them are truly progressive, indicating that many patients suffer from overtreatment [4]. On the other hand, late treatment of a larger tumor can also damage the prognosis after treatment, which requires a timely clinical decision [5]. Hence, to avoid overtreatment and sequelae associated with the treatment of large tumors, early and precise prediction of tumor growth based on longitudinal imaging is highly desirable.

Early studies on image-driven tumor growth prediction typically utilized biomechanical models, such as reaction-diffusion equations, to derive physiological parameters related to tumor progression [6, 7]. However, most of the models require specific imaging modalities that are unfortunately not available in clinical routine for VS. Recently, deep learning models have shown promising performance for longitudinal tumor shape modeling. Inspired by the neural process framework, Petersen et al. proposed to learn a distribution of possible future shapes of glioma using a self-attention mechanism [8]. Instead of generative models, Zhang et al. [9] applied a spatio-temporal ConvLSTM for pancreatic tumor growth modeling. Elazab et al. [10] proposed a 3D GP-GAN that utilizes multiple stacked generative adversarial networks to predict glioma growth. Subsequently, Wang et al. [11] applied a Transformer model to longitudinal CT for 4D lung cancer tumor modeling. Although promising results were demonstrated, most models assume unified time intervals between consecutive scans, which is unfortunately uncommon in the clinic. Moreover, future prediction in high-dimensional image space has large memory requirements, which could limit application [9], and may also introduce spatial redundancy that potentially damage performance [12].

One way to tackle this problem is compressing the input into a low-dimensional latent code utilizing an autoencoder and performing predictions in the latent space [13, 14]. In line with this, we propose to perform future tumor prediction with neural field representations [15, 16]. The key idea of neural fields is to represent a function describing an image or object in the spatial or spatio-temporal domain as a neural network with trainable weights [17]. The neural network can be conditioned on latent

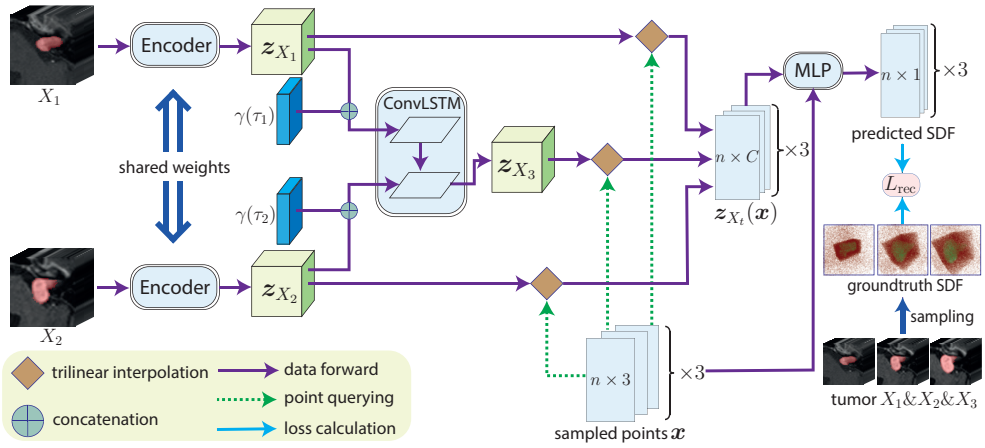


Figure 5.1: The overall architecture of DeepGrowth ($N = 3$). Prior scans are encoded into latent codes, which are concatenated with temporal encoding. The MLP reconstructs the future tumor as an SDF conditioned on the output of the ConvLSTM. L_{rec} is calculated between the predictions and SDF sampled from all three tumor masks.

codes to represent a distribution of objects. Recently, Agro et al. [18] successfully predicted future occupancy maps using spatio-temporal neural fields. However, the method requires sufficient frames over time, while longitudinal medical imaging usually contains only few measurements.

To address these limitations, we propose DeepGrowth, a model that incorporates neural fields and recurrent neural networks for tumor growth prediction. Specifically, DeepGrowth encodes prior images and tumor masks into latent codes and parameterizes the tumor as a signed distance function (SDF). To deal with irregular time intervals between scans, we apply a time-conditioned recurrent module to predict the latent code, on which the reconstruction of the future tumor shape is conditioned. The main contributions of this work are: (1) In contrast to previous studies that perform tumor prediction directly in image space, for the first time, we represent tumor shapes as neural fields and predict the future based on learned latent codes. (2) We introduce a time-conditioned recurrent module with a novel temporal encoding strategy that enables us to query tumor shapes at specific time intervals. (3) The proposed model was evaluated on an in-house longitudinal VS dataset, showing a significantly better performance than other models, in particular for relatively fast growing tumors.

5.2 Methods

Given a patient with N longitudinal images with corresponding segmentations, denoted as $X_t = \{I_t, M_t, D_t\}$, $t = 1, 2, ..., N$, where I_t is the image at time t, M_t is the corresponding tumor mask and D_t the normalized scan date ranging from 0 to 1, our goal is to find a function Φ:

$$\Phi : \{X_1, X_2, ..., X_{N-1}, D_N\} \rightarrow M_N. \tag{5.1}$$

Instead of performing prediction directly in image space, we encode X_t into a low-dimensional latent code and predict future by a time-conditioned recurrent module. See Figure 5.1 for an overview of the model architecture when $N = 3$.

3D tumor shape as signed distance function

In the proposed model, each tumor is encoded into a low-dimensional latent code, which can be used to condition a neural field for tumor shape reconstruction. More specifically, we concatenate $I_t \in \mathbb{R}^{D \times H \times W}$ and $M_t \in \mathbb{R}^{D \times H \times W}$, and encode them via a convolution-based encoder with a downsampling factor s. The latent code is denoted as $\boldsymbol{z}_{X_t} \in \mathbb{R}^{C \times d \times h \times w}$, where $d = D/s$, $h = H/s$, $w = W/s$, and C is the feature dimension. Unlike studies that use a single vector to represent the entire object [15, 16], our latent code contains $d \times h \times w$ vectors, encoding the local information in a more expressive representation [19].

To reconstruct tumor shapes from the latent code, we represent each tumor shape using an SDF [15]. For clarity, we use c_t to denote the tumor contour of M_t, which is a closed 2D manifold embedded in 3D space. Hence, for each M_t, the SDF of the tumor can be defined as:

$$\mathrm{SDF}_{M_t}(\boldsymbol{x}) = \begin{cases} \min_{u \in c_t} \|x - u\|_2, & \text{if } x \text{ inside } c_t \\ 0, & \text{if } x \text{ belonging to } c_t \\ -\min_{u \in c_t} \|x - u\|_2, & \text{if } x \text{ outside } c_t \end{cases} \tag{5.2}$$

where $\boldsymbol{x} = (x, y, z) \in \mathbb{R}^3$. Different from voxelized or meshed representations, the SDF and therefore $\boldsymbol{x}$ is defined over the entire space. In the proposed model, we approximate the SDF by a multilayer perceptron (MLP) f. Similar to [19, 20], we apply a local conditioning strategy, in which $\mathrm{SDF}_{M_t}(\boldsymbol{x})$ is conditioned on the local latent code $\boldsymbol{z}_{X_t}(\boldsymbol{x})$. $\boldsymbol{z}_{X_t}(\boldsymbol{x})$ is a vector of size C queried from the entire latent code $\boldsymbol{z}_{X_t}$ using trilinear interpolation [19]. For each point $\boldsymbol{x}$, we concatenate the coordinates $\boldsymbol{x}$ with $\boldsymbol{z}_{X_t}(\boldsymbol{x})$ as the input of the MLP, which can then be denoted as:

$$\mathrm{SDF}_{M_t}(\boldsymbol{x}) \approx f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}_{X_t}(\boldsymbol{x})), \tag{5.3}$$

where $\boldsymbol{\theta}$ are the parameters of the MLP. Hence, each tumor contour is described by the zero-level set of the SDF estimated by the MLP.

Time-conditioned recurrent module

Earlier studies on tumor prediction usually assume unified time intervals between consecutive scans [11] and predict a more distant future with additional recurrent steps [10]. However, patients frequently receive follow-up scans with irregular time intervals. We therefore introduce a time-conditioned recurrent module, which consists of temporal encoding and a small 3D ConvLSTM, to predict future tumor shapes. The 3D ConvLSTM takes the input of $\boldsymbol{z}_{X_t}$, $t = 1, 2, ..., N-1$ with the study dates D_t, $t = 1, 2, ..., N$ and predicts $\boldsymbol{z}_{X_N}$. To better encode the temporal information, we apply sinusoidal functions to the time intervals similar to positional encoding [21], which we call temporal encoding. Given the time interval $\tau_i = D_{i+1} - D_i$, where $i = 1, 2, ..., N-1$, the temporal encoding is expressed as follows:

$$\gamma(\tau_i) = [\sin(2^0 \pi \tau_i), \cos(2^0 \pi \tau_i), \ldots, \sin(2^{l-1} \pi \tau_i), \cos(2^{l-1} \pi \tau_i)], \tag{5.4}$$

where l is the order of the temporal encoding. To avoid overfitting, a dropout layer is added to the temporal encoding. We then concatenate $\gamma(\tau_i)$ to all vectors of $\boldsymbol{z}_{X_i}$ as the input of the ConvLSTM. Given the output $\boldsymbol{z}_{X_n}$ of the ConvLSTM, we can obtain SDF_{M_n} of the future tumor via Eq. (5.3).

End-to-end network training

All components are optimized together end-to-end. For training, we randomly sample n points from each tumor volume with 80 % of the points sampled near the contour and the rest sampled from the entire space. We apply an ℓ_1 reconstruction loss that maximizes the similarity between the real SDF and the estimations, as suggested in [15], for all N tumors:

$$L_{\mathrm{rec}} = \frac{1}{nN} \sum_{t=1}^{N} \sum_{i=1}^{n} \| f_\theta(\boldsymbol{x}_i, \boldsymbol{z}_{X_t}(\boldsymbol{x}_i)) - \mathrm{SDF}_{M_t}(\boldsymbol{x}_i) \|_1, \tag{5.5}$$

where $\boldsymbol{x}_i$ are the sampled points. To stabilize the training, we apply the ℓ_2 norm to the latent codes as the regularization: $L_{\mathrm{reg}} = \frac{1}{N} \sum_{t=1}^{N} \| \boldsymbol{z}_{X_t} \|_2$. As a result, the overall loss function of the proposed model is $L = \lambda_{\mathrm{rec}} L_{\mathrm{rec}} + \lambda_{\mathrm{reg}} L_{\mathrm{reg}}$, where λ_{rec} and λ_{reg} are the weights of each loss function.

Table 5.1: Quantitative comparison results on a vestibular schwannoma dataset using 5-fold cross-validation. The mean and standard deviation of Dice, 95 % HD, and RVD are reported. The highest values per column are indicated in bold; † indicates a significant difference ($p < .05$) compared to the proposed method.

Method	#params	Dice ↑	95 % HD (mm) ↓	RVD ↓
Stable tumor	N/A	$0.766 \pm 0.143^{\dagger}$	$1.95 \pm 2.55^{\dagger}$	$\mathbf{0.490 \pm 2.99}$
ST-ConvLSTM [9]	$0.6M$	$0.758 \pm 0.141^{\dagger}$	$2.07 \pm 2.65^{\dagger}$	$0.611 \pm 3.62^{\dagger}$
3D ConvLSTM [22]	$4.4M$	$0.784 \pm 0.139^{\dagger}$	$1.91 \pm 2.50^{\dagger}$	$0.564 \pm 3.47^{\dagger}$
DeepGrowth (proposed)	$4.9M$	$\mathbf{0.800 \pm 0.115}$	$\mathbf{1.71 \pm 2.23}$	0.521 ± 3.48

Table 5.2: Quantitative comparison results of the top 20% fastest growing or shrinking VS using 5-fold cross-validation. The mean and standard deviation of Dice, 95 % HD, and RVD are reported. The highest values per column are indicated in bold; † indicates a significant difference ($p < .05$) compared to the proposed method.

Method	#params	Dice ↑	95 % HD (mm) ↓	RVD ↓
Stable tumor	N/A	$0.697 \pm 0.182^{\dagger}$	$4.18 \pm 3.30^{\dagger}$	$0.413 \pm 0.323^{\dagger}$
ST-ConvLSTM [9]	$0.6M$	$0.707 \pm 0.188^{\dagger}$	$4.28 \pm 3.42^{\dagger}$	0.398 ± 0.470
3D ConvLSTM [22]	$4.4M$	$0.736 \pm 0.176^{\dagger}$	$3.87 \pm 3.22^{\dagger}$	0.366 ± 0.332
DeepGrowth (proposed)	$4.9M$	$\mathbf{0.782 \pm 0.120}$	$\mathbf{3.14 \pm 2.22}$	$\mathbf{0.321 \pm 0.315}$

5.3 Experiments

Dataset

To evaluate the proposed method, 131 vestibular schwannoma patients were selected from our previous study [23, 2]. Each patient in the dataset has three consecutive contrast-enhanced T1 (T1ce) scans, separated by 87 to 2157 days. The spatial resolution of the T1ce ranges from $0.254\,\text{mm} \times 0.254\,\text{mm} \times 0.81\,\text{mm}$ to $1.17\,\text{mm} \times 1.17\,\text{mm} \times 1.20\,\text{mm}$, and the in-plane resolution ranges from 256×192 to 640×520. Of all scans, the tumor masks were generated using a segmentation model developed in our previous study based on nnUNet [24, 23]. We aligned all scans of each patient by rigid registration using elastix [25]. All images were then resampled to an isotropic resolution of $0.58\,\text{mm} \times 0.58\,\text{mm} \times 0.58\,\text{mm}$. To avoid the influence of background, $64 \times 64 \times 64$ cropping was performed around the centroid of the tumor. The intensities of T1ce were normalized to $[-1, 1]$ and D_t was normalized to $[0, 1]$ within the original range of 0 to 10 years.

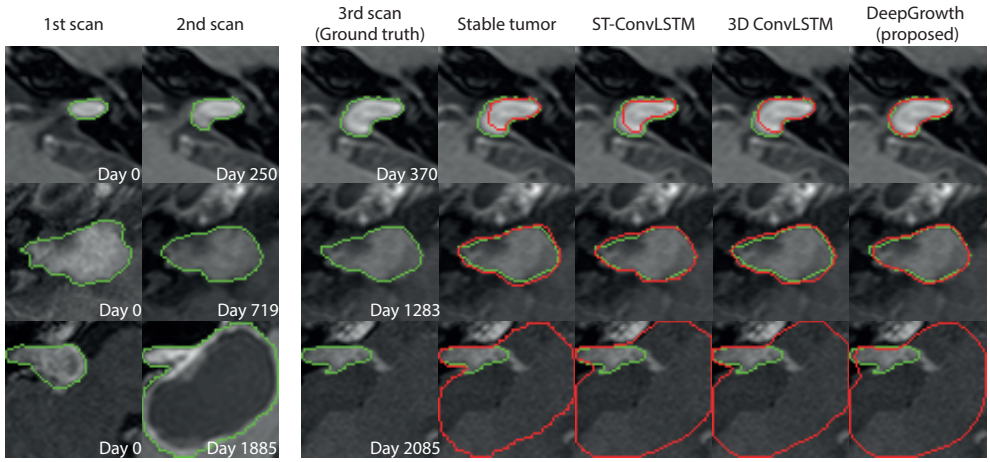


Figure 5.2: Example results of the different models. The first two columns are the input of the models, followed by the ground truth in the third column, and model predictions in subsequent ones. Predicted tumors are depicted in red and the ground truths in green. The dates are the study dates. The last row depicts a tumor that suddenly shrank after the second scan, which was difficult to predict for all models.

Implementation details

We adapted a 3D U-Net from [19] as the encoder with two extra convolutional blocks for downsampling. The ConvLSTM in the time-conditioned recurrent module consists of three 32-channel layers, and the MLP contains five 64-channel layers with sine as the activation function [26]. Due to the limited dataset size and diversity in tumor growth trends, we perform five-fold cross-validation and report the average results of the five folds to avoid bias. We set the downsampling factor $s = 4$ and temporal encoding order $l = 6$ for best performance (see Section 5.3). Little difference was observed between different loss weights, which were set to $\lambda_{\text{rec}} = 1.0$ and $\lambda_{\text{reg}} = 0.1$. The model was optimized using Adam with an initial learning rate of $1e-4$. During inference, the tumor masks were generated from the zero level-set of the predicted SDF and evaluated using the Dice, 95 % Hausdorff distance (95 % HD), and relative volume difference (RVD). All experiments were conducted using Python 3.10 and PyTorch 1.12.1 on a machine equipped with Nvidia Quadro RTX 6000 and Nvidia Tesla V100 GPUs.

Future tumor shape prediction

We first evaluate the proposed model by predicting the third future tumor shape from the first two scans and time intervals. We compare our model against three baselines.

Table 5.3: Quantitative results of top 20% growers when varying temporal encoding.

methods	order	Dice ↑	95 % HD (mm) ↓	RVD ↓
w/o time	N/A	0.765 ± 0.143	3.37 ± 2.49	0.341 ± 0.314
with time	N/A	0.774 ± 0.126	3.23 ± 2.50	0.316 ± 0.278
time + temporal encoding	$l = 4$	0.773 ± 0.144	3.33 ± 2.53	0.316 ± 0.306
	$l = 6$	$\mathbf{0.782 \pm 0.119}$	$\mathbf{3.14 \pm 2.22}$	0.321 ± 0.315
	$l = 8$	0.773 ± 0.142	3.23 ± 2.39	$\mathbf{0.315 \pm 0.363}$

The first baseline assumes the tumor remains stable after the second scan, which is reasonable due to the slow growth of VS, so we simply take the tumor mask from the second time point as a prediction, which we call "stable tumor" in the experiments. The second and third baselines are two ConvLSTM-based models: ST-ConvLSTM [9] and 3D ConvLSTM [22]. ST-ConVLSTM is a smaller 2D model where we use the same architecture as described in the original paper. 3D ConvLSTM, which contains a comparable number of parameters to the proposed model, consists of three layers with (64, 128, 64) channels respectively. Unlike the original papers that use the ℓ_1 loss to train the model to generate binary maps, we used a weighted sum of Dice loss and binary cross-entropy loss, which performed better on our data, to train the baselines. Wilcoxon signed rank tests were performed between the proposed model and each baseline.

The quantitative results are listed in Table 5.1 with visualizations in Figure 5.2. The proposed model performed significantly better than all baselines in terms of Dice and 95 % HD. The proposed model obtained a higher RVD due to an extreme outlier (see last row in Figure 5.2). When removing this outlier, the proposed method obtained an RVD of 0.218 ± 0.248, which outperformed all baselines (0.229 ± 0.230, 0.296 ± 0.304, and 0.261 ± 0.266, respectively).

We noticed that the stable tumor method obtained comparable quantitative scores, which is on par with the fact that many VS grow slowly or even remain stable. Focusing on the top 20% of tumors that grow (or shrink) the most, see Table 5.2, we observe a larger gap between the proposed model and the baselines, indicating the improved capability of modeling tumor growth.

Ablation study

To examine the impact of temporal encoding, we trained two additional models: one without time factors at all, and one using time intervals τ_i directly as suggested in [9]. We also compare the models using different orders l for temporal encoding. The results of the top 20% growers are shown in Table 5.3. Direct use of τ_i barely improved results, while temporal encoding improved the results for all metrics. Best results

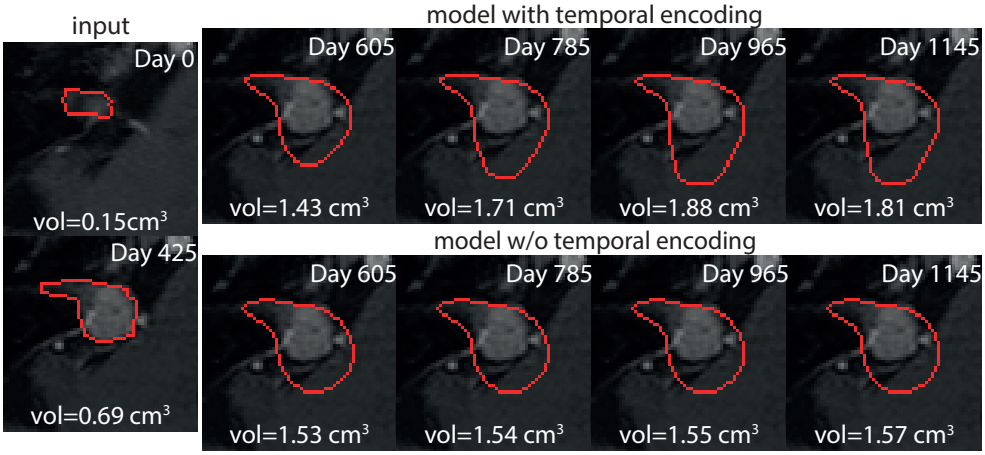


Figure 5.3: Querying the proposed model at different time points (increments of 180 days). We overlaid predictions on I_2 for visualization in columns 3-6. The proposed model can output varied tumor shapes given different time intervals, while the model without temporal encoding outputs almost the same results regardless of the time intervals.

Table 5.4: Quantitative results of DeepGrowth using different downsampling factors.

downsampling factor s	Dice ↑	95 % HD (mm) ↓	RVD ↓
$s = 1$	0.788 ± 0.127	1.87 ± 2.39	0.544 ± 3.68
$s = 2$	0.796 ± 0.122	1.78 ± 2.40	0.577 ± 4.18
$s = 4$	$\mathbf{0.800 \pm 0.115}$	$\mathbf{1.71 \pm 2.23}$	$\mathbf{0.52 \pm 3.48}$
$s = 8$	0.784 ± 0.125	1.85 ± 2.35	0.598 ± 4.10

were obtained for $l = 6$, with higher l leading to overfitting.

As our model allows us to query arbitrary future time points, we show predictions given different τ_2 (with a step of 180 days) in Figure 5.3. We can see that the model using τ_i without temporal encoding outputs almost the same results regardless of the time intervals. On the contrary, by using temporal encoding, the proposed model can output varied tumor shapes given different time intervals, from which we can view how tumors grow over time.

Models using high-resolution feature maps were more difficult to train, while lower-resolution feature maps potentially degraded performance due to lowered expressive capability [13, 14]. We, therefore, varied the downsampling factors s, see Table 5.4, and concluded that $s = 4$ resulted in the best performance.

5.4 Discussion and Conclusion

In this paper, we proposed DeepGrowth, a deep learning model that incorporates neural fields and recurrent neural networks for tumor growth prediction. Unlike conventional models that predict image or segmentation masks directly in the image space [9, 10], we encode tumors into a latent space and predict future latent codes. The future tumor shape is reconstructed as the zero-level set of an SDF conditioned on the predicted latent code via an MLP. A comparison on a longitudinal VS dataset showed improved performance of the proposed model, in particular for more challenging growing or shrinking tumors. We applied temporal encoding to the study intervals, which helped the model to encode time information and output varied tumor shapes given different time intervals. However, it remains to be investigated if tumor growth derived from our predictions can be used to aid clinical decision making. In conclusion, we showed that neural fields hold great promise for information compression, which can facilitate longitudinal tumor modeling.

5.5 Acknowledgements

This study was supported by the China Scholarship Council (grant 202008130140), and by an unrestricted grant of Stichting Hanarth Fonds, The Netherlands (project MLSCHWAN).

References

[1] M. L. Carlson and M. J. Link. "Vestibular schwannomas". In: *New England Journal of Medicine* 384.14 (2021), pages 1335–1348.

[2] O. M. Neve, S. R. Romeijn, Y. Chen, et al. "Automated 2-Dimensional Measurement of Vestibular Schwannoma: Validity and Accuracy of an Artificial Intelligence Algorithm". In: *Otolaryngology–Head and Neck Surgery* 169.6 (2023), pages 1582–1589.

[3] J. Kanzaki, M. Tos, M. Sanna, and D. A. Moffat. "New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma". In: *Otology & neurotology* 24.4 (2003), pages 642–649.

[4] J. P. Marinelli, M. J. Link, and M. L. Carlson. "Size Threshold Surveillance-A Revised Approach to Wait-and-Scan for Vestibular Schwannoma". In: *JAMA Otolaryngology–Head & Neck Surgery* 149.8 (2023), pages 657–658.

[5] D. Li, A. Tsimpas, and A. V. Germanwala. "Analysis of vestibular schwannoma size: A literature review on consistency with measurement techniques". In: *Clinical neurology and neurosurgery* 138 (2015), pages 72–77.

[6] Y. Liu, S. M. Sadowski, A. B. Weisbrod, et al. "Patient specific tumor growth prediction using multimodal images". In: *Medical image analysis* 18.3 (2014), pages 555–566.

[7] N. Meghdadi, M. Soltani, H. Niroomand-Oscuii, and N. Yamani. "Personalized image-based tumor growth prediction in a convection–diffusion–reaction model". In: *Acta Neurologica Belgica* 120.1 (2020), pages 49–57.

[8] J. Petersen, F. Isensee, G. Köhler, et al. "Continuous-time deep glioma growth models". In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. 2021, pages 83–92.

[9] L. Zhang, L. Lu, X. Wang, et al. "Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data". In: *IEEE Transactions on Medical Imaging* 39.4 (2019), pages 1114–1126.

[10] A. Elazab, C. Wang, S. J. S. Gardezi, et al. "GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images". In: *Neural Networks* 132 (2020), pages 321–332.

[11] H. Wang, N. Xiao, J. Zhang, et al. "Static-dynamic coordinated transformer for tumor longitudinal growth prediction". In: *Computers in Biology and Medicine* 148 (2022), page 105922.

[12] B. Liu, Y. Chen, S. Liu, and H.-S. Kim. "Deep learning in latent space for video prediction and compression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pages 701–710.

[13] T. Hu, F. Chen, H. Wang, et al. "Complexity matters: Rethinking the latent space for generative modeling". In: *Advances in Neural Information Processing Systems*. Volume 36. 2024.

[14] R. Rombach, A. Blattmann, D. Lorenz, et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pages 10684–10695.

[15] J. J. Park, P. Florence, J. Straub, et al. “DeepSDF: Learning continuous signed distance functions for shape representation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pages 165–174.

[16] D. Wiesner, J. Suk, S. Dummer, et al. “Generative modeling of living cells with SO(3)-equivariant implicit neural representations”. In: *Medical image analysis* 91 (2024), page 102991.

[17] Y. Xie, T. Takikawa, S. Saito, et al. “Neural fields in visual computing and beyond”. In: *Computer Graphics Forum*. Volume 41. 2. 2022, pages 641–676.

[18] B. Agro, Q. Sykora, S. Casas, and R. Urtasun. “Implicit Occupancy Flow Fields for Perception and Prediction in Self-Driving”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pages 1379–1388.

[19] S. Peng, M. Niemeyer, L. Mescheder, et al. “Convolutional occupancy networks”. In: *European Conference on Computer Vision*. 2020, pages 523–540.

[20] Y. Chen, M. Staring, O. M. Neve, et al. “CoNeS: Conditional neural fields with shift modulation for multi-sequence MRI translation”. In: *Machine Learning for Biomedical Imaging* 2 (Special Issue for Generative Models 2024), pages 657–685.

[21] B. Mildenhall, P. P. Srinivasan, M. Tancik, et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pages 99–106.

[22] X. Shi, Z. Chen, H. Wang, et al. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”. In: *Advances in Neural Information Processing Systems*. Volume 28. 2015.

[23] O. M. Neve, Y. Chen, Q. Tao, et al. “Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study”. In: *Radiology: Artificial Intelligence* 4.4 (2022), e210300.

[24] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pages 203–211.

[25] S. Klein, M. Staring, K. Murphy, et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pages 196–205.

[26] V. Sitzmann, J. Martel, A. Bergman, et al. “Implicit neural representations with periodic activation functions”. In: *Advances in neural information processing systems*. Volume 33. 2020, pages 7462–7473.

6

A deep learning model for data-driven vestibular schwannoma growth prediction

This chapter was adapted from:

Chen, Y., Wolterink, J.M., Neve, O.M., Makarevich, Y., Hensen, E.F., Verbist, B.M., Tao, Q., Staring, M., 2025. A deep learning model for data-driven vestibular schwannoma growth prediction (under review)

Abstract

Purpose

To develop and evaluate a deep learning model for data-driven vestibular schwannoma (VS) growth prediction using longitudinal MRI.

Materials and methods

In this retrospective study, a deep learning model using a recurrent neural network and an implicit neural representations-based decoder was developed to predict tumor growth by generating future tumor masks. Qualitative and quantitative evaluation, including Dice, 95 % Hausdorff distance (95HD), and relative volume error (RVE), was performed. Moreover, tumor diameters were measured and compared to reference values using Bland-Altman plots. Tumor progression was determined by dichotomizing diameter or volume changes, and values were evaluated using the Receiver Operating Characteristic curve and Cohen's kappa.

Results

A total of 316 VS patients with 1488 longitudinal contrast-enhanced T1-weighted MRIs were retrospectively collected from multiple centers. In the volumetric tumor growth prediction, the proposed model outperformed benchmark methods, resulting in a Dice of 0.788 ± 0.090, a 95HD of $1.86 \pm 0.92\,\mathrm{mm}$, and an RVE of $26.7 \pm 27.2\,\%$, respectively. For the samples with a surveillance interval shorter than two years, the Bland-Altman plots showed strong agreement on the diameter measurement, with 95 % limits of agreement of $-0.134 \pm 2.620\,\mathrm{mm}$. Using the diameter measurement, the model determined tumor progression with an accuracy of 92.6 %, a specificity of 96.2 %, and a sensitivity of 62.5 %, showing moderate agreement with the ground truth (Cohen's kappa $= 0.604$).

Conclusion

The deep learning model accurately predicted the future shape of VS within two years using longitudinal MRI, highlighting its potential to forecast tumor progression.

6.1 Introduction

Vestibular schwannomas (VS) are rare, benign tumors arising from vestibulocochlear nerves [1, 2]. Large and progressive VS that are left untreated may compress the brainstem and cerebellum, potentially causing life-threatening complications [1, 3]. Because many VS grow slowly or remain stable, patients with small- to medium-sized tumors are often (initially) managed by active surveillance through repeated MRI, with microsurgery or radiosurgery considered once tumor progression is detected [4, 5, 6]. However, this wait-and-scan approach may lead to intervention on larger tumors, which carries higher surgical or radiation risks [1, 7, 8]. Therefore, prediction of tumor growth is highly desirable, as it can assist in selecting patients who need intervention, help optimize follow-up MRI intervals, and possibly avoid interventions on large tumors.

As symptom progression correlates poorly with tumor growth [9], treatment recommendations are primarily determined by follow-up MRI assessment [1, 6, 10]. In practice, tumor diameters are manually measured, and clinical decisions are made based on either current diameters [6] or the diameter changes over time [11]. However, this assessment is prone to significant error and exhibits high intra- and inter-observer variability [12, 13]. Alternatively, volume measured from tumor segmentation provides more reliable growth assessment but is labor- and time-intensive, limiting its feasibility in routine practice. Moreover, these methods can only detect tumor growth in retrospect, at a point in time where the tumor has already increased, whereas clinical decisions ideally are made before progression has occurred. These limitations highlight the need for a more accurate and reliable yet time-efficient approach for VS growth prediction.

To address this challenge, biophysics-informed models were introduced to model the spatio-temporal evolution of tumor cells [14, 15, 16, 17, 18]. However, these models typically require specific imaging modalities such as dynamic contrast-enhanced MRI (DCE-MRI) and dual-phase CT, which are not available in the clinical routine of VS care. Recently, deep learning models have been developed as data-driven approaches for tumor growth prediction [19, 20, 21]. A previously developed DeepGrowth [22] model predicts VS growth using implicit neural representations combined with a recurrent neural network. While the model has demonstrated promising performance, it remains unclear how well the model helps enhance VS management.

This study aimed to develop and evaluate a deep learning model for data-driven VS growth prediction using longitudinal MRI. We proposed an extension of the DeepGrowth model by incorporating feature-wise linear modulation for time conditioning. The model' s performance and applicability in determining tumor

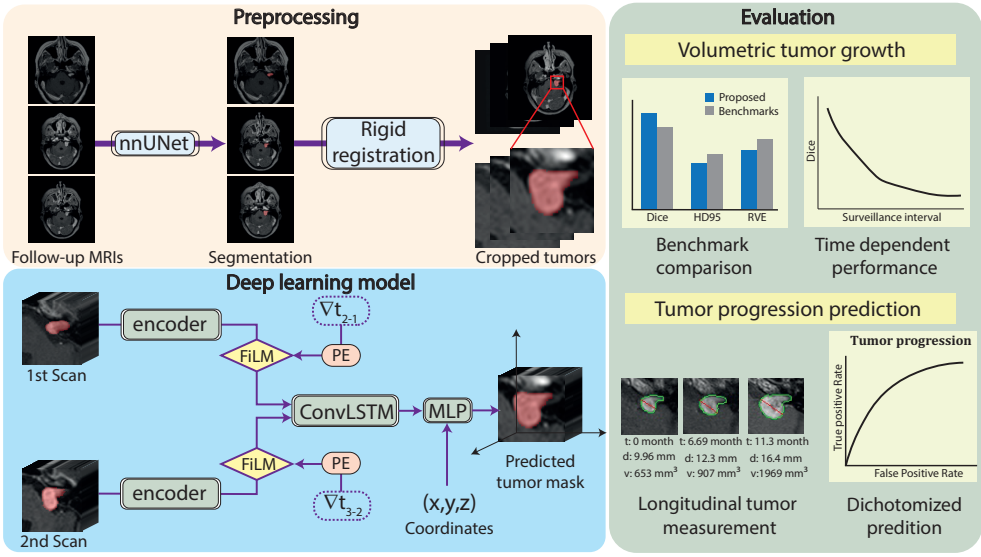


Figure 6.1: The overall pipeline of this study. In the preprocessing step, longitudinal MRI scans and corresponding tumor masks, obtained by an nnU-Net segmentation model, were aligned via rigid registration and cropped for model training and evaluation. The proposed model predicts the future tumor shape from encoded prior scans and time conditioning via positional encoding (PE), feature-wise linear modulation (FiLM), and a Multilayer Perceptron (MLP) decoder. The performance on volumetric tumor growth prediction and determining tumor progression was subsequently evaluated.

progression were evaluated using a retrospective, longitudinal, multi-center dataset.

6.2 Materials and methods

The study protocol (G19.115) was approved by the institutional review board, which granted a waiver of informed consent, as obtaining written informed consent was not reasonably feasible in this retrospective study. The overall pipeline of the study is shown in Figure 6.1.

Data collection and preprocessing

A total of 708 patients with 2380 longitudinal contrast-enhanced T1-weighted (T1ce) MRI were retrospectively collected from over 30 different hospitals between 1994 and 2022, including 134 patients who had been reported previously [10, 11, 22]. Patients with unilateral VS were included, while those with multiple cerebellopontine angle (CPA) tumors or other CPA pathologies were excluded. Cohort and scanner details

Table 6.1: Scanners involved in data collection

Scanner	MF(T)	No. of scan		Scanner	MF(T)	No. of scan	
		M	F			M	F
Philips				Siemens			
Achieva	1.5	134	170	Aera	1.5	54	92
Achieva	3.0	152	145	Avanto	1.5	98	97
Achieva dStream	1.5	15	23	Avanto fit	1.5	22	27
Gyroscan NT	0.5	1	2	Espree	1.5	71	101
Gyroscan NT	1.0	20	18	Harmony	1.0	6	22
Gyroscan NT Intera	1.5	0	1	Magnetom Altea	1.5	0	1
Ingenia	1.5	82	87	Magnetom Expert	1.0	5	11
Ingenia	3.0	31	30	Magnetom Harmony	1.0	1	0
Ingenia Ambition S	1.5	2	1	Magnetom Sola	1.5	6	5
Ingenia Elition S	3.0	3	3	Magnetom Sola fit	1.5	0	1
Ingenia Elition X	3.0	32	31	Magnetom Vida	3.0	10	6
Intera	0.5	1	1	Magnetom Vida fit	3.0	1	0
Intera	1.0	15	6	Magnetom Essenza	1.5	15	18
Intera	1.5	185	149	Skyra	3.0	12	16
Intera	3.0	1	0	Sonata	1.5	2	0
NT Intera	1.0	10	19	SonataVision	1.5	0	2
NT Intera	1.5	1	2	Symphony	1.5	28	41
Panorama	1.0	4	3	Symphony Tim	1.5	0	1
Panorama HFO	1.0	8	17	Verio	3.0	9	17
GE				Hitachi			
Discovery MR450	1.5	5	1	OASIS	1.2	4	0
Discovery MR750	3.0	11	3	Toshiba			
Genesis Signa	1.5	2	5	MRT200PP3	1.5	2	1
Optima MR450w	1.5	1	8	Titan3T	3.0	1	2
Signa Artist	1.5	2	0				
Signa Excite	1.5	9	3				
Signa HDxt	1.5	76	48				
Signa HDxt	3.0	3	0				

Note: MF=Magnetic field, M=male, F=female

are shown in Table 6.1. Scans with imaging issues ($n = 64$) and those obtained after interventions such as microsurgery or radiotherapy ($n = 323$) were excluded. 392 patients with fewer than three follow-up scans were also excluded, as the model requires at least two prior scans and one future scan for validation. The eligible data was then split at the patient level into training (70 %), validation (10 %), and test (20 %) sets. The flowchart is shown in Figure 6.1a. Tumor masks were segmented using a previously validated segmentation model based on nnU-Net [11, 23]. For each patient, longitudinal scans were aligned via rigid registration using Elastix [24] and resampled to the median spacing of the entire dataset. Tumor subregions were then cropped to $64 \times 64 \times 64$ voxels to alleviate the influence of the background.

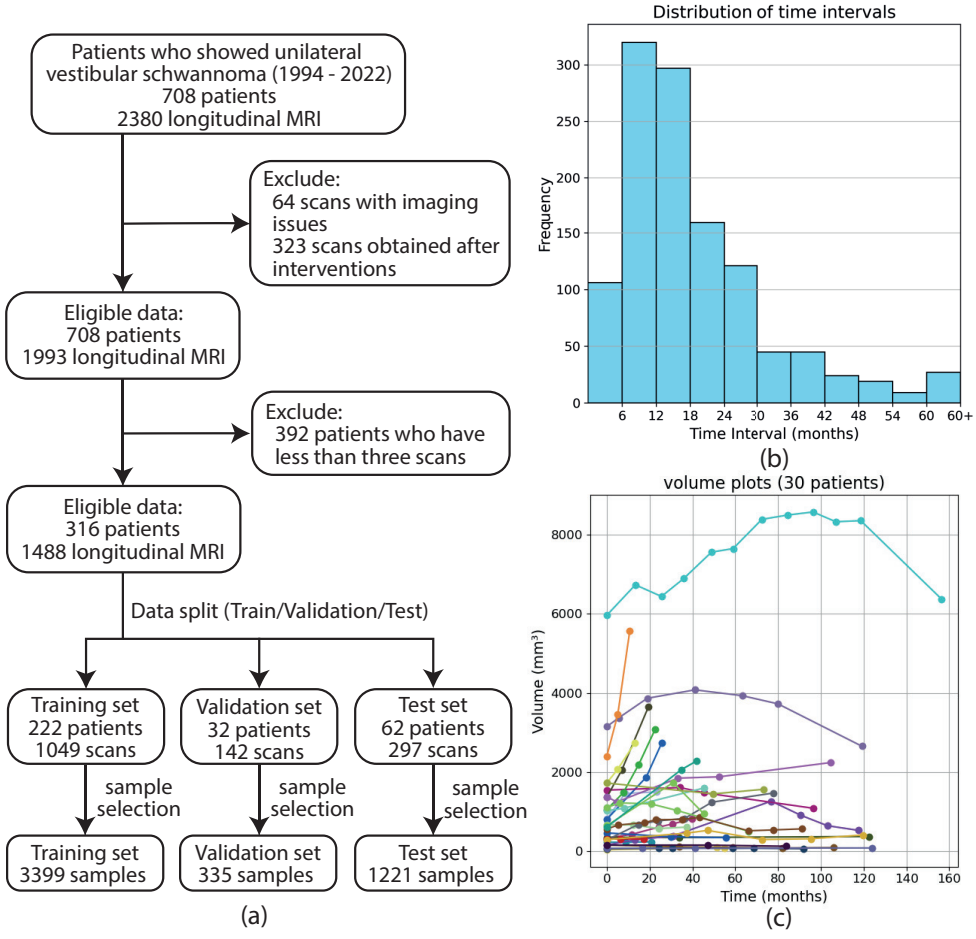


Figure 6.2: Flowchart and the statistics of the dataset. (a) Flowchart of the dataset. (b) The distribution of time intervals between two follow-up MRI studies. (c) Tumor volume trajectories for 30 randomly chosen patients.

Deep learning model for volumetric tumor growth prediction

DeepGrowth [22] consists of a convolutional neural network (CNN), a Convolutional LSTM (ConvLSTM), and a Multilayer Perceptron (MLP). The model concatenates T1ce and corresponding tumor masks of prior scans as the input, and outputs the future tumor mask as a 3D signed distance field (SDF):

$$\mathrm{SDF}(\boldsymbol{x}) = \begin{cases} \min_{u \in C} \|x - u\|_2, & \text{if } x \text{ inside } C \\ 0, & \text{if } x \text{ belonging to } C \\ -\min_{u \in C} \|x - u\|_2, & \text{if } x \text{ outside } C \end{cases} \tag{6.1}$$

where $x = (x, y, z) \in \mathbb{R}^3$ represents 3D coordinates and C denotes the tumor contour, represented by the zero-level set of $\mathrm{SDF}(\boldsymbol{x})$.

To better model the tumor progression over time, we extended the DeepGrowth model by introducing the feature-wise linear modulation (FiLM) mechanism for time conditioning [25]. Specifically, sinusoid function is applied to the time intervals between scans (Δt) for positional encoding (PE). These encodings are then used to condition the model by modulating the feature maps via a learnable transformation. More details can be found in the Appendix.

The model was trained to predict the future tumor mask from two prior MRIs. Each sample comprised three MRIs randomly selected from a patient. Data augmentation, including random translation, flipping, and rotation, was performed. The model achieving the best performance on the validation set was selected for evaluation on the test set. Both training and inference were performed on NVIDIA Quadro RTX6000/RTXA6000 using Python v3.6.8 and PyTorch v1.10.2. The source code is available at `https://github.com/cyjdswx/DeepGrowth`.

Determining tumor progression

The clinical assessment of VS progression typically relies on the changes in tumor diameter or volume measured from follow-up MRI. To resemble clinical procedure, we measured the volume and diameters of the predicted and most recent input tumor masks. Following the method described by Neve et al. [10], on each slice, 2D diameters were measured in planes parallel to the petrous bone, which was automatically estimated by the border between the intra- and extrameatal mask. The largest 2D diameter across all slices was recorded as the tumor diameter. Tumor progression was dichotomized based on changes in volume or diameter between scans.

Model evaluation and statistical analysis

All evaluations were performed on the test set. The model's performance on volumetric tumor growth prediction was quantified using Dice, 95th Hausdorff distance (95HD), and relative volume error (RVE). Comparisons were made with two non-data-driven approaches: (a) Patient-wise linear extrapolation of tumor volume, and (b) Constant tumor, using the most recent prior mask as prediction, and three deep learning models:

Table 6.2: Patient and technical characteristics

Clinical characteristic	
Number of patients	316
Age at diagnosis (years)	57 ± 11
Gender (male/female)	160/156
Number of MRI studies	1488
Intrameatal only	380 (25.5%)
Small (0-10 mm)	316 (21.2%)
Medium (11-20 mm)	617 (41.5%)
Moderately large (21-30 mm)	145 (9.7 %)
Large (31-40 mm)	27 (1.8 %)
Giant (>40 mm)	3 (0.2%)
Technical MRI features	
In-plane resolution (mm)	0.47×0.47 ($0.20 \times 0.20 - 1.09 \times 1.09$)
TE (ms)	7 ($1.82 - 34.0$)
TR (ms)	450 ($6.04 - 2250.0$)
Section thickness (mm)	1.4 ($0.6 - 6.0$)

Note: The age of the patients was presented as mean ± std. The number of MRI studies was presented as a number with percentages in parentheses. The technical MRI features were presented as a median with a range in parentheses. TE = echo time, TR = repetition time.

3D ConvLSTM [26], ST-ConvLSTM [19], and DeepGrowth [22]. Differences between the models were tested by the Wilcoxon Signed Rank Test. A post-hoc analysis of model performance with respect to surveillance interval (future scan time minus last input scan time) was presented as scatter plots. In addition, for patients with four scans, we quantitatively compared the predicted mask at the fourth time point using the first and second scans versus using the second and third scans.

Agreement between predicted and actual diameter measurements was evaluated using the interclass correlation coefficient (ICC) and Bland-Altman plots. To evaluate the model' s performance in determining tumor progression, progressive tumors were defined as those with a growth rate ≥2 mm per year [1]. Receiver operating characteristic (ROC) curves were plotted for diameter change-based and relative volume change-based progression determination. Accuracy, sensitivity, specificity, and Cohen's kappa were calculated for both results, using thresholds of 2 mm per year and 20 % per year, respectively. All analyses were performed in Python v3.6.8, with Numpy v1.21.5, SciPy v1.7.3, and Scikit-learn v1.0.2.

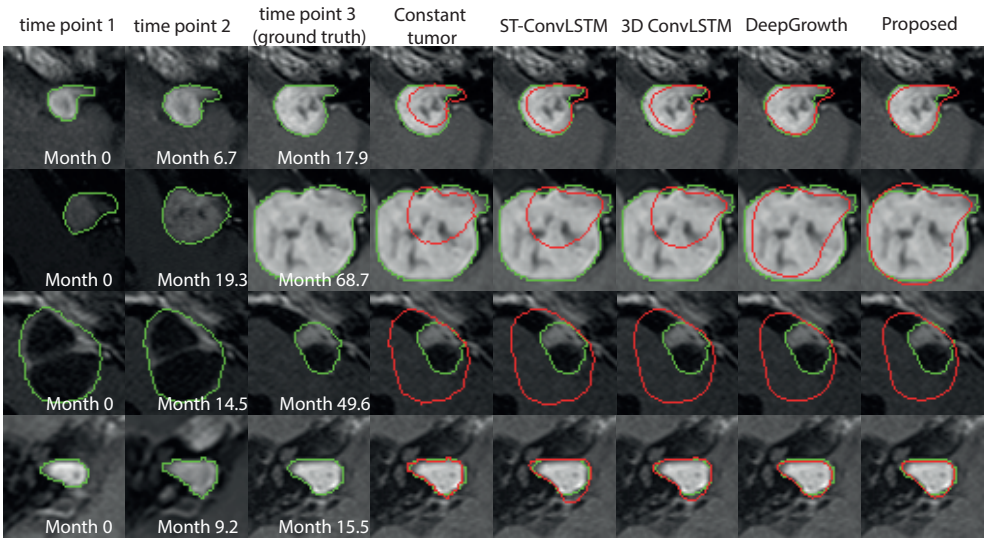


Figure 6.3: Four examples of tumor growth predictions. The first three columns show consecutive scans overlaid with the real tumor masks (green contour), with the acquisition date displayed on each image. Columns four to eight show the real tumor masks (green contour) and the masks predicted by each model (red contour). The first two rows show patients (male, age 73 at first scan; male, age 57 at first scan) with apparent growth in size during surveillance, illustrating the superior performance of the proposed model. The third row shows a patient (female, age 74 at first scan) with a large cystic component. The tumor shrank rapidly, and none of the approaches could predict it. The fourth row shows a patient (female, age 55 at first scan) with a stable tumor, and all approaches performed equally well. The Dice of these patients achieved by the proposed model were 0.858, 0.830, 0.309, and 0.846, respectively, and the 95th Hausdorff distances (95HD) were 1.49 mm, 3.50 mm, 9.42 mm, and 1.18 mm, respectively.

6.3 Results

Longitudinal MRI data

A total of 316 patients (160 male, mean age 57 ± 11; 156 female, mean age 57 ± 11) with 1488 T1ce scans were involved in this study. Table 6.2 shows clinical and technical characteristics of included patients and corresponding scans. According to the tumor classification by Kanzaki et al. [27], the proportion of intrameatal (only), small, medium, moderately large, large, and giant tumors is 25.5% (380/1488), 21.2% (316/1488), 41.5% (617/1488), 9.7% (145/1488), 1.8% (27/1488), and 0.2% (3/1488), respectively. Figure 6.2b shows the distribution of the time intervals between follow-up studies, which range from 0.5 to 124.3 months (mean, 18.9 ± 14.4 months). Across all

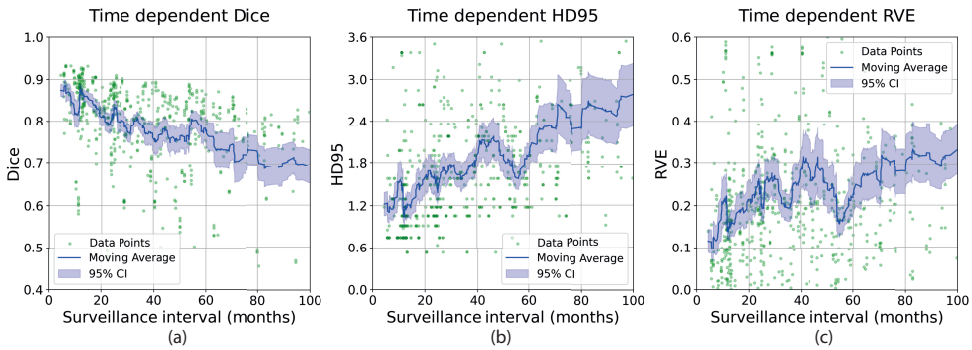


Figure 6.4: Quantitative metrics with respect to surveillance intervals. Green dots represent the data points, and the moving average is plotted as a blue curve with 95 % confidence intervals. The moving average and 95 % confidence intervals were calculated based on a window size of 50 samples. (a), (b) and (c) are the plots of surveillance interval dependent Dice, surveillance intervals dependent 95th Hausdorff distance (95HD), and surveillance interval dependent relative volume error (RVE), respectively.

patients, approximately 58% (184/316) of tumors remained stable (diameter changes <2 mm per year) throughout the entire surveillance. Figure 6.2c shows tumor volume trajectories for randomly selected patients. After data partitioning, the training, validation, and test sets consist of 222 patients with 1049 scans, 32 patients with 142 scans, and 62 patients with 297 scans, respectively. Through exhaustive sample selection, these sets include 3399, 335, and 1221 distinct samples (three MRI studies), respectively.

Volumetric tumor growth prediction

Table 6.3 lists the qualitative results of volumetric tumor growth prediction. Among non-data-driven methods, the Constant tumor approach achieved a Dice of 0.746 ± 0.129, HD95 of 2.25 ± 1.69 mm, and RVE of 32.8 ± 34.2 %, which was better than the patient-wise linear extrapolation (an RVE of 62.2 ± 112.5 %). All deep learning-based models outperformed the non-data-driven approaches, and the proposed model performed best among all benchmarks, yielding 0.788 ± 0.090 in Dice, 1.86 ± 0.92 mm in HD95, and 26.7 ± 27.2 % in RVE, respectively. Note that on average, only DeepGrowth and the proposed model achieved a mean HD95 below 2 mm, which is the threshold commonly used in clinical practice for tumor progression assessment. Representative examples are shown in Figure 6.3. Despite overall strong performance, some outliers were observed, especially among patients with large cystic components (the third row of Figure 6.3), which sometimes exhibit substantial volume change between consecutive

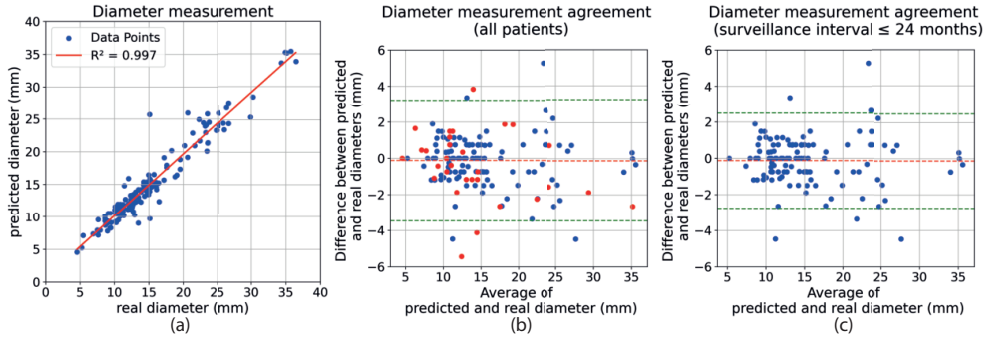


Figure 6.5: Diameter measurements-Deep learning vs. ground truth. (a) Predicted diameter – real diameter plots. The red line shows the results of linear regression. (b) The Bland-Altman Plot of all samples. The blue dots are the data samples with a surveillance interval of less than two years, and the red dots are the remaining samples. (c) Bland-Altman Plot of samples with a surveillance interval of less than two years.

scans.

Figure 6.4 shows how quantitative metrics vary with the surveillance interval. The moving average and 95 % confidence intervals (CIs) were calculated using a window size of 50 samples. As the surveillance interval increased from 4.4 to 99.7 months, the moving average of the Dice decreased from 0.874 [95 % CI: 0.859, 0.889] to 0.693 [95 % CI: 0.653, 0.735], the HD95 increased from 1.22 [95 % CI: 1.06, 1.39] mm to 2.80 [95 % CI: 2.35, 3.22] mm, and the RVE increased from 11.6 % [95 % CI: 8.8 %, 14.4 %] to 33.4 % [95 % CI: 27.4 %, 39.4 %], respectively. The model demonstrated particularly strong performance for the patients with surveillance intervals shorter than two years, for whom the moving averages of Dice mostly fall in $0.80 - 0.89$, HD95 mostly fall in 1.1-1.7 mm, and RVE mostly fall in 10-26 %. Table 6.4 shows the performance comparison using different combinations of input scans. The prediction using early scans (first and second scans, with an average surveillance interval of 33.7 ± 19.9 months) as input yields a mean Dice of 0.750, a mean 95HD of 2.22 mm, and a mean RVD of 32.5 %. These metrics improved to 0.796, 1.80 mm, and 25.3 % respectively, when the later scans (second and third scans, with an average surveillance interval of 17.7 ± 12.5 months) were used for prediction.

Determining tumor progression

To avoid bias toward patients with more scans, only samples comprising three successive scans were involved in the evaluation of tumor progression determination, resulting in 149 samples. As shown in Figure 6.5a, the predicted tumor diameters exhibited strong agreement with the diameters measured from real future scans, with an ICC of

Table 6.3: Quantitative results of volumetric tumor growth prediction

Model	Dice			HD95(mm)			RVE(%)		
	Mean ± Std	median	p	Mean ± Std	median	p	Mean ± Std	median	p
Patient-wise linear extrapolatio	N/A	N/A	N/A	N/A	N/A	N/A	62.2 ± 112.5	35.6	$< .001$
Constant tumor	0.746 ± 0.129	0.771	$< .001$	2.25 ± 1.69	1.83	.009	32.8 ± 34.2	29.2	.010
ST-ConvLSTM [19]	0.753 ± 0.117	0.770	$< .001$	2.31 ± 1.50	1.87	$< .001$	38.1 ± 39.8	28.2	$< .001$
3D ConvLSTM [26]	0.765 ± 0.118	0.794	.006	2.19 ± 1.56	1.82	$< .001$	35.2 ± 35.7	26.4	$< .001$
DeepGrowth [22]	0.783 ± 0.091	0.811	.052	1.95 ± 1.05	1.75	.005	29.5 ± 29.4	23.4	.009
Proposed model	0.788 ± 0.090	0.821	N/A	1.86 ± 0.92	1.67	N/A	26.7 ± 27.2	20.4	N/A

Note: The mean, standard deviation, median, and p-value (Wilcoxon Signed Rank Test conducted between each benchmark method and the proposed model) are reported. Std. = standard deviation, HD95 = 95th Hausdorff distance, RVE = relative volume error

Table 6.4: Quantitative results on different input scans

Input data	Surveillance interval (months)	Dice		HD95(mm)		RVE(%)	
	Mean ± Std.	Mean ± Std.	Median	Mean ± Std.	Median	Mean ± Std.	Median
Early scans	33.7 ± 19.9	0.750 ± 0.104	0.769	2.22 ± 1.25	2.11	32.5 ± 27.7	26.7
Later scans	17.7 ± 12.5	0.796 ± 0.120	0.822	1.80 ± 1.32	1.49	25.3 ± 29.4	15.3

Note: Early scans represent the prediction using the first and second scans as input. Later scans represent the prediction using the second and third scans as input. The mean, standard deviation, and median are reported. Std. = standard deviation, HD95 = 95th Hausdorff distance, RVE = relative volume error

0.98 [95 % CI: 0.97, 0.99] and an $R^2 = 0.997$ based on linear regression. According to the Bland – Altman plot in Figure 6.5b, the predicted diameters were, on average, slightly smaller than the real diameters, resulting in a mean difference of 0.124 mm. The 95 % limits of agreement were -0.124 ± 3.360 mm, which can be further narrowed to -0.134 ± 2.620 mm when we excluded samples with a surveillance interval greater than two years, as shown in Figure 6.5c.

Figure 6.6a and Figure 6.6c present the ROC curves of the performance in determining tumor progression across varying diameter and volume thresholds, respectively. Both ROC curves demonstrate good discriminative performance with an area under the curve (AUC) of 0.82. When using a threshold of diameter increase ≥2 mm per year, the deep learning model detected tumor progression with 62.5% (10/16) sensitivity, 96.2% (128/133) specificity, and 92.6% (138/149) overall accuracy. The agreement on VS progression between deep learning prediction and assessment using real MRI studies resulted in a Cohen's kappa of 0.604. When using a threshold of relative volume increase ≥20 % per year, the sensitivity, specificity, accuracy, and Cohen's kappa were 62.5% (10/16), 91.7% (122/133), 88.6% (132/149), and 0.477, respectively. Both confusion matrices are shown in Figure 6.6b and Figure 6.6d, respectively.

6.4 Discussion

In this study, we extended DeepGrowth by incorporating a time conditioning module based on the FiLM mechanism for VS growth prediction. The proposed model was retrospectively evaluated on a multicenter, longitudinal dataset to assess its performance in predicting volumetric tumor growth and determining tumor progression.

Early attempts at tumor growth prediction focus on biophysics-informed models and image biomarkers. For instance, Roque et al. [15] and Wong et al. [16, 18] initialized reaction-diffusion equations using tumor cellularity estimated from DCE-MRI and dual-phase CT, respectively. Schouten et al. [28] applied logistic regression to imaging biomarkers extracted from DCE-MRI and diffusion-weighted imaging (DWI) for VS growth prediction. However, in the normal clinical VS workflow, advanced MRI sequences like DCE-MRI or DWI are usually not available. To utilize more image modalities, deep learning-based methods were introduced. Zhang et al. [19] proposed a 2D ConvLSTM for pancreas tumor growth prediction, and Elazab et al. [29] proposed a stacked generative adversarial network for glioma growth prediction. Both studies have demonstrated the applicability of deep learning approaches in longitudinal tumor modeling, but only a limited number of patients were included. Recently, Ma et al. [21]

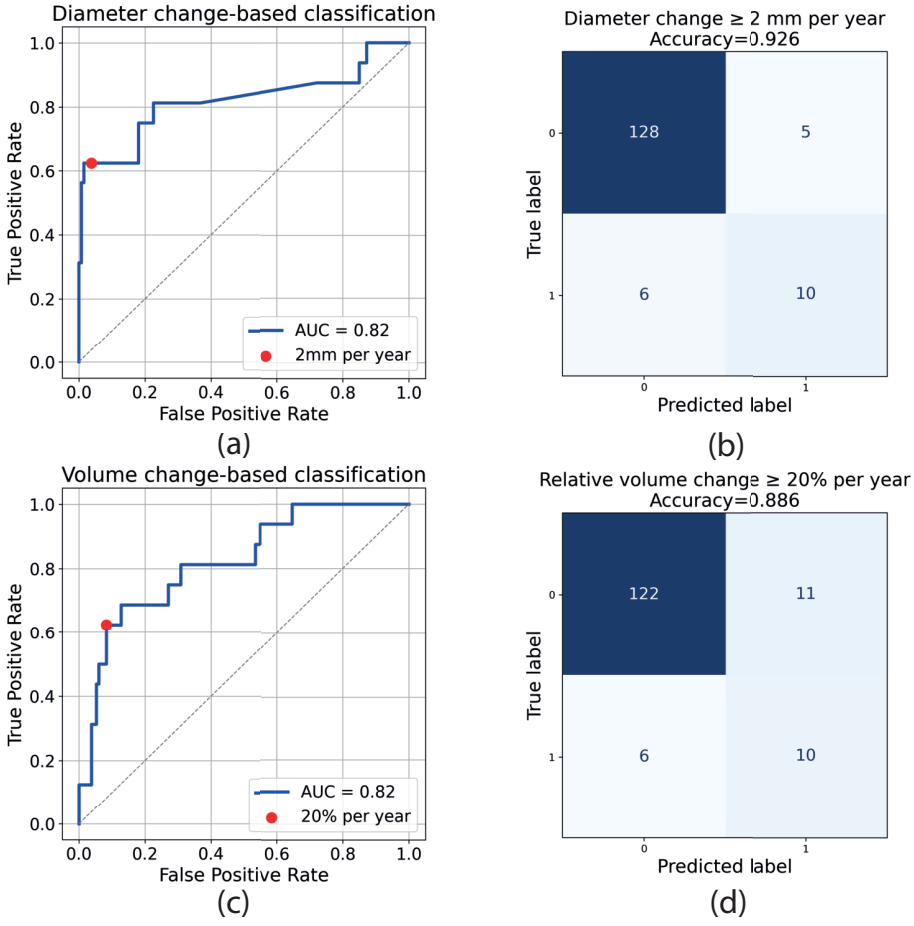


Figure 6.6: Determining tumor progression based on volume or diameter changes. (a) Receiver operating characteristic (ROC) curve showing the classification performance based on tumor diameter change. The red dot represents the results using a threshold of diameter increase ≥2 mm per year to define tumor growth. (b) The confusion matrix of the classification results represented by the red dot in (a). (c) ROC curve showing the classification performance based on tumor volume change. The red dot represents the results using a threshold of relative volume increase ≥20 % per year to define tumor growth. (d) The confusion matrix of the classification results represented by the red dot in (c).

and Wang et al. [20] applied Transformer architectures to a large longitudinal lung CT dataset for lung cancer growth prediction. These models, however, were evaluated solely using volumetric overlap and surface distance metrics. In contrast, our study involved comprehensive clinical validation on a longitudinal dataset, providing a more thorough assessment of model performance.

In our experiments, all deep learning methods have outperformed non-data-driven methods, and the model proposed in the current study performed best. Although some tumors show linear volume growth over time, the linear extrapolation approach performs poorly in RVE on average. The Constant tumor yielded acceptable results due to the slow-growing nature of VS. ST-ConvLSTM [19] performs the worst among all deep learning models, probably because VS can grow in any direction, while a 2D network cannot model growth in the z-axis well. Compared to the DeepGrowth model, the introduction of the FiLM mechanism, which merges image features and learnable features extracted from surveillance intervals, improved performance significantly in all the metrics. This conditioning module also has the potential to be used for incorporating relevant clinical factors, including the patients' age or symptoms.

To show the application scope of the deep learning model for VS growth prediction, we further examined its performance with respect to the surveillance interval. Our analysis revealed that as the surveillance interval increases, the prediction performance in all quantitative metrics declines. Notably, in patients with surveillance intervals shorter than two years, the average HD95 (1.1-1.7 mm) was mostly smaller than the previously reported inter-observer measurement error (1.7 mm) [10]. Not surprisingly, this suggests that our deep learning model offers more reliable predictions over shorter surveillance intervals.

Previous studies have barely evaluated models in terms of tumor measurement, which chiefly drives treatment recommendations in VS management [1]. In our experiment, the proposed model has demonstrated excellent agreement in diameter measurements with the corresponding future scans. With regard to determining tumor progression, the proposed model yields excellent specificity but moderate sensitivity. The agreement between the prediction and the ground truth (a Cohen's kappa of 0.604) is lower than the previously reported inter-observer agreement (a Cohen's kappa of 0.74) using the same data source [10], but comparable to other research (a Cohen's kappa of 0.56) using a different dataset [30]. It is worth noting that these inter-observer agreements were assessed on real clinical scans across the entire follow-up study, while we determined tumor progression solely based on prior MRI studies. Another possible reason for the disagreement is that data-driven models are usually biased towards the majority of samples, which are slow-growing tumors.

Although we retrospectively collected the dataset from multiple centers, the overall dataset size remained limited, which is the common constraint in studies of rare or malignant tumors [31, 32]. This poses a particular challenge for deep learning models, which typically exhibit reduced generalization performance when trained on small datasets. The lack of biomechanical inspiration potentially lowers reliability and makes it difficult to interpret prediction errors. Another limitation comes from the ground

truth of diameter measurements. Although tumor diameters are commonly used in the clinic due to their applicability, human measurement suffers from high inter-observer variability and may also be affected by the restriction of image resolution. This could result in a suboptimal ground truth for the evaluation of progression determination.

In conclusion, we have presented a deep learning model for data-driven VS growth prediction. The proposed model was developed and evaluated using a dataset retrospectively acquired from multiple centers. Our results show that the proposed model achieves strong performance, particularly in predicting tumor growth within two years, highlighting the potential and feasibility of applying deep learning for data-driven VS growth prediction.

6.5 Acknowledgements

This study was supported by the China Scholarship Council (grant 202008130140) and by an unrestricted grant of Stichting Hanarth Fonds, The Netherlands (project MLSCHWAN).

Appendix. Details of the proposed model

The proposed model contains a CNN-based encoder shared by all input scans, a ConvLSTM, feature-wise Linear modulation (FiLM), and an MLP-based decoder.

Time conditioning based on feature-wise linear modulation

To better encode the temporal information, we first perform positional encoding that maps the temporal information into a high-dimensional space [33]. Given the time intervals Δt, the positional encoding is expressed as follows:

$$\gamma(\Delta t) = [\sin 2^0 \pi \Delta t, \cos 2^0 \pi \Delta t, ..., \sin 2^{l-1} \pi \Delta t, \cos 2^{l-1} \pi \Delta t]\,, \tag{6.2}$$

where l is the order that should be optimized based on the specific task. These encodings are then used to condition the model by the FiLM mechanism [25]. More formally, FiLM learns a function of time interval encodings ($\gamma(\Delta t)$) which outputs a linear transformation:

$$y_c = \alpha_c(\gamma(\Delta t))x_c + \beta_c(\gamma(\Delta t))\,, \tag{6.3}$$

where x_c refers to the cth channel of latent features and y_c refers to the cth channel of the modulated features. α_c and β_c are the learnable scale and bias for linear transformation and can be arbitrary neural networks using encoded time interval Δt as input. In practice, FiLM is applied within a residual block, in which the input of FiLM is added to the output. Thus, each layer of the feature map is conditioned on the temporal information independently, which enhances the model in a computationally efficient way.

Network architecture

The encoder of the proposed model is adapted from [34] with two extra convolutional blocks for downsampling, and the ConvLSTM consists of three 32-channel layers. The MLP contains five 64-channel layers and uses the Leaky ReLU function with a negative slope of 0.2 as the activation function after all intermediate layers. Adapted from [35], the last layer of the MLP is followed by a sine function, which can constrain the range of the output to $[-1, 1]$. A dropout layer is added to the positional encoding to avoid overfitting. The FiLM residual block contains a fully connected layer that outputs linear modulation parameters from temporal information and two convolutional layers

before the linear modulation. A ReLU activation function is applied to the FiLM features after linear modulation.

Implementation details

Signed distance field (SDF) points were sampled from the tumor segmentation to train the proposed model. 80 % of the points were sampled near the tumor contour, and the rest were sampled from the entire space. We set the order of positional encoding $l = 6$, which works best for the experiments. The mode was trained by a weighted sum of l_1-based reconstruction loss and l_2 regularization as $L = \lambda_{rec} L_{rec} + \lambda_{reg} L_{reg}$. Little differences were observed between different loss weights, and we set $\lambda_{rec} = 1.0$ and $\lambda_{reg} = 0.1$ for the experiments. The model was initialized by the Xavier method and optimized by the Adam with an initial learning rate of 1×10^{-4}. The training was performed for a total of 450 epochs, and the learning rate was decayed using a step learning rate scheduler with a step size of 50 epochs.

References

[1] M. L. Carlson and M. J. Link. "Vestibular schwannomas". In: *New England Journal of Medicine* 384.14 (2021), pages 1335–1348.

[2] B. J. Arthurs, R. K. Fairbanks, J. J. Demakas, et al. "A review of treatment modalities for vestibular schwannoma". In: *Neurosurgical review* 34.3 (2011), pages 265–279.

[3] A. Harati, K.-M. Scheufler, R. Schultheiss, et al. "Clinical features, microsurgical treatment, and outcome of vestibular schwannoma with brainstem compression". In: *Surgical neurology international* 8 (2017), page 45.

[4] D. H. Coelho, Y. nnU-NetTang, B. Suddarth, and M. Mamdani. "MRI surveillance of vestibular schwannomas without contrast enhancement: clinical and economic evaluation". In: *The Laryngoscope* 128.1 (2018), pages 202–209.

[5] L. Lassaletta, L. A. Cervera, X. Altuna, et al. "Clinical practice guideline on the management of vestibular schwannoma". In: *Acta Otorrinolaringologica (English Edition)* 75.2 (2024), pages 108–128.

[6] J. P. Marinelli, M. J. Link, and M. L. Carlson. "Size Threshold Surveillance-A Revised Approach to Wait-and-Scan for Vestibular Schwannoma". In: *JAMA Otolaryngology–Head & Neck Surgery* 149.8 (2023), pages 657–658.

[7] M. Bailo, N. Boari, A. Franzin, et al. "Gamma Knife radiosurgery as primary treatment for large vestibular schwannomas: clinical results at long-term follow-up in a series of 59 patients". In: *World neurosurgery* 95 (2016), pages 487–501.

[8] G. Grinblat, M. Dandinarasaiah, I. Braverman, et al. "Large and giant vestibular schwannomas: overall outcomes and the factors influencing facial nerve function". In: *Neurosurgical Review* 44.4 (2021), pages 2119–2131.

[9] N. S. Patel, A. E. Huang, E. M. Dowling, et al. "The influence of vestibular schwannoma tumor volume and growth on hearing loss". In: *Otolaryngology–Head and Neck Surgery* 162.4 (2020), pages 530–537.

[10] O. M. Neve, S. R. Romeijn, Y. Chen, et al. "Automated 2-Dimensional Measurement of Vestibular Schwannoma: Validity and Accuracy of an Artificial Intelligence Algorithm". In: *Otolaryngology–Head and Neck Surgery* 169.6 (2023), pages 1582–1589.

[11] O. M. Neve, Y. Chen, Q. Tao, et al. "Fully Automated 3D Vestibular Schwannoma Segmentation with and without Gadolinium-based Contrast Material: A Multicenter, Multivendor Study". In: *Radiology: Artificial Intelligence* 4.4 (2022), e210300.

[12] J. J. Cross, D. M. Baguley, N. Antoun, et al. "Reproducibility of volume measurements of vestibular schwannomas–a preliminary study". In: *Clinical Otolaryngology* 31.2 (2006), pages 123–129.

[13] E. A. Vokurka, A. Herwadkar, N. A. Thacker, et al. "Using Bayesian tissue classification to improve the accuracy of vestibular schwannoma volume and growth measurement". In: *American journal of neuroradiology* 23.3 (2002), pages 459–467.

[14] Y. Liu, S. M. Sadowski, A. B. Weisbrod, et al. “Patient specific tumor growth prediction using multimodal images”. In: *Medical image analysis* 18.3 (2014), pages 555–566.

[15] T. Roque, L. Risser, V. Kersemans, et al. “A DCE-MRI driven 3-D reaction-diffusion model of solid tumor growth”. In: *IEEE transactions on medical imaging* 37.3 (2017), pages 724–732.

[16] K. C. Wong, R. M. Summers, E. Kebebew, and J. Yao. “Tumor growth prediction with reaction-diffusion and hyperelastic biomechanical model by physiological data fusion”. In: *Medical Image Analysis* 25.1 (2015), pages 72–85.

[17] N. Meghdadi, M. Soltani, H. Niroomand-Oscuii, and N. Yamani. “Personalized image-based tumor growth prediction in a convection–diffusion–reaction model”. In: *Acta Neurologica Belgica* 120.1 (2020), pages 49–57.

[18] K. C. Wong, R. M. Summers, E. Kebebew, and J. Yao. “Pancreatic tumor growth prediction with elastic-growth decomposition, image-derived motion, and FDM-FEM coupling”. In: *IEEE transactions on medical imaging* 36.1 (2016), pages 111–123.

[19] L. Zhang, L. Lu, X. Wang, et al. “Spatio-temporal convolutional LSTMs for tumor growth prediction by learning 4D longitudinal patient data”. In: *IEEE Transactions on Medical Imaging* 39.4 (2019), pages 1114–1126.

[20] H. Wang, N. Xiao, J. Zhang, et al. “Static-dynamic coordinated transformer for tumor longitudinal growth prediction”. In: *Computers in Biology and Medicine* 148 (2022), page 105922.

[21] M. Ma, X. Zhang, Y. Li, et al. “ConvLSTM coordinated longitudinal transformer under spatio-temporal features for tumor growth prediction”. In: *Computers in Biology and Medicine* 164 (2023), page 107313.

[22] Y. Chen, J. M. Wolterink, O. M. Neve, et al. “Vestibular schwannoma growth prediction from longitudinal MRI by time-conditioned neural fields”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pages 508–518.

[23] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pages 203–211.

[24] S. Klein, M. Staring, K. Murphy, et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pages 196–205.

[25] E. Perez, F. Strub, H. De Vries, et al. “FiLM: Visual reasoning with a general conditioning layer”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 32. 1. 2018.

[26] X. Shi, Z. Chen, H. Wang, et al. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”. In: *Advances in Neural Information Processing Systems*. Volume 28. 2015.

[27] J. Kanzaki, M. Tos, M. Sanna, and D. A. Moffat. “New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma”. In: *Otology & neurotology* 24.4 (2003), pages 642–649.

[28] S. M. Schouten, D. Lewis, S. Cornelissen, et al. “Dynamic contrast-enhanced and diffusion-weighted MR imaging for predicting tumor growth of sporadic vestibular schwannomas: a prospective study”. In: *Neuro-oncology* 27.4 (2025), pages 1116–1127.

[29] A. Elazab, C. Wang, S. J. S. Gardezi, et al. “GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images”. In: *Neural Networks* 132 (2020), pages 321–332.

[30] A. M. Tolisano, C. C. Wick, and J. B. Hunter. “Comparing linear and volumetric vestibular schwannoma measurements between T1 and T2 magnetic resonance imaging sequences”. In: *Otology & Neurotology* 40.5S (2019), S67–S71.

[31] Z. Haouari, J. Weidner, I. Ezhov, et al. “Efficient deep learning-based forward solvers for brain tumor growth models”. In: *BVM Workshop*. 2025, pages 57–62.

[32] J. Weidner, I. Ezhov, M. Balcerak, et al. “A learnable prior improves inverse tumor growth modeling”. In: *IEEE Transactions on Medical Imaging* (2024).

[33] B. Mildenhall, P. P. Srinivasan, M. Tancik, et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pages 99–106.

[34] S. Peng, M. Niemeyer, L. Mescheder, et al. “Convolutional occupancy networks”. In: *European Conference on Computer Vision*. 2020, pages 523–540.

[35] V. Sitzmann, J. Martel, A. Bergman, et al. “Implicit neural representations with periodic activation functions”. In: *Advances in neural information processing systems*. Volume 33. 2020, pages 7462–7473.

7

Summary, discussion and future work

The diagnosis and management of vestibular schwannoma (VS) rely on longitudinal imaging surveillance and precise tumor measurement. Conventional approaches imply manually measuring every patient, adding a huge workload to the entire healthcare system. This thesis investigates the application of artificial intelligence (AI) in head and neck MRI analysis for data-driven VS care. Based on several deep learning techniques, we developed quantitative approaches for VS segmentation, completion and restoration of missing MRI sequences, and prediction of tumor growth. Comprehensive validations were conducted in both technical and clinical settings.

7.1 Summary

In **Chapter 1**, we provided a general introduction about the role of MRI in head and neck disease management, data-driven vestibular schwannoma care, and deep learning techniques applied in medical image computing.

In **Chapter 2**, we developed an automated VS segmentation tool based on the 3D nnUNet. The tool demonstrated highly accurate tumor volume measurements on both contrast-enhanced T1-weighted and T2-weighted MRIs. The segmentation results were quantitatively and qualitatively evaluated using a multicenter, multivendor setting. The results showed that integrating deep learning models into clinical practice has the potential to improve the accuracy of volume measurements of VS at diagnosis and during follow-up, while reducing the workload of radiologists. In line with the fact that T1ce is the preferred image modality for VS diagnosis and management, CNN performance on T1ce MRI was slightly more accurate compared with T2 MRI, particularly when the tumor border was hard to distinguish from the cerebrospinal fluid solely on T2. We have also designed an observer study that allows a blind, qualitative comparison between the model and human delineation, showing that the CNN tool performs comparably to human observers in clinical practice.

Due to clinical restrictions on the use of contrast agents and diverse imaging protocols across different medical centers, acquiring the most desirable MRI sequences

is often challenging. This may damage the generalization and performance of deep learning models. In **Chapter 3**, we introduced CoNeS, a novel MRI translation model based on conditional neural fields. In the proposed model, the MRI translation problem was formulated as neural fields conditioned on the source images. Benchmark comparison showed that the proposed model achieves superior performance across the entire image scope as well as in the tumor region, which is more clinically relevant. To better understand the gain, we conducted Fourier-domain spectral analysis and a comprehensive ablation study, which together clarify the underlying factors that drive performance improvements.

Contrast-enhanced T1-weighted MRI (T1ce) with a gadolinium-based contrast agent (GBCA) is the gold standard for VS diagnosis and postoperative assessment. However, given the growing concerns about the negative impact of GBCA, there is an increasing interest in reducing the use of GBCA. In **Chapter 4**, we applied the image translation model presented in **Chapter 3** to restore standard-dose T1ce from low-dose MRI scans. We simulated low-dose T1ce with various dose reductions and evaluated the impact of different dose reductions on the deep learning model. The experiments showed that the proposed model can substantially restore the missing contrast and improve image quality, with the most pronounced benefits observed when the input dose is substantially reduced (to 10-30 % of the standard dose). According to the reader study from radiologists, AI-restored images are equally good at tumor detection and diagnosis as the full-dose images. Consistent with visual improvement, the downstream analysis experiments in **Chapter 3** and **Chapter 4** demonstrated that AI-synthetic/restored images can significantly improve segmentation performance.

Vestibular schwannomas are commonly managed through active surveillance with MRI examinations. To avoid overtreatment and sequelae associated with the treatment of large tumors, a precise prediction of tumor growth based on longitudinal imaging is highly desirable. In **Chapter 5**, we proposed DeepGrowth, a model that incorporates neural fields and recurrent neural networks for tumor growth prediction. We evaluated the proposed model on an in-house longitudinal VS dataset, demonstrating superior performance over benchmark models, with particularly significant improvement in predicting fast-growing tumors.

In **Chapter 6**, we extended the model presented in **Chapter 5** by incorporating a novel time-conditioning mechanism based on feature-wise linear modulation. From the predicted tumor masks, we measured the morphological tumor characteristics (diameters and volumes), which can be used to determine tumor progression. The model demonstrates particularly good performance when the surveillance interval is less than two years and shows excellent agreement on the tumor measurements with those measured from real future MRI scans. Despite the moderate sensitivity,

the model demonstrates high specificity with few false positives. The Cohen's kappa between the prediction and ground truth is comparable to the previously reported inter-observer agreement on tumor progression assessment [1].

7.2 Discussion

By helping clinicians interpret population-level data and trends, AI techniques have the potential to improve diagnostic procedures and reduce workload. In this thesis, we have demonstrated the promising performance of AI models across multiple tasks while also revealing certain limitations.

Since the development of deep learning and the successful application of U-Net in the area of medical image segmentation [2], thousands of models, including the most recent Transformer- and Mamba-based models [3, 4], have been proposed. While all new segmentation models claimed state-of-the-art performance, recent research questioned the validation procedure of these studies and argues that none of the advanced architectures truly improve the performance [5]. This implies that a comprehensive, rigorous validation is more important than technical novelty for the safe deployment of AI tools in the clinic. In this thesis, rather than using the latest models that have demonstrated decent results with less rigorous validation, we adopted nnUNet [6], a segmentation framework that has been extensively validated in large-scale clinical tasks and datasets, and focused on the validation that mirrors the clinical setting for VS care. With minimal technical modifications, the proposed model was effectively adapted to the VS segmentation task. In addition to the quantitative evaluation using potentially suboptimal ground truth (due to the inter-annotator variability, as discussed in (**Chapter 2**), the results were assessed by experienced clinicians, who represent the gold standard for clinical diagnosis.

Beyond segmentation, convolutional neural networks (CNNs) have also long been the predominant choice for other medical image computing tasks. In this thesis, we explored the use of implicit neural representations (INRs), where a multilayer perceptron is trained to model complex signals over continuous spatial or spatiotemporal domains, in multiple tasks. Compared to CNN, INR offers several benefits: (1) resolution-agnostic nature [7], (2) memory efficiency [8], and (3) better high frequency preservation (**Chapter 3**). On the other hand, INR solely relies on positional encoding to capture spatial information, showing inferior performance to CNNs in capturing contextual relationships. Inspired by this, we proposed hybrid architectures that combine a CNN encoder with an INR decoder in **Chapter 3** and **Chapter 5**. The results highlighted the potential of INR-based models as practical and effective alternatives for a wide range of medical image computing tasks.

Generative models have gained substantial attention in medical image computing in recent years, owing to their ability to learn data distribution and generate new images [9]. This thesis contributes to this area by developing a generative model based on INRs to synthesize MRI for missing image completion and low-dose MRI enhancement. Unlike models that produce unlimited plausible outputs, our model focuses on paired image translation for deterministic synthesis. This is particularly beneficial when (1) patient data is partially available, and (2) reproducible results are preferred. Our experiments also raised a challenge in the quantitative evaluation of synthetic images. In fact, we observed that common quantitative metrics do not always align with the image quality of synthetic images. For instance, the introduction of adversarial learning enhanced fine details in the generated images but can also degrade similarity scores (**Chapter 3**). This contradiction urges more appropriate metrics to evaluate synthetic medical images.

The ultimate goal of this thesis is to support clinical decision-making and enhance the VS management. In this thesis, we developed a deep learning model that predicts the future shape of the tumor and measured changes in tumor size, which chiefly drive the clinical recommendation. Compared to early attempts, including biophysics-informed models and image biomarkers extracted from advanced imaging, which are not directly applicable in VS management, the proposed model predicts future tumor shape purely based on prior MRIs. As shown in **Chapter 5** and **Chapter 6**, the model showed promising potential in VS growth prediction with only two previous contrast-enhanced T1-weighted MRI scans available. However, similar to the challenges in the image generation problem, quantitatively assessing tumor growth remains difficult. To align with clinical guidelines, tumors exhibiting a diameter increase of more than $2\,\mathrm{mm}$ per year were labeled as progressive. Although widely accepted in practice, progression assessment based on diameter changes is subject to a high inter-observer variability. Moreover, the median image resolution of the dataset is approximately $0.50\,\mathrm{mm} \times 0.50\,\mathrm{mm}$, which is close to the $2\,\mathrm{mm}$ clinical threshold. This resolution limitation introduces uncertainty in the ground truth labels and further impedes reliable evaluation.

Finally, some of the underperformance can be attributed to limitations of the dataset. Although we retrospectively collected VS patients from multiple centers across the Netherlands, the size of the datasets used in this thesis remains limited due to the rarity of VS, and the data distribution is inherently biased. Fast-growing tumors and tumors with cysts are particularly underrepresented, even though the former are of greater clinical concern. In addition, patients who have large tumors typically undergo early interventions (radiotherapy or microsurgery) to prevent further growth, which leads to fewer follow-up scans being available. Last but not least, as all

involved data are retrospective MRI scans before intervention, the proposed model cannot be used for follow-up scans after surgery or radiotherapy without retraining.

7.3 Future perspective

This thesis opens up several promising directions for future investigation. As aforementioned, the scarcity and imbalance of the dataset pose a significant challenge for AI models, whose results are typically biased according to the data distribution. To improve the diversity of the dataset, future studies could consider including more complex cases, including patients with bilateral lesions, VS with large cystic components, and scans acquired after intervention. Additionally, we encourage public data sharing and international collaboration, which is especially valuable for research on rare diseases like VS. The involvement of more complex conditions requires more powerful AI tools. Emerging conditioning approaches, such as prompt learning [10], show strong potential in this direction. Currently, MRI of the cerebellopontine angle CPA is typically performed at magnetic field strengths between 1.0-3.0 T. The introduced 7 T MRI, which provides exceptionally high spatial resolution (up to 0.1 mm [11]), holds great potential for more precise volumetric measurements of vestibular schwannoma. Finally, prospective studies will also be essential to validate generalizability and clinical utility before practical deployment.

We have demonstrated that AI technologies can support precise VS measurement with minimal or no human correction. However, we found that longitudinal tumor growth prediction remains difficult, with significant challenges persisting in both purely data-driven and biomechanical-inspired approaches. Recent studies suggest that combining these two methodologies may offer complementary advantages and improve performance [12]. Before applying these techniques to VS management, it is important to move beyond general partial differential equation models that try to describe any tumor and focus on a VS-specific model due to the diverse tumor pathology. What is equally crucial is the clear formulation of the tumor progression question, including a precise definition of the gold standard. Besides the research topics involved in this thesis, the quality of life (QoL) is another crucial outcome of VS management. While it plays an important role in the treatment recommendation, the QoL of VS patients still lacks quantitative studies due to the significant methodological weaknesses [13]. AI-based techniques have the potential to assist in obtaining optimal QoL by associating it with patient symptoms and treatment strategies.

One main drawback of AI approaches, particularly deep learning, is the lack of interpretability and reliability [14, 15]. These limitations can hinder troubleshooting and impede deployment in high-risk clinical settings. Early attempts addressing this

challenge include model visual explanation [16, 17], the integration of biomarkers [18], and additional supervision with extra clinical information [19]. The newly emerging physics-informed neural networks (PINNs) are also another promising direction for developing more reliable and trustworthy AI tools [20, 21]. This technique holds great potential to solve the reaction-diffusion equation for tumor growth modeling [22]. In addition, while most results in this thesis are presented as deterministic, the problems under consideration, including multi-modality image translation and tumor growth prediction, are inherently uncertain. Compared to traditional deterministic models that tend to produce overconfident predictions, the introduction of uncertainty estimation can improve the reliability of AI models and fill the gap between algorithm research and clinical deployment [23].

Lastly, another key challenge in applying AI techniques to healthcare is how the models are incorporated into clinical workflows. Major vendors like Philips and Siemens have already embedded AI models into commercial scanners for fast MRI reconstruction [24, 25]. For more customized AI applications, it is, however, important to interface the developed tools with the clinical data archive system, such as picture archiving and communication systems (PACS), for a smoother integration. At LUMC, the VS segmentation model developed in **Chapter 2** is currently being implemented following the Medical Device Regulation (MDR) procedure. Despite the promising results the AI tools have demonstrated, we believe human validation or correction is still necessary. This encourages the development of a user-friendly graphical user interface (GUI) to support potential manual annotation.

7.4 Conclusions

In this thesis, we proposed a series of quantitative AI methods for vestibular schwannoma care using head and neck MRI. Our experiments have demonstrated the potential utility of AI models in clinical vestibular schwannoma care, including automatic tumor measurement, missing image completion, low-dose contrast-enhanced image restoration, and data-driven tumor growth prediction. These methods can assist radiologists in interpreting longitudinal MRI, enhance image acquisition, and eventually improve the vestibuar schwannoma care.

While our goal is not to replace clinical expertise, the tools developed in this thesis can meaningfully reduce manual workload and support clinicians in decision-making. The automatic segmentation model, currently being deployed at LUMC, enables precise volumetric assessment of vestibular schwannoma. When certain MRI sequences are not consistently available, the proposed image translation model enhances the robustness and generalizability of the deep learning-based segmentation.

In addition, we demonstrated a potential solution for data-driven VS growth prediction, representing an important step toward reliable automated VS management.

In conclusion, our findings highlight the promise of AI tools in enabling personalized and efficient vestibular schwannoma management. Further research toward developing robust and reliable AI approaches is warranted to facilitate clinical translation.

References

[1] O. M. Neve, S. R. Romeijn, Y. Chen, et al. “Automated 2-Dimensional Measurement of Vestibular Schwannoma: Validity and Accuracy of an Artificial Intelligence Algorithm”. In: *Otolaryngology–Head and Neck Surgery* 169.6 (2023), pages 1582–1589.

[2] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. 2015, pages 234–241.

[3] A. Hatamizadeh, V. Nath, Y. Tang, et al. “Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images”. In: *International MICCAI brainlesion workshop*. 2021, pages 272–284.

[4] Z. Xing, T. Ye, Y. Yang, et al. “Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation”. In: *International conference on medical image computing and computer-assisted intervention*. 2024, pages 578–588.

[5] F. Isensee, T. Wald, C. Ulrich, et al. “nnu-net revisited: A call for rigorous validation in 3d medical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2024, pages 488–498.

[6] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pages 203–211.

[7] Q. Wu, Y. Li, Y. Sun, et al. “An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation”. In: *IEEE Journal of Biomedical and Health Informatics* 27.2 (2022), pages 1004–1015.

[8] Z. Hao, A. Mallya, S. Belongie, and M.-Y. Liu. “Implicit neural representations with levels-of-experts”. In: *Advances in Neural Information Processing Systems* 35 (2022), pages 2564–2576.

[9] W. H. Pinaya, M. S. Graham, E. Kerfoot, et al. “Generative ai for medical imaging: extending the monai framework”. In: *arXiv preprint arXiv:2307.15208* (2023).

[10] H. Yang, X. Liang, Z. Li, et al. “Prompt Mechanisms in Medical Imaging: A Comprehensive Survey”. In: *arXiv preprint arXiv:2507.01055* (2025).

[11] B. L. Edlow, A. Mareyam, A. Horn, et al. “7 Tesla MRI of the ex vivo human brain at 100 micron resolution”. In: *Scientific data* 6.1 (2019), page 244.

[12] J. Weidner, I. Ezhov, M. Balcerak, et al. “A learnable prior improves inverse tumor growth modeling”. In: *IEEE Transactions on Medical Imaging* (2024).

[13] A. Gauden, P. Weir, G. Hawthorne, and A. Kaye. “Systematic review of quality of life in the management of vestibular schwannoma”. In: *Journal of Clinical Neuroscience* 18.12 (2011), pages 1573–1584.

[14] X. Li, H. Xiong, X. Li, et al. “Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond”. In: *Knowledge and Information Systems* 64.12 (2022), pages 3197–3234.

[15] D. Mahapatra, A. Poellinger, and M. Reyes. “Interpretability-guided inductive bias for deep learning based medical image”. In: *Medical image analysis* 81 (2022), page 102551.

[16] R. R. Selvaraju, M. Cogswell, A. Das, et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pages 618–626.

[17] J. R. Clough, I. Oksuz, E. Puyol-Antón, et al. “Global and local interpretability for cardiac MRI classification”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019, pages 656–664.

[18] P. Gamble, R. Jaroensri, H. Wang, et al. “Determining breast cancer biomarker status and associated morphological features using deep learning”. In: *Communications medicine* 1.1 (2021), page 14.

[19] S. Shen, S. X. Han, D. R. Aberle, et al. “An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification”. In: *Expert systems with applications* 128 (2019), pages 84–95.

[20] M. Sarabian, H. Babaee, and K. Laksari. “Physics-informed neural networks for brain hemodynamic predictions using medical imaging”. In: *IEEE transactions on medical imaging* 41.9 (2022), pages 2285–2303.

[21] R. L. van Herten, A. Chiribiri, M. Breeuwer, et al. “Physics-informed neural networks for myocardial perfusion MRI quantification”. In: *Medical Image Analysis* 78 (2022), page 102399.

[22] R. A. Gatenby and E. T. Gawlinski. “A reaction-diffusion model of cancer invasion”. In: *Cancer research* 56.24 (1996), pages 5745–5753.

[23] A. Mehrtash, W. M. Wells, C. M. Tempany, et al. “Confidence calibration and predictive uncertainty estimation for deep medical image segmentation”. In: *IEEE transactions on medical imaging* 39.12 (2020), pages 3868–3878.

[24] *MR SmarSpeed fast imaging technology*. `https://www.philips.co.uk/healthcare/technology/smartspeed-ai`.

[25] *Deep resolve in MRI*. `https://www.siemens-healthineers.com/nl-be/infrastructure-it/artificial-intelligence/ai-campaign/deep-resolve-in-mri`.

List of publications

Journal articles

Neve, O.M.[*], **Chen, Y.[*]**, Tao Q., Romeijn, S.R., de Boer, N.P., Grootjans, W., Kruit, M.C., Lelieveldt, B.P., Jansen, J.C., Hensen, E.F., Verbist, B.M., Staring, M., 2022. Fully automated 3D vestibular schwannoma segmentation with and without gadolinium-based contrast material: a multicenter, multivendor study, Radiology: Artificial Intelligence 4, no. 4 (2022): e210300.

Neve, O.M., Romeijn, S.R., **Chen, Y.**, Nagtegaal, L., Grootjans, W., Jansen, J.C., Staring, M., Verbist, B.M., and Erik F. Hensen E.F., Automated 2-dimensional measurement of vestibular Schwannoma: validity and accuracy of an artificial intelligence algorithm, Otolaryngology – Head and Neck Surgery 169, no. 6 (2023): 1582-1589.

Chen, Y., Staring, M., Neve, O.M., Romeijn, S.R., Hensen, E.F., Verbist, B.M., Wolterink, J.M., Tao, Q., 2024. CoNeS: Conditional neural fields with shift modulation for multi-sequence MRI translation. The journal Machine Learning for Biomedical Imaging 2, 657 – 685

Chen, Y.[*], Weber, R.[*], Neve, O.M., Romeijn, S.R., Hensen, E.F., Wolterink, J.M., Tao, Q., Staring, M., Verbist, B.M., 2025. A deep learning model to reduce agent dose for contrast-enhanced MRI of the cerebellopontine angle cistern (submitted)

Chen, Y., Wolterink, J.M., Neve, O.M., Makarevich, Y., Hensen, E.F., Verbist, B.M., Tao, Q., Staring, M., 2025. A deep learning model for data-driven vestibular schwannoma growth prediction (submitted)

International conference proceedings

Chen, Y., Staring, M., Wolterink, J.M. and Tao, Q., 2023. Local implicit neural representations for multi-sequence MRI translation. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5)

Chen, Y., Wolterink, J.M., Neve, O.M., Romeijn, S.R., Verbist, B.M., Hensen, E.F., Tao, Q. and Staring, M., 2024. Vestibular schwannoma growth prediction from

longitudinal MRI by time-conditioned neural fields. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 508-518)