Title
A Neural Network Approach to Deformable Image Registration

Author
McKenzie, Elizabeth

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Neural Network Approach to Deformable Image Registration

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy in Physics and Biology in Medicine

by

Elizabeth MaryAnn McKenzie

2021

ABSTRACT OF THE DISSERTATION

A Neural Network Approach to Deformable Image Registration

by

Elizabeth MaryAnn McKenzie

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2021

Professor Ke Sheng, Chair

Deformable image registration (DIR) is an important component of a patient's radiation therapy treatment. During the planning stage it combines complementary information from different imaging modalities and time points. During treatment, it aligns the patient to a reproducible position for accurate dose delivery. As the treatment progresses, it can inform clinicians of important changes in anatomy which trigger plan adjustment. And finally, after the treatment is complete, registering images at subsequent time points can help to monitor the patient's health. The body's natural non-rigid motion makes DIR a complex challenge. Recently neural networks have shown impressive improvements in image processing and have been leveraged for DIR tasks. This thesis is a compilation of neural network-based approaches addressing lingering issues in medical DIR, namely 1) multi-modality registration, 2) registration with different scan extents, and 3) modeling large motion in registration. For the first task we employed a cycle consistent

generative adversarial network to translate images in the MRI domain to the CT domain, such that the moving and target images were in a common domain. DIR could then proceed as a synthetically bridged mono-modality registration. The second task used advances in network-based inpainting to artificially extend images beyond their scan extent. The third task leveraged axial self-attention networks' ability to learn long range interactions to predict the deformation in the presence of large motion. For all these studies we used images from the head and neck, which exhibit all of these challenges, although these results can be generalized to other parts of the anatomy.

The results of our experiments yielded networks that showed significant improvements in multi-modal DIR relative to traditional methods. We also produced network which can successfully predict missing tissue and demonstrated a DIR workflow that is independent of scan length. Finally, we trained a network whose accuracy is a balance between large and small motion prediction, and which opens the door to non-convolution-based DIR.

By leveraging the power of artificial intelligence, we demonstrate a new paradigm in deformable image registration. Neural networks learn new patterns and connections in imaging data which go beyond the hand-crafted features of traditional image processing. This thesis shows how each step of registration, from the image pre-processing to the registration itself, can benefit from this exciting and cutting-edge approach.

The dissertation of Elizabeth MaryAnn McKenzie is approved.

Dan Ruan

Minsong Cao

Robert Kaida Chin

Daniel Abraham Low

Fabien Scalzo

Ke Sheng, Committee Chair

University of California, Los Angeles

2021

DEDICATION

To my parents Laura and Devoy McKenzie and my cat Merlin

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Getting a PhD is never easy, and I could have never made it through these years of graduate school without the support of family, friends, and mentors. For this I am eternally grateful. My family was always just a phone call away whenever life seemed particularly difficult. I could never have gotten to where I am today without their loving support of my education and a desire to instill in me a curiosity about the world that would later lead to a career in science. I love you guys.

Along with the amazing opportunities for research, I am grateful that my time in graduate school introduced me to so many wonderful people. These friendships kept me grounded in ways nothing else could. Thank you.

Friendships were made both outside and inside my lab. I am so happy that I was surrounded by such kind and brilliant lab mates. You guys are amazing, and I look forward to working with you guys in the field.

While not human, I am also grateful to my cat Merlin. When the stress of the pandemic seemed overwhelming and I was quarantined at home, he would snuggle up next to me and purr or ask me to play with him. His love helped to relieve anxiety and I can't imagine going through 2020 without him.

The most important decision of one's PhD is the selection of one's advisor. Ke Sheng embodies the title of advisor. He has given me direction when I felt lost, guided my efforts towards rewarding discoveries, offered endless support, and gave me inspiration for what a scientist in medical physics can be. Ke supported not only my professional growth, but also my personal one. Ke's excitement for life and research is contagious, and I am extremely grateful for having the opportunity to work with him these last 5 years. Thank you, Ke.

Chapter 2 is a version of a published journal article:

McKenzie EM, Santhanam A, Ruan D, O'Connor D, Cao M, Sheng K. Multimodality image registration in the head-and-neck using a deep learning-derived synthetic CT as a bridge. *Med Phys*. 2020;47(3):1094-1104. doi:10.1002/mp.13976

Chapter 3 is a version of a published journal article:

McKenzie EM, Tong N, Ruan D, Cao M, Chin RK, Sheng K. Using neural networks to extend cropped medical images for deformable registration among images with differing scan extents. *Med Phys*. 2021;48(8):4459-4471. doi:10.1002/mp.15039

VITA

EDUCATION

M.S. University of Texas MD Anderson Cancer Center Graduate School of Biological Sciences, Medical Physics

B.S. Purdue University, Physics, minor in Mathematics and French

AWARDS

Robert J Shalek Award (2011-2012)
AAPM Scholarship for Future Leaders (2016)
AAPM Expanding Horizons Travel Grant (2018)

PEER REVIEWED PUBLICATIONS

1. **McKenzie, Elizabeth M**., Nuo Tong, Dan Ruan, Minsong Cao, Robert K. Chin, and Ke Sheng. "*Using neural networks to extend cropped medical images for deformable registration among images with differing scan extents*." Medical Physics 48, no. 8 (2021): 4459-4471

2. **McKenzie, Elizabeth M**., Anand Santhanam, Dan Ruan, Daniel O'Connor, Minsong Cao, and Ke Sheng. "*Multimodality image registration in the head-and-neck using a deep learning-derived synthetic CT as a bridge*." Medical physics 47, no. 3 (2020): 1094-1104.

3. Yue, Yong, Zhaoyang Fan, Wensha Yang, Jianing Pang, Zixin Deng, **Elizabeth McKenzie**, Richard Tuli, Robert Wallace, Debiao Li, and Benedick Fraass. "*Geometric validation of self-gating k-space-sorted 4D-MRI vs 4D-CT using a respiratory motion phantom*." Medical physics 42, no. 10 (2015): 5787-5797.

4. Yang, Wensha, **Elizabeth M. McKenzie**, Michele Burnison, Stephen Shiao, Amin Mirhadi, Behrooz Hakimian, Robert Reznik, Richard Tuli, Howard Sandler, and Benedick A. Fraass. "*Clinical experience using a video-guided spirometry system for deep inhalation breath-hold radiotherapy of left-sided breast cancer*." Journal of Applied Clinical Medical Physics 16, no. 2 (2015): 251-260

5. **McKenzie, Elizabeth M.**, Peter A. Balter, Francesco C. Stingo, Jimmy Jones, David S. Followill, and Stephen F. Kry. "*Toward optimizing patient-specific IMRT QA techniques in the accurate detection of dosimetrically acceptable and unacceptable patient plans*." Medical physics 41, no. 12 (2014): 121702.

6. **McKenzie, Elizabeth M.**, Peter A. Balter, Francesco C. Stingo, Jimmy Jones, David S. Followill, and Stephen F. Kry. "*Reproducibility in patient-specific IMRT QA*." Journal of applied clinical medical physics/American College of Medical Physics 15, no. 3 (2014): 4741.

7. Huang, Jessie Y., Kiley B. Pulliam, **Elizabeth M. McKenzie**, David S. Followill, and Stephen F. Kry. "*Effects of spatial resolution and noise on gamma analysis for IMRT QA*." Journal of applied clinical medical physics/American College of Medical Physics 15, no. 4 (2014): 4690.

SELECTED CONFERENCE PRESENTATIONS

1. **McKenzie, Elizabeth M.,** Nuo Tong, Dan Ruan, Minsong Cao, Robert K. Chin, and Ke Sheng. "*Using Neural Networks to Extend Cropped Medical Images for Deformable Registration Among Images with Differing Scan Extents."* Medical physics 48, no. 6 (2021): MO-IePD-TRACK 4-7.

2. **McKenzie, Elizabeth M.,** Anand Santhanam, Dan Ruan, Daniel O'Connor, Minsong Cao, and Ke Sheng. "*Multimodality image registration in the head-and-neck using a deep learning-derived synthetic CT as a bridge*." Medical physics 47, no. 3 (2020): 1094-1104.

3. **McKenzie, Elizabeth M.**, Anand Santhanam, Daniel O'Connor, Minsong Cao, Dan Ruan, and Ke Sheng. " *From Multimodality to Monomodality: Head and Neck MR-CT Registration with Synthetic Image Bridge*." International Journal of Radiation Oncology• Biology• Physics, (2019)

4. **McKenzie, Elizabeth M.**, Dan Ruan, Percy Lee, Ke Sheng. "*A Rigidity Penalty to Improve MR-CT Registration*." ISMRM Meeting (2018)

5.
6. **McKenzie, Elizabeth M**., Peter Balter, Jimmy Jones, David Followill, Francesco Stingo, Kiley Pulliam, and Stephen Kry. "*Evaluation of the sensitivities of patient specific IMRT QA dosimeters*." Medical Physics 40, no. 6 (2013): 240-240.

SELECTED INVITED TALKS

McKenzie, Elizabeth M., Nuo Tong, Anand Santhanam, Minsong Cao, Dan Ruan, and Ke Sheng. "*Hands Off Deep Learning Derived Synthetic CT and Contour Driven Multimodal Registration in the Head and Neck*." Medical Physics, 46, no. 6, pp. E502-E502. Wiley, 2019.
Presented at the Annual AAPM Meeting, San Antonio TX

Boehnke, M. Elizabeth. "*Characterization of a novel detector using ROC*" (2017). ROC: A New Method in Radiotherapy QA
Presented at the Annual AAPM Meeting, Denver CO

McKenzie, Elizabeth M., John Demarco, Jennifer Steers, and Benedick Fraass. "*Assessing the Sensitivity and False Positive Rate of the Integrated Quality Monitor (IQM) Large Area Ion Chamber to MLC Positioning Errors*" (2016)
Presented at the Annual AAPM Meeting, Washington D.C.

CHAPTER 1: Introduction

Image registration allows us to establish a correspondence between imaging datasets, and is a critical component of many medical applications. This correspondence allows for tracking dynamic processes across imaging studies and combining the strengths of different imaging modalities. However, since the human body experiences predominantly non-rigid motion, the most accurate correspondence can only be established through *deformable image registration (DIR)*. Accurately modeling the body's non-rigid motion is not a trivial task. The challenge of finding correspondence is compounded when multiple modalities are involved, as these can have drastically different representations of the same tissues and different scan extents. Extensive research on this topic has been carried out within the classical image processing literature, however the recent renaissance of deep learning has completely changed the landscape of image registration research. One important advantage of deep learning approaches is their ability to recognize complex patterns. Previous research has sought to explore these strengths through the development of registration frameworks and similarity metrics; however, their abilities in the presence of large motion has yet to be proven. Since patients (especially oncology patients) commonly experience large motion, different postures, weight changes, and even surgical resection across the course of their medical care, it is important to discover and develop DIR methods that are able to robustly account for these extreme changes.

Research in image registration has blossomed in recent years with the advent of neural networks. Lingering challenges such as large non-rigid motion, changes in patient weight, surgical alteration, and multi-modality imaging are beginning to see rapid progress as investigators start to harness neural networks' abilities to model complex and emergent phenomenon[1].

Overcoming these lingering challenges is fundamental to the progression of informed medical

decision-making since images taken with different geometries, modalities, or time points require registration in order to make them useful for evaluating and monitoring the patient's progress[2]. This is of particular importance in the treatment of cancer. More specifically, radiation therapy treatment relies heavily on imaging to develop effective and safe treatments. Accurately mapping the anatomy is essential for correctly targeting the tumor with radiation. Recently, adaptive radiotherapy (ART) has emerged as a powerful tool to shape the treatment to the patient's daily anatomical changes, allowing the tumor to be treated with a laser-like focus. Recent studies have shown that ART has the power to provide significant dosimetric benefits to both the tumor and the surrounding normal tissues[3–6]. However, to institute an ART program safely and effectively, the daily imaged anatomy must be registered accurately to the initial planning image[7]. Given the current inadequacies of clinical image registration in terms of *large motion*, *mass change*, *different scan extent,* and *multi-modal imaging*, the registration requirement is a significant hurdle to the implementation of potentially life-saving and novel ART. Thus, finding a solution to the current shortcomings of DIR is highly motivated.

Recent research has shown that it's possible to use networks to translate MRI images into CT images to a high degree of detail in a process called domain transfer[8–11]. Translating one imaging modality into another helps to bridge the difficulties traditionally found in multi-modal DIR[12–17]. This augments the applicability of traditional iterative, intensity-based registration. **In this thesis we improve multi-modal deformable image registration in the head and neck through a style transfer bridge**. For this we apply a generative adversarial network (GAN) with cycle consistency[18] to a training dataset of CT and MR images. The result is an MR-

derived synthetic CT which can be registered to a CT image using well-understood traditional metrics and B-spline registration.

Registering images taken at different times and with different modalities is additionally challenging due to often differing scan lengths. Imaging can be taken with different scan extents capture differing amounts of anatomy. It is challenging for a registration algorithm to rectify anatomy that lies within one image and not the other[19]. This often results in unrealistic deformation of the anatomy at the scan borders. Given that most deformable image registration algorithms are regularized for smooth deformation, the erroneous extreme contraction or expansion near the border can propagate into non-border portions of the image[20]. Networks have been used to artificially extend natural images by predicting what lies beyond their borders[21]. **In this thesis, we develop a network-based image extension technique to rectify images of differing scan extents prior to registration, and demonstrate its capability to improve registration outcomes**.

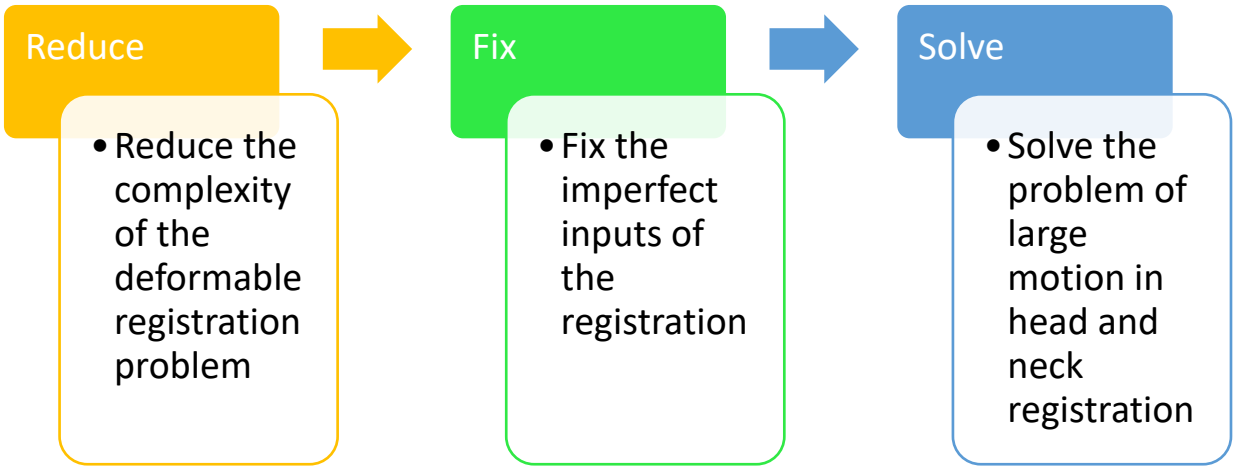
Neural networks can also be trained to directly model a DIR algorithm[1]. These networks replace the conventional iterative process of registration with a one-step transformation estimation. Initial work trained these registration networks using ground truth deformation vector fields (DVF)[22–25]. Such a supervised approach showed great promise but is limited by the availability of true ground truth DVFs, and is trained to replicate, not surpass, current registration algorithms' accuracy. Some have overcome the limitation of procuring ground truth DVFs by simulating known deformations[26–30]. While this has the potential to provide nearly limitless training data, one is limited by the realism of the deformation models. These difficulties

motivated research in unsupervised training, where no ground truth DVF is needed for the loss function. Foundational work in this realm used a neural network such as a U-net architecture to output an estimated DVF[31,32]. These previous advances have relied on convolutional neural networks, which, while powerful, are an inherently local approach. This makes it more challenging to model long range dependencies in anatomical motion. Self-attention networks recently emerged as a powerful way to model global relationships in an image[33]. **This thesis develops a novel DIR network constructed using a self-attention backbone.** This results in a network which can use information from the entire image in its prediction of a DVF for head and neck deformable registration.

Due to the prevalence of large motion, mass change, and differing scan extents in longitudinal medical imaging of cancer patients and the complementary information that multi-modality imaging brings, we believe that focusing on solving the challenges of multi-modal, large motion DIR is of essential importance to the treatment and evaluation of cancer patients. This research focuses on using neural networks to improve the accuracy and efficiency of deformable registration in the head and neck. We select head and neck patients to prove our developed techniques work in a highly challenging arena, with 1) large motion, 2) a large number of small, poorly visualized organs, 3) significant changes in weight/surgical resection, and 4) differing scan extents. The clinical significance of our researched techniques can extend to other anatomical sites and modality pairings, as our results pave the way for research to a more generalized DIR approach.

By developing a methodology using neural networks, we take advantage of their ability to represent complex, learned functions to automatically register images in a single forward pass,

while incorporating empirical constraints. We therefore offer an approach which can be used to not only combine the complimentary information of multiple imaging modalities (e.g. CT, PET, and MR) for oncological treatment planning and response assessment, but also for guidance and treatment adaptation during a patient's treatment course. This thesis focuses on a sequential approach, replacing steps in the traditional image registration pipeline with optimized neural network steps, thus carefully improving DIR part by part. We visualize this thesis' outline in the flowchart below as a Reduce→Fix→Solve strategy. Using this approach, we lay the groundwork for a novel DIR technique, whose hands-off nature could transform the way we approach image registration in our care of cancer patients, in such applications as treatment planning and adaptive radiation treatment, which intrinsically depends on fast, accurate DIR.



References

1. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. *Mach Vis Appl*. 2020;31(1):1-18. doi:10.1007/s00138-020-01060-x

2. Kessler ML. Image registration and data fusion in radiation therapy. *Br J Radiol*. 2006;79(SPEC. ISS.):99-108. doi:10.1259/bjr/70617164

3. Capelle L, Mackenzie M, Field C, Parliament M, Ghosh S, Scrimger R. Adaptive Radiotherapy Using Helical Tomotherapy for Head and Neck Cancer in Definitive and Postoperative Settings: Initial Results. *Clin Oncol*. 2012;24(3):208-215. doi:10.1016/j.clon.2011.11.005

4. Zeidan OA, Langen KM, Meeks SL, et al. Evaluation of image-guidance protocols in the treatment of head and neck cancers. *Int J Radiat Oncol Biol Phys*. 2007;67(3):670-677. doi:10.1016/j.ijrobp.2006.09.040

5. Zeidan OA, Huddleston AJ, Lee C, et al. A Comparison of Soft-Tissue Implanted Markers and Bony Anatomy Alignments for Image-Guided Treatments of Head-and-Neck Cancers. *Int J Radiat Oncol Biol Phys*. 2010;76(3):767-774. doi:10.1016/j.ijrobp.2009.02.060

6. Foroudi F, Wong J, Kron T, et al. Online adaptive radiotherapy for muscle-invasive bladder cancer: Results of a pilot study. *Int J Radiat Oncol Biol Phys*. 2011;81(3):765-771. doi:10.1016/j.ijrobp.2010.06.061

7. König L, Derksen A, Papenberg N, Haas B. Deformable image registration for adaptive radiotherapy with guaranteed local rigidity constraints. *Radiat Oncol*. 2016;11(1):1-9. doi:10.1186/s13014-016-0697-4

8. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT synthesis using unpaired data. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2017;10557 LNCS:14-23. doi:10.1007/978-3-319-68127-6_2

9. Hiasa Y, Otake Y, Takao M, et al. Cross-Modality Image Synthesis from Unpaired Data Using CycleGAN. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. Vol 1. Springer; 2018:31-41. doi:10.1007/978-3-030-00536-8_4

10. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. 2017. doi:10.1002/mp.12155

11. Armanious K, Jiang C, Fischer M, et al. MedGAN: Medical image translation using GANs. *Comput Med Imaging Graph*. 2020;79(8):1-17. doi:10.1016/j.compmedimag.2019.101684

12. Cao X, Yang J, Gao Y, Guo Y, Wu G, Shen D. Dual-core Steered Non-rigid Registration for Multi-modal Images via Bi-directional Image Synthesis. *Med Image Anal*. 2017;0:1-14. doi:10.1016/j.media.2017.05.004

13. Cao X, Yang J, Gao Y, Wang Q, Shen D. Region-Adaptive Deformable Registration of CT/MRI Pelvic Images via Learning-Based Image Synthesis. *IEEE Trans Image Process*. 2018;27(7):3500-3512. doi:10.1109/TIP.2018.2820424

14. Chen S, Quan H, Qin A, Yee S, Yan D. MR image-based synthetic CT for IMRT prostate treatment planning and CBCT image-guided localization. *J Appl Clin Med Phys*. 2016;17(3):236-245.

15. Chen M, Carass A, Jog A, Lee J, Roy S, Prince JL. Cross contrast multi-channel image registration using image synthesis for MR brain images. *Med Image Anal*. 2017;36(3):2-14. doi:10.1016/j.media.2016.10.005

16. Roy S, Carass A, Jog A, Prince JL, Lee J. MR to CT registration of brains using image synthesis. *Med Imaging 2014 Image Process*. 2014;9034:903419. doi:10.1117/12.2043954

17. McKenzie EM, Santhanam A, Ruan D, O'Connor D, Cao M, Sheng K. Multimodality image registration in the head-and-neck using a deep learning-derived synthetic CT as a bridge. *Med Phys*. 2020;47(3):1094-1104. doi:10.1002/mp.13976

18. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proc IEEE Int Conf Comput Vis*. 2017;2017-Octob:2242-2251. doi:10.1109/ICCV.2017.244

19. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report. *Med Phys*. 2017;44(7):e43-e76. doi:10.1002/mp.12256

20. McKenzie EM, Tong N, Ruan D, Cao M, Chin RK, Sheng K. Using neural networks to extend cropped medical images for deformable registration among images with differing scan extents. *Med Phys*. 2021;48(8):4459-4471. doi:10.1002/mp.15039

21. Teterwak P, Sarna A, Krishnan D, et al. Boundless: Generative Adversarial Networks for Image Extension. 2019. http://arxiv.org/abs/1908.07007.

22. Yang X, Kwitt R, Niethammer M. Fast predictive image registration. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2016;10008 LNCS:48-57. doi:10.1007/978-3-319-46976-8_6

23. Rohé MM, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: learning deformable image registration using shape matching. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2017;10433 LNCS:266-274.

doi:10.1007/978-3-319-66182-7_31

24. Jun L V., Yang M, Zhang J, Wang X. Respiratory motion correction for free-breathing 3D abdominal MRI using CNN-based image registration: a feasibility study. *Br J Radiol*. 2018;91(1083):1-9. doi:10.1259/bjr.20170788

25. Cao X, Yang J, Zhang J, et al. Deformable Image Registration Based on Similarity-Steered CNN Regression. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, eds. *Miccai 2017*. Vol 10433. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2017:300-308. doi:10.1007/978-3-319-66182-7_35

26. Uzunova H, Wilms M, Handels H, Ehrhardt J. Training CNNs for image registration from few samples with model-based data augmentation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2017;10433 LNCS:223-231. doi:10.1007/978-3-319-66182-7_26

27. Ito M, Ino F. An automated method for generating training sets for deep learning based image registration. *BIOIMAGING 2018 - 5th Int Conf Bioimaging, Proceedings; Part 11th Int Jt Conf Biomed Eng Syst Technol BIOSTEC 2018*. 2018;2(Biostec):140-147. doi:10.5220/0006634501400147

28. Eppenhof KAJ, Pluim JPW. Pulmonary CT Registration Through Supervised Learning With Convolutional Neural Networks. *IEEE Trans Med Imaging*. 2019;38(5):1097-1105. doi:10.1109/TMI.2018.2878316

29. Sokooti H, de Vos B, Berendsen F, Lelieveldt BPF, Išgum I, Staring M. Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks. In: *Asian Journal of Pharmaceutical and Clinical Research*. Vol 4. ; 2017:232-239. doi:10.1007/978-3-319-66182-7_27

30. Sun Y, Moelker A, Niessen WJ, van Walsum T. Towards Robust CT-Ultrasound Registration Using Deep Learning Methods. In: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Vol 11038. Springer International Publishing; 2018:43-51. doi:10.1007/978-3-030-02628-8_5

31. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca A V. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Trans Med Imaging*. 2019;38(8):1788-1800. doi:10.1109/TMI.2019.2897538

32. Ghosal S, Ray N. Deep deformable registration: Enhancing accuracy by fully convolutional neural net. *Pattern Recognit Lett*. 2017;94:81-86. doi:10.1016/j.patrec.2017.05.022

33. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In: *Lecture Notes in Computer Science*

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol 12349 LNCS. ; 2020:108-126. doi:10.1007/978-3-030-58548-8_7

CHAPTER 2: Multimodality Image Registration in the Head-and-Neck using a Deep Learning Derived Synthetic CT as a Bridge

Introduction

Image registration is often used in medicine for diagnostic and therapeutic purposes[1]. The registration can take place between a single image modality, or different modalities (multimodal registration), which aggregates complementary data from different sources into a spatially unified context[2]. A common multimodal registration problem is magnetic resonance (MR) and computed tomography (CT) registration[2]. MR imaging has superior soft tissue contrast while computed tomography (CT) has better bone contrast and spatial integrity[3]. Specifically, CT is the foundation of modern radiotherapy by providing anatomical information as well as the electron density for treatment planning and dose calculation[4]. In image guided radiation therapy, CT or cone beam CT is instrumental to guide patient set-up. Because of their complementary strengths, MR-CT registration is often needed for accurate tumor and organ-at-risk (OAR) delineation, targeting and sparing[5–9].

Relevant to the current study, head-and-neck radiotherapy benefits from the superior soft tissue contrast provided by the MR images. Studies have demonstrated that MR images in addition to CT improve delineation of head-and-neck target volumes, and reduce interobserver variation[10–14]. Consensus guidelines recommend that MRI be used for primary tumors of the nasopharynx, oral cavity, and oropharynx to contour head-and-neck normal tissues[15]. However, these guidelines also acknowledge the challenges associated with MR-CT registration. While MR-CT

registration is a common practice in head-and-neck radiotherapy, the process and results are not satisfactory due to the different imaging mechanisms and contrast, as well as the unavoidable patient non-rigid motion between scans, such as neck flexion[2]. Deformable registration using commercial algorithms can produce difficult-to-validate distortion and is regarded as unreliable in clinical practice. Instead, rigid registration is performed as a trade-off to avoid the uncertain deformation[16]. Subsequently, delineation based on rigid MR-CT registration is limited to a small volume of interest, without using other information about OARs and lymph nodes from the MR images due to the increasing misalignment with the distance from the volume of interest.

Efforts have been made to address some of the technical issues in multi-modality image registration[17,18]. For example, instead of directly matching the image voxel values, mutual information is used to determine the image similarity based on joint entropy[17]. However, mutual information based on an image histogram cannot resolve tissue types with similar image intensities, such as the bones and air cavities in MR and various soft tissues in CT[2]. The problem is further complicated by the common presence of MR shading and susceptibility artifacts[19]. Some have endeavored to overcome this difference by translating one image into the other, or a third domain. For example, Heinrich et al. used a new image descriptor describing similarities between adjacent patches as features for registration[2]. Researchers have previously used an atlas-based synthetic CT to replace the MR in MR-CT registration in the brain[20,21]; however, the brain experiences considerably less deformation than the head and neck, and thus a domain-translating deformable registration has yet to be proven in this challenging location. Recently, Cao et al. proposed using a patchwise random forest to translate MR and CT into the others' domains for improved pelvic registration[22]. We propose to build on these existing domain-

translating registration techniques by incorporating recent advances in deep learning imaging synthesis.

Specifically, Generative Adversarial Networks (GAN)[23] are capable of converting images of one modality into another[24–26]. For example Wolterink et al. used the CycleGAN implementation[27] to convert brain MR images into synthetic CT images[26]. In the current study, GANs' capability to create synthetic images with the geometry of one image modality and the contrast of the other is used to improve multimodality registration in the head-and-neck. A GAN trained to generate head and neck images learns to estimate realistic anatomy in its image synthesis, and can leverage image features to determine low-confidence regions such as bone air interfaces, which are otherwise invisible in standard T1 weighted MR images. The head and neck is a particularly challenging site in this regard, as there are many bone and air regions in close proximity that can move several centimeters with neck flexion. Deep learning patient-specific image synthesis takes the field beyond atlas-based approaches which try to fit a patient to a standard anatomical layout. By extending modality-translating registration techniques with patient specific deep learning image synthesis, we provide a valuable new technique and prove its performance in the challenging head and neck region.

Methods and Materials

Data

In order to train a network capable of generating synthetic CT's and subsequently test registration accuracy, 25 head-and-neck patients were selected, each with a paired MR and CT

volume acquired on the same day with the same immobilization mask and headrest. The original dataset before processing had 126 to 336 512×512 axial slices with voxel sizes of 1×1×3(or 1.5) mm^3, and 288 334×300 axial slices with 1.5×1.5×1.5mm^3 voxel sizes for CT and MR images, respectively. The MR images were acquired on an MR guided radiotherapy system with a 0.35T B0 using a balanced steady state free-precession sequence. Due to the same rigorous immobilization being used for CT and MR acquisition, deformation between the two image sets was small, providing a unique opportunity for comparison and validation. In this study, we will refer to this MR-aligned CT as $CT_{aligned}$. A five-fold cross validation technique was employed for machine learning purposes. 20 patients were used for training the synthetic CT generating network, and 5 were left out to test replacing an MR image with a synthetic CT during MR-CT registration. This split was rotated through the data, allowing us to include all patients in our analysis.

In addition to the paired images, each test patient also had a diagnostic CT from another time point ranging from 4 days prior to almost 3 years after. One patient did not have a usable separate CT volume, leaving us with 24 patients total for registration testing. The goal was to have patient positions substantially different from the paired images to challenge the registration. For the remainder of this paper, we will refer to these images as $CT_{non\text{-}aligned}$.

Data Processing

All $CT_{aligned}$ datasets were first automatically rigidly registered in Elastix[28,29] to their corresponding MR image using mutual information as the similarity metric. To better show the

posture during network training, volumes were resliced into 2D sagittal slices. All images were resampled to slices of size 256×256, and each voxel was 1.76×1.76×1.5 mm^3. The images were quantized to 256 greyscale values. For CT, this was accomplished by renormalizing and quantizing the image into 256 levels between intensity values -600 and 1400, and between 0 and 600 for MR. In all, this gave 8,350 CT and 8,350 MR sagittal slices to be used for training and testing.

Deep Learning Networks

In a conventional GAN, two networks are used; a generator network attempts to generate realistic images, while a discriminator network attempts to distinguish between real images and those created by the generator. When successful training is complete, the generator is able to create an image that appears to come from the domain of the training set. This study used an adversarial network utilizing cycle-consistency (CycleGAN)[27] with two GAN's: one attempted to generate a realistic synthetic CT (CT_{synth}) slice given a real MR slice, and the other attempted to generate a realistic synthetic MR slice given a real CT slice. The generators were then switched and applied to the synthetic outputs, so that the synthetic MR was translated back into a CT slice, and vice versa. Ideally, the original CT or MR slice should be recovered, and hence this network architecture has cycle consistency. The loss function for CycleGAN therefore has an adversarial loss term for generating realistic CT images, an adversarial loss term for generating realistic MR images, and a cycle consistency loss term to prevent the network from assigning any random realistic-looking image from the other domain.

Overall, the full loss function can be written as:

$$L(G_{CT\to MR}, G_{MR\to CT}, D_{CT}, D_{MR}) = L_{GAN}(G_{CT\to MR}, D_{MR}, CT, MR) +$$
$$L_{GAN}(G_{MR\to CT}, D_{CT}, MR, CT) + \lambda L_{cyc}(G_{CT\to MR}, G_{MR\to CT'}), \quad (1)$$

where λ (set to 10 in this work) is a relative weighting coefficient, and G and D are the generator and discriminator networks with subscripts describing the direction of image translation and discrimination domain, respectively. The adversarial loss is given by:

$$L_{GAN}(G_{CT\to MR}, D_{MR}, I_{CT}, I_{MR}) = E_{MR\sim p_{data(MR)}}[logD_{MR}(I_{MR})] + E_{CT\sim p_{data(CT)}}[log(1 -$$
$$D_{MR}(G_{CT\to MR}(I_{CT}))] \quad (2)$$

where the discriminator gives an output between 0 (image determined to be fake) and 1 (image determined to be real). In the minmax optimization problem, the generator attempts to create a realistic image by minimizing the second term towards a large negative value while the discriminator is trained to maximize the objective by correctly differentiating real images from fake. A similar loss is used for $L_{GAN}(G_{MR\to CT}, D_{CT}, MR, CT)$. The cycle consistency loss using L1 norm is then given as:

$$L_{cyc}(G_{CT\to MR}, G_{MR\to CT'}) = E_{CT\sim p_{data(CT)}}\left[\left\|G_{MR\to CT'}\left(G_{CT\to MR'}(I_{CT})\right) - I_{CT}\right\|_1\right] +$$
$$E_{MR\sim p_{data}(MR)}[\left\|G_{CT\to MR'}\left(G_{MR\to CT'}(I_{MR})\right) - I_{MR}\right\|_1] \quad (3).$$

The generator networks follow the Resnet architecture described in Johnson et al[30]. The discriminator uses a patch-based network described in[31]. Because it is patch-based, this allows greater flexibility for different sized images, as well as forces the discriminator to focus on smaller-scale details. The network hyperparameters used in our study are the same as those in the pytorch-CycleGAN-and-pix2pix repository[1].

Registration

We first used the trained CycleGAN to generate a synthetic CT given a head and neck MR image. We then registered $CT_{non\text{-}aligned}$ to the synthetic MR-derived CT_{synth}, which reduced the multimodal registration problem to a mono-modal one. Conversely, we registered the MR to $CT_{non\text{-}aligned}$ by first registering the CT_{synth} to $CT_{non\text{-}aligned}$, then applying the resulting deformation vector field (DVF) to the original MR. For comparison, direct registrations between MR and CT were also performed. $CT_{non\text{-}aligned}$ was additionally registered to $CT_{aligned}$ to characterize the behavior of a typical mono-modality (CT vs CT) registration. This paper will denote deformable registration with an arrow ($\rightarrow$) pointing from source to target. Figure 1 gives an overview of the registrations performed in this study.

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

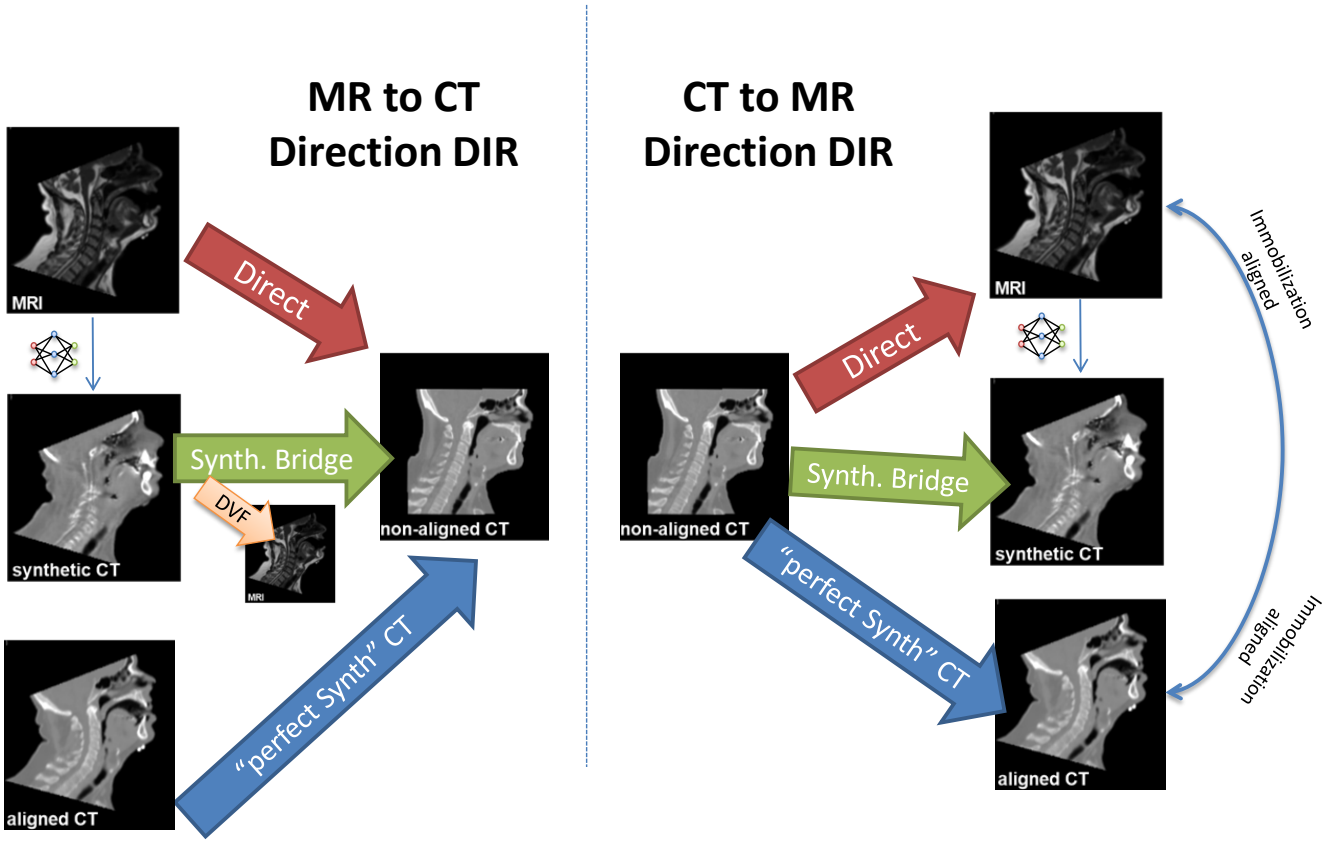


Figure 1 The deformable registrations performed in this study. Registrations were performed in both CT-to-MR and MR-to-CT directions to test for inverse consistency, as well as both directly (multi-modal) and with a synthetic CT bridge (synthetic mono-modal). The DVF from CT_{synth}→ $CT_{non\text{-}aligned}$ was applied to the MR to generate a deformed MR image. The non-aligned CT was also registered to the aligned CT (and vice versa) to see an approximate best-case synthetic mono-modal registration.

The multi-resolution registration using B-splines and mutual information was performed using Elastix[28,29] with six Gaussian blurring levels, repeated. These levels allow for a hierarchical approach to the registration, starting at a coarse resolution with large-scale deformations, and gradually progressing towards a finer resolution for fine-detailed deformations. The B-spline grid spacings for each resolution level were 128, 64, 32, 8, and 4 mm, sequentially. The

Gaussian sigmas were 8, 4, 2, 1, 0.5, and 0.5 voxels isotropically. The registration was optimized using gradient descent[32]. The gradient descent gain factor, a_k, was set to: $a_k = \frac{a}{(50+k+1)^{0.6}}$, where k is the iteration number, and *a* is set for each resolution level to be: 50000, 10000, 2000, 500, 100, 100. Large values of *a* in the coarse resolutions allow the registration to capture large deformations, which were necessary when registering to $CT_{non\text{-}aligned}$. The maximum number of iterations at each resolution level was: 500, 500, 500, 500, 100, and 100.

Analysis

To evaluate the registration, the following tests were performed. The spinal cord was manually contoured on the original MR, $CT_{aligned}$, and $CT_{non\text{-}aligned}$ image volumes for all patients. The cord is an appealing anatomical landmark structure, as it is present throughout the head-and-neck region, reflective of the neck flex, and conspicuous in both modalities. The resulting cord contours from the deformable registration were compared to their respective target volumes' contours using 95% Hausdorff Distance[33], measured in mm.

The Euclidean distance between a set of 11 landmarks (Dens of C2, center of the vertebral bodies of C2-C7, center of left and right eyes, the mental protuberance of the mandible, and the tip of the nose) was evaluated between deformed and target images. We performed a 2-way repeated measures anova with a post hoc Tukey's multiple comparison to test the null hypothesis that all registrations had the same mean error, and identify the significantly different registrations if the null hypothesis was rejected. The two factors in the anova analysis were registration direction and landmark.

Additionally, the quality of the registration itself was evaluated by calculating the Jacobian determinant from each resulting transformation. This was calculated using the Insight Toolkit's implementation, and is given as the scalar determinant of the derivative of the deformation vector field at each point ($\det\left(\frac{dT}{dx}\right), where\ T\ is\ the\ transform$). A reasonable registration in the head and neck anatomical region should have most voxels experiencing small shrinking and expansion with the average Jacobian determinant close to 1.

The quality of the registration is also reflected in its inverse consistency, which was evaluated by comparing the composition of transformation pairs in the opposite direction on a standard CT image to reduce input from the background, then calculating the mean square error (MSE) between that image and the initial image. The MSE was calculated using the Insight Toolkit's implementation, and is the sum of squared differences between intensity values between the images. A lower MSE indicates better inverse consistency. We elected to use this method since we did not expect true inverse consistency in the DVF due to the occasional appearance and disappearance of tissue with different patient positioning (e.g. arms up versus arms down). This is a known challenge in head and neck image registration. However, we can compare directional bias between direct and our proposed registration techniques. Therefore we emphasize that while a 0 MSE would indicate a perfect recovery of the original image, our inverse consistency study was a relative comparison.

Results

Synthetic CT

Figure 2I is a typical synthetic CT achieved in this work. The CT_{synth} image preserves bulk anatomy, distinguishes bones and sinuses, but is missing certain anatomical details of a real CT (e.g. accurate description of individual vertebrae).

Registration Accuracy: Qualitative Evaluation

Figure 2 A-E illustrate an example of a $CT_{non\text{-}aligned}$ (green) registered to an MRI (red) using a synthetic CT. The 3D surface rendering in Figure 2C shows a large initial discrepancy in the head pose. Figure 2D shows how the $CT_{non\text{-}aligned}$ pose was deformed (purple) to match the MRI when the CT is directly registered to the MRI (red). Fig 2E shows how closely the registered $CT_{non\text{-}aligned}$'s pose (blue) matches the target MRI (red) when using a synthetic CT bridge. The head tilt matches better when using a synthetic CT, as can be seen by the improved match in the nose.

Figures 2 J-N show the interior anatomy of registration results in sagittal slices. Looking at the gridlines, Figure 2K shows that the $CT_{non\text{-}aligned}$ matches the MR's pose when CT_{synth} is used as the target. $CT_{non\text{-}aligned}$ registered to MR matched the pose but produced slight unrealistic tissue deformation, as in the stretched sinuses indicated by the red arrow in Figure 2J. Comparing Figure 2N and Figure 2M, the CT_{synth} registered to $CT_{non\text{-}aligned}$ shows even better improvement over direct registration. This is evident in the MR to $CT_{non\text{-}aligned}$ registration's relatively greater stretching in the skull and brain anatomy indicated by the red arrow in Figure 2M, and also the better positioning of the orbit. While the registrations including CT_{synth} matched overall pose,

there is a residual discrepancy due to different mouth opening with or without a bite block. Also, the registration accuracy is similar for CT_{synth} registered to $CT_{non\text{-}aligned}$ and $CT_{non\text{-}aligned}$ registered to CT_{synth}, while there is a noticeable decline in quality for MR registered to $CT_{non\text{-}aligned}$ relative to $CT_{non\text{-}aligned}$ registered to MR, showing improved inverse consistency using a synthetic CT bridge.

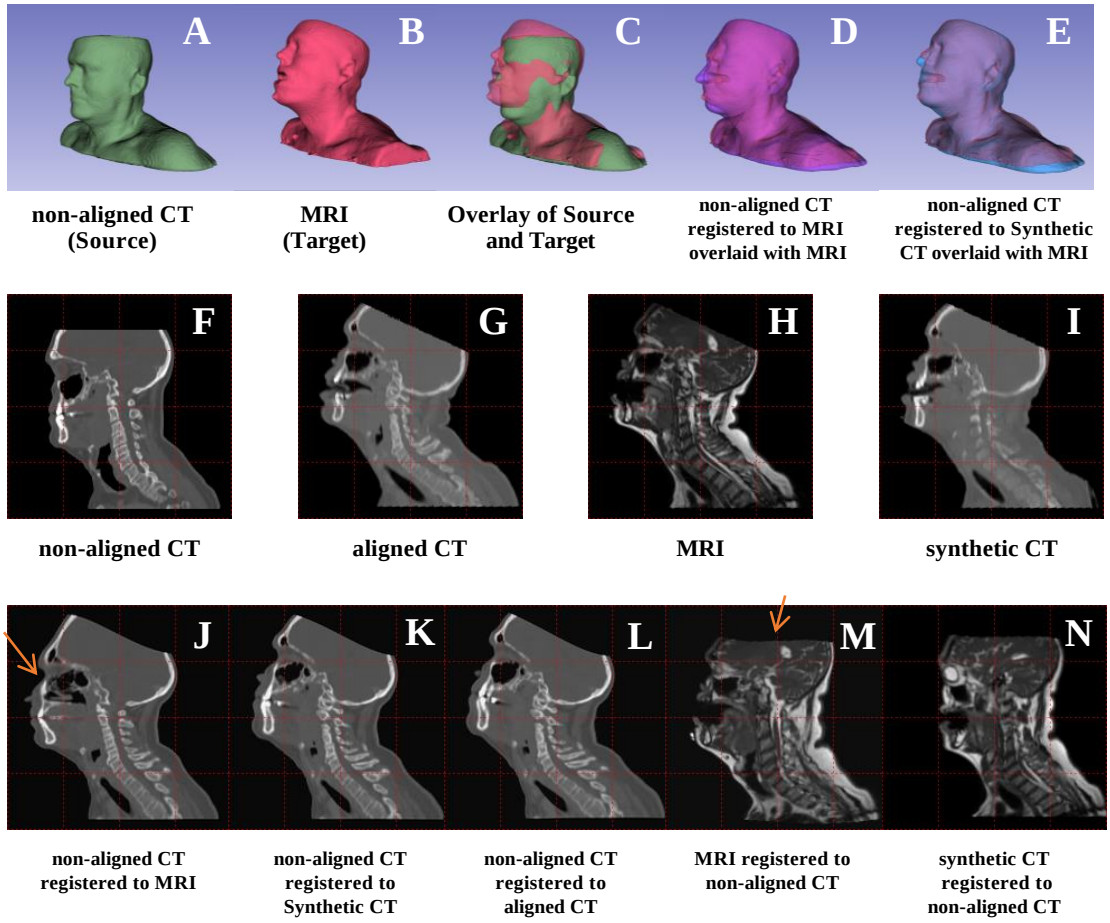


Figure 2 An example patient shows the various registrations studied in this paper. First row: registration of non-aligned CT (green) to an MR image (red). Box D shows the directly registered non-aligned CT (purple); Box E shows registration of the corersponding synthetic CT to the nonaligned CT. Second row: sagittal view. Boxes F-I are the non-deformed

volumes. The third row shows the results for various registrations. Note that the images used in the registration were downsampled to match the 256x256 resolution output from the neural network. All slices shown were in the same location. Arrows denote unanatomical deformation in direct multimodel registration.

Registration Accuracy: Spinal Cord Contour Comparisons

The spinal cord contour comparison results are shown in Figure 3. In Figure 3A-B, the initial 95% Hausdorff distance of the MR and $CT_{non\text{-}aligned}$ rigid alignment is on the horizontal axis, and the vertical axis shows their 95% Hausdorff distance[33] after registration, to reflect the quality of registration and its dependency on the original level of rigid misalignment. The fitted lines are a result of Deming linear regression, where a lower slope means the registration is more robust to initial misalignment. The dotted reference line indicates equal rigid and deformable cord error. For small initial misalignment in the bottom left, the deformable results using CT_{synth} and MR are similar to the rigid results. With larger initial misalignments, there is an increasing divergence in the results with and without CT_{synth}. In the presence of large initial misalignment, compared with the direct registration, our proposed method results in a larger improvement. Using a Deming linear regression and comparing estimated slopes, we found that the error using our method is lower than that of the direct registration in both the MR to CT direction ($p=0.0002$) and the CT to MR direction ($p=0.08$). Unsurprisingly, replacing the MRI with the aligned CT results in the lowest contour misalignment but the difference with our proposed method is not statistically significant ($p=0.34$ and $p=0.65$, respectively).

The slopes of the fitted lines and their 95% confidence intervals are plotted in Figure 3C. The direct registrations show a clear increase in sensitivity to initial misalignment. The wide confidence intervals on the non-aligned CT registered to the synthetic CT reveal the larger

spread when using our method in the CT to MR direction. In fact, all of the CT to MR direction registrations show increased sensitivity relative to their opposite-direction pairs. The registrations between MR and the aligned CT represent the residual sensitivity to initial misalignment inherent to our registration algorithm, as the MR and aligned CT should already have overlapping cord contours.

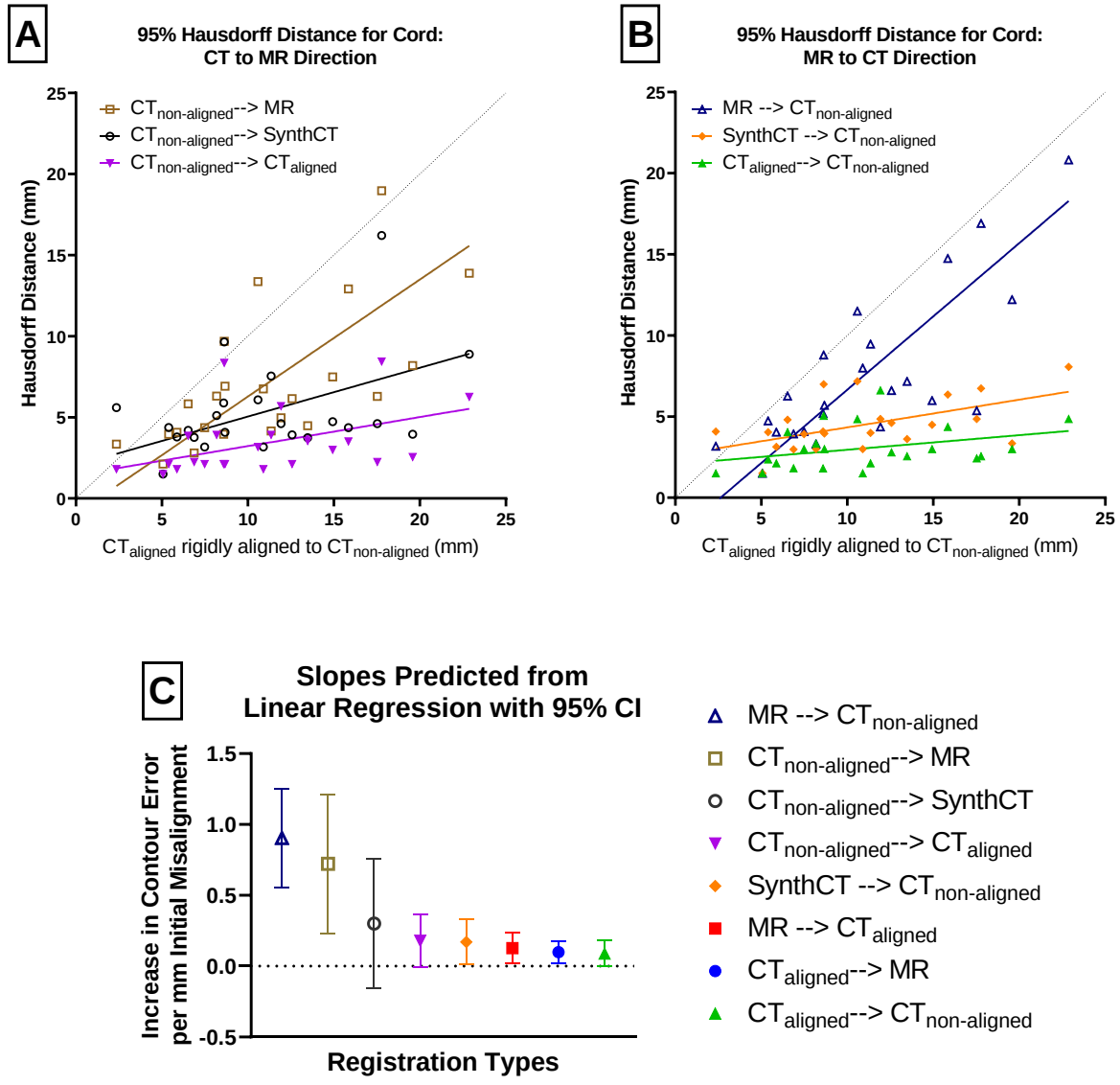


Figure 3 The spinal cord 95% Hausdorff distance for the deformations is plotted in the top two figures as a function of the initial rigid alignment Hausdorff distance. Thus, the error in cord alignment can be evaluated in terms of how misaligned the images were initially. The diagonal line shows where the deformable image registration's cord error would equal the

initial rigid alignement's cord error. The top figure is divided in the CT-to-MR direction on the left and the MR-to-CT direction on the right. The bottom figure shows the slopes of the best fit lines with their 95% confidence intervals.

Registration Accuracy: Landmark Analysis

Figure 4 summarizes the landmark analysis. The plot is ordered from the lowest average landmark error to the highest. The vertebrae landmarks tended to be lower than those on the head. This is caused by the distances of landmarks to the head rotating motion axis, as a small head tilt could lead to large distances in the eyes, nose, and mandible. The registrations between the aligned CT and MR are consistent across all landmarks, as these images were already closely aligned. Interestingly, when the MR was replaced with the aligned CT in the non-aligned CT registrations, the landmark error also stays relatively consistent. This aligned CT and MR registration error defines the performance upper bound of the proposed method if we were able to generate perfect synthetic CTs.

The direct registration shows the second largest variation in performance after the rigid alignment with respect to landmark type. Figure 5 shows the landmark error by registration type and the statistically significant groupings from the anova post-hoc Tukey test. There are four groups in order of increasing average error,

Group 1: deformable registrations between the $CT_{aligned}$ and MR, and the $CT_{aligned}$ and $CT_{non\text{-}aligned}$;

Group 2: the registrations of our proposed method between $CT_{non\text{-}aligned}$ and the Synthetic CT;

Group 3: the direct $CT_{non\text{-}aligned}$ and MR registrations;

Group 4: the rigid alignment between $CT_{non\text{-}aligned}$ and $CT_{aligned}$.

The outliers in the different registration groups were either nose or mandible landmarks. Figure 5B tabulates the average and standard deviation of the landmarks per registration type. There is an average reduction of 3.8mm in average landmark error (from 9.8mm to 6.0mm) by replacing the MR in the MR registered to $CT_{non\text{-}aligned}$ registration with a synthetic CT.

If the synthetic CT were replaced with a $CT_{aligned}$, the error decreases further by 2.2mm (from 6.0mm to 3.8mm). The trend is similar in the CT to MR direction. The average landmark error is reduced by 3.4 mm (from 10.0mm to 6.6mm) when replacing MR with a synthetic CT in the $CT_{non\text{-}aligned}$ registered to MR registration. The error is further reduced by 2.7mm (from 6.6mm to 3.9mm) with registration to a $CT_{aligned}$. The error reduction from direct registration to synthetic CT bridged registration is not only significant (2way ANOVA with Tukey's multiple comparisons test, $p<0.001$) but also greater than half the potential improvement with registration to $CT_{aligned}$, which is typically unavailable.

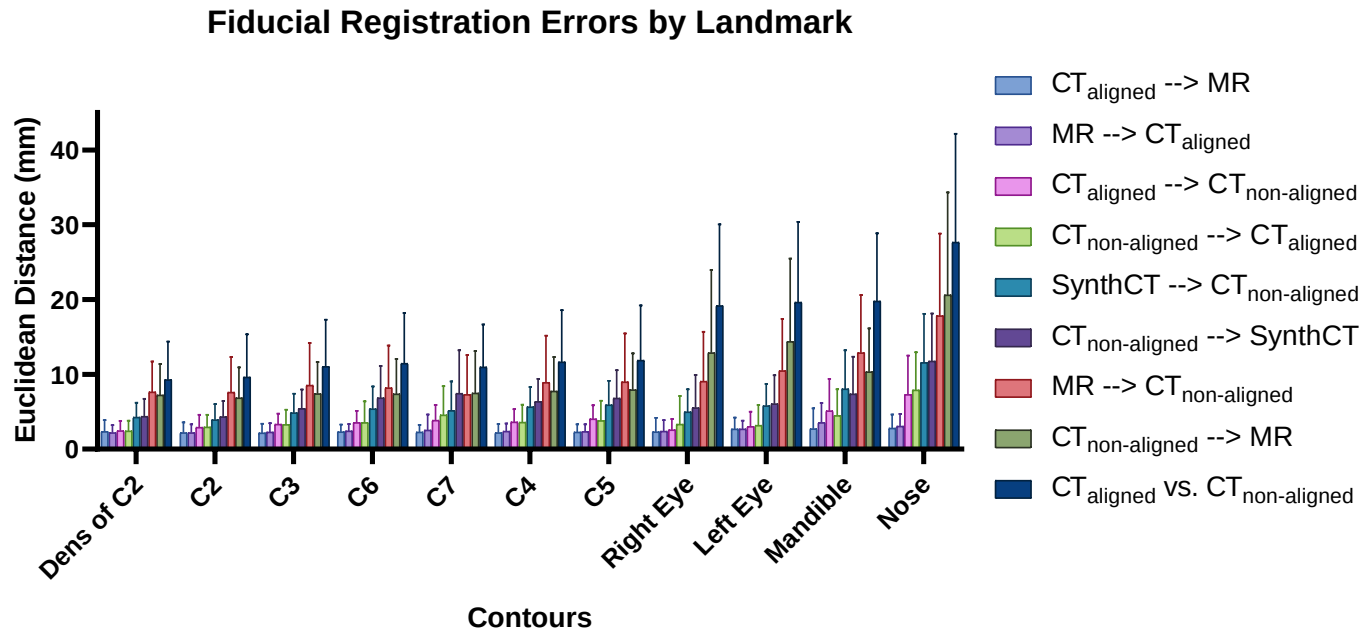
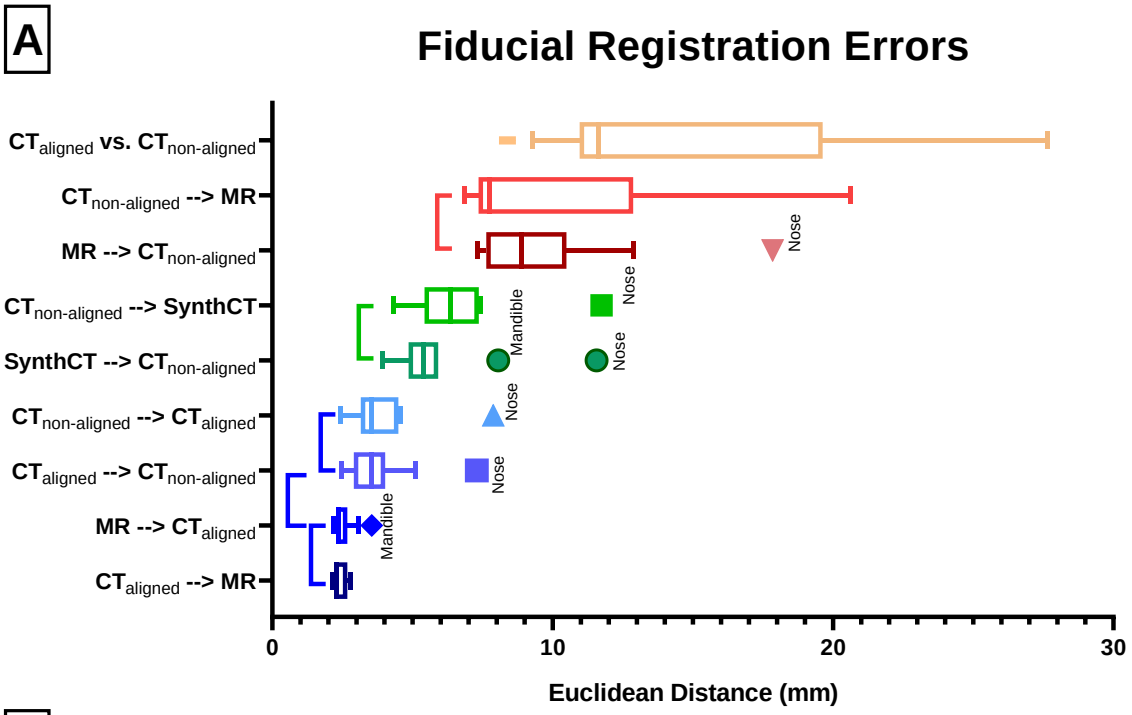


Figure 4 The patient-average Euclidean landmark error for the various registrations investigated in this study. The bars are ordered by the average error across all landmarks. From this figure, we can see that for the large $CT_{non\text{-}aligned}$ registrations our proposed method has an overall lower landmark error than the direct registration method. The rigid alignment between $CT_{aligned}$ and $CT_{non\text{-}aligned}$ is denoted by $CT_{aligned}$ vs. $CT_{non\text{-}aligned}$.

A



B

Registration	Average Landmark Error ± Standard Deviation (mm)
MR → $CT_{non\text{-}aligned}$	9.8 ± 3.1
SynthCT → $CT_{non\text{-}aligned}$	6.0 ± 2.1
$CT_{aligned}$ → $CT_{non\text{-}aligned}$	3.8 ± 1.4
$CT_{non\text{-}aligned}$ → MR	10.0 ± 4.3
$CT_{non\text{-}aligned}$ → SynthCT	6.6 ± 2.0
$CT_{non\text{-}aligned}$ → $CT_{aligned}$	3.9 ± 1.5

Figure 5 Combining all landmarks together to better visualize the post hoc results by registration type. The vertical bars to the left of the boxplots indicate registrations which were not significantly different. The rigid alignment between $CT_{aligned}$ and $CT_{non\text{-}aligned}$ is denoted by $CT_{aligned}$ vs. $CT_{non\text{-}aligned}$. Below is a table to more easily display the decrease in average landmark error with the proposed method. The bottom rows in each section show the average error if $CT_{aligned}$ is used as a surrogate for the MR. This represents a "best-case" scenario.

Registration Accuracy: Jacobian Determinant

The Jacobian determinant was calculated for each transformation. The mean across each resulting 3D matrix was found. Figure 6 shows the descriptive statistics averaged across all 24 test patients, for each registration investigated. All of the registrations have mean Jacobian determinants around 1.0. This shows that the majority of the deformed images did not experience large expansion or shrinking, consistent with the head-and-neck anatomy.

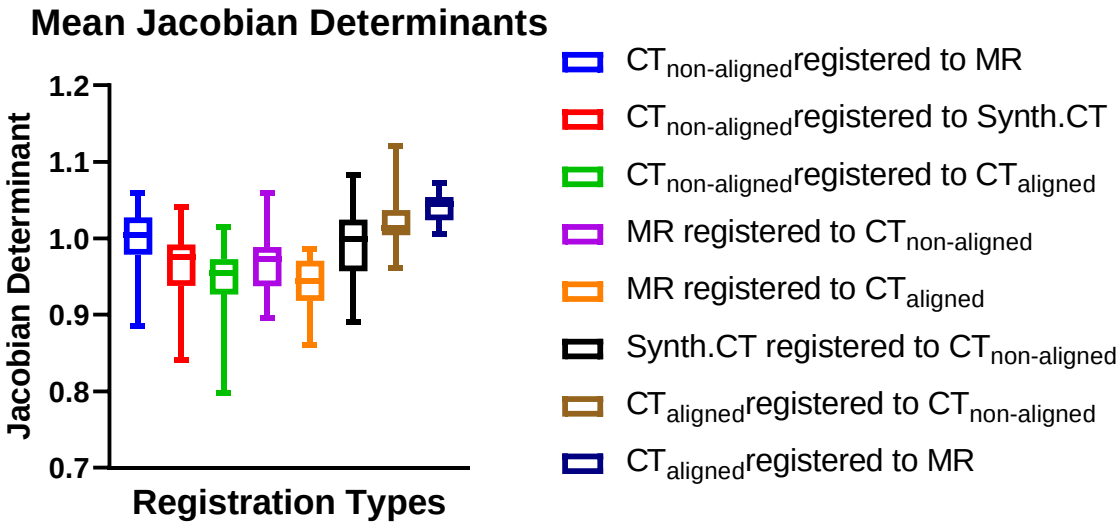


Figure 6 Patient-averaged descriptive statistics of Jacobian determinants across different deformable registration types. Error bars show the range.

Inverse Consistency

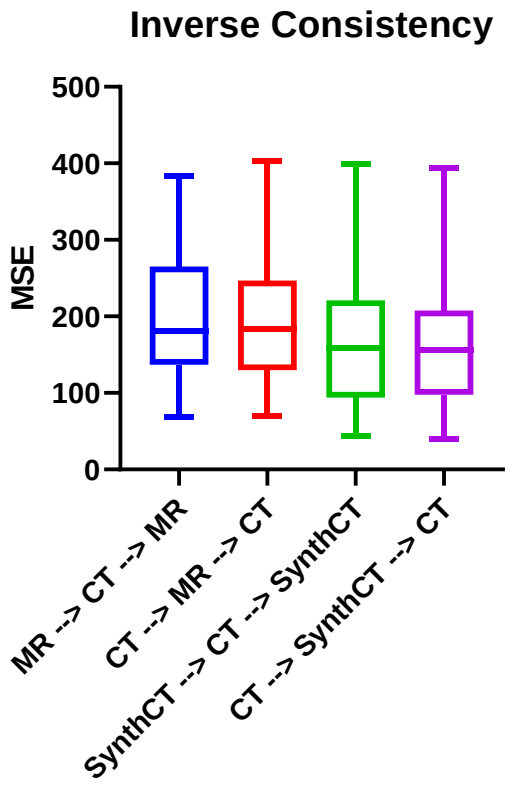


Figure 7 Transformation pairs in the opposite direction (e.g. non-aligned CT→MR and MR → non-aligned CT) were composed together and a single baseline CT was transformed under this composition. Given perfect inverse consistency, the original image should be recovered. The mean square error (MSE) was calculated between the original and transformed CT for the direct and synthetic CT bridged registrations, and in both directions. The boxplot bars show 5-95% range. A paired T-test was performed to evaluate significant difference.

There is an improved inverse consistency when using the synthetic CT bridge in both the CT to MR direction (MSE of 193.9 to 165.1, p=0.04) and the MR to CT direction (MSE of 197 to 168, p=0.04). Statistical comparisons were made using a paired T-test.

Discussion

Multi-modal deformable registration is an important technique yet a challenging problem due to its ill-conditioned nature. It is made more difficult by different material-to-imaging-value mapping and large deformations. Both compounding problems are seen in head-and-neck MR- and CT-scanned cancer patients and have hampered the utility of multimodal imaging for their radiotherapy. In the current study, we showed that the difficulty can be effectively mitigated using deep-learning generated synthetic images, with which the multimodel registration problem is reduced to a monomodal one. For large deformation, this novel registration pipeline is able to significantly improve the deformable registration results versus direct registration. The anatomy is more accurately morphed to the target images as shown in the quantitative results of spinal cord contours and the landmark tests. An additional benefit of this method is that the pipeline can be fully automated.

Current clinical practice often uses rigid registration to align CT and MR images in the head-and-neck, which is clearly suboptimal given the discrepancy in patient posture as shown in Figure 2. We show a method that offers significant improvement over the current standard of care. In addition, we show that our method is more robust than traditional, direct deformable image registration (DIR) methods. A significant challenge in this work was ensuring accurate deformation even in the presence of large head motion. Through careful parameter tuning, a

balance was able to be struck which was both accurate and robust. It was noted during this tuning process that the direct multimodal registration results were more sensitive to choice of parameters and initial conditions relative to our CT_{synth} method. Future work will endeavor to discover better ways to automatically choose these parameters for both head-and-neck, as well as other anatomical sites.

It was also observed that some of the artifacts in the MR would disappear during the process of generating synthetic CT's. Artifact reduction in MR using deep learning has been previously studied[34,35] therefore, while not the focus of this study, we are unsurprised at this result. It is well known that the variability in MRI intensity values can make image registration more challenging, and numerical techniques exist to mitigate these issues[36]. A possible added benefit of our technique may be an implicit correction in MRI intensity variations during the process of generating a synthetic CT, thus further aiding the registration. Our study was not designed to pursue this question, but would be an interesting future pursuit.

This study's analysis process closely follows the recommendations for DIR quality assurance put forward by the AAPM Task Group 132[1]. They recommend evaluating registrations with landmark error, contour error, the Jacobian determinant, and inverse consistency. Our proposed method demonstrates superior landmark error and cord contour conformation, while also showing reasonable Jacobian determinant values and improved inverse consistency. These results make us confident that a CT_{synth} based deformable registration in the head-and-neck is a valuable tool, even in the setting of large neck flexion. We saw average landmark improvements of 3-4mm for our method, which is more than half of the 6mm improvement seen in registrations between the MR and aligned CT. The aligned CT acts as a surrogate for a more realistic

synthetic CT since its anatomy closely matches the MR's. The improvements are on the same order of margins (~3mm) used in head-and-neck radiotherapy, thus they are considered clinically relevant. Note that the utilized registration algorithm does not explicitly penalize a registration for violating inverse consistency, thus the non-zero MSE. The result indicates that the registration is biased with the choice of the target image, which is also seen in the directionality-dependent differences in registration. In fact, most registration algorithms used are asymmetric[17] and this is an ongoing avenue of research.

There are a few limitations in the current study. First, although the data is unique to offer rigidly aligned CT/MR for validation, the patient number is relatively small, which has limited the power of statistical analysis. It may also have limited the quality of generated synthetic images. We performed five-fold cross validation to allow all cases to contribute to the performance analysis. Second, the MR images are from a low field scanner for MR-guided radiation therapy that provides inferior quality to the diagnostic images for head-and-neck registration. It is possible that the quality of the synthetic images can be improved based on diagnostic and multiparametric MR images. Improvements to the CT_{synth} generation could lead to further accuracy, as seen in the $CT_{aligned}$ registrations. Additionally, it is important to note that the MR and CT images were acquired with different resolutions, although they were resampled to be the same. Changing this additional variable could lead to different registration accuracies.

In our work, we used PET attenuation correction CT's as the source of our large misalignement CT's. In this way, the positioning would be very different from the immbolized planning CT. Previous work[37] examined the difficulty of registering a PET attenuation correction CT with a

treatment planning CT. They found large variability in alignment of the spinal cord (5.3mm) and mandible (5.4mm) post DIR, which were still superior to rigid registration (10.6mm and 5.5mm, for spinal cord and mandible, respectively). Our direct registration from the PET CT ($CT_{non-aligned}$) to the planning CT ($CT_{aligned}$) resulted in an average landmark error of 4.5mm for the mandible and 4.7mm for the spinal cord, while the CT_{synth} bridge method had a 7.4 mm mandible error and 6.6mm error for the cord. These values are consistent with what was shown in the referenced study. While we only have one landmark and one contour in common with this study, it shows that even in the setting of CT-CT registration, large deformations in the head-and-neck can be difficult to register. In synergy with using a CT_{synth} bridge, improvements in mono-modality registration would also lead to better multimodal registration in the head-and-neck. Currently, research using neural networks offers some exciting new avenues in this regard, including completely learning-based unsupervised DVF generation[38–41]. However, the performance of these methods depends on the availability and quality of training sets, which are particularly challenging for multimodel registration. The proposed synthetic image bridge can work well with new deformable registration techniques optimized for single modality registration.

Conclusion

Multi-modality deformable registration is challenging, especially in regions of large deformation. CT and MR are important, complementary modalities in the treatment of head-and-neck cancer. By first transforming the MR into a CT_{synth} and running a synthetic mono-modal registration, we showed that we were able to produce improved registration results in the form of lower landmark error and more accurate contour warping. Furthermore, we showed that our DIR method

improves inverse consistency and has realistic Jacobian determinant values. Continued efforts to improve CT_{synth} generation could advance this technique further.

References

1. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report. *Med Phys*. 2017;44(7):e43-e76. doi:10.1002/mp.12256
2. Heinrich MP, Jenkinson M, Bhushan M, et al. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Med Image Anal*. 2012;16(7):1423-1435. doi:10.1016/j.media.2012.05.008
3. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol*. 2017;12(1):28. doi:10.1186/s13014-016-0747-y
4. Khan FM. *The Physics of Radiation Therapy*. 4th ed. (Pine J, Murphy J, Panetta A, Larkin J, eds.). Baltimore, MD: Lippincott Williams & Wilkins; 2010.
5. Ulin K, Urie MM, Cherlow JM. Results of a multi-institutional benchmark test for cranial CT/MR image registration. *Int J Radiat Oncol Biol Phys*. 2010;77(5):1584-1589. doi:10.1016/j.ijrobp.2009.10.017
6. Roberson PL, McLaughlin PW, Narayana V, Troyer S, Hixson G V., Kessler ML. Use and uncertainties of mutual information for computed tomography/magnetic resonance (CT/MR) registration post permanent implant of the prostate. *Med Phys*. 2005;32(2):473-482. doi:10.1118/1.1851920
7. Dean CJ, Sykes JR, Cooper RA, et al. An evaluation of four CT-MRI co-registration techniques for radiotherapy treatment planning of prone rectal cancer patients. *Br J*

Radiol. 2012;85(1009):61-68. doi:10.1259/bjr/11855927

8. Daisne JF, Sibomana M, Bol A, Cosnard G, Lonneux M, Grégoire V. Evaluation of a multimodality image (CT, MRI and PET) coregistration procedure on phantom and head and neck cancer patients: Accuracy, reproducibility and consistency. *Radiother Oncol*. 2003;69(3):237-245. doi:10.1016/j.radonc.2003.10.009
9. Nyholm T, Nyberg M, Karlsson MG, Karlsson M. Systematisation of spatial uncertainties for comparison between a MR and a CT-based radiotherapy workflow for prostate treatments. *Radiat Oncol*. 2009;4(1). doi:10.1186/1748-717X-4-54
10. Weltens C, Menten J, Feron M, et al. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiother Oncol*. 2001;60(1):49-59. doi:10.1016/S0167-8140(01)00371-1
11. Rasch C, Keus R, Pameijer FA, et al. The potential impact of CT-MRI matching on tumor volume delineation in advanced head and neck cancer. *Int J Radiat Oncol Biol Phys*. 1997;39(4):841-848. doi:10.1016/S0360-3016(97)00465-3
12. Emami B, Sethi A, Petruzzelli GJ. Influence of MRI on target volume delineation and IMRT planning in nasopharyngeal carcinoma. *Int J Radiat Oncol Biol Phys*. 2003;57(2):481-488. doi:10.1016/S0360-3016(03)00570-4
13. Chung NN, Ting LL, Hsu WC, Lui LT, Wang PM. Impact of magnetic resonance imaging versus CT on nasopharyngeal carcinoma: Primary tumor target delineation for radiotherapy. *Head Neck*. 2004;26(3):241-246. doi:10.1002/hed.10378
14. Chuter R, Prestwich R, Bird D, et al. The use of deformable image registration to integrate diagnostic MRI into the radiotherapy planning pathway for head and neck cancer. *Radiother Oncol*. 2017;122(2):229-235. doi:10.1016/J.RADONC.2016.07.016

15. Brouwer CL, Steenbakkers RJHM, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol*. 2015;117(1):83-90. doi:10.1016/j.radonc.2015.07.041

16. Fortunati V, Verhaart RF, Angeloni F, et al. Feasibility of multimodal deformable registration for head and neck tumor treatment planning. *Int J Radiat Oncol Biol Phys*. 2014;90(1):85-93. doi:10.1016/j.ijrobp.2014.05.027

17. Sotiras A, Davatzikos C, Paragios N. Deformable Medical Image Registration: A Survey. *IEEE Trans Med Imaging*. 2013;32(7):1153-1190. doi:10.1109/TMI.2013.2265603

18. Schnabel JA, Heinrich MP, Papież BW, Brady SJM. Advances and challenges in deformable image registration: From image fusion to complex motion modelling. *Med Image Anal*. 2016;33:145-148. doi:10.1016/j.media.2016.06.031

19. Maes F, Vandermeulen D, Suetens P. Medical image registration using mutual information. *Proc IEEE*. 2003;91(10):1699-1722. doi:10.1109/JPROC.2003.817864

20. Roy S, Carass A, Jog A, Prince JL, Lee J. MR to CT registration of brains using image synthesis. *Med Imaging 2014 Image Process*. 2014;9034:903419. doi:10.1117/12.2043954

21. Chen M, Carass A, Jog A, Lee J, Roy S, Prince JL. Cross contrast multi-channel image registration using image synthesis for MR brain images. *Med Image Anal*. 2017;36(3):2-14. doi:10.1016/j.media.2016.10.005

22. Cao X, Yang J, Gao Y, Guo Y, Wu G, Shen D. Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. *Med Image Anal*. 2017;41:18-31. doi:10.1016/j.media.2017.05.004

23. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks.

2014:1-9. doi:10.1001/jamainternmed.2016.8245

24. Hiasa Y, Otake Y, Takao M, et al. Cross-Modality Image Synthesis from Unpaired Data Using CycleGAN. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. Vol 1. Springer; 2018:31-41. doi:10.1007/978-3-030-00536-8_4

25. Tanner C, Ozdemir F, Profanter R, Vishnevsky V. Generative Adversarial Networks for MR-CT Deformable Image Registration. *arXiv Prepr arXiv180707349*. 2018:1-11.

26. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT synthesis using unpaired data. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2017;10557 LNCS:14-23. doi:10.1007/978-3-319-68127-6_2

27. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proc IEEE Int Conf Comput Vis*. 2017;2017-Octob:2242-2251. doi:10.1109/ICCV.2017.244

28. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix : A Toolbox for Intensity-Based Medical Image Registration. 2010;29(1):196-205.

29. Shamonin D, Bron E, Lelieveldt B, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinform*. 2014;7(January):1-15. doi:10.3389/fninf.2013.00050

30. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2016;9906 LNCS:694-711. doi:10.1007/978-3-319-46475-6_43

31. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. *Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR*

2017. 2017;2017-Janua:5967-5976. doi:10.1109/CVPR.2017.632

32. Klein S, Staring M, Pluim JPW. Comparison of gradient approximation techniques for optimisation of mutual information in nonrigid registration. 2005;(2):192. doi:10.1117/12.595277

33. Dubuisson M-P, Jain AK. A modified Hausdorff distance for object matching. 2002;(1):566-568. doi:10.1109/icpr.1994.576361

34. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*. 2019;29(2):102-127. doi:10.1016/j.zemedi.2018.11.002

35. Higaki T, Nakamura Y, Tatsugami F, Nakaura T, Awai K. Improvement of image quality at CT and MRI using deep learning. *Jpn J Radiol*. 2019;37(1):73-80. doi:10.1007/s11604-018-0796-2

36. Bağci U, Udupa JK, Bai L. The role of intensity standardization in medical image registration. *Pattern Recognit Lett*. 2010;31(4):315-323. doi:10.1016/j.patrec.2009.09.010

37. Hwang AB, Bacharach SL, Yom SS, et al. Can Positron Emission Tomography (PET) or PET/Computed Tomography (CT) Acquired in a Nontreatment Position Be Accurately Registered to a Head-and-Neck Radiotherapy Planning CT? *Int J Radiat Oncol Biol Phys*. 2009;73(2):578-584. doi:10.1016/j.ijrobp.2008.09.041

38. Krebs J, Mansi T, Delingette H, et al. Robust Non-rigid Registration Through Agent-Based Action Learning. *Med Image Comput Comput Assist Interv − MICCAI 2017*. 2017;10433:344-352. doi:10.1007/978-3-319-66182-7

39. Simonovsky M, Gutierrez-Becker B, Mateus D, Navab N, Komodakis N. A Deep Metric for Multimodal Registration. *Med Image Comput Comput Interv -- MICCAI 2016*. 2016;9902:10-18. doi:10.1007/978-3-319-46726-9

40. Dalca A V., Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal*. 2019;57:226-236. doi:10.1016/j.media.2019.07.006

41. Balakrishnan G, Zhao A, Sabuncu MR, Dalca A V., Guttag J. An Unsupervised Learning Model for Deformable Medical Image Registration. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2018:9252-9260. doi:10.1109/CVPR.2018.00964

CHAPTER 3: Extending Cropped Medical Images With Neural Networks for Deformable Registration Among Images with Differing Scan Extents

Introduction

Deformable image registration (DIR) is a topic of intense research and clinical interests in radiation therapy. DIR establishes correspondence between medical images for imaging information synthesis, dose accumulation and adaptive treatment planning. For these applications, DIR is frequently performed on image pairs that exhibit non-rigid motion. On the other hand, the usefulness of DIR can be limited by low accuracy and robustness. The process of matching one image to another can introduce erroneous or unrealistic tissue deformation[1], requiring practice of caution when DIR is involved in clinical decision for interventions. In cases where the DIR accuracy is unsatisfactory or the accuracy cannot be verified, rigid registration is used instead as a compromise[2–4].

Besides differences in multimodal image intensity and large deformation, a common factor contributing to the DIR difficulty is the mismatch in the image scan extents or the field-of-view. Because the boundary conditions are not explicitly available, unrealistic deformation is often introduced in DIR. The unrealistic stretch or compression of tissues is most severe near the edges of an image but can propagate through the entire image volume with smoothness constraints in the DIR DVF. The scan extent mismatch is common in retrospective analysis, where images were acquired with varying scanning protocols, as well as in multimodal registration problems. In image guided radiotherapy, cone beam CT (CBCT) images are used to help with patient set up

but CBCTs have a substantially more limited coverage in both the axial and longitudinal dimensions compared with the planning CT. Imaging volume mismatch is also common in MR to CT registration. MR provides superior soft tissue visualization that is helpful for tumor and normal tissue delineation, but the MR imaging volume is often smaller than the planning CT. MR images acquired on oblique orientations further complicate the imaging volume mismatch issues. We previously demonstrated that the challenges in registering MR to CT due to differences in imaging intensities can be mitigated via synthetic image bridge[5] but the issues due to mismatched imaging volumes persist. According to TG-132, differences in scan extent are a major source of deformable registration error[6].

Research to mitigate the adverse impact due to imaging volume mismatch has been reported. A straightforward approach to reduce the registration error due to mismatched imaging extent is to manually crop the larger imaging volume to match the scan length of the shorter image[7,8]. Manually cropping the images not only reduces the workflow efficiency, but also introduces error because the image matching lines are not explicitly available to the operator. The error can be substantial when large patient pitch correction or deformation is involved. Periaswamy and Farid used an expectation maximization algorithm to simultaneously segment the more complete image volumes and register partial images[9]. The method effectively contained the registration error due to image artifacts but its ability to handle both large deformation and mismatched scan volume was not demonstrated. To address differing scan lengths in CBCT and planning CT registration, researchers and then relied on the DVF smoothness constraint to outside the effective field of view[4,10]. The method was shown to reduce the DIR error, but the registration accuracy was still limited by the lack of contextual information due to missing volumes.

Aside from their specific algorithms, the existing methods share the strategy of using the intersection of the two images as the starting point of DIR. By doing so, the imaging information in the more complete image is discarded despite its potential value for the overall registration accuracy. In this study, we take a fundamentally different approach. Instead of cropping the images, we propose to fill the missing portion of the anatomy using neural networks. The registration can then proceed using the artificially extended image. We design the study to answer two questions: (1) Do registrations with artificially extended images perform as well as registration pairs with equal extent (2) How does the quality of the registration with artificially extended images vary as a function of the initial amount of missing tissue?

Materials and Methods

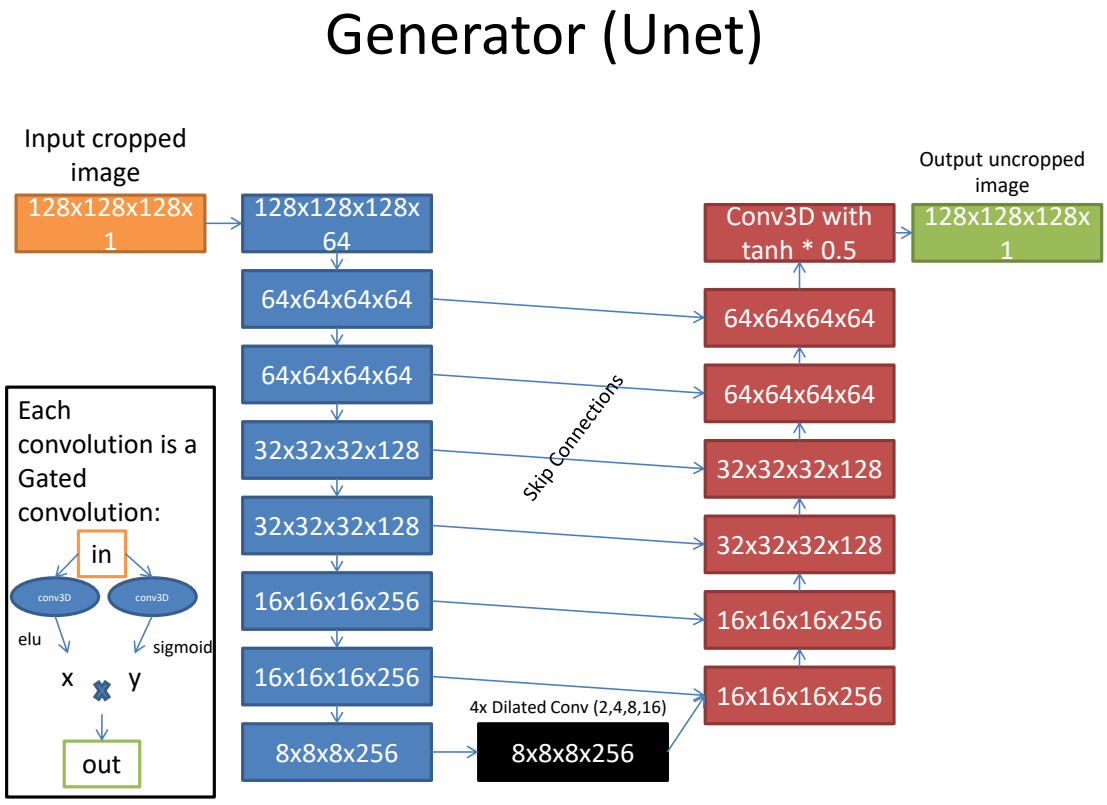
Dataset

Head and Neck CT images were acquired from The Cancer Imaging Archive (TCIA) dataset[11]. We had a total of 409 training, 53 validation, and 53 testing images. Scan extent went from the top of skull to approximately the carina. Scanning beds and immobilization equipment were masked out of the images. For input into the network, all images were rigidly registered to a template image and downsized to 128x128x128 with 4mm isotropic voxels. Image intensity value were clipped to a range of [-1024, 3000], then normalized to [-1, +1]. For analysis, volumes were automatically segmented using a neural network approach[12,13]. This resulted in 16 contours per patient.

Network

For this work, we used a Generative Adversarial Network (GAN) approach to extend the cropped volume[14]. We term the network **CropGAN**. A GAN consists of a generator to create synthesized data, and a discriminator to judge if data is synthesized or real. The input to our generator was a cropped volume, and the output was a volume with the missing portion replaced with synthesized data. The cropped region was randomly created each iteration of training, where the angle was randomly varied between 0 and 45 degrees in the superior-inferior direction, and between -5 and 5 degrees in the other 2 dimensions; and the amount of cropping was varied on both the superior and inferior edges from 120 to 210mm. We chose to vary the cut angle of the crop to simulate two common scenarios in DIR. First, as a preprocessing step, rigid registration is performed prior to DIR. Correction of the patient pitch and yaw will lead to oblique cutting planes relative to the target image. The second scenario is registration of the MR acquired in oblique orientations. In CropGAN, the generator was a 3D U-net with skip connections[15]. At the bottom of the U-net, we used 4 dilated convolutions to increase the amount of contextual information for prediction[16]. All of the convolutions in the U-net used instance normalization with elu activation and were gated so the network could adaptively learn feature selection, as was done by Yu et al[17]. The discriminator had 3 inputs: the cropped image, either the original full image (uncropped) or synthesized output from the generator (synthetically uncropped), and the mask used to crop the image. The cropped and uncropped (either full or synthetic) inputs were first concatenated together. The Discriminator was dual branching, with one branch operating on the entire concatenated image, while the other branch applied the mask

for cropping to only use the data from within the mask to focus on the fidelity of the synthesized portion. The discriminator used spectral normalization, which has been shown to add stability to discriminator training[18]. The output of the discriminator was a concatenation of the two branches. Figure 1 shows details of the networks.



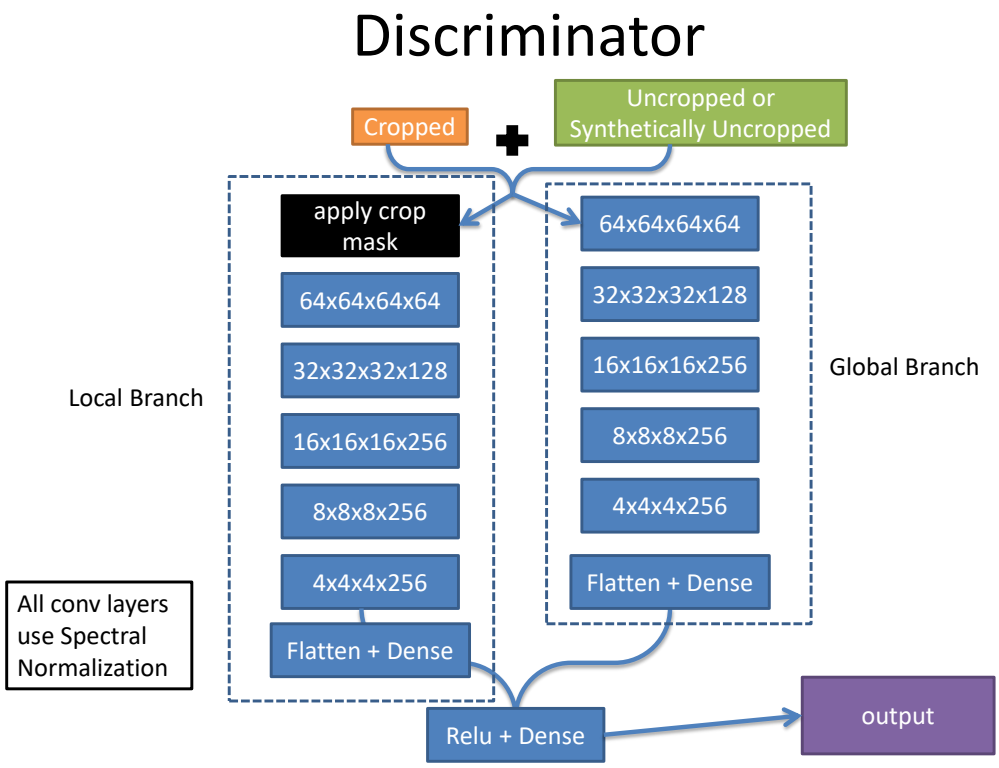


Figure 1 Architecture of Generator (top) and Discriminator (bottom).

For the loss function we followed the formulation of Hui et al[19], which uses several deep feature-based losses. We passed the generated and target uncropped image through a previously trained VGG network[20]. This network was trained to classify CT and MR imaging sites from patches and had learned activations pertinent to these modalities' features[21]. An example showing the first 5 activation layers for the generator output and ground truth target is given in Figure 2.

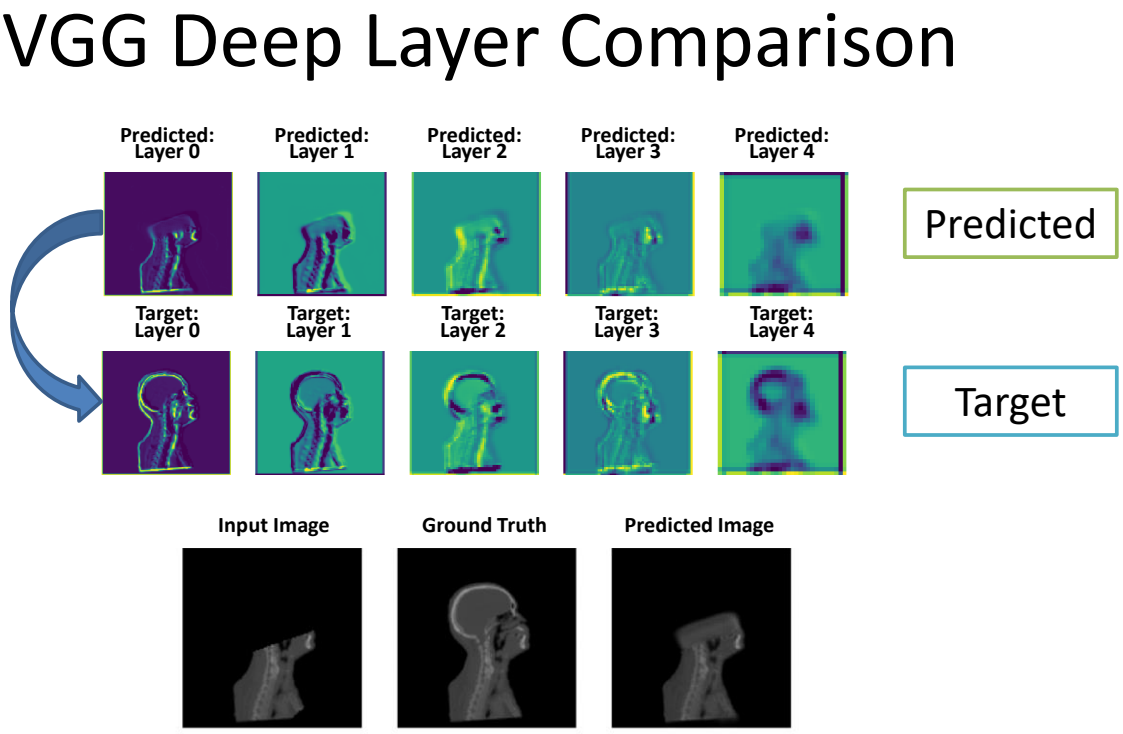


Figure 2 visualizes the activations in the first 5 layers of the VGG network for a predicted and uncropped ground truth image. These activations are used as features, which are compared using equations 1, 3, and 5 to produce a similarity metric, driving the predicted image to resemble the target.

We compared the activations between the generated and target images in two ways. First, we compared the activations from the first 5 convolutional VGG layers [Equation 1].

Equation 1

$$loss_{vgg} = \sum_{l=1}^{5} w^l \frac{\left\| \Psi^l_{I\,gt} - \Psi^l_{I\,output} \right\|_1}{N_{\Psi^l_{I_{gt}}}}$$

Where $\Psi^l_{I_*}$ is the activation map of the l^{th} layer, for image volume I_* (gt = ground truth, output=output from generator). $N_{\Psi^l_{I_{gt}}}$ is the number of elements in the ground truth image's l^{th} layer. w^l weights each addend as a function of the channel size of the l^{th} layer of the ground truth image [Equation 2]. $C^l_{\Psi^l_{I_{gt}}}$ in Equation 2 is the channel size of $\Psi^l_{I_{gt}}$.

Equation 2

$$w_l = \frac{1e6}{C^l_{\Psi^l_{I_{gt}}}}$$

Second, to focus on more challenging areas of the image, we compared the error map weighted activations from the first two VGG layers. [Equation 3]

Equation 3

$$loss_{vgg\_challenge} = \sum_{l=1}^{2} w^l \frac{\left\| M^l_{guidance} \odot \left(\Psi^l_{I_{gt}} - \Psi^l_{I_{output}} \right) \right\|_1}{N_{\Psi^l_{I_{gt}}}}$$

Where $M^l_{guidance}$ is the error map associated with layer l and is used to give more weight to the VGG layer differences which are more challenging to match. For each layer, $M^l_{guidance}$ is given by $M^{l+1}_{guidance} = average\ pooling\left(M^l_{guidance}\right)$. $M^1_{guidance}$ is equal to $M_{guidance,p}$ which is the error map value at position p [Equation 4]. $M_{guidance,p}$ is derived from the generated image and its corresponding ground truth. Average-pooled guidance maps give a spatial correspondence between the differences seen in the images, and the differences seen at deeper layers.

Equation 4

$$M_{guidance,p} = \frac{M_{error,p} - \min(M_{error})}{\max(M_{error}) - \min(M_{error})}$$

$$\text{where, } M_{error} = \left(I_{out} - I_{gt}\right)^2$$

Mean absolute error was used to assess the fidelity between the generated and target uncropped images. This is given as $L_{fidelity}$. For the adversarial loss, we used a Wasserstein Hinge loss[22]. In addition to the adversarial loss, the discriminator also used a deep feature-based loss. The activation layers of the cropped-area discriminator branch were used to compare the generator output and ground truth [Equation 5].

Equation 5

$$Loss_{disc\_features} = \sum_{l=1}^{6} w^l \frac{\left\| D^l(I_{gt}) - D^l(I_{output}) \right\|_1}{N_{D^l(I_{gt})}}$$

The total loss function is thus:

Equation 6

$$Total\ loss = L_{adverarial} + \lambda_1 L_{fidelity} + \lambda_2 L_{vgg} + \lambda_3 L_{vgg\_challenge} + \lambda_4 L_{disc\_features}$$

We searched for a stable training result by iteratively varying the weights (λ_*) using the validation set. This led to empirically selected weights of 20, 10, 10, and 5, respectively. Further

tuning may possibly lead to improved results. Our generator and discriminators used an RMSprop optimizer with a learning rate of 0.00005. We used a batch size of 2 and trained for 2000 epochs.

The output from the network was a 128x128x128 image with 4mm^3 voxels. The synthesized portion was resized to 512x512x512 (1mm^3 voxels) to match the size of the original image. The final synthetically extended image only had synthesized voxels in the cropped region. The non-cropped portion was copied to the final image.

Registration

We tested how well deformable registration with the synthetic cropped images compared to uncropped (ground truth) registration and cropped registration. To do this, we deformably registered the moving images (cropped, uncropped, and synthetic uncropped) to the same target images (cropped, uncropped, and synthetic uncropped) in all unique combinations. It is worth noting that for fair comparison the synthetic image volumes are only used to assist DIR. When the moving image was synthetically extended, we applied the resulting deformation vector field to the cropped image such that the final result only included actual scanned data.

Without losing generality, we performed registrations using an open source B-spline method (Elastix[23,24]). We used a multi-resolution deformable registration scheme and mutual information as the cost function, as in a previous publication[5]. This method was selected due to its competitive performance in registering head and neck images[25], open-source nature to

facilitate comparison, and flexible registration parameter settings; however, the CropGAN images are expected to work with other registration algorithms.

Analysis

We tested our hypothesis that synthetically extending cropped images would lead to the same registration quality as a registration performed with the full, ground truth images by evaluating the similarity between deformed and target contours. To avoid being skewed by the organ size, instead of the Dice index, the similarity was calculated using the 95% Hausdorff distance surface matching metric[26,27]. We analyzed our results using a one-way ANOVA amongst registration pairs, as well as a linear regression between the pre- and post- registration contour similarity. All analyses were performed using GraphPad Prism.

We tested our secondary hypothesis that the registration quality using synthetically extended images would be the same as using full images independent of the initial cropping amount (range of approximately the superior apex of the skull to the inferior nose: 120 to 210 mm superior cropping) by cropping the same image by variable amounts, synthesizing the missing portion using CropGAN, then performing the same registration tests. For this experiment the angle of cropping was kept a constant 23 degrees in the superior-inferior direction. This angle was selected to be the middle of the angle ranges used during training. The other two cropping planes had angles of zero. The results were averaged over 3 patients and analyzed using linear regression.

Results

Execution

Training the CropGAN network took 6 days on one Nvidia Quadro RTX 8000 GPU. Once completely trained, inference took 0.04 seconds to synthesize the missing part of the cropped image.

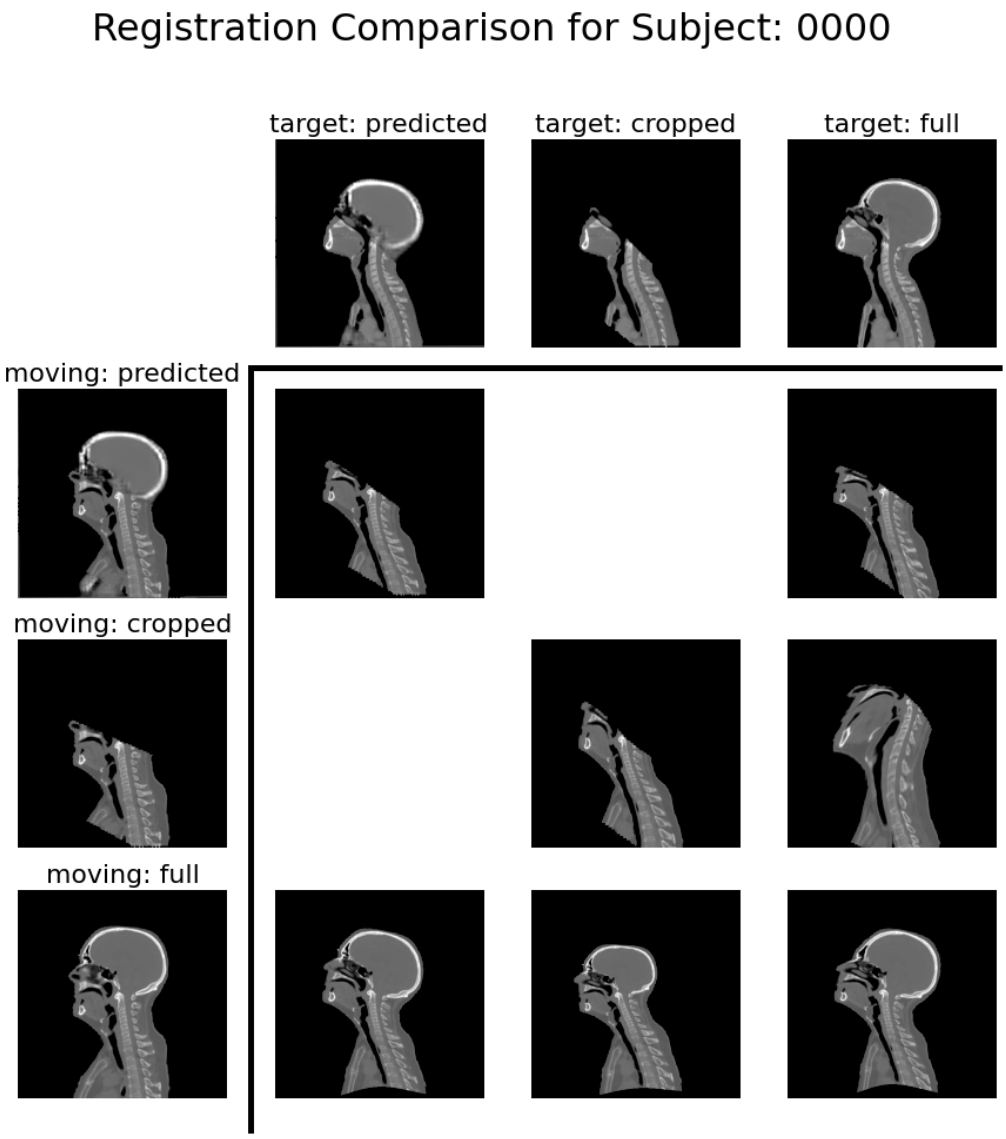


Figure 3 Example registration of one of the 53 test patients. Columns are for a given target image; Rows are for a given moving image. The intersection shows the deformable registration result. Our method (top row of central grid) has applied the deformation vector field to the original cropped image, so only real data is included in the final registration result. Cropped-to-predicted, and predicted-to-cropped are not shown, since if one cropped image could be predicted, the other could feasibly be predicted as well.

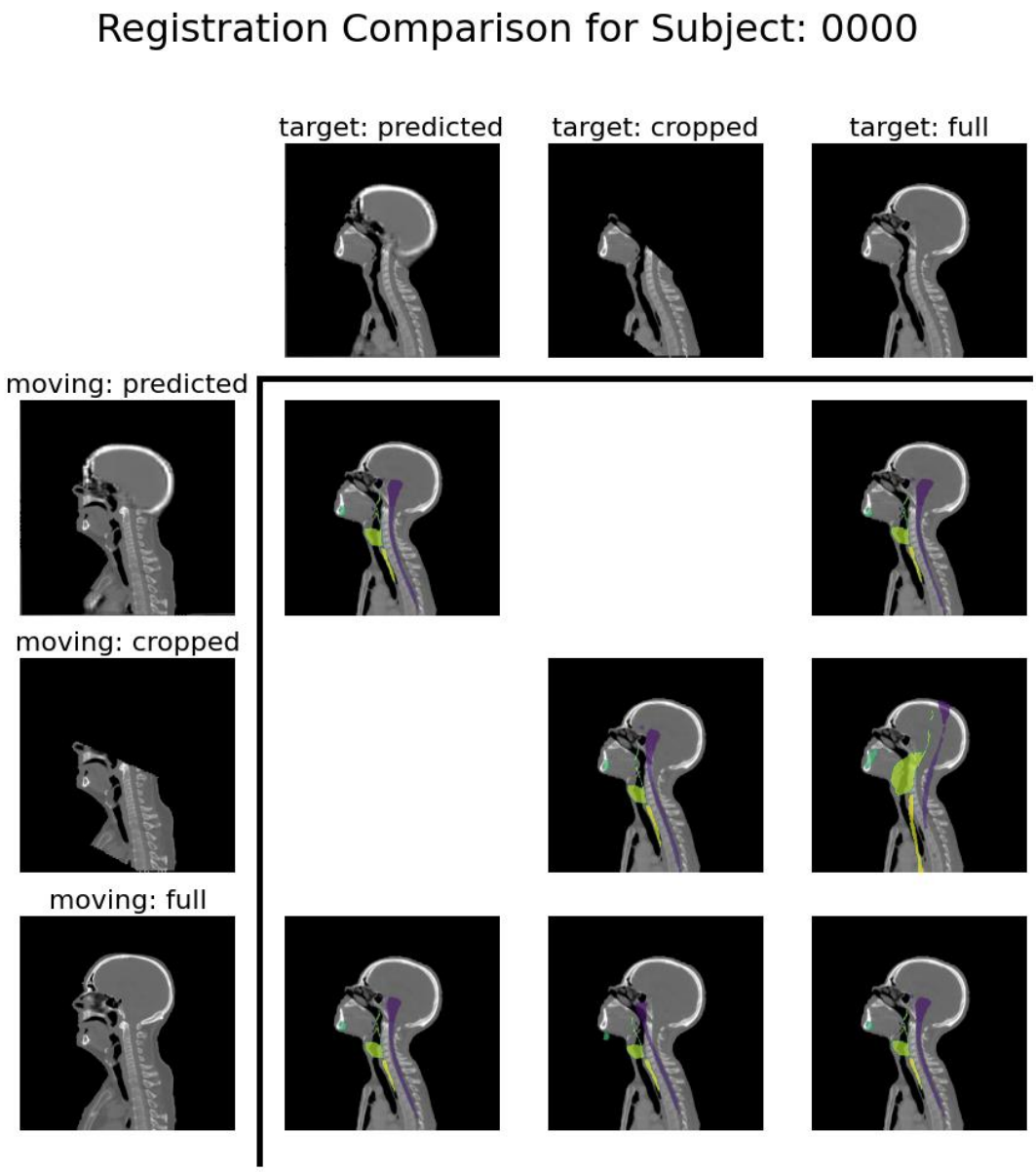


Figure 4 Example registration of one of the 53 test patients. Columns are for a given target image; Rows are for a given moving image. The intersection shows the deformable registration result applied to the moving image's contours.

Registration Comparisons

We performed 364 deformable image registrations to compare all 7 combinations of source and target volumes across the 52 test images (one of the 53 test images was held back as the target image), with random amounts of induced cropping from 120 to 210mm. An example showing one such set of registrations is given in Figure 3 and Figure 4. The columns are for a given target image (synthesized, cropped, or full) and the rows are for a given source image (synthesized, cropped, or full). The intersection of a row and column show the registration result (Figure 4 shows the respective deformed contours overlaid on the target). From the provided example, it is clear that deformable registration between images with different scan extent leads to unrealistic distortion. Compared with the worst case when the moving image is cropped and the target image is full, registering a full volume to the cropped volume results in less distortion, though it is still worse than registering between two uncropped images. Registering with the synthesized images in all three cases leads to close performance to the registration of uncropped images as shown in the deformed contours of Figure 4.

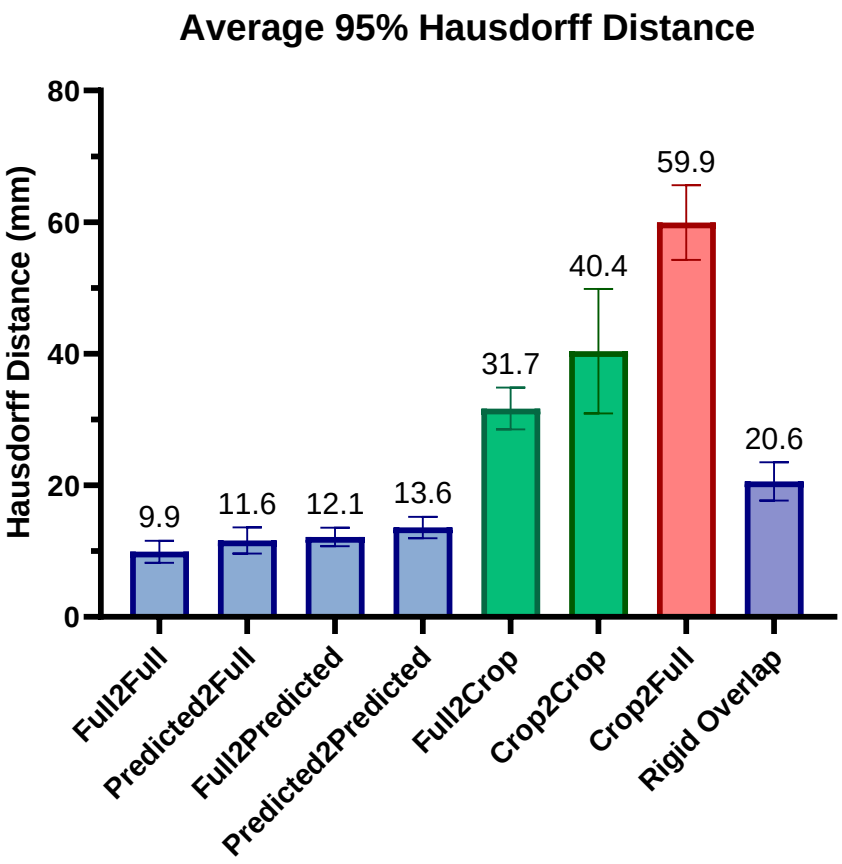


Figure 5 The average 95% Hausdorff Distance between deformed and target contours for each registration pair, averaged across all 16 contours, and all 53 test patients. The best-case registration is "Full2Full" (leftmost bar), while registrations using our method are shown in the next 3 bars. These 4 leftmost bars are not significantly different (represented by being the same color blue). Full2Crop and Crop2Crop were not significantly different, while Crop2Full was the highest of all. While the difference between our method and rigid overlap did not reach significance, we did see a near halving of the error (20.6mm average error down to ~12mm error). The error bars are the 95% confidence interval.

In the quantitative analysis, the 95% Hausdorff distance averaged across all contours is displayed with 95% confidence intervals in Figure 5. Using a one-way ANOVA with a post-hoc Tukey multiple comparison test, registration using CropGAN synthesized images in all three cases is not statistically different from the best-case full image registration ($p>0.9$), while it is

significantly different from registrations using a cropped image in either the source or target ($p<0.0001$). While the average contour distance of registrations using synthesized images was approximately half that of a simple rigid alignment, this difference was not statistically significant. We strengthened these conclusions by testing for equivalence between our proposed method and full image registration using a two-one-sided t-test. We chose our equivalency delta to be the average error reported for the automatic contouring algorithm (3.39mm 95% Hausdorff distance[13]). We concluded with 95% confidence that both synthesized-to-full and full-to-synthesized registrations were equivalent to a full-to-full image registration within the error of contouring. Registrations using synthesized images for both the source and target had confidence limits 1.6mm beyond this contouring error threshold. Thus, while it may not be significantly different from full image registration, it is advantageous to have either the source or target image be full. Interestingly, the full-to-cropped registration had confidence intervals 20.7mm beyond the contour error.

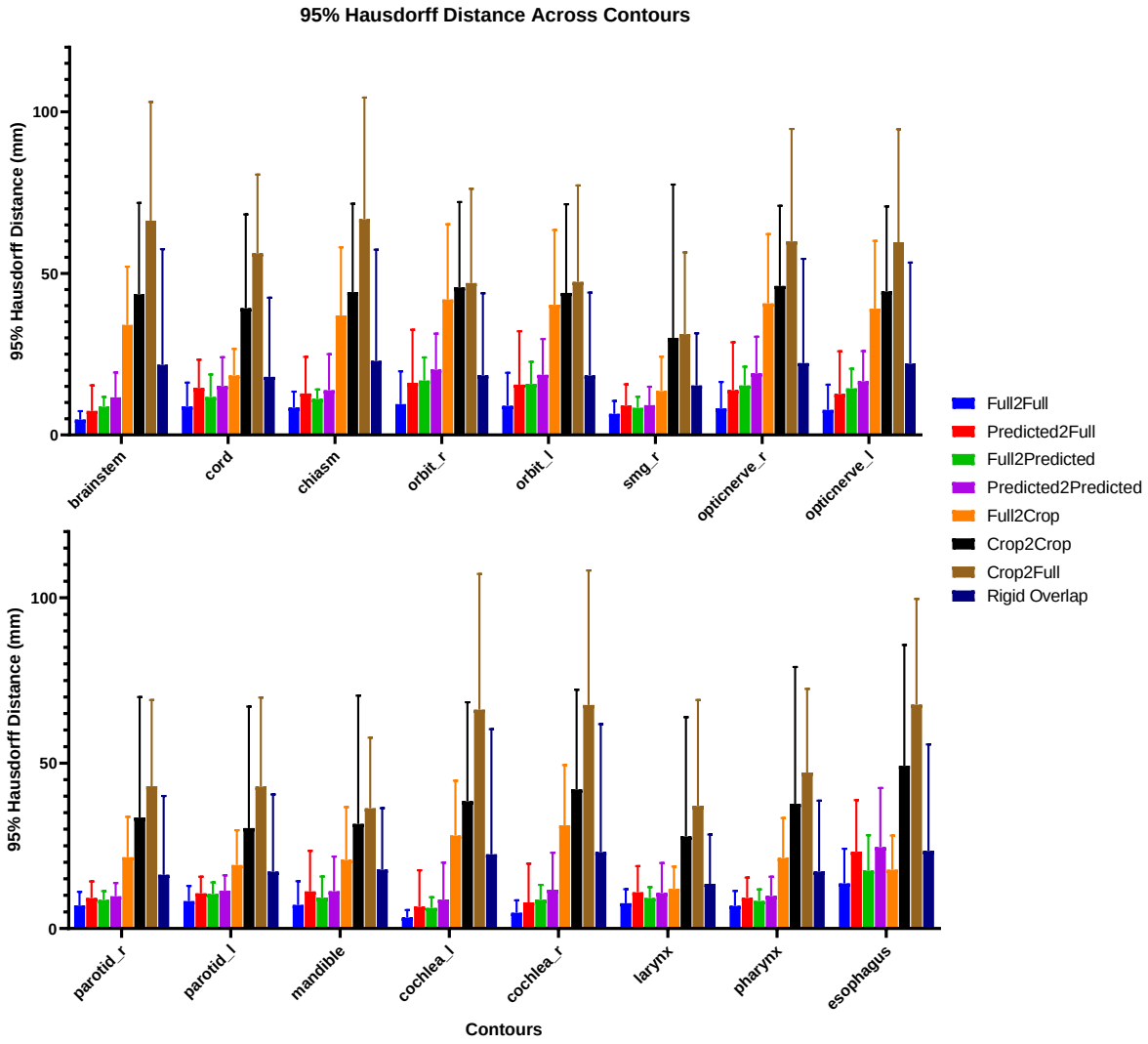


Figure 6 shows the average registration error broken down by each contour used in this study. The error bars are standard deviation. The horizontal axis has been split into to levels to improve readability.

Figure 6 shows the 95% Hausdorff distance of individual contours. The trends are consistent with the average distance analysis. The esophagus showed the greatest error across registrations due to inconsistencies in the inferior range of this lower contrast contour.

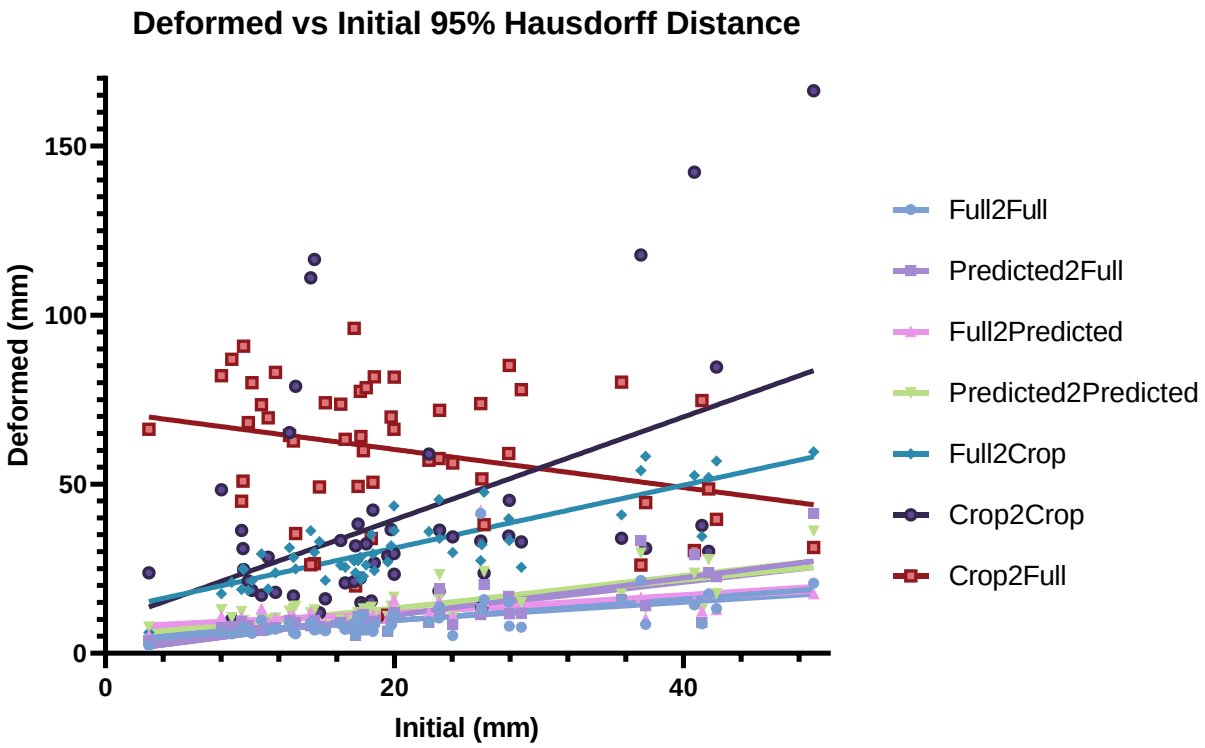


Figure 7 demonstrates how the registration error varies as a function of initial rigidly aligned 95% Hausdorff distance. The intial overlap value is a measure of the difficulty of the registration. The full-to-full registrations and the ones using our proposed CropGAN technique all have relatively flat lines, showing their robustness to the registration's initial conditions.

To analyze the dependence of our results on the degree of registration difficulty, we plotted the average 95% Hausdorff distance of the registrations as a function of the original degree of error (Figure 7). A simple linear regression was used to obtain best fit lines. The slope for the proposed method closely follows that of the best-case full image registration. Specifically, the full-to-full image registration and full-to-synthesized slopes were not significantly different ($p=0.29$), with a shared slope of 0.27. Synthesized-to-full and synthesized-to-synthesized registrations were also not significantly different ($p=0.10$) with a shared slope of 0.48; however,

they were significantly different from a full-to-full registration. Additionally, the relatively flat nature of these lines indicates that our proposed method performs well across a broad range of registration difficulty. When both the source and target images are cropped, the more challenging registrations can have errors exceeding 10cm. When only the source or target are cropped errors are in the 3-5cm range. This is consistent with the qualitative results shown in Figure 3.

Dependence on Cropping Amount

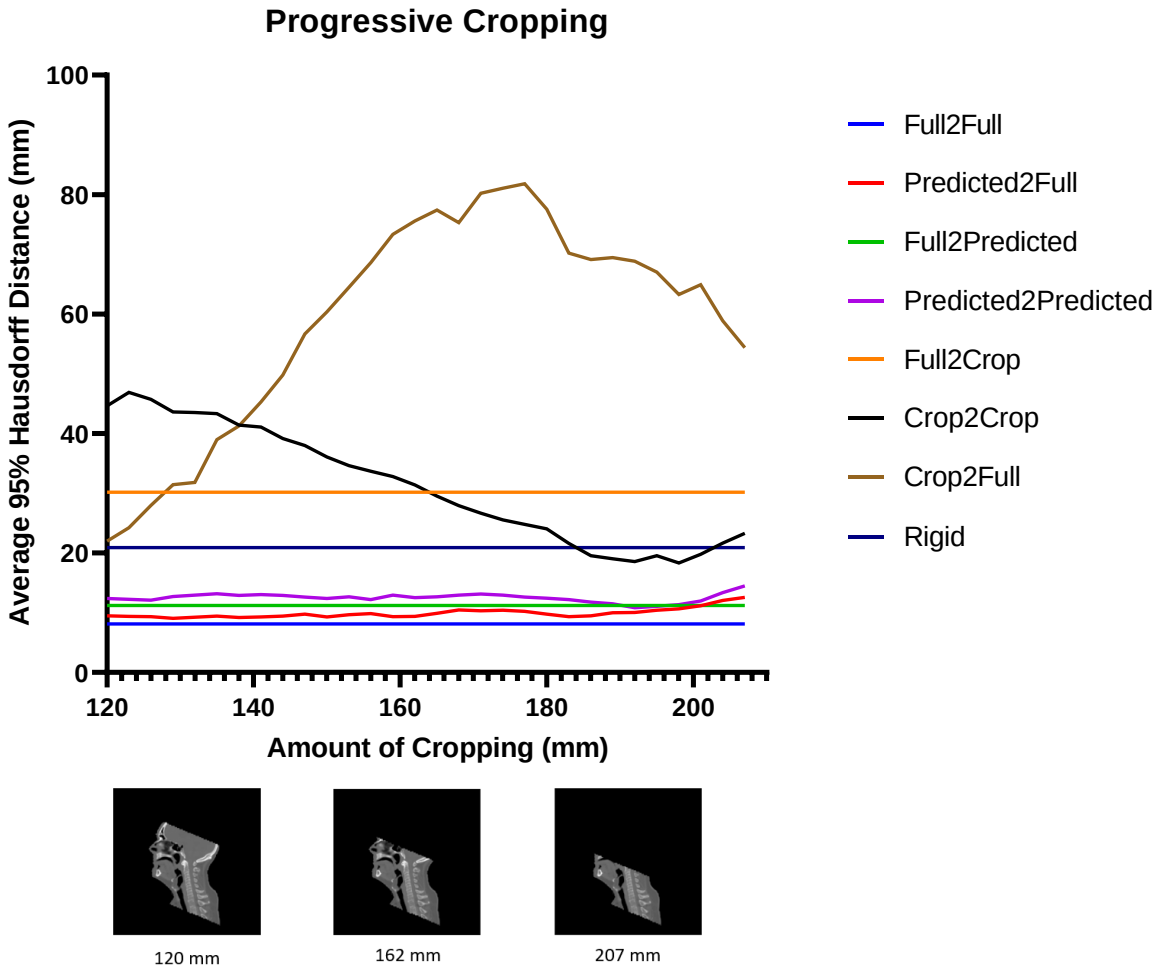


Figure 8 shows the registration error as a function of missing tissue averaged across three patients. The images below the plot help to visualize a given amount of cropping. We see that for our method (purple and red lines) the amount of superior cropping does not have much of an effect before 18.7cm, where nearly all the superior half of the tissue is missing. This demonstrates the robustness of our technique. Our method is also superior to rigid alignment for all cropping amounts investigated. When the moving image is left cropped (brown and black lines), the registration quality varies wildly.

To analyze the effect the amount of cropping had on the registration result, we progressively increased the amount superiorly cropped in the moving image, while keeping the angle of

cropping constant (S-I plane=23˚, A-P=L-R=0˚). These increasingly cropped images were passed through the trained CropGAN network to synthesize the missing regions. The target images (full, cropped, and synthesized) were kept the same for this experiment. The results averaged across three patients, along with a visualization of the cropping extent, are shown in Figure 8. The lines for registrations where the moving image was full are flat, as these are not changing in this experiment; however, they offer useful reference lines. Our proposed method performs as well as a full image registration until approximately 18.7cm of scan is missing from the superior edge. The synthesized-to-full registration was closest to the full registration result (average difference of 1.5mm). When both the moving and target images were synthesized, the average difference was 4.5mm from the full registration. This is contrasted with the registrations including cropped images, which had an average difference of 22mm and higher. The registrations with cropped moving images show larger error, however there is a noticeable decrease around 170mm. For cropped-to-cropped registration (black line), the error decreases until the cropped moving image extent matches the cropped target image's extent. As the moving image is cropped further, the extents again become mismatched, and the error increases. For the cropped-to-full registration (brown line), the error increases until the regularization of the registration algorithm prevents further stretching of the small moving image to the larger full image. While the average error for cropped-to-full registration appears to decrease slightly in Figure 8, observing the registration results directly reveals extreme distortion for these larger crop amounts.

For our proposed method, the synthesized-to-full and synthesized-to-synthesized registrations are *independent* of crop extent (slope was not significantly different from 0, p=0.2887 and 0.8556, respectively) until this extreme cut point of 18.7cm. This point roughly corresponds to the

region of the nose, suggesting this to be an important landmark for synthesis. This is in sharp contrast to the large, varying results with the original cropped registrations. These results show that our method is robust across a wide range of scan extents.

CropGAN Synthesis Visual Performance Variation

While it not the intent to recover the accurate anatomy for individual patients, it is interesting to visually exam the potential for creating missing tissues. Figure 9 shows two representative cases for good (top row) and poor (bottom row) synthesis of missing imaging volumes. In the good case, the network synthesized realistic anatomies including sinuses, sternum, and heart. In the poor case, the network failed to generate the patient nasal and skull base anatomies possibly due to the low number of training images and large variation of metal artifacts. In any case, the anatomies generated using CropGAN in its current form is not actual and cannot be used as such. For registration purposes, however, the quality of image synthesis achieved using CropGAN appears to provide adequate contextual information for DIR.

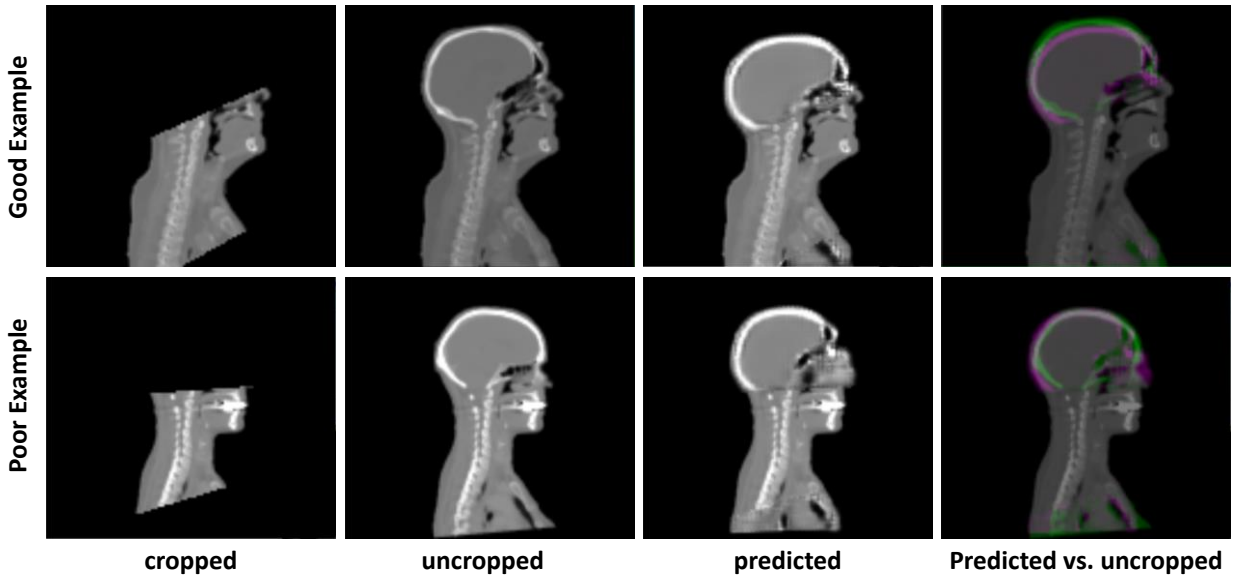


Figure 9 An example showing a good (top) and poor (bottom) predicted completion of a cropped image. The first column shows the cropped image, the second column shows the uncropped ground truth, and the third column shows the predicted result from our network. A difference image is shown in the right-most column (best seen in color). The poor prediction occurs near an artefact in the mouth.

Discussion

We present here a novel solution to directly address adverse effects due to inadequate or mismatched scan extent in deformable registration. DIR between images of insufficient extents is a major source of registration error. Existing approaches focus on cropping the larger or more complete images to better match the cropped images, which results in loss of information that could have benefited the registration. This is particularly problematic when both the moving and target images have inadequate scan extent for registration. We were able to artificially extend cropped images using a method which is fully 3D. The method is fully automated and able to handle a broad range of scan extent differences. Once trained, our method fills the missing

volume in 0.04 seconds, making it an amenable addition to a clinical workflow. To our knowledge, the current study is the first to synthesize the missing or cropped imaging volumes to improve the registration performance. It is worth emphasizing that the synthesized anatomy cannot represent the actual patient anatomy in the missing volumes. It serves the purpose of assisting DIR of the actual imaging volume. Therefore, when the moving image is cropped, we apply the DVF from the synthesized moving image registration to the cropped image, thereby only including the real imaging data in the final result.

The task to synthesize missing image slices itself is also novel. Neural networks have been used to inpaint a missing patch inside a 2D medical image slice[28–30], which is a considerably less challenging problem that is analogous to interpolation with known boundary conditions around the missing patch. In contrast, synthesizing data in a cropped image is analogous to extrapolation with undefined boundary conditions. A study looking at network-based image extension in 2D landscape photos was able to successfully extend natural 2D images, however they caution that their results did not apply well to photos of human faces.[22] Our proposed CropGAN uses generative adversarial networks (GAN) to synthesize missing data, a technique that has been well tested in network-based inpainting tasks[17,31]. Specifically, we based our method on the winner of the AIM 2020 Challenge on Extreme Inpainting[19,32], which used deep features in the generator, discriminator, and a VGG net as terms in the loss function. That study showed impressive results filling in holes of 2D color photos, yet has not yet been pursued for image extension nor for 3D medical images.

Our proposed method of using a neural network to synthetically extend 3D cropped images improves deformable registration between images of differing scan extents. It creates a bespoke synthesis in the cropped region that takes cues from each image's anatomy. In most cases, it synthesizes realistic anatomies even far beyond the line of cropping. CropGAN creates details such as sinuses, lungs, orbits, and heart that continue smoothly from the available anatomical information. These large details help anchor the registration algorithm while it optimizes the correspondence within the real portions of the image. This advantage was seen even with extreme differences in scan extent (e.g., a cranium to carina scan and a scan only including the neck).

It has been observed that deforming a full image to a cropped image is more robust than the reverse. Therefore, implementing inverse consistency[33] in DIR could conceivably improve the registration of the former case. However, as shown in the study, CropGAN synthesized images still significantly outperforms the case of full-to-crop registration.

We noticed that the synthesized images were poorer when metal imaging artefacts were near the cropping boundary. This may be due to a lack of accurate anatomical cues near the boundary which the network can use to make its prediction. Interestingly, once further away from the artefact, the network can still create realistic anatomies, and the registration result is not significantly different from cases without such artifacts, as well as to the result using full images. Therefore, while the study was not designed to quantify the effects of artefacts, current results suggest our technique is robust to this effect.

We chose to use a b-spline based registration algorithm for our study, but this technique is generalizable for other algorithms. Our preliminary results suggest that CropGAN similarly improves demons based registration. Preprocessing with CropGAN could also aid in other medical imaging neural network tasks, since it can help to standardize the data.

One limitation of this technique is that it requires training data for each region one wishes to extend. We focused this study on CT images in the head and neck since this region can include large non-rigid motion. For other anatomical sites or imaging modalities, one would need a large dataset with similar scan extents to provide supervised training between an induced cropped image and its full image ground truth. For example, cone beam CT deformable registration may benefit from our method due to its limited field of view relative to the simulation CT target. However, the network would need to be trained and verified on this different modality. Improvement may also be seen by better selecting the trained VGG net used for deriving the deep features in the loss function. The network we used was trained for the unrelated task of classifying small 3D patches of CT and MR images by their scan site. The features learned from this network may not be optimal for our task, which used full image volumes. While having a VGG net trained on full images may lead to a better result, recent research has suggested that using deep features to assess image similarity can be surprisingly effective even when the network was trained for an unrelated task[34]. An additional limitation is the inherent uncertainty in the contours used for this study's analyses. We used a previously developed in-house segmentation network to increase reproducibility[13]. While this network demonstrated impressive dice scores, there is still an inherent amount of uncertainty. Despite the abovementioned limitations, we have provided a foundation upon which other studies can extend our work. The

code which was used for this manuscript can be found at https://github.com/emckenzi123/CropGAN .

Conclusion

Differences and inadequacy in scan extent is a difficult problem in medical image deformable registration. We proposed a solution using a neural network to synthesize the missing portions of the scan. These syntheses were able to successfully create realistic anatomy for the missing volume with details such as sinuses, orbits, skull, lungs, and heart. It was also robust to the amount of cropping in the inferior and superior directions. After filling the cropped volumes using CropGAN synthesis, the two images can then be deformably registered as though they had the same full scan length. Using 95% Hausdorff distance on a selection of head and neck contours, we found that our registration workflow was able to match contours equally well to a registration with complete scans. CropGAN performance for DIR as a function of cropped tissue is robust up to until 20cm of the superior end of the head was missing. By using CropGAN as a preprocessing step to deformable registration, we have provided an intuitive solution to the challenge of registration with different scan extents.

References

1. Kirby N, Chuang C, Ueda U, Pouliot J. The need for application-based adaptation of deformable image registration. *Med Phys*. 2012;40(1):011702. doi:10.1118/1.4769114
2. Geets X, Daisne JF, Tomsej M, Duprez T, Lonneux M, Grégoire V. Impact of the type of imaging modality on target volumes delineation and dose distribution in pharyngo-

laryngeal squamous cell carcinoma: comparison between pre- and per-treatment studies. *Radiother Oncol*. 2006;78(3):291-297. doi:10.1016/j.radonc.2006.01.006

3. Geets X, Tomsej M, Lee JA, et al. Adaptive biological image-guided IMRT with anatomic and functional imaging in pharyngo-laryngeal tumors: Impact on target volume delineation and dose distribution using helical tomotherapy. *Radiother Oncol*. 2007;85(1):105-115. doi:10.1016/j.radonc.2007.05.010
4. Veiga C, McClelland J, Moinuddin S, et al. Toward adaptive radiotherapy for head and neck patients: Feasibility study on using CT-to-CBCT deformable registration for "dose of the day" calculations. *Med Phys*. 2014;41(3):031703. doi:10.1118/1.4864240
5. McKenzie EM, Santhanam A, Ruan D, O'Connor D, Cao M, Sheng K. Multimodality image registration in the head-and-neck using a deep learning-derived synthetic CT as a bridge. *Med Phys*. 2020;47(3):1094-1104. doi:10.1002/mp.13976
6. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report. *Med Phys*. 2017;44(7):e43-e76. doi:10.1002/mp.12256
7. Zhen X, Yan H, Zhou L, Jia X, Jiang SB. Deformable image registration of CT and truncated cone-beam CT for adaptive radiation therapy. *Phys Med Biol*. 2013;58(22):7979-7993. doi:10.1088/0031-9155/58/22/7979
8. Ottosson W, Lykkegaard Andersen JA, Borrisova S, Mellemgaard A, Behrens CF. Deformable image registration for geometrical evaluation of DIBH radiotherapy treatment of lung cancer patients. *J Phys Conf Ser*. 2014;489:012077. doi:10.1088/1742-6596/489/1/012077

9. Periaswamy S, Farid H. Medical image registration with partial data. *Med Image Anal*. 2006;10(3):452-464. doi:10.1016/j.media.2005.03.006

10. Yang D, Goddu SM, Lu W, et al. Technical Note: Deformable image registration on partially matched images for radiotherapy applications. *Med Phys*. 2009;37(1):141-145. doi:10.1118/1.3267547

11. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

12. Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys*. 2018;45(10):4558-4567. doi:10.1002/mp.13147

13. Tong N, Gou S, Yang S, Cao M, Sheng K. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. *Med Phys*. 2019;46(6):2669-2682. doi:10.1002/mp.13553

14. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks. 2014:1-9. doi:10.1001/jamainternmed.2016.8245

15. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. May 2015. http://arxiv.org/abs/1505.04597.

16. Dinkla AM, Wolterink JM, Maspero M, et al. MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network. *Int J Radiat Oncol Biol Phys*. 2018. doi:10.1016/j.ijrobp.2018.05.058

17. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T. Free-Form Image Inpainting With Gated Convolution. 2020:4470-4479. doi:10.1109/iccv.2019.00457

18. Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. *arXiv*. 2018.

19. Hui Z, Li J, Wang X, Gao X. Image fine-grained inpainting. *arXiv*. 2020;(2):1-11.

20. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc*. September 2014:1-14. http://arxiv.org/abs/1409.1556.

21. Avants B, Greenblatt E, Hesterman J, Tustison N. *Deep Volumetric Feature Encoding for Biomedical Images*. Vol 12120 LNCS. Springer International Publishing; 2020. doi:10.1007/978-3-030-50120-4_9

22. Teterwak P, Sarna A, Krishnan D, et al. Boundless: Generative Adversarial Networks for Image Extension. 2019. http://arxiv.org/abs/1908.07007.

23. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix : A Toolbox for Intensity-Based Medical Image Registration. 2010;29(1):196-205.

24. Shamonin D, Bron E, Lelieveldt B, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinform*. 2014;7(January):1-15. doi:10.3389/fninf.2013.00050

25. Li X, Zhang Y, Shi Y, et al. Comprehensive evaluation of ten deformable image registration algorithms for contour propagation between CT and cone-beam CT images in adaptive head & neck radiotherapy. Zhang Q, ed. *PLoS One*. 2017;12(4):e0175906. doi:10.1371/journal.pone.0175906

26. Huttenlocher DP, Rucklidge WJ, Klanderman GA. Comparing images using the Hausdorff distance under translation. In: *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 1992-June. IEEE Comput.

Soc. Press; 1992:654-656. doi:10.1109/CVPR.1992.223209

27. Raudaschl PF, Zaffino P, Sharp GC, et al. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Med Phys*. 2017;44(5):2020-2036. doi:10.1002/mp.12197

28. Armanious K, Mecky Y, Gatidis S, Yang B. Adversarial Inpainting of Medical Image Modalities. *ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc*. 2019;2019-May:3267-3271. doi:10.1109/ICASSP.2019.8682677

29. Wei D, Ahmad S, Huo J, et al. Synthesis and Inpainting-Based MR-CT Registration for Image-Guided Thermal Ablation of Liver Tumors. In: Shen D, Liu T, Peters TM, et al., eds. *Medical Image Computing and Computer Assisted Intervention -- MICCAI 2019*. Cham: Springer International Publishing; 2019:512-520.

30. Zhang S, Wang L, Zhang J, et al. Consecutive Context Perceive Generative Adversarial Networks for Serial Sections Inpainting. *IEEE Access*. 2020;8:190417-190430. doi:10.1109/access.2020.3031973

31. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS. Generative Image Inpainting with Contextual Attention. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2018:5505-5514. doi:10.1109/CVPR.2018.00577

32. Ntavelis E, Romero A, Bigdeli S, Timofte R. AIM 2020 Challenge on Image Extreme Inpainting. October 2020. http://arxiv.org/abs/2010.01110.

33. Yang D, Li H, Low DA, Deasy JO, Naqa I El. A fast inverse consistent deformable image registration method based on symmetric optical flow computation. *Phys Med Biol*. 2008;53(21):6143-6165. doi:10.1088/0031-9155/53/21/017

34. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of

Deep Features as a Perceptual Metric. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2018;(1):586-595. doi:10.1109/CVPR.2018.00068

CHAPTER 4: Predictive Head and Neck Registration Using Self-Attention with Positional Encoding

Introduction

Deformable image registration (DIR) is an unsolved problem in medical imaging. Despite its importance in diagnoses and treatment, there are still many opportunities for improvement, especially in the presence of large motion. The image processing capabilities of neural networks have shown great promise for medical imaging registration[1–9]. Most of these network-based approaches use learned convolution filters, which is an inherently local approach. This means that any distant information must pass through several layers of the network[10]. In the application of DIR, this can limit the long-range propagation of information in determining the appropriate deformation. To overcome this challenge, we propose the use of self-attention networks. Self-attention determines the response at a point in the image as a learned weighted sum of all the points in the image[11]. For example, cues in the cervical spine position could inform the network about deformation at a point in the shoulders. A self-attention network coordinates information from the entire image to determine its prediction. This study leverages the strengths of self-attention networks to predict fully 3D deformations of medical images.

Image registration is of particular importance in the medical field of radiation oncology. Spatial correspondences between images need to be established to track disease progression, combine complimentary imaging modalities to better visualize anatomical structures, and to setup patients in reproducible positions[12]. Additionally, due to the fractionated nature of radiation oncology treatments, the patient's anatomy may change over the course of treatment, necessitating an

adaptive approach. Adaptive radiation therapy uses deformable registration to propagate contours and accumulate dose across a potentially changing anatomy. All these medical applications require a high degree of accuracy to safely guide the patient's treatment. Consensus guidelines state that registered contours should agree within 2-3 mm[13]. However, recent investigations have shown that our current commercial DIR algorithms fall short of this goal, especially in the presence of large deformation[14].

Another application of DIR is the alignment of images from different patients, or inter-patient registration. This is used in important applications such as combining data to study patterns of recurrence and metastases, where one needs to register many patients to a common template[15–17]. An additional application of inter-patient registration is atlas-based segmentation[18,19]. The contours from an atlas must be deformed to the target patient's anatomy to be correctly mapped. Both intra-patient (registration of images from the same patient) and inter-patient registration play important roles in the treatment and study of radiation therapy.

Current clinical DIR uses iterative techniques, requiring several optimization steps before an acceptable solution is obtained[20]. The full optimization can take minutes. Additionally, these traditional methods rely on hand crafted metrics and regularizations to drive the optimization. Machine learning research has shown impressive results in image processing tasks[21,22]. These capabilities have recently been applied to image registration[9]. A moving and target image are input into a network, and the network is trained to predict a deformation vector field (DVF) to deform the moving image to match the target[4]. To the best of our knowledge, all current neural network medical DIR approaches use convolutional neural networks (CNN's) in their backbone[23]. This is the most popular neural network architecture for image processing. It relies on non-linearly combining learned abstract filters to represent complex functions. This is a

powerful technique, but an inherently local one. Attention networks were originally developed for natural language processing[10], but recent advances in image segmentation have shown that self-attention networks model long range dependencies in images well[11]. This improves segmentation because the network can look for context throughout the image to determine the boundaries, shapes, and occlusions of an object. Recently, a model was developed which use attention to augment CNN-based DIR[24]. The input was first passed through several convolution layers, then passed through a self-attention stage before continuing as a fully convolutional network. As of writing, the only purely self-attention-based DIR model uses the moving image as input to an encoder and the target image to a decoder, in a fashion similar to transformer language translation models[25]. This pioneering work was only able to predict on small 2D MNIST images due to the computational complexity of self-attention. We propose leveraging computationally efficient axial attention to build the first fully self-attentional DIR for 3D images. We predict that strength of long-distance data connections will prove advantageous in DIR.

Unlike CNN's, attention networks do not learn a static filter. Instead, they aggregate information from the entire input and produce weights that dynamically change with the input. These weights are determined from learned keys, queries, and values. More specifically, given an input image $x \in \mathbb{R}^{h \times w \times d_{in}}$, with height h, width w, and d_{in} channels, the output y_o at position o is computed by:

$$y_o = \sum_{p \in N} softmax_p\left(q_o^T k_p\right) v_p$$

Where the queries $q_o = W_Q x_o$, keys $k_o = W_K x_o$, and values $v_p = W_V x_o$ are functions of learnable weights applied to the input, and N is the entire image lattice. This formulation, while powerful is computationally expensive. It can become prohibitively so with larger images,

especially 3D image volumes ($\mathcal{O}(h^2w^2d^2)$). This has forced previous research with large images to apply self-attention only on down sampled feature maps or smaller images[24]. All previous applications of attention to medical DIR have used it only as an augmentation to a CNN backbone at deeper layers[24]. A recent study overcame the issue of computation cost by introducing axial attention[26]. Axial attention breaks down an image into its 1D components (e.g., rows and columns) and computes the attention along each. By sequentially computing and combining 1D components, a given point in the output can be sparsely connected to the entire input. This reduces the complexity to $\mathcal{O}(hwdm)$ for 3D images with a feature map span m. Wang et al took axial attention and added relative position sensitive attention[11]. This allows for the keys, queries, and values to learn context-dependent relative position encoding to further guide the attention. In our present work, we build upon this position sensitive axial self-attention model and extend it from a 2D segmentation model to a 3D model which can predict deformation vector fields given a pair of images. By leveraging self-attention's ability to model long-range dependencies, we demonstrate a DIR network that densely incorporates cues from the entire 3D volume.

Long range dependencies can be especially important in the context of large motion. The setting of DIR in the head and neck is an example where large motion exists (e.g. neck flexion), and could benefit from our self-attention approach. We train a network to deformably register both inter- and intra- patient pairs of head and neck CT's as a proof of concept for a fully attentional DIR model, and compare our results to both CNN-based DIR models, as well as the more traditional B-spline registration approach.

Methods

Dataset

Head and Neck CT images were acquired from The Cancer Imaging Archive (TCIA) dataset[27]. We had a total of 409 training, 53 validation, and 44 inter-patent testing images. Additionally, 14 intra-patient testing images were collected from our institution. These consisted of 7 CT scans for radiotherapy simulation and 7 patient-matched CT scans for PET attenuation correction. Scan extent went from the top of skull to approximately the carina. Scanning beds and immobilization equipment were masked out of the images. For input into the network, all images were rigidly registered to a template image and downsized to 128x128x128 with 4mm isotropic voxels. Image intensity value were clipped to a range of [-1024, 3000], then normalized to [-1, +1]. For segmentation matching in the loss function, volumes were automatically segmented using a neural network approach[28,29]. This resulted in 17 contours per patient.

Network Design

The network was constructed to take in a pair of 3D images (source and target) and output a predicted DVF, along with the deformed source volume. Each input volume underwent two layers of convolution prior to concatenation and max pooling, resulting in a tensor with 64 features and a size of 32x32x32 voxels. The overall architecture was U-shaped, with an encoding and decoding side. The sides of the U were connected via skip connections with concatenation and convolutions of kernel size 1. The number of features going down the encoder branch were 256, 512, 1024, 2048 for each level respectively. The decoder branch reversed these feature numbers. Each level of the U-net was constructed using a self-attention block with positional encoding. Each self-attention block was constructed using a residual approach, where the block was divided into an attention branch and an identity branch which were added together and followed by a non-linearity. Each attention branch consisted of a 1x1x1

kernel down-sampling 3D convolution, axial attention sequentially calculated in the sagittal, coronal, and axial anatomical planes, then an up-sampling 1x1x1 kernel 3D convolution.

The axial attention was accomplished by reshaping the input into a 1D tensor along the selected dimension. For example, for a given tensor with shape [*batch size, features, height, width, depth*], the 1D transformed axial *depth* tensor would have a shape [*batch size* x *height* x *width*, *features*, *depth*]. The analogous operations were performed for the *height* and *width*. This enables us to factorize the 3D space into sequential operations along the height, width, and depth of the imaging volume. This computationally efficient approach allows us to have a global receptive field for attention.

After transformation to 1D, learnable weights are multiplied by the input, and yield learned keys, queries, values, and positional encoding. The keys, queries, values, and position encoding are thus linear projections of the input. To compute the attention block output at a given location, that output location's query tensor is multiplied by the key tensors for every point in the input. This product gives a high value when a query is more parallel with a given key, which signals that the input at the key's location is important to the output location, and thus should be "paid attention to." Additionally, importance of an input point is further weighted by its learned positional encoding. That is to say, not only does the value at a given point determine its importance, but also its relative position. Position encoding is learned for each of the keys, queries, and values. Therefore, not only is the relative positional importance of the different input points learned, but the input point's key can additionally indicate important locations. With a large receptive field, the output at a given output location can be unaware of the region from which an input's value tensor came. Therefore, the value is also given a learned positional encoding. Formally:

$$y_o = \sum_{p \in \mathcal{N}_{1xm}(o)} softmax_p\left(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k\right)\left(v_p + r_{p-o}^v\right)$$

where m is the span of the entire input at that layer, and r_{p-o}^q , r_{p-o}^k, and r_{p-o}^v are the relative learned positional encodings for the queries, keys, and values, respectively. Note that the subscript is $p - o$ for the positional encodings, indicating that these positions are relative relationships between the input and output locations. The final result is a map of attention that looks to the entire input for guidance, incorporating information about all of the input values and their positions. For each attention block we linearly project the input into keys, queries, values, and positional encoding 8 times and compute the attention in parallel along each projection. This multi-head attention[10] formulism allows our network to jointly attend to multiple representations of the inputs and learn to perform different tasks. The above formulation was first proposed by Wang et al[11], and we have adapted it for 3D inputs and outputs in DIR.

After proceeding through the U-net, the output is up-sampled back to size 128x128x128, with 128 features. A convolution with kernel size 1 reduces the channels to 3 to yield a final deformation vector field, where a vector of length 3 is defined at every 3D point. This DVF can then be applied to the input moving volume to give a final deformed image to match the target. We use the spatial transformer from Balakrishnan et al[4] to efficiently and differentiably deform images.

Our network, named ***Amorwarp***, is shown in Figure 1. After convolutions we used instance normalization and a leaky relu activation. We used an optimizer with stochastic gradient descent, 10 warm-up epochs[30], and an intial learning rate of 0.0257 with cosine learning rate schedule.

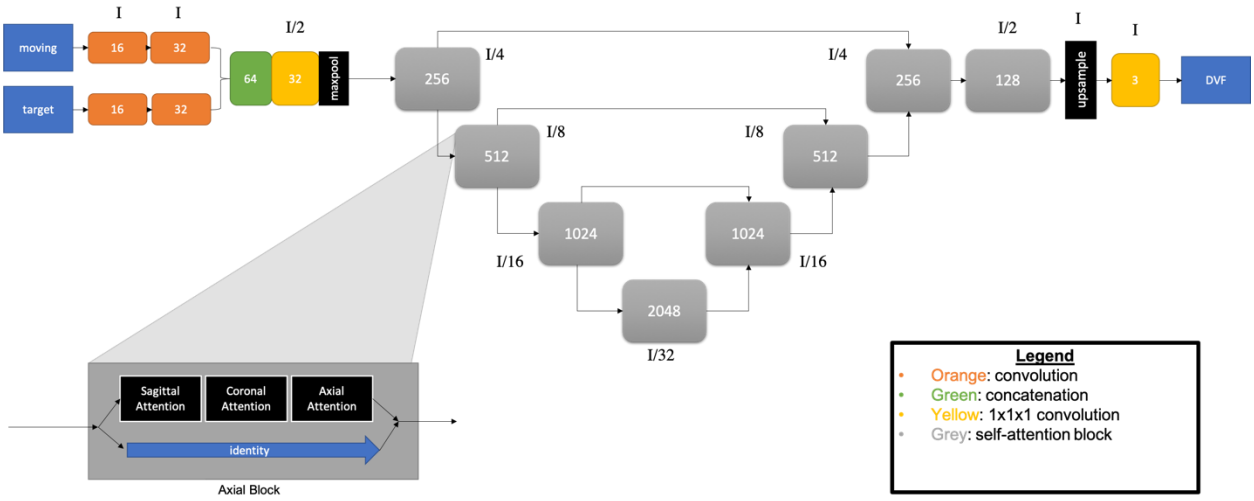


Figure 1 Network Design of Proposed Model (Amorwarp). The moving and target image each undergo two layers of convolutional encoding prior to concatenation and input into the U-net. The U-net is built on a self-attention backbone, with each attention block consisting of a residual design and axial attention applied sequentially in the sagittal, coronal, and axial planes. The U-net possesses skip-connections, and the final DVF prediction is obtained through an upsampling with trilinear interpolation and a final 1x1x1 convolution to transform the features into a 3-term vector (x,y,z). In this figure, the number of features is numbered inside the boxes, while the scale relative to the input size is shown outside each box.

Our loss function consisted of a mean square error fidelity term between the deformed and target image, a bending energy term on the DVF to promote smooth deformations, an average dice loss term between the deformed and target contours to encourage anatomical alignment, and a deep feature VGG loss term to improve detail matching between deformed and target images. The deep feature loss term was constructed to weight more dissimilar portions of the image more[31,32].

Analysis

The novelty of our proposed method lies in its self-attention backbone. Therefore, to evaluate our network's performance, we trained a convolution-based U-net to accept the same moving and target image pairs and predict a DVF[4]. This network is labeled "Voxelmorph" in our results. Since our network was trained using a Dice loss for contour matching, we also trained a U-net

with the same Dice loss for comparison. This network is labeled "Voxelmorph (Dice)" in our results. We additionally used an established non-neural network approach based on B-splines for comparison[33,34]. These B-spline based results are labeled "Elastix". The registrations were compared using 95% Hausdorff distance on a set of 17 pre-calculated automatic head and neck contours. The overall mean contour match was compared across methods, as well as a function of pre-registration contour match, which was used as a surrogate for registration difficulty. These analyses were computed for both inter-patient and intra-patient registrations. In addition to contours, we also manually selected landmarks for the intra-patient dataset. For each registration method, the target registration error (TRE) was calculated between the deformed and target landmarks. Qualitatively, we also investigated deformed moving image visualizations as well as deformation vector field visualizations.

Results

The network was trained using a Nvidia Quadro RTX 8000. The moving and target images were randomly sampled from the 409-volume training dataset for a total of 167,281 possible pairs. 409 samples were run per epoch, for a total of 280 epochs. Once trained, inference took 0.6 seconds per registration. During evaluation, the network was tested on 22 pairs of inter-patient registrations, and an additional 7 pairs of intra-patient registration.

Qualitatively, the learned model was able to predict DVF's that led to well-matched deformed moving images (see Figure 2). Difficult deformations, such as closing and opening the mouth were achieved without gross distortion of the surrounding anatomy. Upon inspection of the DVF, we see that the network was able to learn discontinuities around the motion of connected components, demonstrating the model's ability to learn to associate disparate parts of the image.

For comparison to an iteration-based traditional registration approach, we also registered the test set using Elastix. An example deformation using the same volumes is shown in Figure 3. Of note, the algorithm struggled with the mouth. This is due to the inherent regularization in B-spline registrations, which ensures smoothness across the DVF. This smoothness is readily seen in the DVF overlay of Figure 3. For a comparison with another neural network approach, Figure 4 shows the same registration pair but with a convolutional neural network model[2]. This approach also shows a smoother DVF, yet the deformed image matches the target much better than with Elastix. While the deformed image using a CNN approach is very similar to our proposed approach, we notice lingering issues with the buccal cavity, as well as the fact that the CNN network compressed tissue superiorly to match the inferior image's edge.

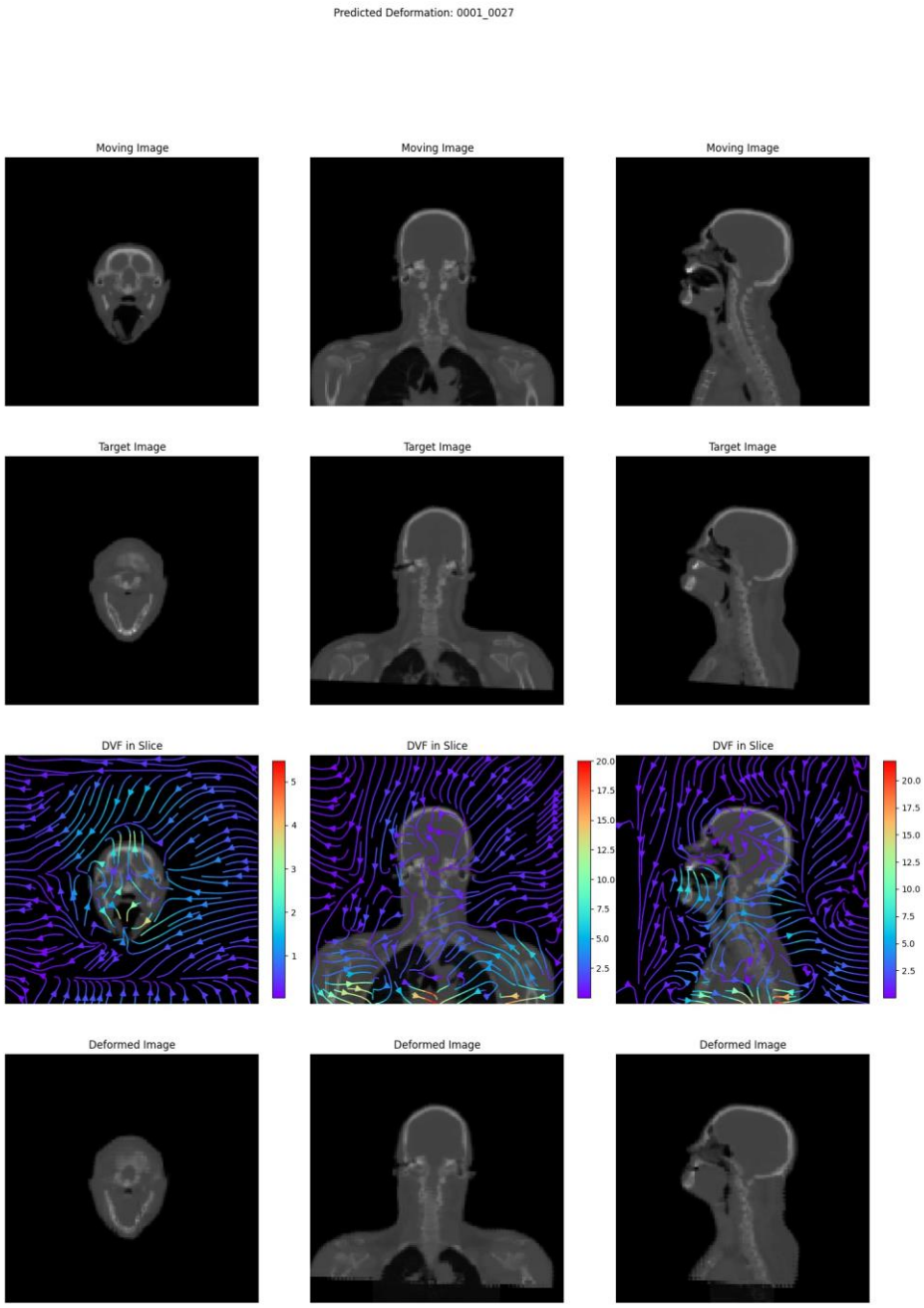


Figure 2 Example Deformation using AmorWarp. The top row is the moving image, the second row is the target image. The third row shows the predicted DVF overlaid on the moving image. The fourth row shows the resulting deformation.

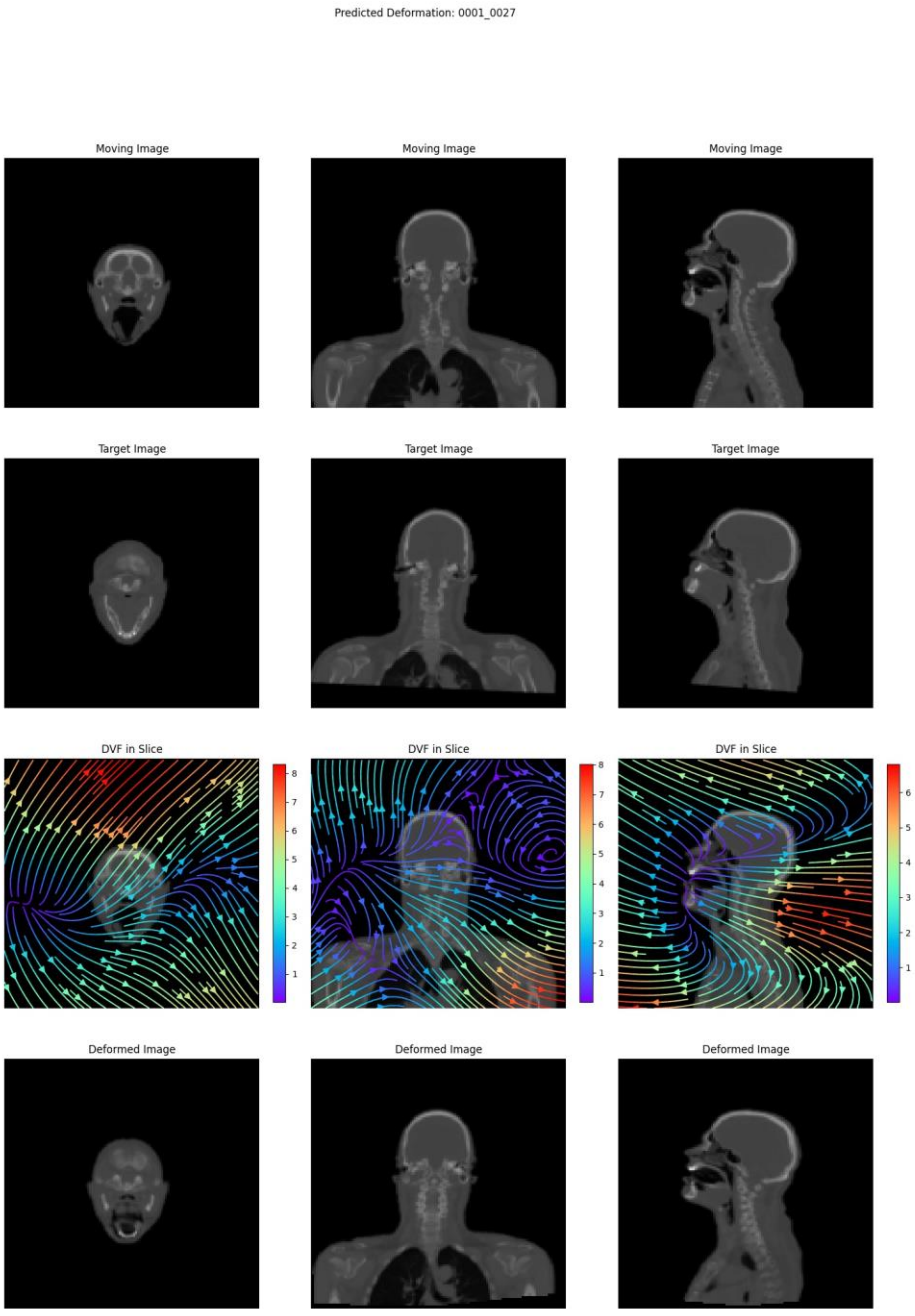


Figure 3 Example Deformation using Elastix. The moving and target image are the same as in Figure 2. Here the B-spline approach struggles with the open and closed mouth. The DVF (third row) is notably smoother than in a network-based approach.

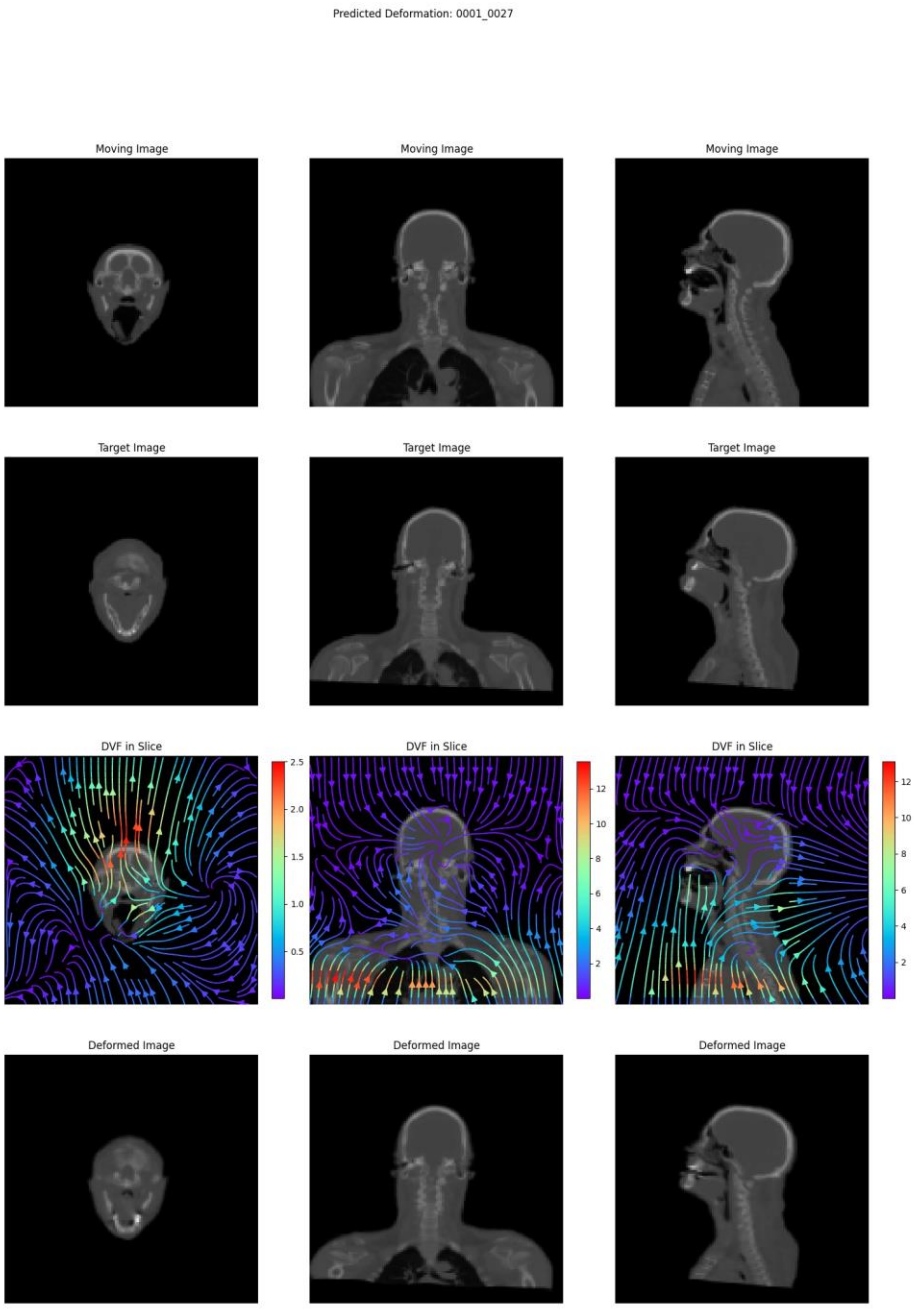


Figure 4 Example Deformation using Voxelmorph. The moving and target image are the same as in Figure 2. This CNN based registration shows a smoother DVF (third row), yet there is close agreement between deformed (fourth row) and target (second row) images. Relative to our approach, the mouth displays some residual mismatch, and the inferior portion of the image has been compressed up into the body.

Intra-patient registrations are visualized similarly in Figure 5, Figure 6, and Figure 7, for our proposed method, Voxelmorph, and Elastix, respectively. Qualitatively, the Elastix registration appears to have maintained many of the features of the pre-registered source image instead of deforming to the target. Our method and Voxelmorph show similar registration results, however our method again ejects cropped voxels out of the image, while Voxelmorph attempts to squish the inferior voxels superiorly to match the scan edge.

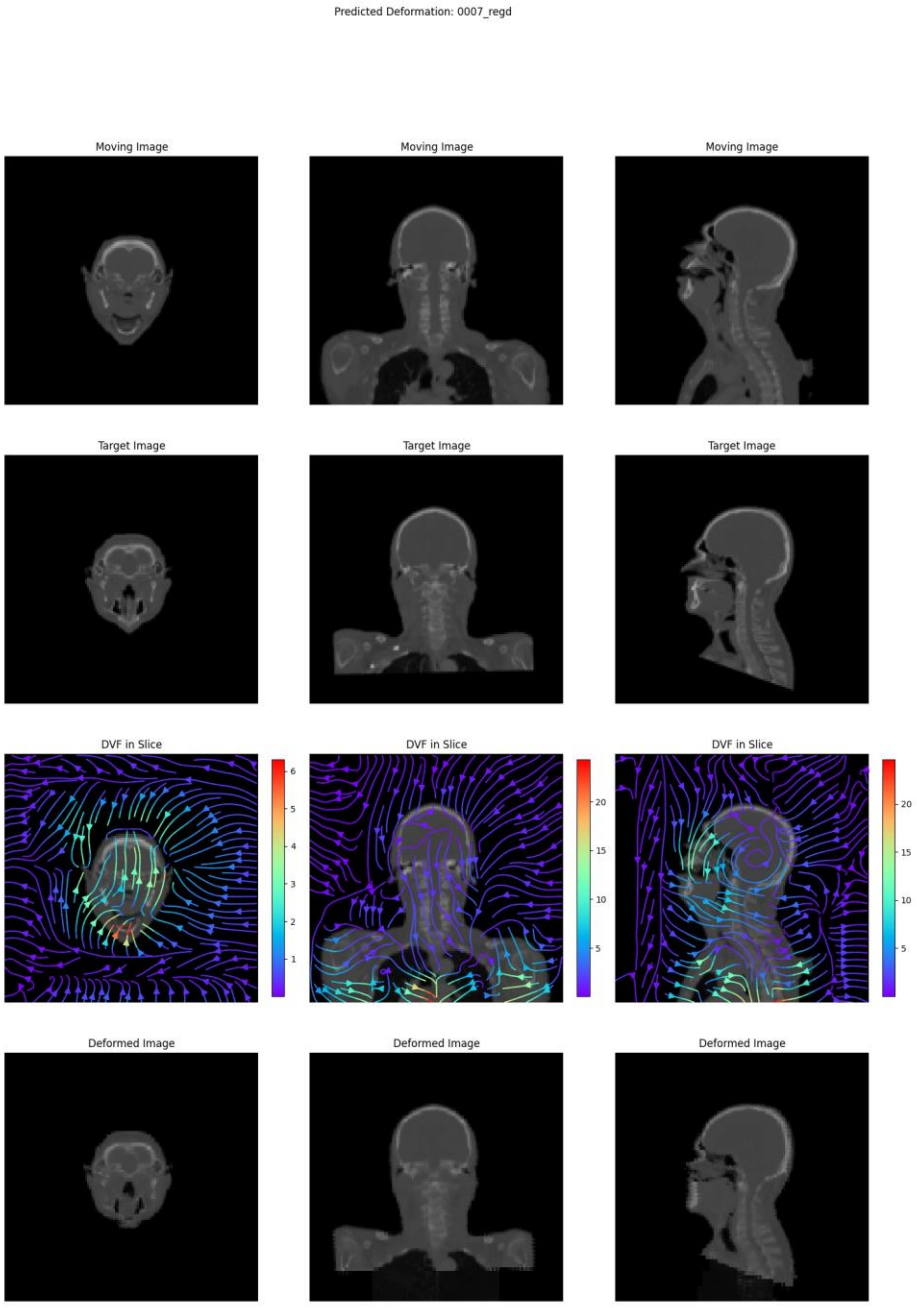


Figure 5 Example Intra-patient Deformation using AmorWarp. The top row is the moving image, the second row is the target image. The third row shows the predicted DVF overlaid on the moving image. The fourth row shows the resulting deformation.

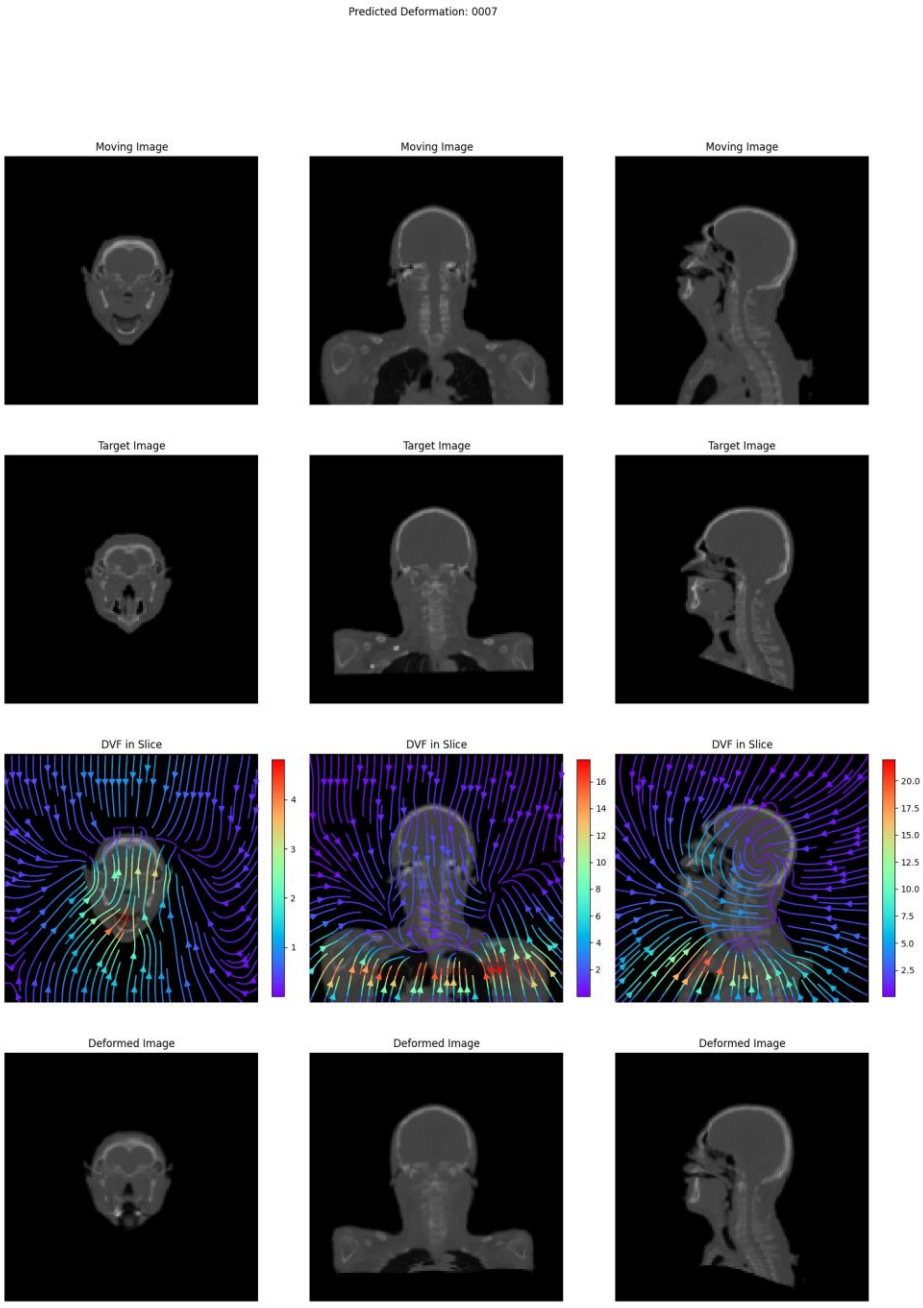


Figure 6 Example Intra-patient Deformation using Voxelmorph. The moving and target image are the same as in Figure 5. This CNN based registration again shows a smoother DVF (third row), yet there is close agreement between deformed (fourth row) and target (second row) images. Relative to our approach, the nose displays unnatural distortion, and the inferior portion of the image has again been compressed up into the body.

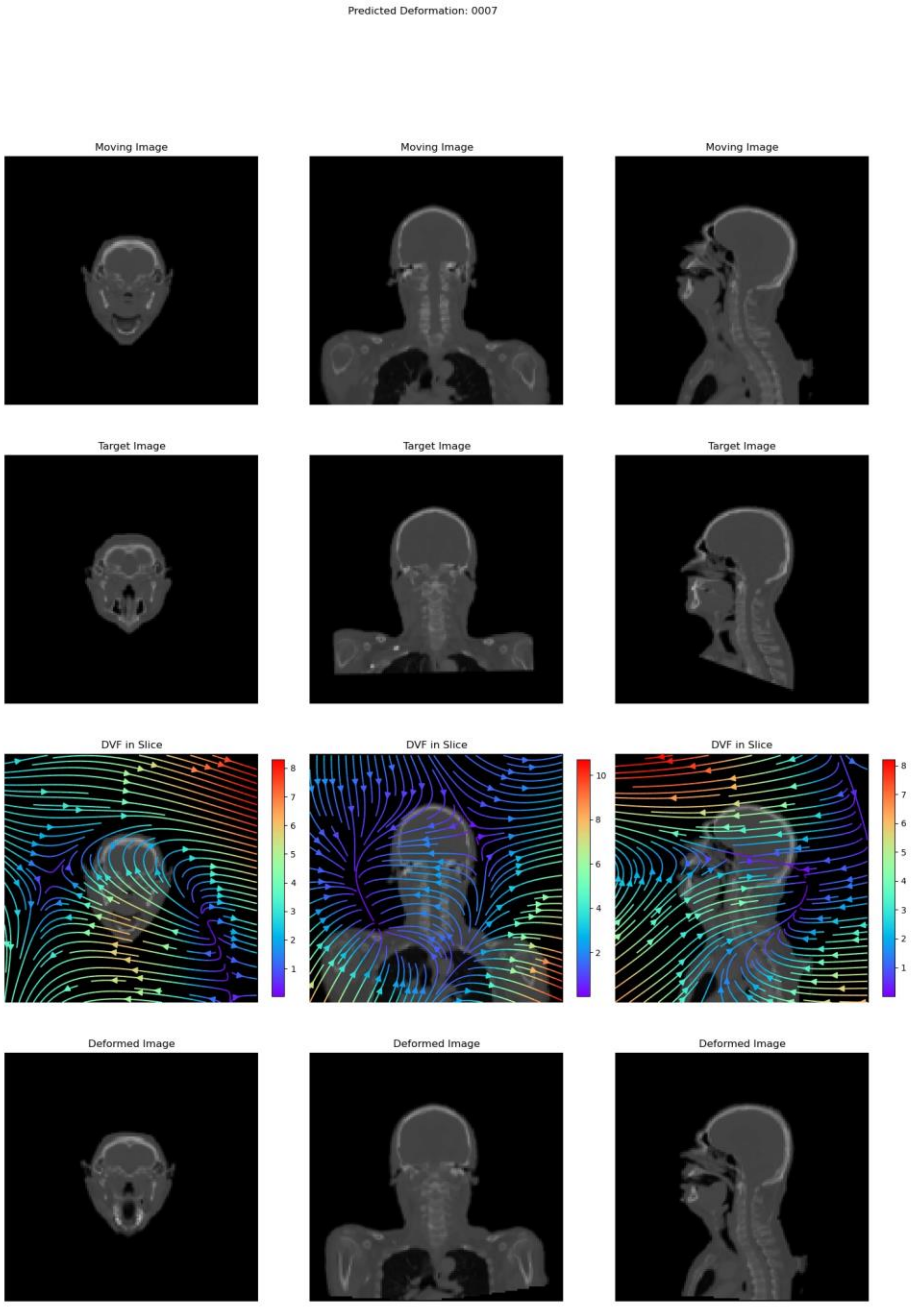


Figure 7 Example Intra-patient Deformation using Elastix. The moving and target image are the same as in Figure 5. Here the B-spline approach gives a final image (fourth row) that is more similar to the pre-moved image than the target. The DVF (third row) is again notably smoother than in a network-based approach.

Figure 8 takes the same example registration displayed in figures 5, 6, and 7 and shows the final deformed locations of a selection of landmarks. The top left of the figure shows the landmarks

placed on the source image. The other panels show the landmarks deformed with different DIR algorithms and overlaid on the same target image. One can readily see that the tip of the nose (blue circle) is correctly placed with our method (top right panel), and yet is incorrectly placed for the other methods. Also, a location within the cavity of the mouth (green circle) is correctly kept within the mouth with our method. Voxelmorph pushed the buccal cavity point superiorly into the maxilla, while Elastix pushed it superiorly into the nasal sinuses. The upper teeth (yellow circle) were close to the target location for our method, while Voxelmorph and Elastix pushed them into the nose. Additionally, the inferior points (teal circle, yellow x) were correctly placed for our method and Elastix, yet Voxelmorph pushed them superiorly as it struggled to press the inferior tissue cranially to handle the shorter scan length.

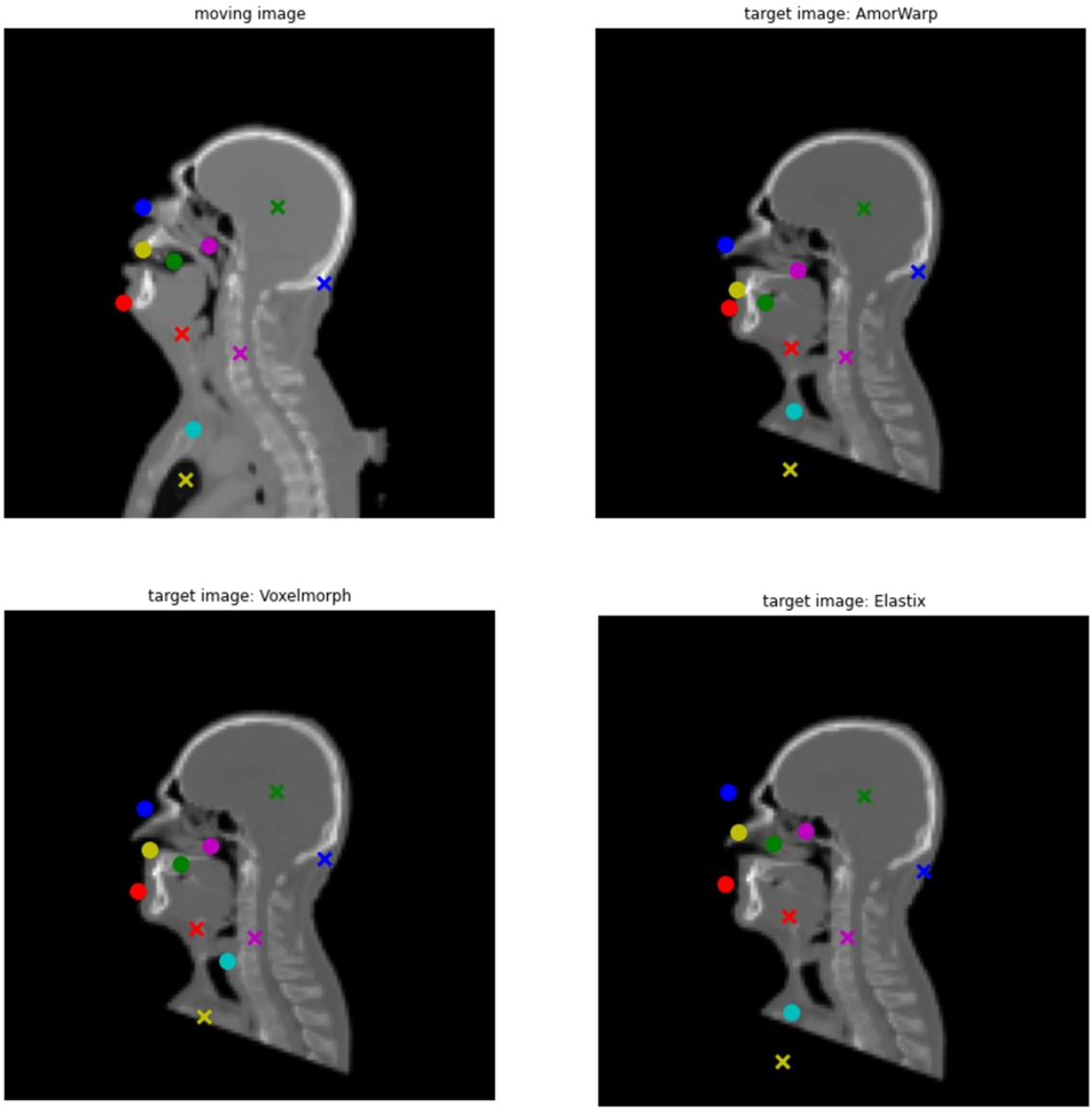


Figure 8 Intra-patient Landmark Registration Example. The top left panel shows a selection of anatomical points on the source image. The other panels show the target image with the same points deformed according to Amorwarp (top right), Voxelmorph (bottom left), and Elastix (bottom right). The nose, teeth, and buccal cavity show large misplacement in Voxelmorph and Elastix, while our method shows close alignment.

Quantitatively, we compared the results using automatically segmented contours. The moving volumes' contours were transformed using the same predicted DVF and nearest neighbor interpolation. The 95% Hausdorff distance was computed between the moving and target contours both before and after registration. The pre-registration results serve as a surrogate for difficulty of the registration.

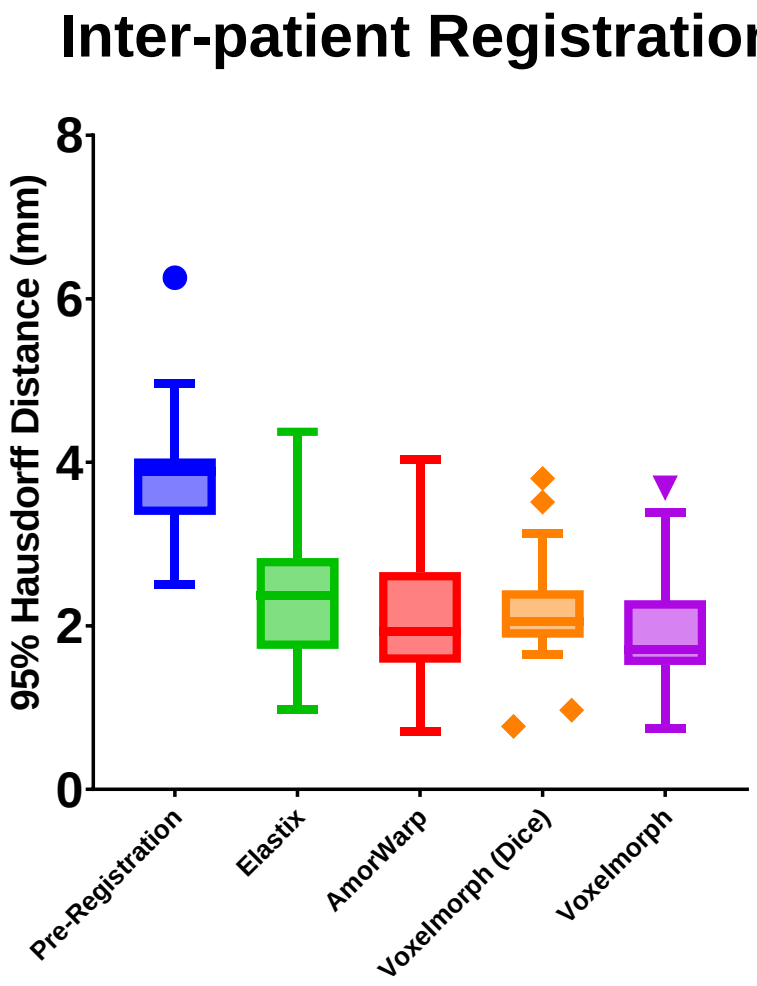


Figure 9 Average Contour Matching Error for Inter-patient Registration. All post-registration methods significantly improved contour matching. Our proposed method (AmorWarp) was not significantly different from the other DIR methods.

Figure 9 displays the overall average of the 17 contours and 22 inter-patient registration pairs. A two-way ANOVA with Tukey multiple comparison analysis showed that while all DIR algorithms were significantly different from the pre-registration results, among the different DIR methods only Elastix and Voxelmorph were significantly different ($p<0.0001$). We further analyzed the results using intra-patient registration pairs (Figure 10). The algorithm with the most DVF regularization (Elastix) gave the lowest error. This was followed by our proposed

model. Intra-patient registrations require less radical deformations than between different patients, and the organ sizes should stay relatively the same. This shows that our method was able to maintain consistency in organs when the patient moved to different positions. When we look at contour error as a function of pre-registration error, this relationship becomes more salient (Figure 11). Our method performs better relative to a CNN registration approach for large intra-patient deformations. However, when comparing the slopes from a simple linear regression using a two-tailed t-test, we did not find a significant difference amongst all algorithms investigated ($p=0.27$). Similarly, we analyzed the post-registration error as a function of pre-registration error for the inter-patient test registrations (Figure 12). Here the trend of our method follows closely to the CNN approach, however the slopes amongst all methods were again not significantly different ($p=0.30$).

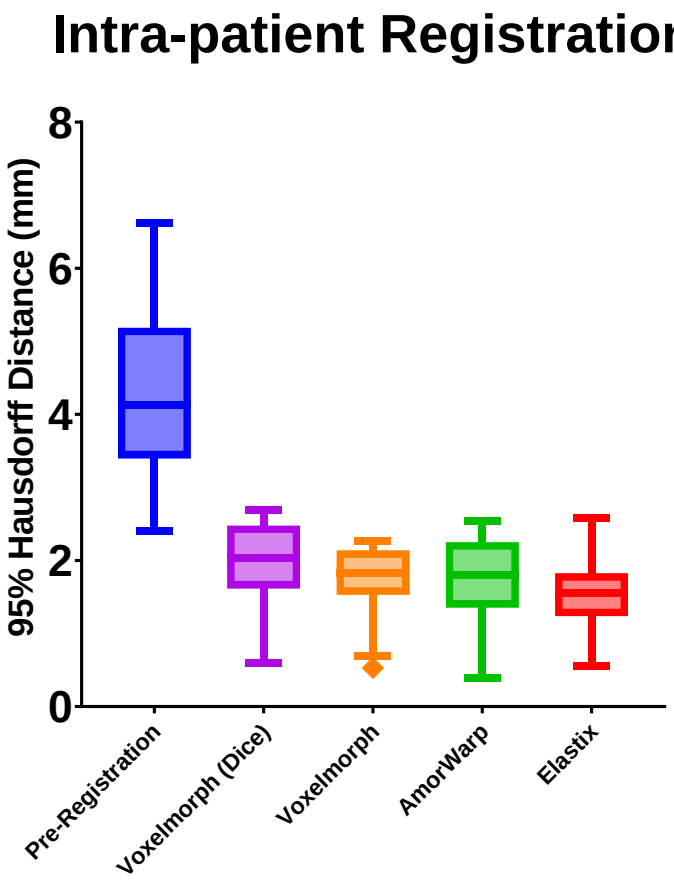


Figure 10 Average Contour Matching Error for Intra-patient Registration. All post-registration methods significantly improved contour matching. With intra-patient registration, more regularized deformation benefits the registration.

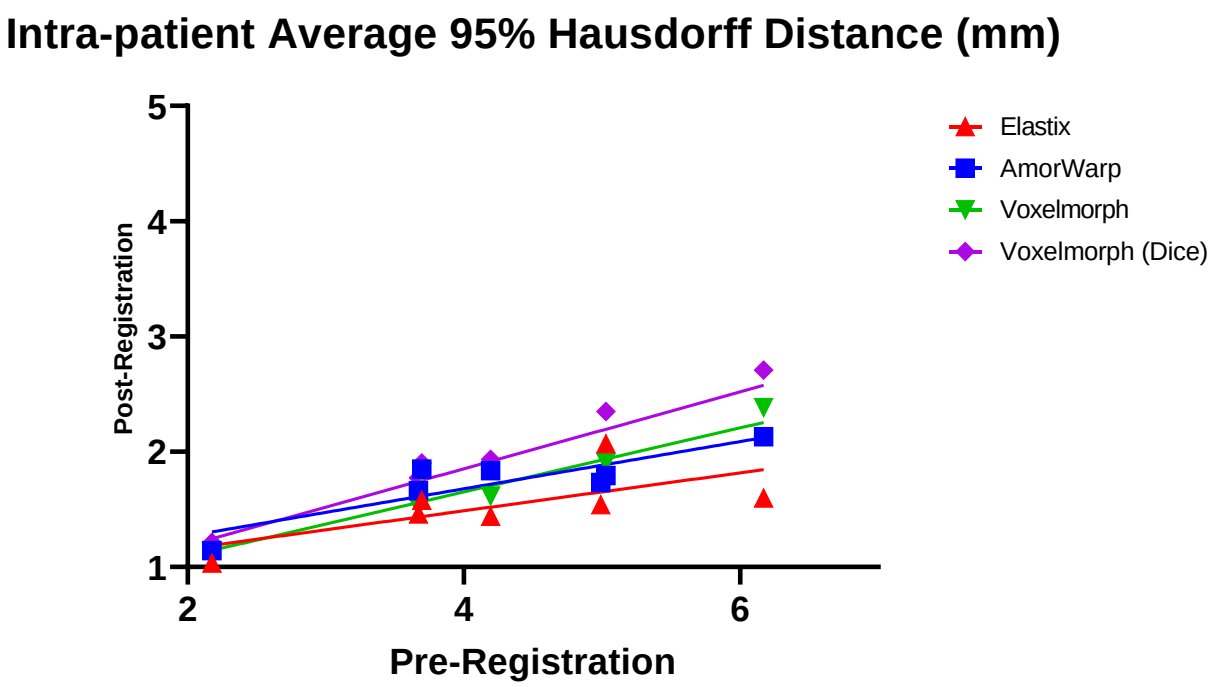


Figure 11 Intra-patient Contour Error as a Function of Registration Difficulty. The 95% Hausdorff distance is averaged across all contours in this plot. The pre-registration 95% Hausdorff distance serves as a surrogate for registration difficulty. Elastix performs the best at the highest registration difficulty, followed by our proposed method.

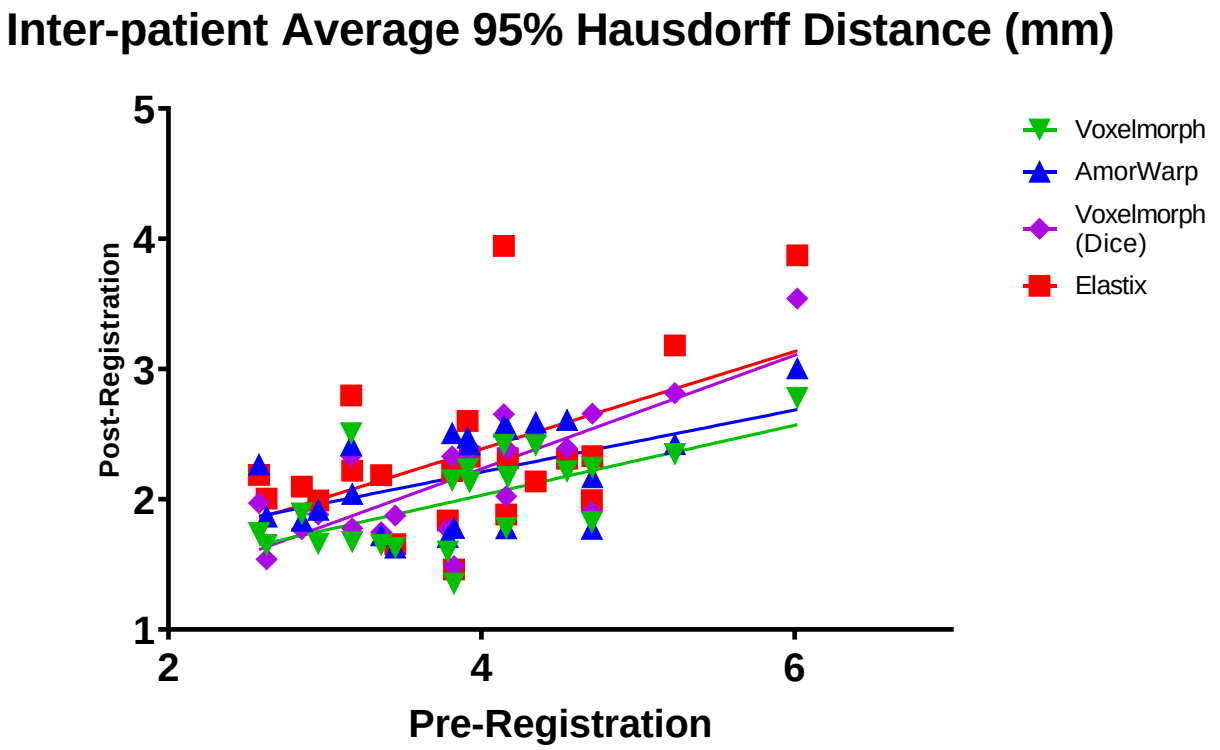


Figure 12 Inter-patient Contour Error as a Function of Registration Difficulty. The pre-registration 95% Hausdorff distance serves as a surrogate for registration difficulty. The neural network methods perform similarly at low difficulty. At higher difficulty, AmorWarp and Voxelmorph show the lowest error, with Elastix and Voxelmorph trained with Dice showing the highest.

Another interesting result is that Voxelmorph with added Dice loss performed worse than Voxelmorph trained without contour guidance. In our experiments in developing AmorWarp, we found that the addition of contour guidance added only a modest quantitative improvement, yet visually, regions such as the mandible underwent less unrealistic distortion.

In addition to contour metrics, we also analyzed the Target Registration Error (TRE) of a selection of well-visualized landmarks, notably the tip of the nose, center of top teeth, mental protuberance (chin), and epiglottis. We discovered that our proposed model performs the best on average, though it was not significantly different from Voxelmorph ($p=0.34$).

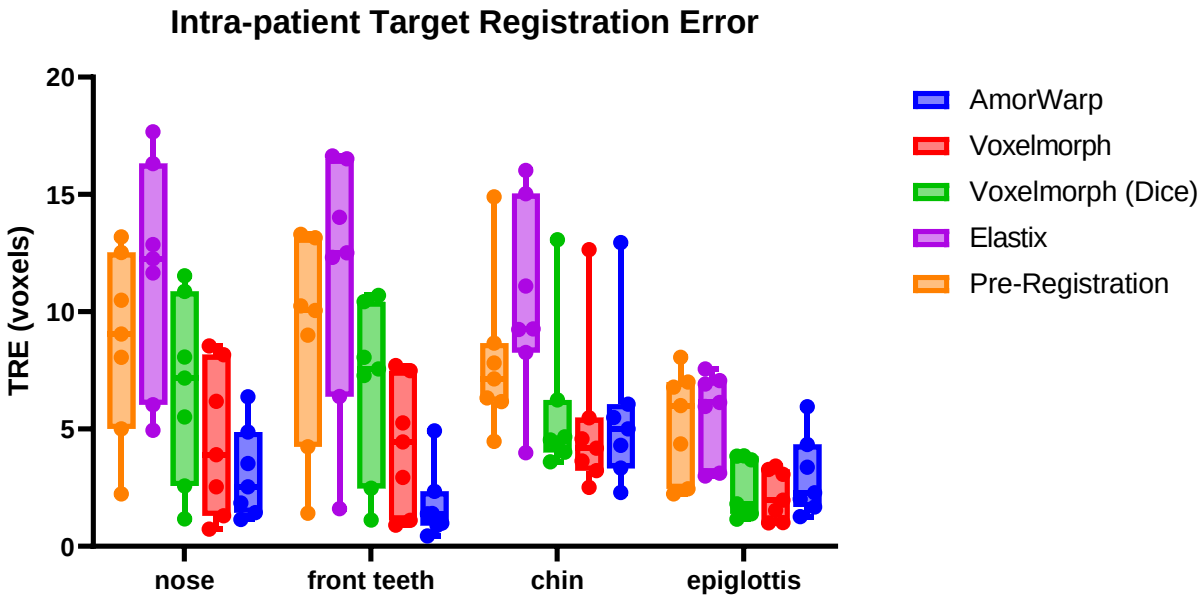


Figure 13 Intra-patient Target Registration Error for a selection of landmarks. Our proposed method (Amorwarp, Blue) demonstrates competitively lower error than the other methods in the nose and front upper teeth. Our method shows similar results to Voxelmorph in the chin and epiglottis. Elastix shows higher error than pre-registration TRE.

Table 1 Average Intra-patient Target Registration Error averaged across landmarks

Registration	Mean	Std. Deviation
AmorWarp	3.4	1.6
Voxelmorph	4.0	1.3
Voxelmorph (Dice)	5.4	2.0
Elastix	9.8	2.8
Pre-Registration	7.7	1.6

When plotting the TRE as a function of pre-registration TRE, our proposed method showed the lowest error, especially for more difficult registrations (Figure 14). This matches what was observed qualitatively in Figure 8.

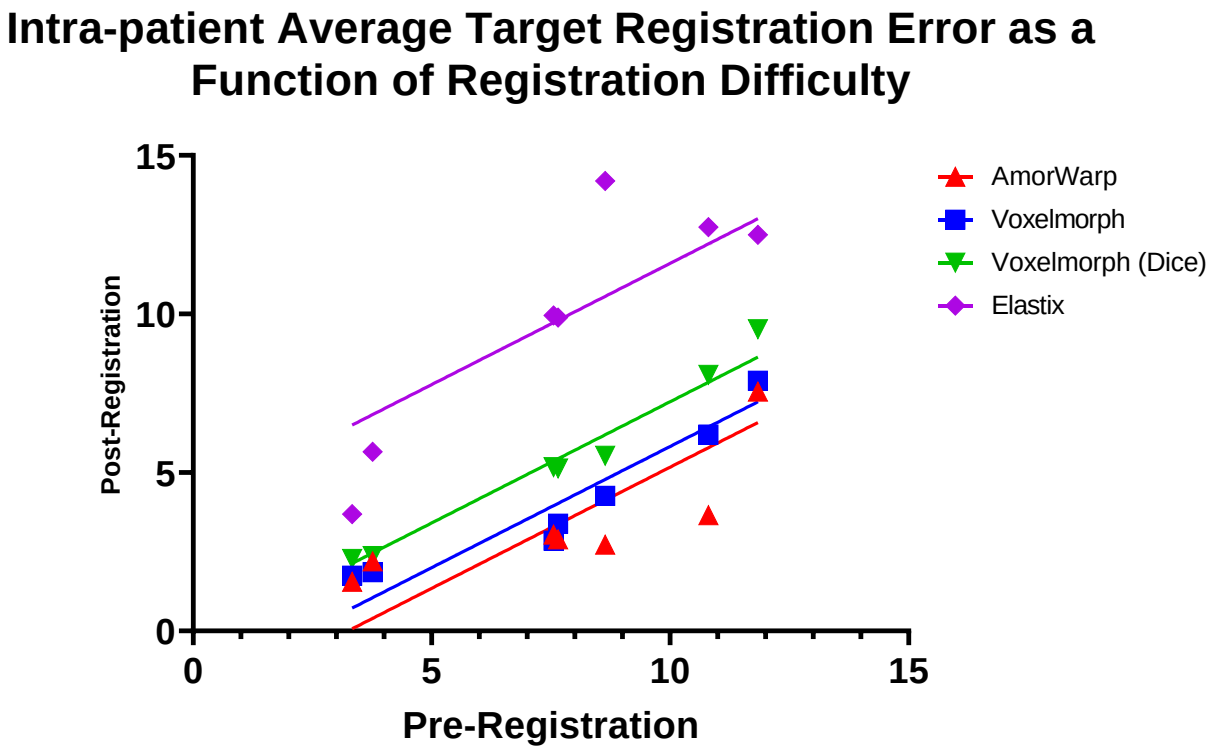


Figure 14 TRE as a Function of Registration Difficulty. Our proposed method (AmorWarp) displays the strongest resilience to registration difficulty.

Discussion

Our novel 3D DIR model relies on a self-attention backbone to compute the DVF as a representation of the relationship between two inputs. Self-attention in the form of transformers have already proven themselves capable in natural language processing to take an input text and output a summary or translation. In our study the combined inputs are "translated" into a deformation vector field, where the DVF is simply another representation of the inputs. This contrasts with the conventional perspective of translating the moving image into a target image. The major advantage of self-attention is its ability to model long range dependencies in the input. Since our input is the combination of moving and target images, understanding long range relationships within this amalgamation is crucial for appropriate deformation prediction.

Recent work shows that convolution can be dispensed with altogether in favor of a solely attentional approach[10]. There are many advantages of attention over convolution. In addition to the previously mentioned long-range dependency learning, attention also provides a unique way to aggregate information from the input. Instead of learning a series of static filters, attention dynamically computes filters based on what it sees in the input. This is due to its elegant system of keys and queries. This system is extended to an additional *spatial* advantage. Whereas convolution kernels spatially represent relationships through relative value placement in their kernels and layer down-sampling, attention with positional encoding learns dynamic representations of relative positional encoding across the entire input and output. This allows for an enormous amount of flexibility for the network to learn spatial relationships.

Our model harnesses these attention advantages for volumetric DIR. While we modeled our architecture off the well-studied U-net as used in [4], we suspect that there is large room for improvement as other network architectures are explored. On the encoder side of our U-net, attention queries come from the output of the previous encoding layer. When going "up" the U-net, queries come from an up-sampling of the previous decoder layer concatenated with the output of the resolution-matched encoder layer. Recall that each position in the output can attend to all positions in the input layer, thus learning the full context of the abstract representation of the input at different sampling resolutions. This allows an enormous amount of encoded information to inform each point in the output.

Previous implementations of self-attention noted that pixel-level detail processing was inferior relative to CNN's[25,35]. During the development of our model, we also noticed this behavior. To improve detail matching, we added a deep feature loss based on matching encoded pre-trained VGG layers between target and deformed image. Addition of a deep layer loss

improved the mean square error between target and deformed image by 46% (0.003023 to 0.001623 at convergence). We suspect that introduction of a loss based on convolutional kernel features donates some of the advantages of a CNN to our attention network.

We also experimented with a semi-supervised contour loss using the Dice metric. The contours were not fed into the network but were used in the loss to guide training. We trained our model with and without Dice loss and noticed a negligible difference in the mean square error, and indeed in the dice contour matching values themselves (0.45 vs 0.43, for training with and without dice loss, respectively). However, upon qualitative inspection of the results, the network trained with dice led to more anatomically realistic results. This suggests that a different contour matching loss may help to drive the network better than the familiar and convenient dice term.

Our experiments suggest that our self-attention model produces results in line with previous DIR methodologies. When registering two different subjects, a DIR algorithm had to predict large deformations in structures that normally would not change (e.g., the skull, overall habitus). This necessity went beyond the regularization in B-spline based registration, and it was often unable to deform the moving image to the target. The CNN approach was able to best handle these distortions from a contour matching perspective. Our proposed method closely corresponds to the contour matching of the CNN approach.

When deforming subjects to themselves at a different time point, the requirements of deformation slightly change. Many aspects of the anatomy remain static, although large articulations can still be present. This time, the more regularized B-spline algorithm best matched the investigated contours. It can move large chunks of the image while maintaining

local anatomy's relative positions. Our proposed method closely matched the results of the B-spline registration for these intra-patient results.

Inter- and intra- patient registration serve different purposes. For example, inter-patient registration can help create atlas-based contours or align images of different subjects for a research study. This type of registration requires the freedom to make large deformations throughout the anatomy without excessive distortion. However, for most medical applications, intra-patient registration is used to combine information for a single patient's treatment. This type of registration requires the ability to predict motion along anatomical constraints. It can still contain large motion, but many aspects of the anatomy remain unchanged. These two types of registration have somewhat conflicting goals, and it is challenging to find an algorithm that meets them all. In this study, we observed that our proposed model was able to match the best performing algorithm in both scenarios. It could well match anatomy that was both realistically and unrealistically distorted. This behavior may be related to its ability to learn matching points at large distances in its input, which is a unique feature to attention-based networks.

While our network exhibited strong results in inter- and intra- patient DIR, we recognize that this is an early implementation of a promising technique. Many design choices of this network could be further optimized for better results. Our proposed method was compared to more established techniques which have been extensively optimized. We hope that with further research into attention-based DIR, the exciting advantages of this technique can lead to a new paradigm in medical 3D deformable image registration.

Conclusion

A novel self-attention based deformable image registration model was developed for 3D medical imaging. We leveraged axial factorization to overcome the computational complexity of attention networks and allow us to apply them to fully 3D input. This work opens the door for further applications of attention techniques in medical DIR.

References

1. de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I. End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. In: Vol 1. ; 2017:204-212. doi:10.1007/978-3-319-67558-9_24
2. Balakrishnan G, Zhao A, Sabuncu MR, Dalca A V., Guttag J. An Unsupervised Learning Model for Deformable Medical Image Registration. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2018:9252-9260. doi:10.1109/CVPR.2018.00964
3. Dalca A V., Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal*. 2019;57:226-236. doi:10.1016/j.media.2019.07.006
4. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca A V. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Trans Med Imaging*. 2019;38(8):1788-1800. doi:10.1109/TMI.2019.2897538
5. Neylon J, Min Y, Low DA, Santhanam A. A neural network approach for fast, automated quantification of DIR performance. *Med Phys*. 2017;44(8):4126-4138. doi:10.1002/mp.12321
6. Yang X, Kwitt R, Niethammer M. Fast predictive image registration. *Lect Notes Comput*

Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2016;10008 LNCS:48-57. doi:10.1007/978-3-319-46976-8_6

7. Kuppala K, Banda S, Barige TR. An overview of deep learning methods for image registration with focus on feature-based approaches. *Int J Image Data Fusion*. 2020;00(00):1-23. doi:10.1080/19479832.2019.1707720
8. Zhu X, Ding M, Huang T, Jin X, Zhang X. PCANet-based structural representation for nonrigid multimodal medical image registration. *Sensors (Switzerland)*. 2018;18(5). doi:10.3390/s18051477
9. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. *Mach Vis Appl*. 2020;31(1):1-18. doi:10.1007/s00138-020-01060-x
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;2017-Decem(Nips):5999-6009.
11. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 12349 LNCS. ; 2020:108-126. doi:10.1007/978-3-030-58548-8_7
12. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report. *Med Phys*. 2017;44(7):e43-e76. doi:10.1002/mp.12256
13. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report. *Med Phys*. 2017;44(7):e43-e76.

doi:10.1002/mp.12256

14. Shi L, Chen Q, Barley S, et al. Benchmarking of Deformable Image Registration for Multiple Anatomic Sites Using Digital Data Sets With Ground-Truth Deformation Vector Fields. *Pract Radiat Oncol*. 2021;11(5):404-414. doi:10.1016/j.prro.2021.02.012
15. Lao Y, Yu V, Pham A, et al. Quantitative Characterization of Tumor Proximity to Stem Cell Niches: Implications on Recurrence and Survival in GBM Patients. *Int J Radiat Oncol Biol Phys*. 2021;110(4):1180-1188. doi:10.1016/j.ijrobp.2021.02.020
16. Calais J, Czernin J, Cao M, et al. 68 Ga-PSMA-11 PET/CT mapping of prostate cancer biochemical recurrence after radical prostatectomy in 270 patients with a PSA level of less than 1.0 ng/mL: Impact on salvage radiotherapy planning. *J Nucl Med*. 2018;59(2):230-237. doi:10.2967/jnumed.117.201749
17. Kamal M, Mohamed ASR, Fuller CD, et al. Patterns of Failure After Intensity Modulated Radiation Therapy in Head and Neck Squamous Cell Carcinoma of Unknown Primary: Implication of Elective Nodal and Mucosal Dose Coverage. *Adv Radiat Oncol*. 2020;5(5):929-935. doi:10.1016/j.adro.2020.04.025
18. Kim N, Chang JS, Kim YB, Kim JS. Atlas-based auto-segmentation for postoperative radiotherapy planning in endometrial and cervical cancers. *Radiat Oncol*. 2020;15(1):1-9. doi:10.1186/s13014-020-01562-y
19. Paragios N, Duncan J, Ayache N. Handbook of biomedical imaging: Methodologies and clinical research. *Handb Biomed Imaging Methodol Clin Res*. 2015:1-511. doi:10.1007/978-0-387-09749-7
20. Kessler ML. Image registration and data fusion in radiation therapy. *Br J Radiol*. 2006;79(SPEC. ISS.):99-108. doi:10.1259/bjr/70617164

21. Maier A, Syben C, Lasser T, Riess C. A gentle introduction to deep learning in medical image processing. *Z Med Phys*. 2019;29(2):86-101. doi:10.1016/j.zemedi.2018.12.003

22. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42(December 2012):60-88. doi:10.1016/j.media.2017.07.005

23. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: A review. *Phys Med Biol*. 2020;65(20). doi:10.1088/1361-6560/ab843e

24. Chen J, He Y, Frey EC, Li Y, Du Y. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration. 2021:1-9. http://arxiv.org/abs/2104.06468.

25. Wang Z, Delingette H. Attention for Image Registration (AiR): an unsupervised Transformer approach. 2021:1-9. http://arxiv.org/abs/2105.02282.

26. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. CCNet: Criss-Cross Attention for Semantic Segmentation. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Vol 2019-Octob. IEEE; 2019:603-612. doi:10.1109/ICCV.2019.00069

27. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

28. Tong N, Gou S, Yang S, Cao M, Sheng K. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. *Med Phys*. 2019;46(6):2669-2682. doi:10.1002/mp.13553

29. Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys*. 2018;45(10):4558-4567. doi:10.1002/mp.13147

30. Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts. *5th Int Conf Learn Represent ICLR 2017 - Conf Track Proc*. August 2016:1-16.

31. Hui Z, Li J, Wang X, Gao X. Image fine-grained inpainting. *arXiv*. 2020;(2):1-11.

32. McKenzie EM, Tong N, Ruan D, Cao M, Chin RK, Sheng K. Using neural networks to extend cropped medical images for deformable registration among images with differing scan extents. *Med Phys*. 2021;48(8):4459-4471. doi:10.1002/mp.15039

33. Shamonin D, Bron E, Lelieveldt B, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinform*. 2014;7(January):1-15. doi:10.3389/fninf.2013.00050

34. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix : A Toolbox for Intensity-Based Medical Image Registration. 2010;29(1):196-205.

35. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in Transformer. February 2021:1-12. http://arxiv.org/abs/2103.00112.