GADOLINIUM-FREE CONTRAST-ENHANCED MRI (GFCE-MRI) SYNTHESIS VIA DEEP LEARNING FOR RADIOTHERAPY OF NASOPHARYNGEAL CARCINOMA

LI WEN

PhD

The Hong Kong Polytechnic University

2023

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

GADOLINIUM-FREE CONTRAST-ENHANCED MRI (GFCE-MRI) SYNTHESIS VIA DEEP LEARNING FOR RADIOTHERAPY OF NASOPHARYNGEAL CARCINOMA

LI WEN

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

February 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Li Wen (Name of Student)

Abstract

Nasopharyngeal carcinoma (NPC) is a highly infiltrative and radiosensitive malignancy. Radiotherapy is currently the mainstay therapeutic remedy. In radiotherapy of NPC patients, the gadolinium-based contrast enhanced MRI (CE-MRI) plays a critical role in NPC delineation. However, the gadolinium-based contrast agents (GBCAs) associated safety issues have attracted serious attention of clinicians in recent years. To reduce or eliminate the use of GBCAs, deep learning has been proposed to synthesize the gadolinium-free contrast enhanced MRI (GFCE-MRI), aiming at providing an alternative to gadolinium-based CE-MRI. Nevertheless, recent studies mostly focus on novel deep learning algorithms development or feasibility investigations for disease diagnosis in different anatomies, such as brain, liver, and breast. Currently, these is no study has been reported for NPC radiotherapy. In this study, we for the first time developed deep learning algorithm to synthesize GFCE-MRI from contrast-free T1-weighted (T1w) and T2-weighted (T2w) MRI for radiotherapy of NPC patients. Specifically, we achieved three research objectives in this study: (i) to develop a novel multimodality-guided synergistic neural network (MMgSN-Net) that tailored for GFCE-MRI synthesis of NPC patients; (ii) to investigate and improve the GFCE-MRI model generalizability using multi-institutional MRI data; and (iii) to investigate the clinical efficacy of GFCE-MRI in primary NPC tumor delineation. Our experiments showed that the proposed MMgSN-Net is able to generate highly realistic GFCE-MRI images and the quantitative results outperformed three comparing state-of-the-art methods. We also found that the heterogeneity of multi-institutional MRI heavily

affects generalizability of the well-trained single-institutional model. After training the model with multi-institutional data and shorting the multi-institutional data distribution variations, the model generalizability has been significantly improved. The clinical evaluation results also suggest that our synthetic GFCE-MRI is highly promising for clinical use, with Dice Similarity Coefficient (DSC) of 0.762 and Hausdorff Distance (HD) of 1.932mm, respectively. The dosimetric difference of planning target volumes between real patients and synthetic patients was less than 1%, which is acceptable for radiotherapy as reported by two board-certified oncologists.

Publications arising from the thesis

Journal articles

1. Cheng Li, **Wen Li**, Hairong Zheng, Jing Cai, Shanshan Wang. *Artificial intelligence in multiparametric magnetic resonance imaging: A review*. Medical Physics. 2022; 49(10): e1024-e1054.

2. Chenyang Liu, Mao Li, Haonan Xiao, Tian Li, **Wen Li**, Jiang Zhang, Xinzhi Teng, Jing Cai. *Advances in MRI-guided precision radiotherapy*. Precision Radiation Oncology. 2022; 6(1): 75-84.

3. Chenyang Liu, Tian Li, Peng Cao, Edward S. Hui, Yat-Lam Wong, Haonan Xiao, Shaohua Zhi, Ta Zhou, **Wen Li**, Saikit Lam, Andy Lai-yin Cheung, Victor Ho-fun Lee, Michael Ying, Jing Cai. *Respiratory-correlated Four-dimensional Magnetic Resonance Fingerprinting (RC-4DMRF) for Magnetic Resonance Imaging (MRI)-guided Radiotherapy of Liver Cancer*. International Journal of Radiation Oncology* Biology* Physics. Under Review.

4. Haonan Xiao, Ruiyan Ni, Shaohua Zhi, **Wen Li**, Chenyang Liu, Ge Ren, Xinzhi Teng, Weiwei Liu, Weihu Wang, Yibao Zhang, Hao Wu, V.H.F. Lee, Lai Yin Andy Cheung, Hing-Chiu Charles Chang, Tian Li, Jing Cai. *A Dual-supervised Deformation Estimation Model (DDEM) for Constructing Ultra-quality 4D-MRI based on a Commercial Low-quality 4D-MRI for Liver Cancer Radiation Therapy*. Medical Physics. 2022; 49(5): 3159-3170.

5. Haonan Xiao, Xinyang Han, Shaohua Zhi, Yat Lam Wong, Chenyang Liu, **Wen Li**, Weiwei Liu, Weihu Wang, Yibao Zhang, Hao Wu, Ho-Fun Victor Lee, Lai-Yin Andy Cheung, Hing-Chiu Charles Chang, Tian Li, Jing Cai. *D2R Model: A joint MRI reconstruction and registration model for real-time multiparametric 4D-MRI construction*. 2022; Under Review.

6. Shaohua Zhi, Yinghui Wang, Haonan Xiao, Ti Bai, Hong Ge, Li Bing, Chenyang Liu, **Wen Li**, Tian Li, Jing Cai. *Coarse–Super-Resolution–Fine Network (CoSF-Net): A Unified End-to-End Neural Network for 4D-MRI with Simultaneous Motion Estimation and Super-Resolution*. IEEE Transactions on Medical Imaging. Under Review.

7. **Wen Li**, Haonan Xiao, Tian Li, Ge Ren, Saikit Lam, Xinzhi Teng, Chenyang Liu, Jiang Zhang, Francis Kar-ho Lee, Kwok-hung Au, Victor Ho-fun Lee, Amy Tien Yee Chang, Jing Cai. *Virtual Contrast-enhanced Magnetic Resonance Images Synthesis for Patients with Nasopharyngeal Carcinoma using Multimodality-guided Synergistic Neural Network*. International Journal of Radiation Oncology* Biology* Physics. 2022; 112 (4): 1033-1044.

8. **Wen Li**, Saikit Lam, Tian Li, Andy Lai-Yin Cheung, Haonan Xiao, Chenyang Liu, Jiang Zhang, Xinzhi Teng, Shaohua Zhi, Ge Ren, Francis Kar-ho Lee, Kwok-hung Au, Victor Ho-fun Lee, Amy Tien Yee Chang, Jing Cai. *Multi-institutional Investigation of Model Generalizability for Virtual Contrast-Enhanced MRI Synthesis*. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022; 765-773.

9. **Wen Li**, Saikit Lam, Tian Li, Jens Kleesiek, Andy Lai-Yin Cheung, Ying Sun, Francis Kar-ho Lee, Kwok-hung Au, Victor Ho-fun Lee, Jing Cai. *Model Generalizability Investigation for GFCE-MRI Synthesis in NPC Radiotherapy Using Multi-institutional Patient-based Data Normalization.* IEEE Journal of Biomedical and Health Informatics. Under Review.

10. **Wen Li**, Dan Zhao, Zhi Chen, Zhou Huang, Saikit Lam, Andy Lai-Yin Cheung, Haonan Xiao, Chenyang Liu, Francis Kar-Ho Lee, Kwok-Hung Au, Victor Ho-Fun Lee, Jing Cai, Tian Li. *Clinical Evaluation of Artificial Intelligence-assisted Multi-institutional Virtual Contrast Enhanced MRI in Primary Gross Tumor Volume Delineation for Patients with Nasopharyngeal Carcinoma.* International Journal of Radiation Oncology* Biology* Physics. Under Review.

11. Xinzhi Teng, Jiang Zhang, Alex Zwanenburg, Jiachen Sun, Yu-Hua Huang, Saikit Lam, Yuanpeng Zhang, Bing Li, Ta Zhou, Haonan Xiao, Chenyang Liu, **Wen Li**, Xinyang Han, Zongrui Ma, Jing Cai. *Building Reliable Radiomic Models using Image Perturbation*. Scientific Reports. 2022; 12(1): 1-10.

12. Xinzhi Teng, Jiang Zhang, Zongrui Ma, Yuanpeng Zhang, Sai Kit Lam, **Wen Li**, Haonan Xiao, Tian Li, Bing Li, Ta Zhou, Ge Ren, Francis Kar-Ho Lee, Kwok-Hung Au, Victor Ho Fun LEE, Yee Chang, Jing Cai. *Improving Radiomic Model Reliability using Robust Features from Perturbations for Head-and-Neck Carcinoma.* Frontiers in Oncology. 2022; 5681.

Conference abstracts

1. Jiamin Chen, **Wen Li**, Haonan Xiao, SaiKit Lam, Jingtang Chen, Chenyang Liu, Andy Lai-Yin Cheung, Jing Cai. *Virtual Non Contrast Tomography Synthesis for Hepatocellular Carcinoma Patients Using Multimodality-Guided Synergistic Neural Network*. The AAPM 64th Annual Meeting & Exhibition. July 10-14, 2022.

2. **Wen Li**, Ge Ren, Tian Li, Haonan Xiao, Francis Kar-ho Lee, Kwok-hung Au, Jing Cai. CE-Net: multi-inputs contrast enhancement network for nasopharyngeal carcinoma contrast-enhanced T1-weighted MR synthesis. International Society for Magnetic Resonance in Medicine (ISMRM). December 16, 2021.

3. **Wen Li**, Saikit Lam, Haonan Xiao, Tian Li, Ge Ren, Shaohua Zhi, Xinzhi Teng, Chenyang Liu, Jiang Zhang, Francis Kar-ho Lee, Kwok-hung Au, Victor Ho-fun Lee, Amy Tien Yee Chang, Jing Cai. *Gadolinium-free Contrast-enhanced MRI (GFCE-MRI) Synthesis via Generalizable MHDgN-Net for Patients with Nasopharyngeal Carcinoma*. International Society for Magnetic Resonance in Medicine (ISMRM). May 7-12, 2022.

Acknowledgements

I would like to take this opportunity to thank all the people who gave me help and support during my PhD period.

Among all, I would like to first express my sincere thanks to my chief supervisor, Prof. Cai Jing, for his kind support, patient guidance, valuable advice, and precious opportunities during the PhD period. For my project, Prof. Cai provided a lot of valuable suggestions and comments from the clinical and practical perspectives, which helped my project become more comprehensive and made sure my project was in the right direction. To make us well-prepared for future challenges, Prof. Cai tried his best to exercise us by all manner of means, such as help us improve the presentation skills and writing skills throughout the group meeting, as well as encourage us sharing our mind during the group meeting. Due to the lack of clinical background, Prof. Cai kindly provide me an opportunity to conduct a three months' field trip in West China Hospital. I learned a lot of clinical knowledge from this field trip, which helped me to better solve my research problems with clinical perspectives. In my daily research, Prof. Cai also provided me with the largest freedom to explore my project. I truly appreciate Prof. Cai for all the support.

Importantly, I would like to thank collaborators Dr. Zhao Dan and Dr. Huang Zhou from Beijing Cancer Hospital for their kind help in clinical evaluations of the synthetic GFCE-MRI, I would also like to thank collaborators Mrs. Chen Jiamin from Jiangmen Central Hospital, Mr. Wu Xiaohui from Suining Central Hospital, Mrs. Hu

Die from Southwest Medical University Affiliated Hospital of Traditional Chinese Medicine, and Mr. Heng Li from the Second Affiliated Hospital of Chongqing Medical University, as well as the seven invited radiation oncologists from these four hospitals for their generous support of the GFCE-MRI image distinguishability experiment and discussion of the results.

Next, I would like to thank my group members, especially Lam Saikit, Liu Chenyang, Xiao Haonan, Teng Xinzhi, Zhang Jiang, and Chen Zhi for their help in my project. In my first publication during my PhD, Saikit spent a lot of time helping to reorganize my manuscript, and he also provided many valuable suggestions to help improve the quality of the manuscript. During this period, I learned a lot of writing skills from him. When I encountered research problems, Liu Chenyang, Xiao Haonan, Teng Xinzhi and Zhang Jiang were always friendly to share their ideas with me, their suggestions always helped me solve my problems. Chen Zhi is an experienced clinical medical physicist. In the clinical evaluation of the GFCE-MRI, he spent a great deal of his personal time helping to generate radiotherapy plans. My thesis would be not possible without their kind help.

The biggest thanks must go to my family members, especially my parents. They always give me the most freedom, encourage me to do what I want to do, and always give me the best unconditionally. I hope I would not let them down and become someone they can be proud of.

At last, I would like to thank the thesis committee and the external examiners for their valuable time to coordinate and assess my thesis. I hope my research could have some contributions to the medical community and well-being of the NPC patients.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AJCC	American Joint Committee on Cancer
ASCO	American Society of Clinical Oncology
CE-MRI	Contrast-enhanced Magnetic Resonance Imaging
CT	Computed Tomography
DSC	Dice Similarity Coefficient
DL	Deep Learning
DVH	Dose-volume Histogram
DWI	Diffusion-weighted Imaging
EBT	External Beam Radiotherapy
EBV	Epstein-Barr Virus
EDM	External Distribution Matching
EES	Extravascular-extracellular Space

EMA	European Medicines Agency
FDA	The United States Food and Drug Administration
FLAIR	Fluid-attenuated Inversion Recovery
GBCAs	Gadolinium-based Contrast Agents
GFCE-MRI	Gadolinium-free Contrast-enhanced Magnetic Resonance Imaging
GTV	Gross-tumor-volume
HD	Hausdorff Distance
HPV	Human Papillomavirus
IARC	International Agency for Research on Cancer
ICC	Intraclass Correlation Coefficient
IGRT	Image-guided Radiotherapy
IMRT	Intensity-modulated Radiotherapy
JI	Jaccard Index

MAE	Mean Absolute Error
MMgSN-Net	Multimodality Guided Synergistic Neural Network
MRI	Magnetic Resonance Imaging
MRIgRT	Magnetic Resonance Image-guided Radiotherapy
MSE	Mean Squared Error
Nex	Number of Excitation
NPC	Nasopharyngeal Carcinoma
NSF	Nephrogenic Systemic Fibrosis
OOD	Out of Distribution
PBT	Proton Beam Therapy
PSNR	Peak Signal-to-noise Ratio
PTV	Planning Target Volume
SD	Standard Deviation

SGS	Synergistic Guidance System
SRS	Stereotactic Radiosurgery
SSIM	Structural Similarity Index
STIR	Short Tau Inversion Recovery
SWI	Susceptibility-weighted Imaging
T1w	T1-weighted
T2w	T2-weighted
TE	Echo Time
TR	Repetition Time
WHO	The World Health Organization

1. Introduction

1.1 Nasopharyngeal carcinoma

1.1.1 NPC basics

Nasopharyngeal carcinoma (NPC), located in an intricated nose-pharynx ministry, is a highly infiltrative malignancy (Lin et al., 2015). **Figure 1-1 (a)** shows the anatomic position of nasopharynx, where nasopharyngeal carcinoma usually occurred. NPC is a soft-tissue mass, it presents a high tendency to invade nearby healthy soft tissues, neural structures, and bony skull base (Li, Xiao, et al., 2021). A case of the infiltrative NPC is shown in **Figure 1-1 (b)**.

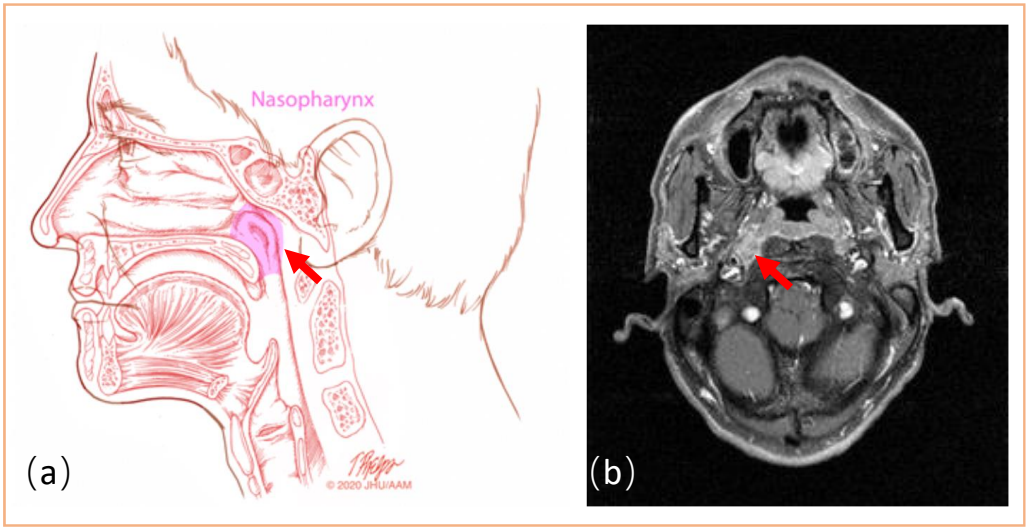

Figure 1-1. Illustration of nasopharynx and NPC. (a): The anatomic position of nasopharynx, where nasopharyngeal carcinoma usually occurred. (b) nasopharyngeal carcinoma in contrast-enhanced MRI. Red arrows show the position of NPC. NPC: nasopharyngeal carcinoma.

As a head and neck cancer, NPC is distributed with distinct geographical characteristics (Chen et al., 2019). In 2018, 129079 new cases and 72987 new deaths were recorded globally, accounting for 0.7% and 0.8% of all cancer types, respectively (Bray et al., 2018). More than 70% of newly diagnosed cases were found in East and Southeast Asia (Chen et al., 2019). In China and Indonesia, 60558 and 17992 new cases were reported in 2018, accounting for 47.7% and 14.2% of all new NPC cases (Chen et al., 2019). Besides East and Southeast Asia, Micronesia, Polynesia, and parts of Africa are also suffered from a high incidence and mortality rate.

In recent years, the incidence of NPC shows an increased tendency. It was reported that the worldwide new cases of NPC were 86500 (Chua et al., 2016), 129079 (Bray et al., 2018), 133354 (Sung et al., 2021) in 2012, 2018 and 2020 respectively. Compared to the reported deaths in 2018, 7021 additional deaths were reported globally in the year of 2020. Male has a higher incidence and mortality rate than female, with a number of 96371 cases of incidence and 58094 cases of mortality for male against 36983 of incidence and 21914 of mortality for female. The incidence and mortality between male and female are 2.61/1 and 2.65/1 in 2020 (Sung et al., 2021). **Figure 1-2** illustrated the worldwide age standardized incidence rates for both male and female in 2020 (Observatory, 2020).

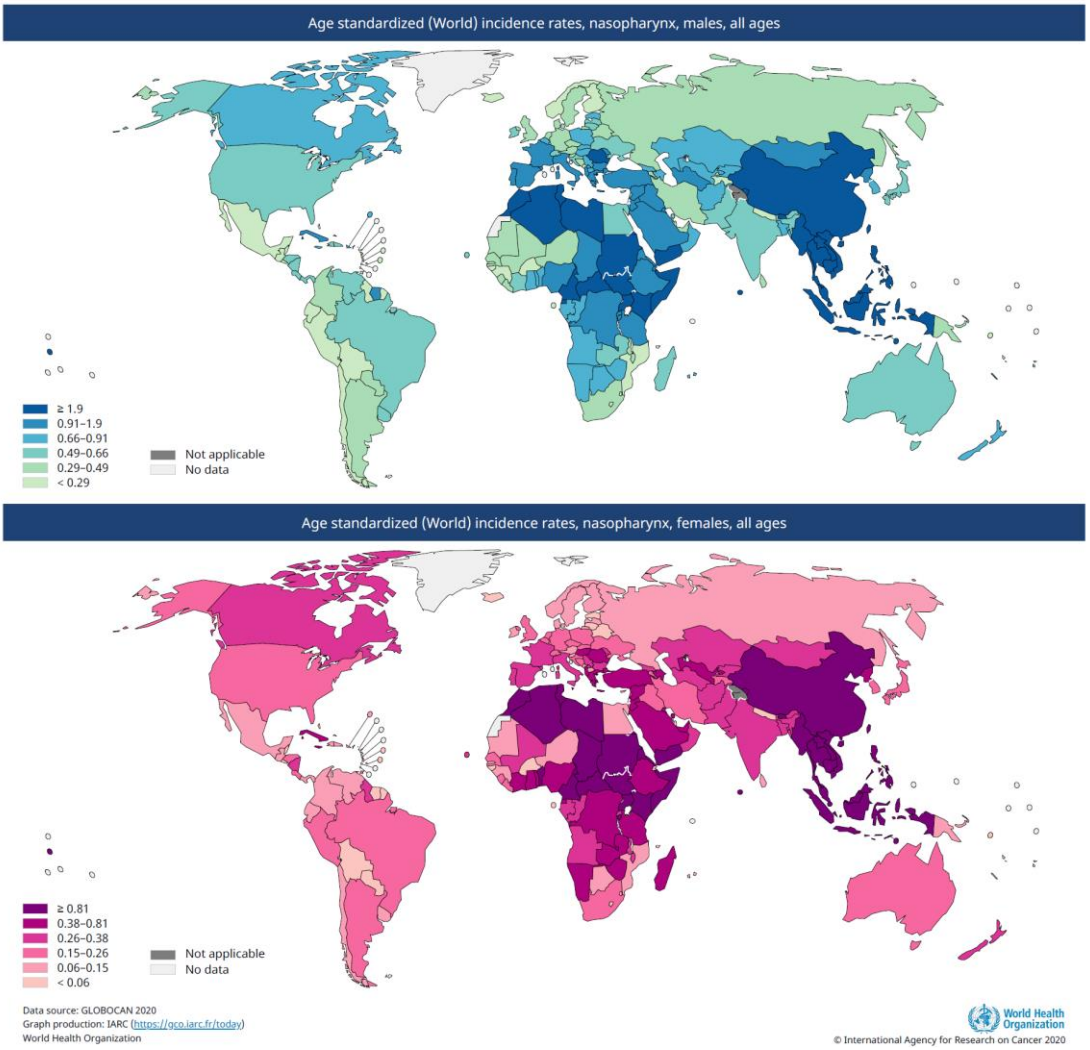

Figure 1-2. Worldwide age standardized incidence rates of male and female in 2020 (Observatory, 2020).

1.1.2 Risk factors of NPC

The occurrence of NPC was considered to be implicated in several etiologic factors. Viral infection, salted fish consumption, alcohol drinking, cigarette smoking (Kamran et al., 2015), and environmental factors were considered to be the most common NPC risk factors. The incidence of NPC is tightly related to Epstein-Barr virus (EBV) (Chua

et al., 2016). In high NPC incidence areas, 90%-100% patients were found to be infected with EBV (Ho et al., 2013; Kamran et al., 2015). Another possible viral risk is human papillomavirus (HPV), but at present, no clear relationship has been established (Chua et al., 2016). Besides EBV and HPV, salted fish was believed to be an important carcinogen (Kamran et al., 2015). It was observed that the boat people in Southern China has a high NPC incidence (Lee et al., 2012). After carefully identification, the Cantonese salted fish has been listed as a group-one carcinogen by International Agency for Research on Cancer (IARC) (Kamran et al., 2015). Excessive smoking and alcohol assumption were also demonstrated to be risk factors for NPC. A study conducted by Nam et al. demonstrated that the NPC incidence rate was found to be 3 times than the normal group in the group with excessive smoking consumption (Nam et al., 1992). They also found the people who with heavy alcohol consumption will have a risk 80% higher than people who without alcohol consumption. The occurrence of NPC is also thought to be related to environmental factors. Buell et al. observed an incidence decline among Southern Chinese after migration to California (Buell, 1974).

1.1.3 NPC stage

According to histological characteristics, the World Health Organization (WHO) classified NPC into three subtypes: keratinizing squamous cell carcinoma (Subtype-1), differentiated non-keratinizing carcinoma (Subtype-2), and undifferentiated non-keratinizing (Subtype-3). It was found that 70%-80% Subtype-1 patients were related to infection of EBV. For Subtype-2 and Subtype-3, however, almost all patients were linked with infection of EBV (Adoga et al., 2018; Blanchard et al., 2018; Fu et al.,

2018). According to the diagnostic results such as physical examinations, biopsies, and imaging, the NPC can be further divided into different stages. The NPC patients with similar stages are tend to accept similar treatment in clinical. The most widely accepted staging method is the TNM system that published by American Joint Committee on Cancer (AJCC), the latest version was taking effect from 2018 (Lee et al., 2019). The TNM system classify NPC patients by three key information: the main tumor extent (T-Stage), the spread to nearby neck lymph nodes (N-Stage), the spread to distant parts (metastasis) such as the bone, lung, and liver (M-Stage). Depending on the severity of each stage, the TNM stages are split into several substages, including Tis, T0, T1, T2, T3, T4, N0, N1, N2, N3, M0, and M1. These substages can further be grouped to Stage-0 (Tis, N0, M0), Stage-I (T1, N0, M0), Stage-II ((T1/T0, N1, M0), or (T2, N0/N1, M0)), Stage-III ((T1/T0, N2, M0), or (T2, N2, M0), or (T3, N0-N2, M0)), Stage-IVA ((T4, N0-N2, M0), or (Tis-T4, N3, M0)), and Stage-IVB (Tis-T4, N0-N3, M1) according the severity of the three stages (R. Guo et al., 2019).

1.1.4 Clinical treatment for NPC

NPC is naturally radiosensitive. Radiotherapy is currently the mainstay therapeutic remedy. Besides radiotherapy, chemotherapy and surgery can be used as a combination to improve the therapeutic outcome (Chen et al., 2019). Radiotherapy is a cancer treatment that uses high dose of radiation to destroy cancer cells and shrink tumors. There are different types of radiotherapy treatment for NPC patients, commonly used radiotherapy types are external-beam radiotherapy (EBT), such as intensity-modulated radiotherapy (IMRT), proton beam therapy (PBT) and stereotactic radiosurgery (SRS).

The EBT is the most commonly applied radiotherapy, which delivers the radiation from a radiotherapy machine outside the body. The intensity-modulated radiotherapy (IMRT) is one typical type of EBT, which allows delivering effective x-rays from different angles by advanced computer programs to reduce side effects for NPC patients. IMRT was recommend by American Society of Clinical Oncology (ASCO) for Stage-II to Stage-IVA patients. Another EBT type is PBT. Instead of using high-energy x-rays, PBT uses high energy protons to kill cancer cells, which can be used for patients with later-stage NPC. SRS delivers precisely-targeted radiation beams to treat the NPC tumor, which helps preserve nearby healthy tissues. SRS can be used to treat tumor that has grown to skull base or brain. Brachytherapy (ASCO, 2020) is a type of internal radiotherapy that is delivered by radioactive implants, which is often used to treat recurrent NPC. Besides radiotherapy, chemotherapy is usually applied before (induction chemotherapy), after (adjuvant chemotherapy), or at the same time with radiotherapy (chemoradiotherapy) to enhance the treatment outcome. Chemotherapy uses drugs to kill cancer cells, always by stopping cancer cells from dividing. Stage-II to Stage-IVA patients were usually recommended for chemotherapy. Occasionally, surgery is conducted when the cancer has spread to lymph nodes, especially for some undifferentiated nasopharynx tumor. However, surgery may cause some severe side effects such as nerve damage, swelling, and facial disfigurement (Chen et al., 2019). In this study, we mainly focus on radiotherapy for NPC treatment, especially the magnetic resonance image-guided radiotherapy (MRIgRT).

1.2 MRIgRT and gadolinium-based contrast-enhanced MRI (CE-MRI)

MRIgRT is an emerging technique that takes advantage of the excellent soft tissue contrast of magnetic resonance imaging (MRI) images (Schmidt et al., 2015). In 2014, the first clinical MRIgRT technique was implemented in Washington (Henke et al., 2018). After that, the MRIgRT technique has been rapidly expanded to multiple institutions and countries (Henke et al., 2018). Compared to traditional x-ray-based image-guided radiotherapy (IGRT), MRI is featured with free of ionizing radiation, superior soft-tissue contrast, any oblique angle imaging, and motion resolving capabilities (Freedman et al., 2018). MRI is particular popular for pediatric populations where ionizing radiation should be carefully managed, especially for those patients who need repeated scan during radiotherapy treatment (Schmidt et al., 2015). In recent years, MRI has been successfully applied to radiotherapy procedures such as tumor delineation, treatment planning, dose calculation, treatment delivery, and outcome assessment (Wen et al., 2020). At present, MRI has become a standard part in radiotherapy planning workflow, which allows higher-quality delineation of tumor and organs at risk (Bahig et al., 2019). The improved quality of delineation, combined with functional imaging techniques, is promising for individualized radiotherapy (Bahig et al., 2019). In this study, we focus on MRI guided tumor delineation for NPC patients.

In a successful radiotherapy, accurate tumor delineation is the foremost prerequisite. However, tumor delineation for NPC patients is particularly challenging in view of the deeply infiltrative nature of NPC, which presents a high tendency to invade nearby normal soft tissues, bony skull base, as well as neural structures, thus obfuscating oncologists for accurate assess and delineate the tumor from healthy tissues. In general, the CE-MRI, which is generated by injection of gadolinium-based contrast

agents (GBCAs), is utilized to enhance the visibility of tumor. The contrast agents that used in MRI are generally paramagnetic, one typical agent is gadoterate (a kind of macrocyclic extracellular fluid agent that approved by both European Medicines Agency (EMA) and The United States Food and Drug Administration (FDA)). After administrated orally or intravenously, the contrast agent flows through the blood. The T1 relaxation time of nearby protons is shortened by interacting with the contrast agent. Compared with healthy tissues, the tumor has a more rapid uptake and washout rate due to the leaky immature vascular system (Schmidt et al., 2015). In T1-weighted (T1w) MRI scanning, the shortened T1 relaxation time tissues appear bright in T1w MRI images compared to surrounding normal tissues that without contacting with the contrast agent. The T1w MRI images that enhanced by contrast agent are the CE-MRI. For the tumor regions, the surrounding blood vessels are disrupted, leading to contrast agent leak out from the blood vessels into the extracellular space, therefore enhancing the signal of tumor regions. As a functional imaging technique, the spatial resolution of CE-MRI allows quantitatively analysis the microenvironment changes and blood perfusion status of tumor at millimeter level (Torheim et al., 2014; Zhao et al., 2019). The most popular CE-MRI quantitative analysis method is the Tofts model (Chikui et al., 2012; Tofts, 2010), which contains four parameters that related to pharmacokinetics of gadolinium: K^{trans} (volume transfer constant), Ve (volume fraction of extravascular extracellular space), Kep (rate constant), and Vp (volume fraction of plasma). K^{trans} is the most commonly used CE-MRI parameter, representing the diffusion rate of gadolinium from plasma to the extravascular-extracellular space (EES) in unit time. According to the gadolinium influx rate from plasma to EES, the K^{trans} reflects the blood

flow status and capillary leakage of the tumor. Ve represents the ratio between the volume of the gadolinium leaking into the EES from plasma and the whole EES volume. It is a number between 0 and 1, reflecting the capacity of the gadolinium in EES. Kep represents the diffusion rate of gadolinium from EES back to the plasma, which can be calculated by K^{trans} / Ve. The increase of Kep may represents the increase of K^{trans} or decrease of Ve, or both. The last parameter is Vp, representing the percentage of gadolinium in plasma. The Vp is very small in many lesions and can be ignored. In tumors with abundant blood supply, however, the contribution of intravascular signal to the total signal may larger than 10%, which cannot be ignored. Based on the CE-MRI, these four parameters can be calculated pixel by pixel and generate different parameter maps (Schmidt et al., 2015). In addition, the relationship between these four parameters, integrated with the time curve, can reflect the microenvironment and the status of the tumor (Cheng et al., 2013; Tofts et al., 1999; Vajapeyam et al., 2017).

Besides the application of tumor delineation in radiotherapy, the CE-MRI also shows the capability to predict the response of tumor and normal organs to radiotherapy (Cao, 2011). By assessing the tumor response to radiotherapy at an early stage, the clinical oncologists can adaptively optimize the treatment plan based on the functional changes of tumor earlier than morphologic alterations, thus achieving a better treatment outcome. In addition, the early assessment of the dose response in normal organs provides the possibility to further reduce the radiation injury to normal organs (Granata et al., 2021; Hylton, 2006; Zahra et al., 2007).

1.3 Safety issues of GBCAs

Despite the valuable applications of CE-MRI in radiotherapy, in recent years, accumulated evidences have demonstrated the GBCAs-related safety issues. The safety issues include adverse reactions, deposition, and toxicities. The adverse reactions can be classified to two categories: physiologic and hypersensitivity-related (Fraum et al., 2017). Physiologic reactions are dose related. According to the severity, the gadolinium caused physiologic reactions are ranged from mild (such as vomiting) to severe (such as refractory vasovagal reactions). The hypersensitivity-related reactions are also termed allergic-like reactions, which are caused by the immune system, such as limited urticaria (mild) and anaphylactic shock (severe) (Fraum et al., 2017). Raisch et al. (Raisch et al., 2014) investigated 614 cases of severe gadolinium-based adverse reactions, they found 53% of these cases were resulted in hospitalization; 31% of the cases were justified lift-threatening; and 7% and 2% cases were caused death and disability, respectively. In 2006, researchers first found the deposited gadolinium in skin of renal failure patients (Grobner, 2006). After that, the presence of gadolinium was also found in bone (Gibby et al., 2004), liver (Maximova et al., 2016), and brain structures such as dentate nucleus and globus pallidus (Kanda et al., 2015b). The gadolinium deposition cases were even found in pediatric patients with normal renal function. A large amount of evidence has shown the potential toxicities in patients. In a study published in 2016, Semelka et al. reported a series of gadolinium-related clinical symptoms, including central and peripheral pain, headache and bone pain, as well as skin thickening (Semelka et al., 2016). Importantly, the fatal nephrogenic systemic fibrosis (NSF) was found to closely connected with administration of GBCAs in end-stage renal failure patients (Mathur et al., 2020). The mechanism of gadolinium

deposition and toxicities in patients is currently remains unknown. The deposition and toxicities of gadolinium have triggered the abolishment of macrocyclic GBCAs in European countries in 2017 (Kleesiek et al., 2019b). For safety consideration, the use of GBCAs were recommended to be eliminated or reduced. To avoid the use of GBCAs, various strategies were proposed, including using contrast-free MRI, CT, and ultrasound to replace the use of gadolinium-based CE-MRI (Diop et al., 2013). Gadolinium-free contrast agents were also been explored. At present, however, none technique was found to have adequate clinical value to replace the use of GBCAs (Kleesiek et al., 2019b).

1.4 Deep learning for GFCE-MRI synthesis

In recent years, deep-learning (DL) assisted image synthesis has been caught in the spotlight of attention in the medical domain (Liang et al., 2019; Ren et al., 2021). The capability of deep neural networks in unraveling complex tumor-related characteristics (Amin et al., 2018; Saba et al., 2020; Shkolyar et al., 2019) has motivated scientists to synthesize gadolinium-free contrast enhanced MRI (GFCE-MRI) images from low-dose or contrast-free MRI images (Gong et al., 2018b; Kleesiek et al., 2019b). In 2018, the first DL assisted technique was developed to synthesize the GFCE-MRI from contrast-free T1w MRI and 10% low-dose MRI. This study demonstrated the possibility to reduce the gadolinium dose by 90% through DL (Gong et al., 2018b). Followed by this study, in 2019, Kleesiek et al. used multiparametric MRI including T1w, T2-weighted (T2w), T2w fluid-attenuated inversion recovery (FLAIR), diffusion-weighted imaging (DWI), and susceptibility-weighted imaging (SWI) images to train a

DL model to synthesize the GFCE-MRI, which validated the feasibility to generate the GFCE-MRI without any administration of GBCAs (Kleesiek et al., 2019b). Both of these two works were targeted on brain cancers. In 2020, Zhao et al. successfully used the T1w MRI to synthesize the GFCE-MRI for liver cancer detection (Zhao et al., 2020b). Followed by these works, in 2021, many different DL assisted techniques were proposed to synthesize the GFCE-MRI images for different anatomies, such as brain (Bône et al., 2021; Calabrese et al., 2021; Chen et al., 2021; Kim et al., 2021; Luo et al., 2021b; Pasumarthi et al., 2021b), liver (Xu et al., 2021b), and breast (Kim et al., 2021).

1.5 Challenges of current studies

Despite the great success that has been achieved by previous works, yet, the existing DL assisted methods still suffer from three major deficiencies/challenges: **(i)** the GFCE-MRI synthesis for NPC patients remains unexplored; **(ii)** existing methods have low or unknown model generalizability; and **(iii)** inadequate clinical evaluations of the synthetic GFCE-MRI for radiotherapy applications. Following are detailed descriptions of these challenges.

Challenges in synthesizing GFCE-MRI for NPC patients

NPC is a highly infiltrative malignancy that originated in the intricated nose-pharynx ministry. Accurate NPC target delineation is a critical step to ensure a good tumor control. Compared with the previous investigated anatomies (brain, liver, and breast), NPC presents a high tendency to invade surrounding soft tissues, neural structures, and

bony skull base, obfuscating physicians for accurate assessment and delineation of tumor extent. In clinical, CE-MRI through injection of GBCAs renders superior discrimination between tumor and the invaded healthy soft-tissue, and hence has become an indispensable technique in NPC delineation for radiotherapy purpose. At present, however, no study has been proposed to eliminate the GBCAs for NPC patients.

Challenges in model generalizability

DL algorithms are data-driven. The performance of DL models largely relies on the homogeneity of training and testing data (Long et al., 2013). Recently, Roberts et al. (Roberts et al., 2021) analyzed 415 CT or X-Ray based studies on detection and prognostication of COVID-19. They found none of the models were of potential clinical use, and the underlying data bias is a key cause of failure. Compared to CT or X-Ray images, MRI images present apparent inter-center heterogeneity due to different scanners, imaging protocols, as well as potential population demographics (Liu et al., 2020b). The inherent discrepancies in multi-center data challenge the wide application of GFCE-MRI models.

Challenges in clinical evaluation of synthetic GFCE-MRI for radiotherapy applications

Although several studies have been proposed to synthesize the GFCE-MRI for various applications (such as tumor detection, diagnosis, and treatment), most of these studies only developed technical methods to synthesize the GFCE-MRI and evaluated the synthetic GFCE-MRI using quantitative metrics such as mean absolute error (MAE),

peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). No study was focused on clinical evaluations of the synthetic GFCE-MRI, which is of vital importance for bench-to-bedside application of the GFCE-MRI in real world.

1.6 Objectives of our study

To tackle the above-mentioned challenges, in this work, we aim to develop and evaluate a clinical applicable DL assisted GFCE-MRI synthesis technique for NPC radiotherapy applications. Specifically, we have three objectives:

Objective 1: to develop a multimodality-guided synergistic neural network (MMgSN-Net) for GFCE-MRI synthesis in patients with NPC.

Objective 2: to assess and improve the MMgSN-Net model generalizability using multi-center data.

Objective 3: to comprehensively evaluate the potential clinical efficacy of the proposed GFCE-MRI technique in radiotherapy applications.

1.7 Thesis layout

This thesis first introduced the background of our research, including the basics of NPC, the safety issues of GBCAs, and previous DL assisted methods to synthesize the GFCE-MRI for providing a CE-MRI alternative to eliminate the use of GBCAs. In section 1, three major challenges (section 1.5) and the objectives (section 1.6) to tackle these challenges were figured out. In second, third and fourth sections, detailed methods that applied to achieve the three objectives and corresponding results will be illustrated.

Next, a discussion section (section 5) will be made to figure out the significance and limitations of our present work. Finally, section 6 is a conclusion to summarize our current research.

2. Development of a GFCE-MRI technique for NPC patients

2.1 Abstract

Purpose: To investigate a novel deep-learning network that synthesizes GFCE-MRI from multimodality contrast-free MRI for NPC patients.

Methods and Materials: This experiment presents a retrospective analysis of multi-parametric MRI, with and without contrast enhancement by GBCAs, obtained from 64 biopsy-proven NPC patients treated at Queen Elizabeth Hospital. A MMgSN-Net was developed to leverage complementary information between contrast-free T1w and T2w MRI for GFCE-MRI synthesis. 35 patients were randomly selected for model training, whereas 29 patients were employed for model testing. The synthetic images generated from MMgSN-Net were quantitatively evaluated against real GBCA-enhanced T1w MR images using a series of statistical evaluating metrics, which include mean absolute error (MAE), mean squared error (MSE), structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR). Qualitative visual assessment between the real and synthetic MRI was also performed. Effectiveness of our MMgSN-Net was compared with three state-of-the-art deep-learning networks, including U-Net, CycleGAN, and Hi-Net, both quantitatively and qualitatively. Further, a Turing test was carried out by seven board-certified radiation oncologists from four hospitals for assessing authenticity of the synthesized GFCE-MRI images against the real GBCA-enhanced T1w MRI.

Results: Results from the quantitative evaluations demonstrated that our MMgSN-Net

outperformed U-Net, CycleGAN and Hi-Net, yielding the top-ranked scores in averaged MAE (44.50 ± 13.01), MSE (9193.22 ± 5405.00), SSIM (0.887 ± 0.042), and PSNR (33.17 ± 2.14). Further, the mean accuracy of the seven readers in the Turing tests was determined to be 49.43%, equivalent to random guessing (i.e., 50%) in distinguishing between real GBCA-enhanced T1-weighted and synthetic GFCE-MRI. Qualitative evaluation indicated that MMgSN-Net gave the best approximation to the ground-truth images, particularly in visualization of tumor-to-muscle interface and the intra-tumor texture information.

Conclusions: Our MMgSN-Net was capable of synthesizing highly realistic GFCE-MRI that outperformed the three comparing state-of-the-art networks.

2.2 Introduction

NPC is a highly infiltrative and radiosensitive malignancy that located in an intricated nose-pharynx ministry (Lin et al., 2015). Radiotherapy is currently the mainstay therapeutic remedy, enabling non-invasive cancer eradication while protecting surrounding healthy tissue. Accurate tumor delineation is the foremost prerequisite for successful radiotherapy treatment, which, however, is particularly challenging in NPC in view of its deeply infiltrative nature. As a soft-tissue mass, NPC presents a high tendency to invade nearby healthy soft tissues, neural structures and bony skull base, obfuscating physicians for accurate assessment and delineation of tumor extent.

CE-MRI through injection of GBCAs renders superior discrimination between tumor and the invaded healthy soft-tissue, and hence has become an indispensable

technique in NPC delineation for radiotherapy purpose. Nevertheless, a number of safety concerns associated with bioaccumulation of GBCAs have recently been raised in the medical community (Broome et al., 2007; Grobner & Prischl, 2007; Kanda et al., 2015b; Kanda et al., 2014; Kleesiek et al., 2019b; Marckmann et al., 2006; Nguyen et al., 2020b; Olchowy et al., 2017; Thomsen, 2006; Wong et al., 2020).

Accumulated evidence in the body of literature since 2006 has indicated that gadolinium exposure has been strongly associated with an elevated risk of nephrogenic systemic fibrosis, which is a serious fibrotic disease of skin, joints, eyes and internal organ, in patients with renal deficiencies (Broome et al., 2007; Grobner & Prischl, 2007; Marckmann et al., 2006; Thomsen, 2006). More recent studies have highlighted bioaccumulation of previously administrated GBCAs in areas of dentate nucleus and globus pallidus within the brain on "contrast-free" T1w MRI images (Kanda et al., 2014; Nguyen et al., 2020b; Olchowy et al., 2017), regardless of patient's kidney function (Kanda et al., 2015b). Of note, these findings triggered abolishment of linear GBCAs in European countries in 2017 (Kleesiek et al., 2019b). Although the use of macrocyclic GBCAs could mitigate the risk of undesirable gadolinium accumulation, the mechanism of gadolinium uptake and deposition in patients is yet to be thoroughly elucidated, and there is a worldwide interest to minimize the administration of GBCAs whenever appropriate (Wong et al., 2020). Further, a portion of cancer patients, particularly the elderly who are at greater risk of developing kidney malfunctions, may be considered ineligible for GBCA injection for safety concerns. Considering all these, it is imperative to provide contrast-agent-free alternatives to the community, in the hope of replacing the use of GBCA-enhanced MRI in the long run.

In recent years, DL assisted image synthesis has been caught in the spotlight of attention in the medical domain (Liang et al., 2019; Ren et al., 2021). The capability of deep neural networks in unraveling complex tumor-related characteristics (Amin et al., 2018; Saba et al., 2020; Shkolyar et al., 2019) has motivated scientists to synthesize GFCE-MRI images from contrast-free MR images for brain cancer patients (Gong et al., 2018b; Kleesiek et al., 2019b). For instance, Gong et al. (Gong et al., 2018b) developed a U-shape DL neural network that concatenated GBCA-free (0% dose) T1w and GBCA-low (10% dose) CE-MRI brain images for synthesizing GFCE-MRI images as if it were generated from full dose of GBCA. Results from their study demonstrated feasibility of DL to capture contrast enhancement information from GBCA-full CE-MRI images and synthesize GFCE-MRI images with adequate image quality. On this ground, Kleesiek et al. (Kleesiek et al., 2019b) subsequently devised a three-dimensional Bayesian neural network that concatenated a total of 10 different MR modalities for generating GFCE-MRI images, confirming the role of DL network in utilizing diverse contrast-free imaging modalities for image synthesis. While these findings were promising, these existing DL networks have deficiencies in leveraging complementary information between input imaging modalities. Impact of this limitation on the network performance can be more prominent in the case of deeply infiltrative NPC due to the intricated relationship of pixel intensity between imaging modalities (C. Li et al., 2019).

In this study, we, for the first time, developed a novel MMgSN-Net that is capable of optimizing complementary features between multiparametric MR modalities, including contrast-free T1w and T2w images, for GFCE-MRI synthesis. Effectiveness

of our MMgSN-Net was compared quantitatively against several state-of-the-art DL models via a series of evaluating metrics. The authenticity of our synthesized GFCE-MRI was assessed by seven board-certified radiation oncologists from four hospitals via the Turing tests. To our best knowledge, we were the first to demonstrate the feasibility of GFCE-MRI synthesis in the context of NPC disease. The success of this study would provide the community with an effective contrast-agent-free alternative for NPC tumor delineation in future.

2.3 Methods and materials

2.3.1 Patient data

Multi-parametric MR images, including T1w, T2w and CE-MRI, were retrospectively retrieved from 64 biopsy-proven (Stage I-IVb) NPC patients who received radiotherapy at Queen Elizabeth Hospital between 2012 and 2016. Patient consent was waived due to the retrospective nature of this study. All MR images were acquired under a 1.5 Tesla MRI scanner (Avanto, Siemens, Germany). Acquisition parameters for the T1w and ceT CE-MRI images include: repetition time (TR): 562–739 ms; echo time (TE): 13–17 ms; matrix: 256–320; slice thickness: 3.3–4.0 mm; voxel size 0.75–0.94 mm. In particular, the CE-MRI images were acquired less than 30 seconds post GBCA injection (Gd-DOTA, 0.2 ml/kg). The T2w MR images were acquired using the short tau inversion recovery (STIR) sequence with the following acquisition parameters: TR: 7640 ms; TE: 97 ms; inversion time: 165 ms; matrix: 320; slice thickness: 4.0 mm; voxel size 0.75 mm.

2.3.2 MMgSN-Net architecture

The proposed MMgSN-Net was configurated for GFCE-MRI synthesis. The MMgSN-Net consists of five key modules: multimodality learning module, synergistic guidance system (SGS), self-attention module, multi-level module, and discriminator. **Figure 2-1** illuminates the overall architecture of the MMgSN-Net. Detailed descriptions of each module are presented as follows:

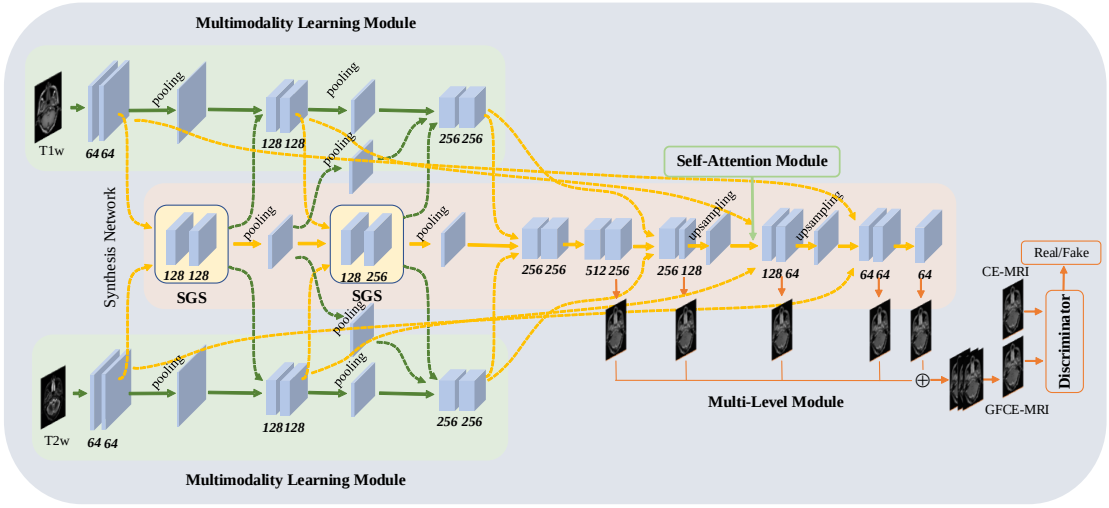

Figure 2-1. The framework of the proposed MMgSN-Net for GFCE-MRI synthesis. It consists of five key components: the multimodality learning module, synergistic guidance system, self-attention module, multi-level module, and discriminator. SGS: synergistic guidance system.

A. Multimodality Learning Module

This module was devised to unravel tumor-related imaging features from each of the input MR modalities, overcoming the limitation of single modality-based GFCE-MRI synthesis. As indicated in **Figure 2-1**, it contains two channels for the two studied imaging modalities (T1w and T2w), each channel consists of three convolution blocks and two pooling layers. The convolution layers inside the convolution blocks are

followed by batch normalization to standardize the extracted features using the mean and standard deviation of the extracted features. After batch normalization, the activation function *LeakyRelu* (Xu et al., 2015) was utilized to introduce non-linearity into the extracted features. The learned features were downsampled using 2 × 2 max-pooling layers. To fuse the extracted information from T1w and T2w modalities, we generated 64, 128, and 256 features from the first, second, and third convolution block, respectively.

B. Synergistic Guidance System

This system was specifically designed for leveraging complementary information between the two studied imaging modalities in a synergistic manner.

To fuse the learned information from multimodality learning modules, a common strategy is to directly concatenate the information to different channels as input. Alternative combination methods include pixel-wise summation, pixel-wise product, and pixel-wise maximization. Inspired by Zhou et al. (Zhou et al., 2020), we first used pixel-wise summation, pixel-wise product, and pixel-wise maximization separately to generate different fused features. Subsequently, we concatenated them as different channels followed by a convolution layer to adaptively select useful complementary information for final GFCE-MRI synthesis. Except that, there are some differences from Zhou's work.

First, in Zhou's work, separate information extractors learn the features from each input modality individually, and the extractors cannot communicate with each

other, which may limit complementary information learning. Inspired by the knowledge distillation concept (Hinton et al., 2015; C. Li et al., 2019), where a master network modulates the learning activity of an assistant network, we used the SGS as supervisor to fuse the learned information from each modality, after fuse operation, the output features from SGS contain the information of both T1w and T2w MR. Then we fed the fused information back to the next convolution block of the multimodality learning module to guide complementary information selection. In this way, the multimodality learning module can aware the information from the other modality, and the power of each individual multimodality learning module was further harnessed by communication and cooperation among the two modules in learning the complementary information for GFCE-MRI synthesis. The fused features were not only fed directly back to the second convolution block of each input channel in the multimodality learning model, but also sent to the third convolution block via the adoption of an additional pooling layer optimizing the size of output features from the first SGS. Second, our MMgSN-Net contains only two SGSs and two pooling layers that fuse and down-sample the extracted features, acting as the encoders of the synthesis network. The size of the SGS filters is 3 × 3, and the number of filters for the first and second SGS is 128 & 128, and 128 & 256, respectively. Third, the extracted features from the multimodality learning module was fed into the SGS without any pooling operation to avoid removal of critical information prior to feature fusion.

C. Self-Attention Module

In a convolutional neural network, the large tissue across intra-slice image regions are

captured by the convolution operator. As the field of the convolution operator is merely locally receptive, optimization algorithms may encounter difficulty in searching for the optimum parameter values when capturing the large size tissues (Zhang et al., 2019). Two possible solutions are either using multiple convolution layers or increasing the size of the convolution kernels. However, both solutions would degrade the computational efficiency. An optimal balance between the ability to capture the large size information and the computational efficiency can be achieved by the self-attention mechanism (Cheng et al., 2016; Parikh et al., 2016; Vaswani et al., 2017), which calculates the response at a position as a weighted sum of the features at all positions.

For GFCE-MRI synthesis, the NPC tumors can be highly aggressive, which presents a high tendency to invade nearby healthy tissues like neural structures and bony skull base. The size of tumor sometimes can be large and exists across different image regions. With limited convolutional kernel size, the algorithms may encounter difficulty in capturing this large structural information, for example, the shape of infiltrative tumor. So, in MMgSN-Net, a self-attention module was introduced to capture the large size information across image regions, enabling MMgSN-Net to faithfully preserve the shape of large anatomic structures. The self-attention module was of the same type as that used in (Zhang et al., 2019), and was inserted between the second and third convolution block of the synthesis network decoder.

D. Multi-Level Module

Multi-level feature integration has been widely applied in areas of image segmentation and edge detection. Several studies (Long et al., 2015; Xie & Tu, 2015; Zhang et al.,

2018) have shown that integrating features from multiple deep layers can improve the performance in image segmentation and, more remarkably, in edge detection. In GFCE-MRI image synthesis, edge information is critical for discriminating the tumor from surrounding normal tissues. Thus, a multi-level module was utilized in this study to aggregate the multi-level features. In our MMgSN-Net, we performed upsampling for the output features on each side of the decoders to the size of the output image. Subsequently, we fused the up-sampled features through a concatenation operation and applied a 1 × 1 convolution layer for final output generation.

E. Discriminator

A discriminator was utilized to distinguish synthetic images from real CE-MRI images, thus to improve the GFCE-MRI synthesis performance through adversarial learning. An overall structure of the discriminator is illustrated in **Figure 2-2**. This is a "PatchGAN"-based (Isola et al., 2017; Zhou et al., 2020; Zhu et al., 2017) discriminator that classifies input images based on whether the image patches are real or fake (i.e. synthetic). Different from regular GAN discriminator that maps an input image to single "real" or "fake" output, the PatchGAN-based discriminator maps an input image P to a M × N size output Q (in this study, M = 16, N = 14), all pixels in Q are labelled with "real" (for real input P) or "fake" (for synthetic input P). For each pixel in Q, we can trace back to its receptive field. Here, the receptive field means the "patch" that needs to be classified (for example, the dotted patches in P). The final image authenticity will be determined by averaging the M × N results in Q. One advantage of the PatchGAN-based discriminator is that it has fewer parameters than a full image discriminator (Isola

et al., 2017). We set the batch normalization momentum as 0.8 and the *LeakyReLu* slope as 0.2. For the first four convolutional layers, we set the filter stride to 2 and padding to 1.

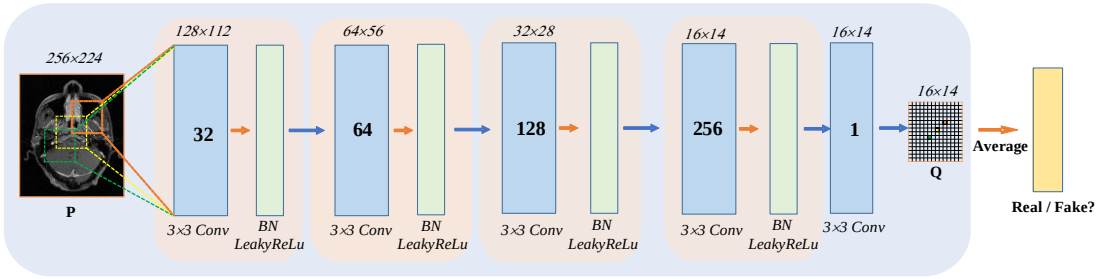

Figure 2-2. Schematic illustration of the PatchGAN-based discriminator, which consists of three iterative operations: 3 × 3 Conv, BN, and LeakyReLu. Numbers in blue box represent output feature numbers, and numbers at the top of the input image P and output Q, and blue box indicate the output feature size. The orange, yellow, and green points in output Q show the output results generated by the orange, yellow, and green dotted patches in input P, respectively. Conv: Convolutional layer; BN: Batch normalization.

2.3.3 Implementation details

All the T1w, T2w and CE-MRI images for each NPC patient were acquired for radiotherapy purpose and were well-aligned. Rigid registration was applied to fine-tune the alignment, when necessary. Triangle thresholding (Zack et al., 1977) was performed to eliminate background noise from all MR images, which may otherwise be mistakenly learned by the deep learning network and lead to model performance degradation. A total of 35 patients were used for model training, whereas 29 patients were employed for model testing. Two-dimensional axial slices with a matrix size of 256 × 224 were adopted to acquire knowledge information from the T1w and T2w images for mapping

the CE-MRI images. Prior to the model training, all images were linearly normalized to a range of (-1,1). The T1w and T2w MR images were used as inputs to the network, and the CE-MRI images were used as learning targets.

The *L1* loss between the synthetic GFCE-MRI and the corresponding real GBCA-enhanced T1w MR images was deployed as the loss function of our synthesis network. MSE loss was used as the loss function of the PatchGAN-based discriminator for distinguishing between real and fake patches. The Adam algorithm was utilized to optimize the generated model. The network was trained under a fixed learning rate of 0.0002 with 200 epochs, with the batch size of 1. The code was implemented in the PyTorch library using an NVIDIA RTX 3090 graphic card.

2.3.4 Model evaluation

The effectiveness of our MMgSN-Net was assessed quantitatively using a series of evaluating and compared against three state-of-the-art image synthesis networks: CycleGAN (Zhu et al., 2017), U-Net (Ronneberger et al., 2015), and Hi-Net (Zhou et al., 2020). Besides, Turing tests were conducted by seven board-certified oncologists for examining authenticity of the synthesized GFCE-MRI images against the real GBCA-enhanced T1w MR images. Furthermore, a qualitative evaluation was carried out by visual inspection of the real and fake images. The three comparing networks are described as follows.

1) **CycleGAN** (Zhu et al., 2017). This network allows for training without the need of paired image data, which can alleviate data shortage problem during image synthesis. However, Li et al. (Li, et al., 2020) reported that the use of a paired dataset,

compared to unpaired dataset, in CycleGAN training led to an improved model performance. In this study, therefore, we utilized a paired dataset for training. The CycleGAN network, which only supports single input channel, was applied for model training using T1w and T2w images separately, referred to as CycleGAN_T1w, CycleGAN_T2w, respectively.

2) **U-Net** (Ronneberger et al., 2015). This network uses a mirrored encoder–decoder architecture to acquire knowledge information for input-to-output image mapping. As a renowned DL neural network, U-Net was applied in the two previous studies on GFCE-MRI synthesis (Gong et al., 2018b; Kleesiek et al., 2019b), which are the only publications found in the literature. In this study, therefore, we compared our MMgSN-Net against this U-Net for GFCE-MRI synthesis. To determine which input imaging modality contributes to more information for GFCE-MRI prediction, we first separately used the T1w and T2w images as input (U-Net_T1w, U-Net_T2w), and combined both the T1w and T2w images through different channels (U-Net_T1w+T2w).

3) **Hi-Net** (Zhou et al., 2020). This network shares similar characteristics of our MMgSN-Net in that it allows for multiple inputs of different modalities and deploys two autoencoder-like structures to extract the modality-specific features. In this study, both T1w and T2w images were used as input for Hi-Net training.

Quantitative evaluation: Four widely-adopted evaluating metrics in areas of medical imaging synthesis (Frangi et al., 2018; Huynh et al., 2015; Nie et al., 2016; Nie et al., 2017), including MAE, MSE, SSIM, and PSNR, were used in this study to

quantitatively evaluate the model performance. These metrics are expressed below:

$$MAE = \frac{1}{N}|y(x) - g(x)|, \quad (2\text{-}1)$$

$$MSE = \frac{1}{N}(y(x) - g(x))^2, \quad (2\text{-}2)$$

$$PSNR = 10log_{10}^{\left(\frac{L^2}{MSE}\right)}, \quad (2\text{-}3)$$

$$SSIM = \frac{(2\mu_{y(x)}\mu_{g(x)}+c1)(2\sigma_{y(x)g(x)}+c2)}{(\mu_{y(x)}{}^2+\mu_{g(x)}{}^2+c1)(\sigma_{y(x)}^2+\sigma_{g(x)}^2+c2)}, \quad (2\text{-}4)$$

where N is the number of pixels in each image slice; $y(x)$ and $g(x)$ denote the ground truth image and synthetic GFCE-MRI image, respectively. $\mu_{y(x)}$, $\mu_{g(x)}$ and $\sigma_{y(x)}$, $\sigma_{g(x)}$ are the means and variances of the ground truth image and the synthetic image, while $\sigma_{y(x)g(x)}$ is the covariance of $y(x)$ and $g(x)$. $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ are two variables used to stabilize the division by the weak denominator, and L is the dynamic range of the pixel values. Here, $L = 4095$, $k_1 = 0.01$, and $k_2 = 0.03$ were set by default.

Qualitative evaluation: To visually evaluate the image quality of the synthetic GFCE-MRI images, qualitative evaluation was conducted by visually analyzing the synthetic GFCE-MRI images against the input T1w, T2w and ground truth CE-MRI images. Tumor regions were zoomed in for better comparison. In addition, difference map between the ground truth CE-MRI and the synthesized GFCE-MRI by our MMgSN-Net is illustrated for visualizing uncertainties in relation to GFCE-MRI synthesis.

Turing test — Clinical evaluation: The Turing test is a long-established test in areas of artificial intelligence for determining the capability of a machine to exhibit intelligent

human behavior (Kleesiek et al., 2019b; McDermott, 2007). In this study, we deployed the Turing test to assess authenticity of the synthetic GFCE-MRI images generated by our MMgSN-Net. Seven board-certified radiation oncologists from four hospitals participated in discriminating the synthetic GFCE-MRI images from the real CE-MRI images. In an attempt to balance the clinical workloads of the participating oncologists, we randomly chose 5 patients from our test set for the Turing test. For each patient, we randomly selected 10 tumor-bearing image slices (5 ground truth CE-MRI images plus 5 paired synthetic GFCE-MRI images) and presented them to the participating oncologists in a random order. The oncologists were blinded with regard to the relative proportions of ground truth and synthetic images. Additionally, they were asked to provide justifications when determining a synthetic case, allowing us to realize potential limitations of our MMgSN-Net.

2.3.5 Ablation study

To identify the importance of the key components in our MMgSN-Net, three ablation studies were conducted. First, to evaluate the importance of the SGS, we replaced it with the concatenation operation. The learned features from individual multimodality learning modules were directly concatenated without performing feature selection. Second, to validate the importance of the multi-level module, we compared the synthesis performance of full MMgSN-Net with that in an absence of the multi-level module. Third, to verify the importance of the self-attention module, we removed it and compared the resulting version of MMgSN-Net with the full version.

2.4 Results

2.4.1 Quantitative evaluation

Table 2-1 summarizes the results of quantitative comparisons between our MMgSN-Net and the comparing state-of-the-art DL networks for both whole image and tumor regions, in aspects of MAE, MSE, PSNR, and SSIM. For MMgSN-Net, the mean (± standard deviation (SD)) of the MAE, MSE, PSNR, and SSIM for the synthesized GFCE-MRI images relative to the ground truth CE-MRI images were calculated to be 44.50 ± 13.01, 9193.22 ± 5405.00, 0.887 ± 0.042, and 33.17 ± 2.14 for whole image and 110.31 ± 20.69, 25924.77 ± 10385.70, 0.706 ± 0.073, 28.74 ± 1.52 for tumor regions, respectively. Of note, our MMgSN-Net significantly outperformed all the comparing networks in all studied aspects ($p < 0.05$). Among the comparing state-of-the-art networks, on the other hand, U-Net obtained the best performance in all four evaluating aspects, while the CycleGAN models (both CycleGAN_T1w and CycleGAN_T2w) underperformed the others.

Overall, in comparison with the state-of-the-art networks, our MMgSN-Net achieved outstandingly, with mean MAE improvements of 13.07% versus the Hi-Net, 3.47% versus the multi-channel U-Net, 31.32% versus the CycleGAN_T1w, and 30.40% versus the CycleGAN_T2w.

Table 2-1. Quantitative error evaluation of different deep learning models for GFCE-MRI synthesis. ↑ indicates that a larger number represents better performance, ↓ indicates that a smaller number represents better performance. MAE, mean absolute error; MSE, mean squared error; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index; SD, standard

deviation.

		MAE ± SD (↓)	MSE ± SD (↓)	SSIM ± SD (↑)	PSNR ± SD (↑)
U-Net_T1w	Whole image	50.39 ± 13.70	11934.18 ± 5878.76	0.864 ± 0.042	31.91 ± 1.91
	Tumor regions	127.20 ± 19.01	34168.37 ± 10137.90	0.637 ± 0.063	27.47 ± 1.23
U-Net_T2w	Whole image	47.32 ± 13.55	10474.32 ± 5591.32	0.877 ± 0.041	32.59 ± 2.18
	Tumor regions	117.47 ± 20.11	29532.56 ± 9824.42	0.679 ± 0.068	28.17 ± 1.47
U-Net_T1w+T2w	Whole image	46.10 ± 13.15	9596.54 ± 5360.18	0.886 ± 0.042	32.95 ± 2.08
	Tumor regions	112.89 ± 18.87	27218.09 ± 9711.72	0.700 ± 0.068	28.46 ± 1.33
CycleGAN _T1w	Whole image	64.79 ± 15.78	18198.07 ± 7790.22	0.799 ± 0.049	30.03 ± 1.83
	Tumor regions	164.18 ± 15.41	53467.99 ± 9147.11	0.495 ± 0.042	25.45 ± 0.76
CycleGAN _T2w	Whole image	63.94 ± 15.48	17445.77 ± 7467.58	0.802 ± 0.042	30.21 ± 1.83
	Tumor regions	156.84 ± 14.80	48520.38 ± 8652.91	0.514 ± 0.038	25.78 ± 0.77
Hi-Net	Whole image	51.19 ± 13.74	12088.02 ± 6098.83	0.862 ± 0.041	31.87 ± 1.94
	Tumor regions	126.38 ± 19.36	34004.66 ± 10066.85	0.648 ± 0.061	27.42 ± 1.13
MMgSN-Net	**Whole image**	**44.50 ± 13.01**	**9193.22 ± 5405.00**	**0.887 ± 0.042**	**33.17 ± 2.14**
	Tumor regions	**110.31 ± 20.69**	**25924.77 ± 10385.70**	**0.706 ± 0.073**	**28.74 ± 1.52**

2.4.2 Qualitative evaluation

Figure 2-3 illuminates visual comparisons between the ground truth CE-MRI images and synthesized GFCE-MRI images obtained by using the studied DL networks. For T1w and T2w input images, the tumor structure and adjacent muscle texture are not clearly discernible in the input T1w MR image **(Figure 2-3 (a))**, while the tumor edge is clearer in the input T2w MR image **(Figure 2-3 (b))**. For tumor delineation, the ground truth CE-MRI image obtained following the injection of GBCAs **(Figure 2-3 (c))** outperforms both the T1w and T2w images, clearly revealing the tumor structure and adjacent muscle texture.

Regarding the synthetic images generated from the three U-Net models, they are relatively blurry throughout the images **(Figure 2-3 (F)-(H))**. The tumor structure predicted by U-Net_T2w is more discernable than that obtained from U-Net_T1w

(Figure 2-3 (g) and (f), respectively). The joint T1w-T2w synthesized U-Net images **(Figure 2-3 (H))** achieves the best discriminability of tumor's morphology against the ground truth, compared to both U-Net_T1w and U-Net_T2w generated images.

With regard to the Hi-Net predicted GFCE-MRI images **(Figure 2-3 (E)),** the overall image quality was visually comparable to the ground truth image **(Figure 2-3 (C))**. Nevertheless, the tumor-to-muscle interface was not in a good agreement compared with the ground-truth images, while our MMgSN-Net **(Figure 2-3 (d))** achieved a satisfying approximation to the ground-truth **(Figure 2-3 (c)).**

For the two CycleGAN models **(Figure 2-3 (i) and (j))**, the tissue structures, such as the temporalis tendon and surrounding muscles, are the least discernable. Notably, the synthetic images predicted by our MMgSN-Net **(Figure 2-3 (D & d))** visually yields the best approximation to the ground-truth images, in particular to the tumor-to-muscle interface and the texture information, outperforming all the comparing networks. These qualitative findings are well in line with the results of quantitative evaluation.

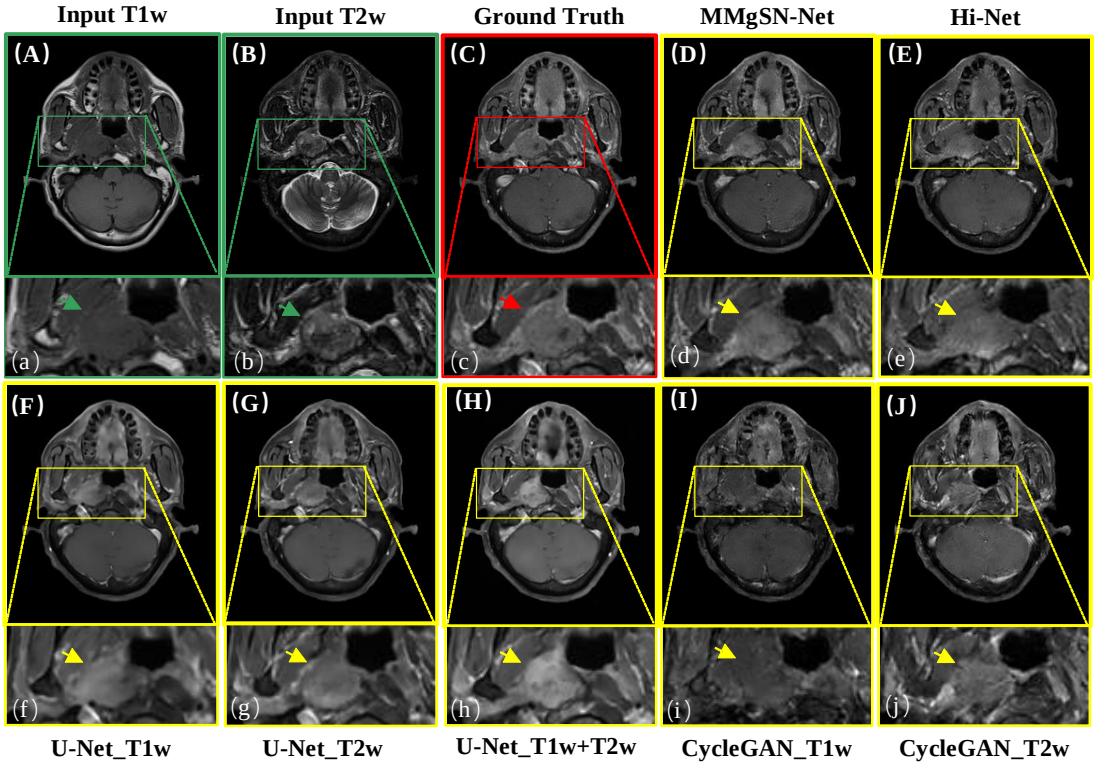

Figure 2-3. Visual evaluation of our MMgSN-Net and the comparing state-of-the-art networks for virtual contrast-enhanced T1-weighted MR synthesis. (A) and (B) are the input T1w MR image and T2w MRI image, respectively; (C) is the ground truth gadolinium-based contrast-enhanced T1-weighted MRI; other images are the synthetic results of different networks.

Figure 2-4 visualizes difference maps between the real CE-MRI images and synthetic GFCE-MRI images from different patients for visualizing uncertainties in relation to GFCE-MRI synthesis. A difference map window with a range of (0, 0.2) was set to clearly visualize the differences. It can be observed that prediction uncertainties most occurred at the edges between anatomic structures. Besides, structures of evenly-changing pixel values (such as the maxillary sinus and cerebellum) could be accurately predicted by our MMgSN-Net.

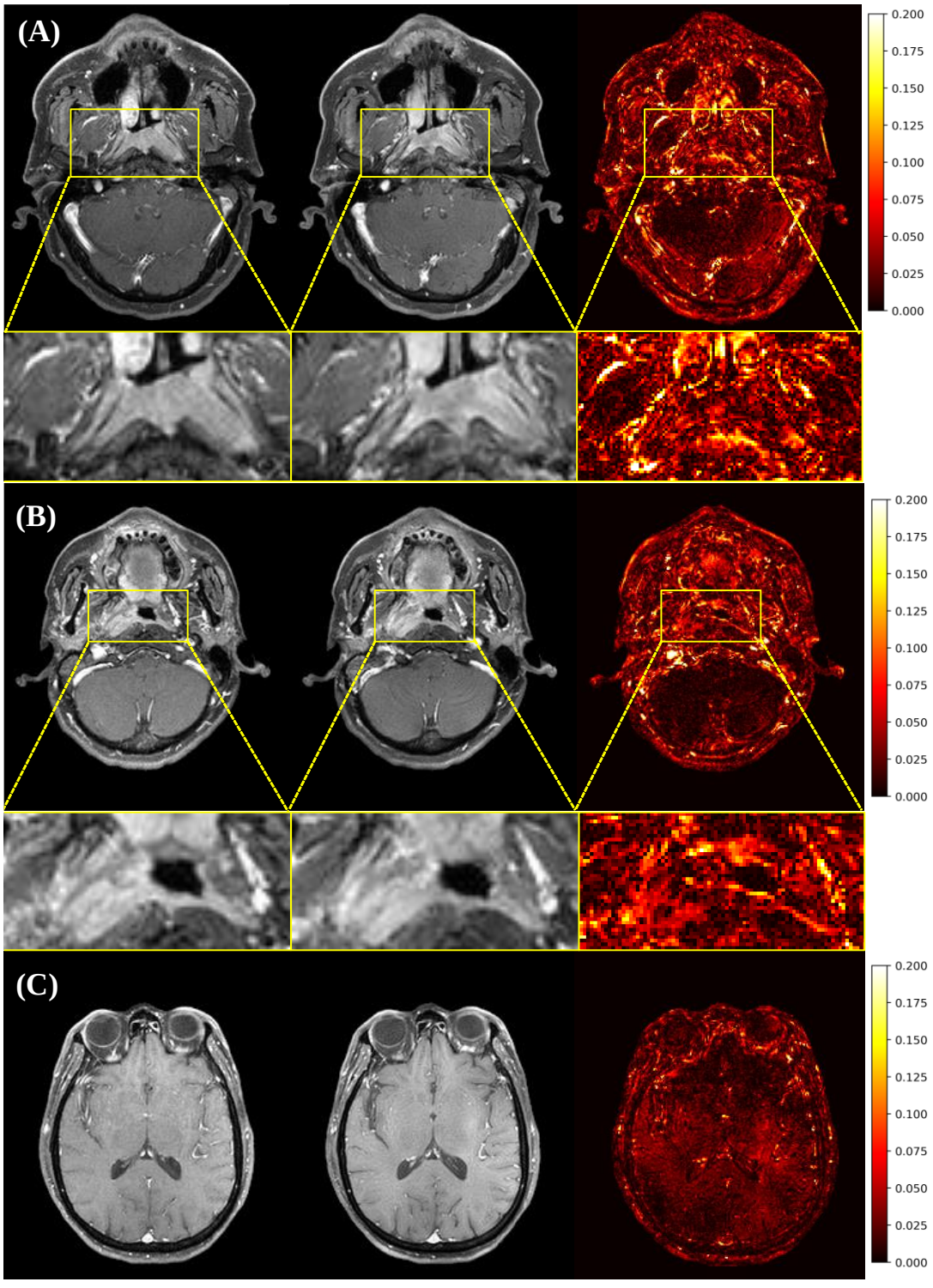
(A)
(B)
(C)
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Real CE-MRI
Synthetic GFCE-MRI
Difference Map

Figure 2-4. Difference Maps (third column) between the real CE-MRI images (first column) and the synthetic GFCE-MRI images predicted by our MMgSN-Net (second column). (A)-(C): different axial slices.

2.4.3 Turing test results

Table 2-2 summarizes quantitative results of the Turing tests from the 7 participating oncologists. In hospital 1, the two oncologists failed to differentiate between the real CE-MRI and GFCE-MRI images in approximately half of the cases, with an accuracy of 52% and 42% for oncologists 1 and 2, respectively. They reported that their decisions were mostly based on the clarity of the alveoli and blood vessels, as well as the texture of the muscles and cerebellum. In hospital 2, the two oncologists raised the difficulties in discriminating the real and fake images based on the irregularly shaped tumor structures. For this reason, they made their decisions according to the anatomical structures and image signal intensities during the Turing test, resulting in an accuracy of 58% and 52% for oncologists 3 and 4 from hospital 2, respectively. In hospital 3, discussion sessions were held between the oncologist 5 and 6, in view of the heavy clinical workload. An overall accuracy of 58% was reported based on their judgements. They reported that their decisions were made based on the differences between the parotid gland and non-vascular tissues. In hospital 4, the oncologist correctly identified only 13, leading to an accuracy of 26%, and was unable to make decisions for another 13 images. Overall, the average accuracy of the 7 oncologists was 49.43%, which is in close approximation to a random guess accuracy (i.e., 50%).

Table 2-2. Results of the Turing test conducted by the 7 clinical radiation oncologists from 4

hospitals.

Hospital	Radiation Oncologist	Evaluation Results		Percentage
Hospital 1	Oncologist 1	Correct:	26	52%
		Incorrect:	21	42%
		Give up:	3	6%
	Oncologist 2	Correct:	21	42%
		Incorrect:	20	40%
		Give up:	9	18%
Hospital 2	Oncologist 3	Correct:	29	58%
		Incorrect:	21	42%
		Give up:	0	0%
	Oncologist 4	Correct:	26	52%
		Incorrect:	24	48%
		Give up:	0	0%
Hospital 3	Oncologists 5 and 6	Correct:	29	58%
		Incorrect:	21	42%
		Give up:	0	0%
Hospital 4	Oncologist 7	Correct:	13	26%
		Incorrect:	24	48%
		Give up:	13	26%
	Average:	Correct:		49.43%
		Incorrect:		43.43%
		Give up:		7.14%

2.4.4 Ablation study

In the ablation studies, the MAE, MSE, SSIM, and PSNR values were found to be inter-correlated. For simplicity, therefore, only results of MAE was described here. First, after replacing the GSG with the concatenation operation, the MAE increased from 44.50 ± 13.01 to 45.43 ± 12.97 ($p < 0.05$), implying that the SGS contributed to accuracy improvement. Second, after excluding the multi-level module, the MAE increased from 44.50 ± 13.01 to 45.22 ± 13.04 ($p < 0.05$), suggesting that this multi-level module enhanced the synthesis performance of CE-Net. Third, after removing the self-attention module, the MAE increased from 44.50 ± 13.01 to 45.89 ± 13.02 ($p < 0.05$), indicating that self-attention module is helpful in capturing long-term

dependencies.

2.5 Discussion

In radiotherapy, GBCA-assisted CE-MRI has been considered essential for delineation of deeply infiltrative NPC neoplasm. A recent growing body of evidence regarding safety issues of GBCAs administration, however, has stimulated awareness of the community to investigate contrast-agent-free alternatives, in the hope of replacing the use of GBCA in the long run. A few DL models have been introduced up to the present, in brain diseases (Gong et al., 2018b; Kleesiek et al., 2019b). While satisfying in brain imaging, their models were deficient in leveraging complementary information between input imaging modalities. Impact of this deficiency in their models could be more detrimental in the case of deeply infiltrative NPC (Li et al., 2019). Herein, we, for the first time, developed a novel MMgSN-Net to compensate for this deficiency and investigated image synthesis in NPC. In this discussion, we attempted to highlight key findings of our results, scrutinize possible underlying reasons, and provide research community with potential directions in future.

Results from the quantitative evaluations demonstrated that our MMgSN-Net outperformed all the comparing networks for both whole image and tumor regions **(Table 2-1)**, yielding the top-ranked scores in averaged MAE (44.50 ± 13.01, 110.31 ± 20.69), MSE (9193.22 ± 5405.00, 25924.77 ± 10385.70), SSIM (0.887 ± 0.042, 0.706 ± 0.073), and PSNR (33.17 ± 2.14, 28.74 ± 1.52) for whole image and local tumor regions, respectively. This is in line with findings of our qualitative evaluation, where the synthetic images predicted by our MMgSN-Net **(Figure 2-3 (D & d))** visually

yielded the best approximation to the ground-truth images, in particular to the tumor-to-muscle interface and the intra-tumoral texture information, outperforming all the comparing networks. Similar to our MMgSN-Net, both U-Net_T1w+T2w and Hi-Net models deployed both T1w and T2w MR images as inputs for model training. A distinct difference of our network from these two comparing networks lies to its capability to leverage complementary information between each of the unique input imaging modalities, rather than using a simple additive concatenation of different input modalities. This may shed some light on the outstanding performance of our MMgSN-Net, compared with these two networks **(Table 2-1)**. Besides, the U-Net yielded the second best-performing model among the studied networks, as indicated in **Table 2-1**. We found that the synthetic images generated by U-Net were over-smoothed, leading to loss of detailed information, for instance, regarding the cerebellum and muscle texture, as illustrated in **Figure 2-3**. It could be partially attributed to the incapability of the L1 loss function for capturing high-frequency signals in MR images of NPC (Isola et al., 2017), where there are complex relationships among an ensemble of fine anatomic tissues in the nose-pharynx ministry. On the contrary, the CycleGAN gave rise to the worst model performance **(Table 2-1)**. To a degree, this may be explained by the limitation of the backward cycle adopted in the CycleGAN network. Although the backward cycle has been used to maintain cycle consistency, it increases number of training parameters, which may result in model underfitting given the small-sized training samples.

Intriguingly, it was observed that inputting single T2w MR images yielded better performance in both U-Net and CycleGAN networks than when using single T1w

MR images **(Table 2-1)**. A possible explanation would be related to the superiority of T2w MR images in revealing hyperintensity or inhomogeneity information on various pathologies (Cheng et al., 2016), such as in peripheral edema and tumor necrosis, which makes T2w MR images contribute to more valuable information on pathology-related contrast enhancement for GFCE-MRI synthesis, compared to contrast-free T1w MR images. This finding is also consistent with a brain tumor study conducted by Kleesiek et al., who reported that T2w MR images provided more useful information for GFCE-MRI synthesis (Kleesiek et al., 2019b). Another interesting observation was that the presence of the self-attention module in our MMgSN-Net architecture enhanced tumor edge detection in the synthetic images during our ablation study, implicating potential of our model in NPC delineation.

Although there are no studies on image synthesis for NPC in the literatures, comparisons between results of our study and previous works on brain diseases highlight the superiority of our MMgSN-Net model. Gong et al. (Gong et al., 2018b) reported a mean SSIM value of 0.85 ± 0.07 using a U-Net model that was trained on 10% GBCA-dose CE-MRI images and contrast-free T1w MR images of 10 patients with brain diseases. Kleesiek et al. (Kleesiek et al., 2019b) trained a 3D BayesUNet using multi-parametric MR modalities of 47 contrast-enhanced samples and obtained a mean SSIM of 0.862 ± 0.029. In models of these two publications (Cheng et al., 2016; Zhang et al., 2019), information in different input modalities was simply concatenated into different channels without emphasis on potential interaction of features between the modalities. In comparison, our MMgSN-Net achieved a higher mean SSIM of 0.887 ± 0.042 after training with 35 samples using both T1w and T2w MR images. To a large

extent, we inferred that this improvement in SSIM was mainly attributable to the capacity of our MMgSN-Net in unraveling complementary information from individual unique imaging modalities for GFCE-MRI synthesis.

The degraded accuracy shown in **Figure 2-4** may be, in part, explained by the imperfect alignment among the T1w, T2w and CE-MRI images. While it should be noted that existing image registration methods are still struggling to achieve one-to-one pixel correspondence and was found to be influential in medical image synthesis tasks (Han, 2017). The misalignment can lead to structural shift between input and target pairs, thus leading to inaccuracy during model training, since the model will be trained to make wrong prediction (Han, 2017). As a comparison, we directly used the data acquired from hospital system as input without any registration fine-tuning, we observed a performance decrease of 18.36%, 54.58%, 5.81% and 5.59% for MAE, MSE, SSIM and PSNR, respectively. An example of the influence of image registration is illustrated in **Figure 2-5**.

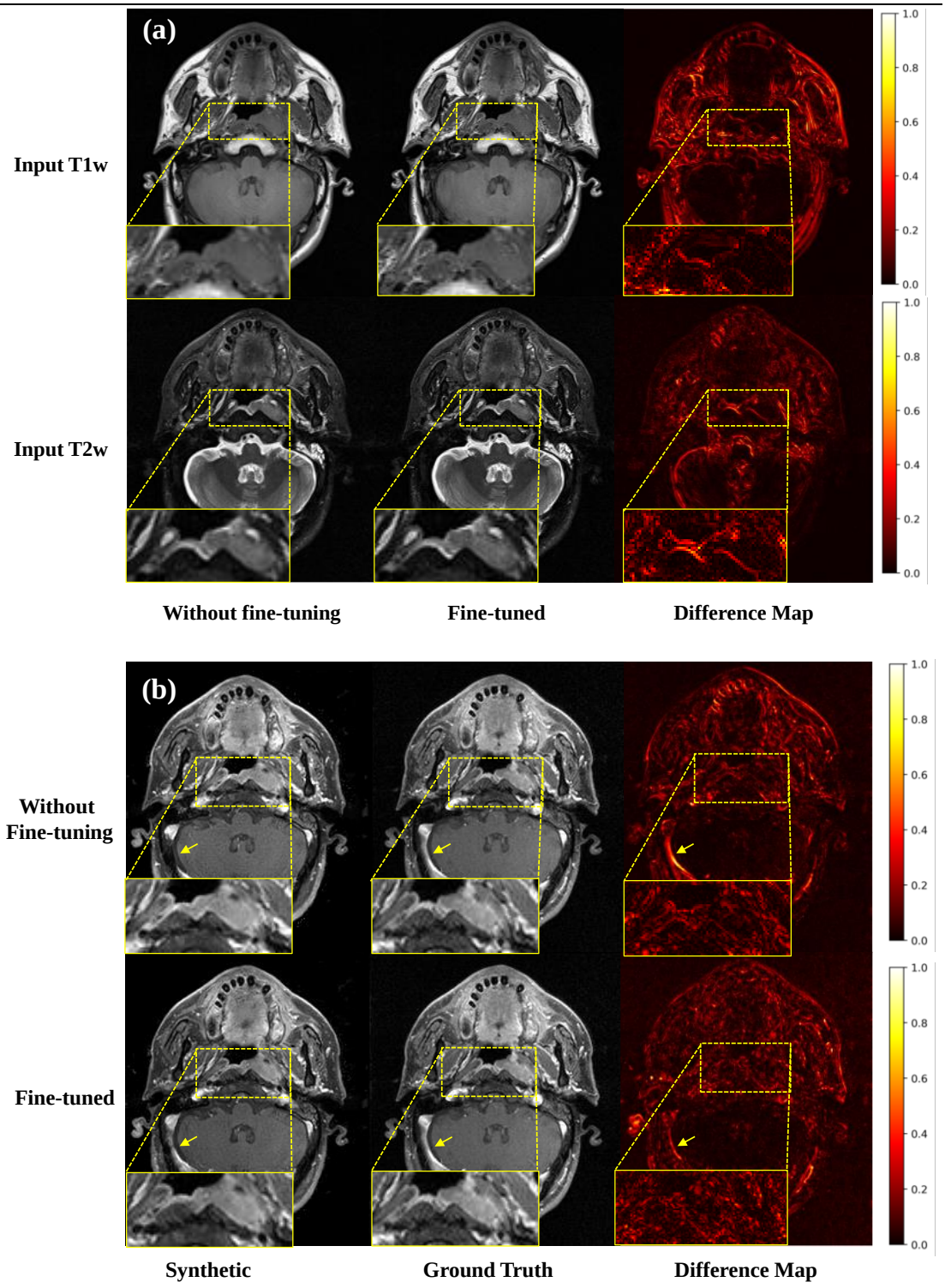

Figure 2-5. An example of the influence of image registration. (a): structural shift of input T1w (first row) and T2w (second row) between two image registration methods: registered from

hospital system without fine-tuning, and fine-tuned with rigid registration. (b): resultant variations caused by image registration. The first row and the second row show the difference between synthetic GFCE-MRI and ground truth CE-MRI of two registration methods.

Furthermore, results of the Turing test underscored the reliability of our MMgSN-Net. In a study conducted by Kleesiek et al. (Kleesiek et al., 2019b), two resident radiologists were invited to distinguish 10 synthetic MR images from another 10 real CE-MRI images, chosen in a random manner. The radiologists correctly discriminated between the real and synthetic images in 80% and 90% of cases, respectively. By contrast, in our work, seven experienced oncologists from multiple hospitals were merely able to correctly classify 49.43% of the presented images, suggesting high authenticity of our synthesized GFCE-MRI images. It is noteworthy that the high authenticity of our model can be observed in both tumor-bearing and tumor-free MR slices. In tumor-bearing slices, our MMgSN-Net model provided comparable tumor visualization as compared with the ground-truth (**Figure 2-3 (c) and (d)**). The degree of contrast enhancement is related to the density of capillary bed around the neoplasm (Mann et al., 2019), which is thought to be absent in normal tumor-free regions. In line with this line of thinking, our model also correctly predicts the non-enhanced information in tumor-free MR slices, as illustrated in **Figure 2-4 (C)**.

In spite of these exciting findings, our study has several limitations. Our network was trained and validated using a small-sized NPC data from the same MRI scanner at a single institution. Synthesis failure is likely to be observed with limited training samples for specific patients. An example of unsatisfactory case is shown in

Figure 2-6. In addition, intratumoral heterogeneity can be another impacting factor to the synthesis results. The intratumoral heterogeneity exists at the cellular level, and is highly influenced by its genetic background and surrounding micro-environment (Just, 2014). It causes heterogeneous tumor signal intensities of MR images (O'Connor et al., 2015), as shown in **Figure 2-6** where the arrows indicate the intratumoral heterogeneity in T1w, T2w, and ground truth CE-MRI images. Furthermore, another factor that may limit the performance of our synthesis network is that our network was trained with T1w and T2w MR images only. It is likely that T1w and T2w MR images may not provide complete information for synthesizing contrast enhancement for some structures such as sinus sigmoideus, as shown in red arrows in **Figure 2-6**. This problem can be potentially addressed by including more MRI modalities (such as diffusion-weighted MRI) as input to our network. While we also believe a homogeneous dataset is advantageous for model development, the generalizability of our results using a larger dataset from different scanners and medical centers is warranted to minimize the so-called "data bias" issue (Kazemifar, et al., 2021). This is currently being undertaken and would be considered as an extension of this study. Apart from this, although we invited a total of seven board-certified radiation oncologists for conducting the Turing test to assess the authenticity of the synthetic images, they were not asked to perform tumor delineation on the synthetic images, restricted by their existing heavy burdens in clinic. Nevertheless, this should be considered in future in order to further contextualize results of this study in aspects of NPC delineation. Since our network is a 2D network, which is likely to limit the performance on coronal and sagittal views (**Figure 2-7**), to extend the application scope of our network, in the long run, we would upgrade our

MMgSN-Net to 3D architecture and incorporate additional MR modalities for GFCE-MRI synthesis in future.

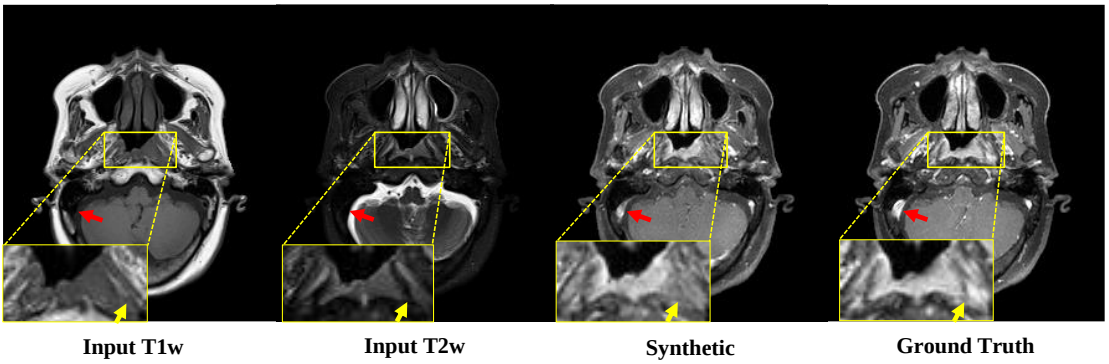

Figure 2-6. An example of a less satisfactory case. The images from left to right show input T1w, input T2w, the synthetic GFCE-MRI and ground truth CE-MRI, respectively. Yellow arrow shows the heterogeneous signal of tumor in different MR modalities.

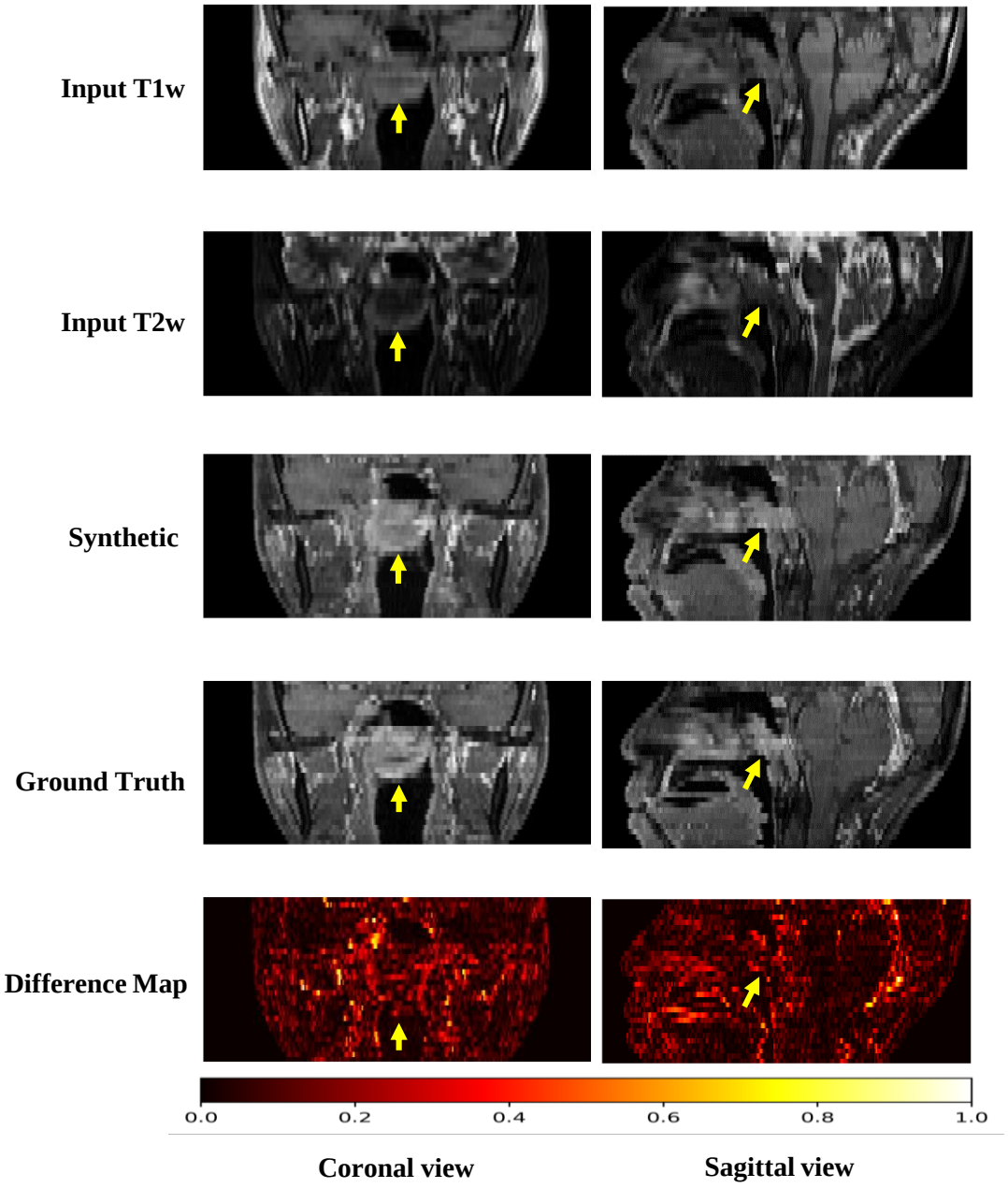

Figure 2-7. Illustration of coronal view (the first column) and sagittal view (the second column) of synthetic GFCE-MRI. from top to bottom: input T1w MR, input T2w MR, synthetic GFCE-MRI from the proposed method, ground truth CE-MRI and the difference map between GFCE

MRI and CE-MRI. Yellow arrows show the position of tumor.

2.6 Conclusion

In this study, we developed and evaluated a novel MMgSN-Net for GFCE-MRI synthesis for NPC patients. Our MMgSN-Net was capable of synthesizing highly realistic GFCE-MRI images in both quantitative and qualitative aspects and outperformed the three studied state-of-the-art networks. Moving forward, a larger multi-center cohort study is warranted to ensure model generalizability. Future works on tumor delineation on the synthetic images are recommended to further contextualize results of this study.

3. Evaluation and improvement of GFCE-MRI model generalizability

3.1 Abstract

Recently, deep learning has been demonstrated to be feasible in eliminating the use of GBCAs through synthesizing GFCE-MRI from contrast-free MRI sequences, providing the community with an alternative to get rid of GBCAs-associated safety issues in patients. Nevertheless, generalizability assessment of the GFCE-MRI model has been largely challenged by the high inter-institutional heterogeneity of MRI data, on top of the scarcity of multi-institutional data itself. Although various data normalization methods have been adopted in previous studies to address the heterogeneity issue, it has been limited to single-institutional investigation and there is no standard normalization approach presently. In this study, we aimed at investigating generalizability of GFCE-MRI model using data from seven institutions by manipulating heterogeneity of training MRI data under two popular normalization approaches. A MMgSN-Net was applied to map from T1w and T2w MRI to CE-MRI for GFCE-MRI synthesis in patients with nasopharyngeal carcinoma. MRI data from three institutions were used separately to generate three uni-institution models and jointly for a tri-institution model. Patient-based Min-Max and Z-Score normalization were applied for data normalization of each model. MRI data from the remaining four institutions served as external cohorts for model generalizability assessment. Quality of GFCE-MRI was quantitatively evaluated against ground-truth CE-MRI using MAE and

PSNR. Results showed that performance of all uni-institution models remarkably dropped on the external cohorts. By contrast, model trained using multi-institutional data with Z-Score normalization yielded improved model generalizability.

3.2 Introduction

NPC is a highly aggressive epithelial carcinoma originating in the mucosal lining of the nasopharynx, has long been prevalent in the population of East and Southeast Asia (Chang et al., 2021). Radiotherapy is currently the mainstay treatment modality for NPC, which achieved 66%-83% 5-year survival rate for NPC patients with radiotherapy alone (Xu et al., 2016). Precise tumor delineation is the most critical prerequisite for a successful radiotherapy treatment. CE-MRI, using GBCAs, has become an indispensable part in accurate NPC tumor delineation (Lee et al., 2018) in routine radiotherapy treatment planning practice. Nevertheless, emerging evidence has shown that NSF, a severe disease that can lead to joint contractures and immobility, has been strongly linked with the administration of GBCAs in renal failure patients (Holowka et al., 2019). Further evidence has shown that gadolinium accumulation in the dentate nucleus and globus pallidus has been observed in paediatric patients (Roberts et al., 2017). Apart from this, gadolinium deposition was also observed in patients with normal renal function (Roberts et al., 2016). The mechanism of gadolinium deposition in patients has not been fully elucidated, and the underlying long-term effects remain unclear. Therefore, there is a global consensus to minimize or avoid GBCA exposure to patients whenever possible (Holowka et al., 2019). Considering this, a GBCA-based CE-MRI alternative is desperately demanded.

Numerous efforts have been made to address the GBCA-associated safety issues. Worldwide interests have sparked recently in synthesizing GFCE-MRI, which serves similar purposes as the CEMRI, through deep learning approaches (Bône et al., 2021; Chen et al., 2022; Gong et al., 2018a; Kleesiek et al., 2019a; Xiao, et al., 2022; Luo et al., 2021a; Pasumarthi et al., 2021a; Xu et al., 2021a; Zhao et al., 2020a). However, current works have focused on model development or feasibility studies at different tumor sites using in-house datasets. It has been reported that the models trained with in-house dataset may perform poorly on datasets from external institutions (Jia et al., 2020; Liu et al., 2020a; Xing et al., 2018), which largely limits the wide application of proposed approaches. Therefore, a generalizable GFCE-MRI model is highly demanded in clinical practice, which extends the GFCE-MRI technique to a considerably wider range of hospitals for use.

Despite the urgent need for generalizable models, limited research has been conducted to investigate the underlying mechanism of model generalizability and the methods to improve the model generalizability, especially for the multi-parametric MRI images, presumably due to two key challenges: 1) high inter-institutional heterogeneity of MRI data; 2) scarcity of multi-institutional MRI data. The MRI images from different institutions often suffer from large domain shifts due to the use of diverse scanning parameters, scanners of different field strengths, as well as different patient demographics, leading to large distribution divergences such as means, standard deviations, and intensity ranges (**Figure 3-1**). These challenges have raised a growing concern of model generalizability developed using deep learning algorithms, which strongly rely on the assumption that the training data and testing data are independent

and identically distributed (i.i.d.) (Wang & Deng, 2018). In reality, however, the external MRI datasets are typically out-of-distribution (OOD) due to the abovementioned domain shift, incurring tremendous performance degradation of the trained models (Wang & Deng, 2018). To tackle this, a potential remedy to improve model generalizability is to integrate multi-institutional MRI images during model training to enlarge view of deep learning models (Dou et al., 2021; Lam, et al., 2022), which has been rarely reported in the literature, probably due to the scarcity of multi-institutional data for patient privacy protection. Another potential solution is to develop a generalizable network architecture by mapping data distributions from source domain to target domain (Wang & Deng, 2018; Wolleb et al., 2022), while these approaches are limited to specific domain datasets. As such, data normalization techniques have been widely used to improve the model performances in a range of application areas. Nevertheless, related research in multi-institutional setting that contain various real-world distributions of MRI data is severely scarce in the body of literature.

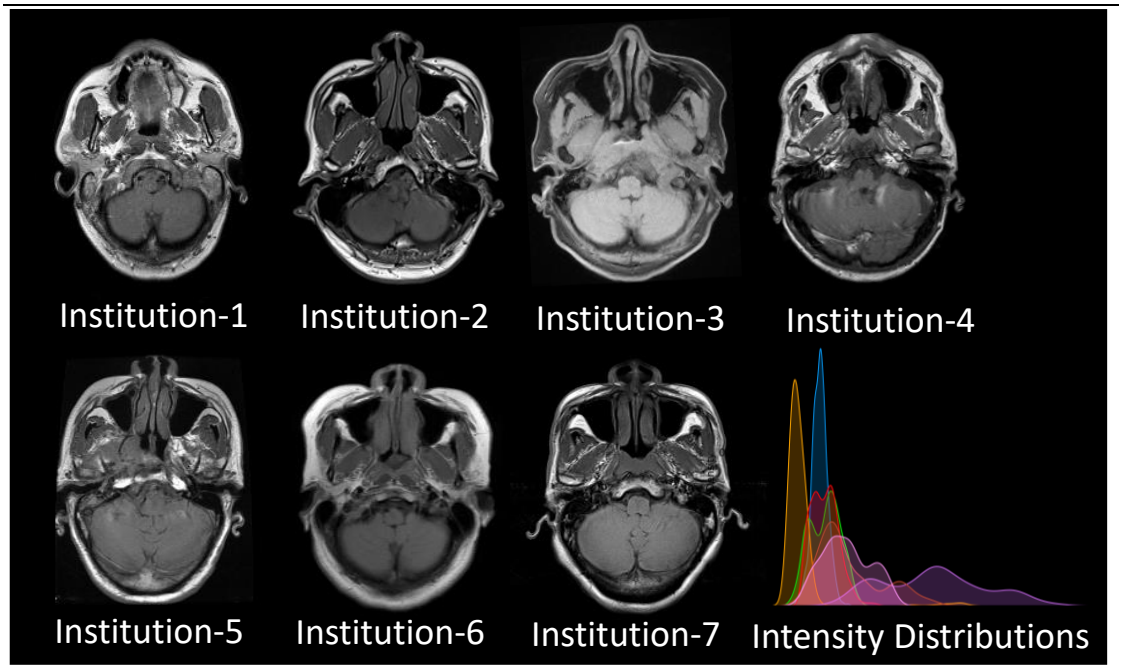

Figure 3-1. Illustration of heterogeneity of multi-institutional MRI data.

We consider minimize the distribution variations between training and external testing MRI data by using data normalization should be a practical approach to improve the model generalizability since it requires no model architecture improvement and retraining the model. In this study, we included MRI data from seven different institutions, aiming at investigating the GFCE-MRI model generalizability influenced by distribution difference between training and external testing data. Specially, we investigated: (i) how significant is the influence of different data normalization methods on the model generalizability; (ii) how significant is the degradation of external performance for models trained with single-institution MRI; and (iii) how significant is the improvement of external performance when using multi-institutional MRI for model development.

Compared to other tumor types such as brain and liver tumors, NPC is highly infiltrative with ill-defined tumor-to-normal tissue interface, which presents challenges to oncologists in NPC contouring. Hence, the success of this study may not only provide the medical community with better insights into the issue of GFCE-MRI model generalizability of NPC patients, but also may potentially be translated to other cancer types as well. To the best of our knowledge, this is the first multi-institutional investigation for GFCE-MRI synthesis. As a result, this study may have a far-reaching impact on the medical community to better understand the issue of model generalizability, establish a standard multi-institutional data normalization method, and further facilitate the development of generalizable GFCE-MRI models in the future.

3.3 Methods and materials

3.3.1 Patient data

A total of 256 NPC patients from seven medical institutions were retrospectively collected in this study. For fair comparisons, same number of patients (71 patients) were retrieved from Institution-1, Institution-2, and Institution-3, respectively for uni-institution and tri-institution model development, 18 patients, 9 patients, 9 patients, and 7 patients were retrieved from Institution-4, . . . , Institution-7, respectively for external testing to evaluate the model generalizability. T1w MRI, T2w MRI, and CE-MRI were collected for each patient. This study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HAHKW IRB), reference number UW21-412 and the Research Ethics Committee (Kowloon Central/Kowloon East), reference number KC/KE-18-0085/ER-

1. Due to the retrospective nature of this study, patient consent was waived. All images were acquired in the same position and automatically aligned. For model training, all images were resampled to the size of 256*224 using bilinear interpolation (Gribbon & Bailey, 2004). For Institution-1, Institution-2, and Institution-3, the 71 patients were randomly divided into 53 and 18 for model training and validation, respectively.

3.3.2 Study design

The overall idea of this study was first using the data collected from three different institutions (i.e., Institution-1, Institution-2, and Institution-3) to develop a series of separately and jointly trained models using different data normalization methods for investigating the GFCE-MRI model generalizability. The separately and jointly trained models were referred to as uni-institution models and tri-institution models, respectively. **Table 3-1** illustrated the overall study design.

Table 3-1. The overall study design. Min-Max and Z-Score normalization were used to normalize the datasets, and the multi-institutional datasets were trained separately and jointly to compare the model generalizability on four external datasets. Ins: Institution.

Normalization	Model name	Training			Testing			
		Ins-1	Ins-2	Ins-3	Ins-4	Ins-5	Ins-6	Ins-7
Min-Max	Uni-m1	✓			✓	✓	✓	✓
	Uni-m2		✓		✓	✓	✓	✓
	Uni-m3			✓	✓	✓	✓	✓
	Tri-M	✓	✓	✓	✓	✓	✓	✓
Z-Score	Uni-z1	✓			✓	✓	✓	✓
	Uni-z2		✓		✓	✓	✓	✓
	Uni-z3			✓	✓	✓	✓	✓
	Tri-Z	✓	✓	✓	✓	✓	✓	✓

1) Neural Network: The MMgSN-Net was used as the base network in this study. The MMgSN-Net is a 2D deep learning algorithm (Xiao, et al., 2022), which consists of five key modules: multimodality learning module, synthesis network, self-attention module, multi-level module, and a discriminator. The structure of the MMgSN-Net is illustrated in **Figure 2-1**. The T1w and T2w MRI were put into the multimodality learning module separately. The multimodality learning module was used to extract the modality-specific features. The extracted modality-specific features were put into the SGS in synthesis network for complementary feature selection and fusion. In the decoder of synthesis network, the fused features and the learned features from

multimodality learning modules were concatenated to different channels. The self-attention module and multi-level module were applied to capture the long-term dependencies and detect the edge information of the high-level features, respectively. A discriminator was utilized to distinguish the synthetic GFCE-MRI from ground-truth CEMRI, thus encouraging the synthesis network to generate more realistic GFCE-MRI.

2) Data Normalization: Data normalization plays a pivotal role in model development (Hu et al., 2022). It minimizes feature bias by transforming the features into a common space so that larger numeric feature values cannot dominate smaller numeric feature values (García et al., 2015). Currently different data normalization methods are applied in medical image translation tasks. The most popular two normalization methods are Min-Max (also called scaling) (Gajera et al., 2016) and Z-Score (Fei et al., 2021). These two normalization methods are also applied to different objects prior to training, i.e., dataset-based, patient-based, and single-image based normalization. In natural image tasks, most studies are 2D-based networks, which always use the statistical values of each single image or the whole dataset for data normalization (Liu et al., 2020a). For medical images, however, image and dataset-based normalization may not appropriate for clinical applications, especially for 3D volumes since the image-based normalization ignores the inter-slice adjacent information within a volume, which leads to contrast bias of generated images between two nearby slices, while dataset-based normalization brings challenge during model inference for a new patient as only statistical values of this specific patient could be used for data normalization. Herein, we consider that patient-based normalization is proper in medical image studies, which is more applicable to clinical setting. In this study, the patient-based Min-Max

normalization and patient-based Z-Score normalization were applied to shorten the distribution variations among training datasets and external unseen datasets using the statistical values of each patient. Then we evaluated the model generalizability affected by these two data normalization methods. The two normalization methods could be mathematically described as

$$x_{min\_max} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3\text{-}1)$$

$$x_{z\_score} = \frac{x - \mu_x}{\sigma_x} \quad (3\text{-}2)$$

Where x represents the intensities of each patient volume, while x_{min}, x_{max}, μ_x, and σ_x are minimum value, maximum value, mean value and standard deviation of the patient. x_{min_max} and x_{z_score} are the patient data after Min-Max and Z-Score normalization, respectively. The Min-Max normalization rescales the intensity range to [0, 1] and preserves the relationship among the original data values due to its linear transformation nature, while Z-Score normalize the mean value and standardization of the patient to 0 and 1 respectively, which enables the comparison of two datasets with different distributions. As shown in **Figure 3-2**, prior to data normalization, severe inter-institutional distribution discrepancy exists. The distribution discrepancy has been shortened after data normalization, especially after the Z-Score normalization.

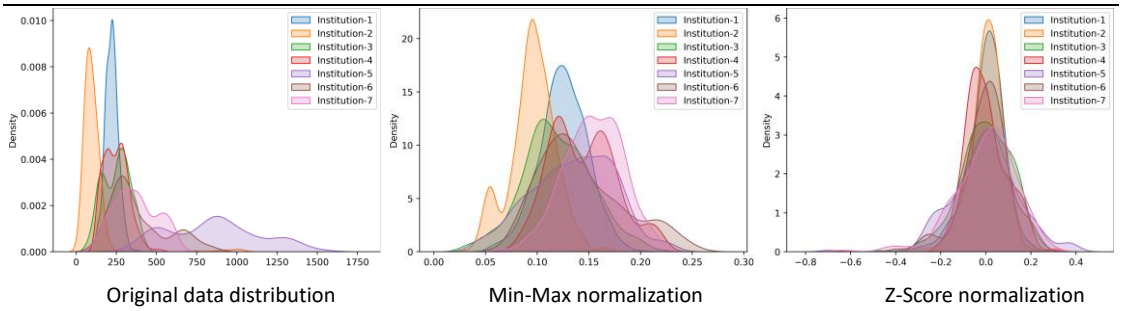

Figure 3-2. Data distribution changes after patient-based Min-Max and Z-Score normalization. From left to right: the original data distribution without data normalization; the MRI distribution after Min-Max normalization and the MRI distribution after Z-Score normalization.

3) Uni-institution Models: To investigate how significant is the external performance degradation for the GFCE-MRI models that trained with single-institution MRI, we first trained three uni-institution models using data from Institution-1, Institution-2, and Institution-3 for each normalization method separately. 53 patients were used for training of each uni-institution model. For each uni-institution model, 18 patients were used for validation to ensure the model performance. Min-Max normalization and Z-Score normalization were applied prior to model training. The three uni-institution models were labeled as Uni-m1, Uni-m2, and Uni-m3 for Min-Max normalization and Uni-z1, Uni-z2, and Uni-z3 for Z-Score normalization, respectively. The generalizability of these models was evaluated using four external datasets (i.e., Institution-4 to Institution-7).

4) Tri-institution Models: To investigate how significant is the external performance improvement for models that trained with diversified multi-institution MRI, we trained the GFCE-MRI model jointly with data from Institution-1 to Institution-3. Considering

that the number of training samples may influence assessment of the tri-institution model since we cannot determine whether the model generalizability improvement is caused by a diverse dataset or an increasement of training samples. Therefore, we randomly selected 18 patients from each institution's training dataset. Then randomly discarded one patient sample to ensure training samples were the same as the number for uni-institution models. The two normalization methods also applied to develop the tri-institution model prior to training. The two tri-institution models with different normalization methods were labeled as Tri-M (with Min-Max normalization) and Tri-Z (with Z-Score normalization), respectively. The four datasets from Institution-4 to Institution-7 were used for external testing to evaluate the model generalizability.

3.3.3 Evaluations

1) Quantitative Evaluation: To quantitatively evaluate the performance of uni- and tri-institution models, MAE and PSNR between the synthetic GFCE-MRI and ground-truth CE-MRI were calculated. The MAE and PSNR have been widely employed for medical image analysis tasks. MAE measures pixel-wise differences while PSNR measures the ratio between the maximum power of a signal and the power of noise (Han, 2017; Li et al., 2020; Xiao, et al., 2022). Smaller MAE and larger PSNR values indicate better quantitative results. Prior to quantitative evaluation, we rescaled the CE-MRI and predicted GFCE-MRI intensities to [0, 1] to compute the percentage differences between GFCE-MRI and CE-MRI. Paired two-tailed t-test (significance level, p=0.05) was performed to analysis if there is significant difference between results from different models.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - f(x_i)|}{n} \quad (3\text{-}3)$$

$$MSE = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{n} \quad (3\text{-}4)$$

$$SSIM = \frac{(2\mu_{y_i}\mu_{f(x_i)}+c1)(2\sigma_{y_i f(x_i)}+c2)}{(\mu_{y_i}{}^2+\mu_{f(x_i)}{}^2+c1)(\sigma_{y_i}^2+\sigma_{f(x_i)}^2+c2)} \quad (3\text{-}5)$$

$$PSNR = 20 \cdot log_{10}^{\frac{max\,(y_i)\cdot\sqrt{n}}{\|y_i - f(x_i)\|_2}} \quad (3\text{-}6)$$

Where y_i and $f(x_i)$ are intensities of real CE-MRI and GFCE-MRI, n is the number of intensities. Here $max\,(y_i)$ is 1 as we have rescaled the CE-MRI and GFCE-MRI intensities to [0, 1]. μ_{y_i}, $\mu_{f(x_i)}$ and σ_{y_i}, $\sigma_{f(x_i)}$ are means and variances of the ground truth image and the synthetic image, while $\sigma_{y_i f(x_i)}$ is the covariance of y_i and $f(x_i)$. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two variables used to stabilize the division by the weak denominator, and *L* is the dynamic range of the pixel values. Here, $L = 4095$, $k_1 = 0.01$, and $k_2 = 0.03$ were set by default.

2) Qualitative Evaluation: To visually assess the performance of the models on external datasets, we applied the trained uni- and tri-institution models to the external datasets without any model-based updating. Prior to results inference, patient-based Min-Max and patient-based Z-Score normalization were applied to uni-institution models and tri-institution model for external results comparison. The input T1w, T2w MRI and ground-truth CE-MRI were shown alongside the GFCE-MRI generated from different models.

3.4 Results

3.4.1 Quantitative results

1) Generalizability of single-institution models: All uni-institution models suffered from severe performance drop on external MRI data for both Min-Max and Z-Score normalizations. **Table 3-2** and **Table 3-3** summarize the quantitative comparisons between the synthetic GFCE-MRI and ground-truth CE-MRI using Min-Max and Z-Score, respectively. As MAE and PSNR have the similar tread, we use the MAE as an indicator to illustrate the results. The average MAE increased from 25.39 ± 3.59 to 43.11 ± 11.74 for Uni-m1, 24.45 ± 3.67 to 51.54 ± 11.53 for Uni-m2, 29.56 ± 6.92 to 41.02 ± 10.86 for Uni-m3, and from 23.03 ± 3.18 to 37.83 ± 8.05 for Uni-z1, 24.87 ± 4.64 to 39.88 ± 10.51 for Uni-z2, 26.84 ± 6.17 to 34.62 ± 7.06 for Uni-z3, respectively. The percentage of uni-institution models' external performance degradation were shown in **Table 3-4**. The average performance drop for MAE were 69.86% and 51.20% for Min-Max and Z-Score normalization respectively, indicting the model trained with single-institution MRI data failed to generalize to external MRI datasets. The largest performance degradation model was Uni-m2 (with 110.80% drop) for Min-Max normalization and Uni-z1 (with 60.35% drop) for Z-Score normalization respectively, indicating that different normalization methods do tremendously influence the uni-institution model generalizability, even the models were trained with same source MRI.

Table 3-2. Internal and external quantitative results using Min-Max normalization.

Model	Testing	MAE ± SD ($\times 10^3$)	MSE ± SD ($\times 10^4$)	SSIM ± SD	PSNR ± SD
Uni-m1	***Institution-1***	***25.39 ± 3.59***	***19.76 ± 4.79***	***0.875 ± 0.019***	***33.45 ± 1.38***
	Institution-4	52.12 ± 10.89	78.37 ± 26.53	0.737 ± 0.030	27.65 ± 1.72

	Institution-5	35.03 ± 6.56	45.57 ± 16.74	0.800 ± 0.028	30.47 ± 1.24
	Institution-6	34.97 ± 4.02	28.84 ± 4.49	0.732 ± 0.059	31.65 ± 0.67
	Institution-7	40.80 ± 9.12	53.41 ± 21.25	0.788 ± 0.045	29.35 ± 1.51
	Overall	***43.11 ± 11.74***	***57.07 ± 28.59***	***0.757 ± 0.049***	***29.35 ± 2.15***
Uni-m2	***Institution-2***	***24.45 ± 3.67***	***26.64 ± 4.95***	***0.864 ± 0.024***	***32.17 ± 0.89***
	Institution-4	50.26 ± 7.11	75.10 ± 15.31	0.730 ± 0.027	27.50 ± 0.95
	Institution-5	51.76 ± 6.28	74.91 ± 16.36	0.747 ± 0.049	27.83 ± 1.02
	Institution-6	58.74 ± 19.93	92.45 ± 52.76	0.651 ± 0.055	27.05 ± 2.13
	Institution-7	45.27 ± 3.83	61.61 ± 10.13	0.761 ± 0.025	28.41 ± 0.73
	Overall	***51.54 ± 11.53***	***76.50 ± 29.06***	***0.722 ± 0.055***	***27.62 ± 1.35***
Uni-m3	***Institution-3***	***29.56 ± 6.92***	***33.99 ± 13.69***	***0.847 ± 0.039***	***31.30 ± 1.72***
	Institution-4	44.53 ± 7.63	61.83 ± 16.81	0.807 ± 0.035	28.51 ± 1.32
	Institution-5	35.67 ± 5.09	44.50 ± 10.26	0.812 ± 0.034	30.09 ± 0.78
	Institution-6	45.36 ± 15.96	54.92 ± 33.18	0.742 ± 0.053	29.41 ± 2.08
	Institution-7	33.30 ± 7.81	38.40 ± 14.44	0.839 ± 0.042	30.69 ± 1.48
	Overall	***41.02 ± 10.86***	***52.94 ± 22.09***	***0.800 ± 0.051***	***29.39 ± 1.68***
Tri-M	Institution-1	26.27 ± 4.01	21.74 ± 5.54	0.867 ± 0.021	33.06 ± 1.30
	Institution-2	26.27 ± 4.19	29.28 ± 5.24	0.855 ± 0.025	31.74 ± 0.86
	Institution-3	28.91 ± 6.38	32.73 ± 12.76	0.847 ± 0.045	31.45 ± 2.05
	Overall	***27.15 ± 5.13***	***27.92 ± 9.73***	***0.856 ± 0.033***	***32.08 ± 1.65***
	Institution-4	41.82 ± 7.82	55.28 ± 14.16	0.809 ± 0.029	28.97 ± 1.20
	Institution-5	41.55 ± 9.04	58.00 ± 21.22	0.807 ± 0.025	29.19 ± 1.51
	Institution-6	46.12 ± 13.55	54.13 ± 28.27	0.743 ± 0.048	29.29 ± 1.84
	Institution-7	33.53 ± 8.21	40.20 ± 16.46	0.821 ± 0.038	30.57 ± 1.56
	Overall	***41.31 ± 10.34***	***53.15 ± 20.59***	***0.797 ± 0.044***	***29.34 ± 1.58***

Table 3-3. Internal and external quantitative results using Z-Score normalization.

Model	Testing	MAE ± SD ($\times 10^3$)	MSE ± SD ($\times 10^4$)	SSIM ± SD	PSNR ± SD
Uni-z1	***Institution-1***	***23.03 ± 3.18***	***16.68 ± 4.27***	***0.879 ± 0.022***	***34.21 ± 1.58***
	Institution-4	43.10 ± 5.91	55.87 ± 13.27	0.736 ± 0.026	28.96 ± 1.20
	Institution-5	32.74 ± 6.27	39.67 ± 14.34	0.788 ± 0.035	31.03 ± 1.16
	Institution-6	32.07 ± 5.05	24.85 ± 5.65	0.741 ± 0.062	32.36 ± 1.07
	Institution-7	38.22 ± 8.77	46.33 ± 16.59	0.769 ± 0.060	29.84 ± 1.42
	Overall	***37.83 ± 8.05***	***44.43 ± 17.57***	***0.753 ± 0.049***	***30.25 ± 1.80***
Uni-z2	***Institution-2***	***24.87 ± 4.64***	***26.18 ± 6.21***	***0.854 ± 0.030***	***32.28 ± 1.10***
	Institution-4	48.47 ± 7.30	74.49 ± 18.11	0.715 ± 0.033	27.62 ± 1.22
	Institution-5	31.35 ± 7.52	38.60 ± 15.35	0.795 ± 0.044	31.33 ± 1.51
	Institution-6	33.27 ± 5.23	29.02 ± 7.37	0.749 ± 0.059	31.68 ± 1.14
	Institution-7	37.27 ± 9.36	47.09 ± 17.84	0.771 ± 0.058	29.76 ± 1.57
	Overall	***39.88 ± 10.51***	***53.00 ± 24.75***	***0.748 ± 0.056***	***29.59 ± 2.23***
Uni-z3	***Institution-3***	***26.84 ± 6.17***	***29.38 ± 12.34***	***0.847 ± 0.042***	***31.97 ± 2.09***
	Institution-4	38.30 ± 5.53	50.53 ± 13.06	0.788 ± 0.039	29.50 ± 1.21
	Institution-5	31.92 ± 7.32	38.99 ± 14.62	0.803 ± 0.036	31.06 ± 1.42
	Institution-6	30.78 ± 4.70	24.32 ± 6.01	0.761 ± 0.066	32.52 ± 1.08
	Institution-7	33.51 ± 8.08	36.38 ± 15.13	0.803 ± 0.059	30.95 ± 1.50
	Overall	***34.62 ± 7.06***	***40.33 ± 16.16***	***0.788 ± 0.051***	***31.01 ± 1.30***
Tri-Z	Institution-1	23.71 ± 3.12	18.68 ± 4.71	0.875 ± 0.020	33.72 ± 1.43
	Institution-2	25.74 ± 4.80	27.90 ± 6.72	0.851 ± 0.027	32.01 ± 1.10
	Institution-3	27.36 ± 6.80	30.30 ± 13.59	0.842 ± 0.049	31.87 ± 2.23
	Overall	***25.60 ± 5.34***	***25.63 ± 10.44***	***0.856 ± 0.037***	***30.69 ± 1.73***
	Institution-4	37.20 ± 5.14	37.20 ± 5.14	0.796 ± 0.029	29.72 ± 1.21

	Institution-5	29.94 ± 6.43	36.79 ± 14.46	0.811 ± 0.034	31.69 ± 1.25
	Institution-6	29.60 ± 4.94	22.94 ± 5.76	0.776 ± 0.062	32.78 ± 1.12
	Institution-7	33.04 ± 8.38	37.35 ± 13.97	0.811 ± 0.053	30.87 ± 1.57
	Overall	***33.41 ± 6.92***	***38.41 ± 14.97***	***0.797 ± 0.045***	***30.96 ± 1.75***

2) Generalizability of tri-institution models: The model generalizability improved when training the model with more diverse MRI data for both Min-Max and Z-Score normalization methods. As shown in **Table 3-5**, the overall external performance obtained 8.65% improvement for Tri-M model and 10.77% improvement for Tri-Z model in MAE.

3) Influence of normalization methods to model generalizability: The quantitative results from **Table 3-4** and **Table 3-5** indicate that Z-Score normalization outperformed the Min-Max normalization on external datasets, with less average performance drop for uni-institution models (51.20% v.s. 69.86% for MAE, 102.03% v.s. 143.91% for MSE, 11.24% v.s. 11.83% for SSIM, and 7.64% v.s. 10.83% for PSNR, respectively) and more average improvement for tri-institution models (10.77% v.s. 8.65% for MAE, 16.35% v.s. 14.51% for MSE, and 2.23% v.s. 1.92% for PSNR). Moreover, as shown in **Table 3-2** and **Table 3-3**, though the overall external performance of Tri-M outperformed Uni-m2 but with comparable external performance with Uni-m1 and slightly worse than Uni-m3, while the Tri-Z model that normalized with Z-Score method outperformed all uni-institution models, suggesting that Z-Score normalization outperforms Min-Max normalization in model generalizability improvement.

Table 3-4. External performance drop of uni-institution models.

	Min-Max	Z-Score

Model	MAE	MSE	SSIM	PSNR	Model	MAE	MSE	SSIM	PSNR
Uni-m1	69.79%	188.82%	13.49%	12.26%	Uni-z1	64.26%	166.37%	14.33%	11.58%
Uni-m2	110.80%	187.16%	16.44%	14.14%	Uni-z2	60.35%	102.44%	12.41%	8.33%
Uni-m3	28.99%	55.75%	5.55%	6.10%	Uni-z3	28.99%	37.27%	6.97%	3.00%
Overall	***69.86%***	***143.91%***	***11.83%***	***10.83%***	***Overall***	***51.20%***	***102.03%***	***11.24%***	***7.64%***

Table 3-5. External performance improvement of tri-institution models.

Model	MAE	MSE	SSIM	PSNR
Tri-M	8.65%	14.51%	4.91%	1.92%
Tri-Z	10.77%	16.35%	4.46%	2.23%

3.4.2 Qualitative results

To visually evaluate the external generalization performance of uni-institution and tri-institution models with different normalization methods, we illustrated the external results of different models in **Figure 3-3**. The generalizability of uni-institution models varies greatly regardless which normalization method was used. All uni-institution models showed worse generalizability to external MRI data with varied contrast enhancement failure in tumor and tumor-to-normal tissue contrast (indicated with red arrows), especially the model trained with Institution-2 data (i.e., Uni-m2 and Uni-z2, with overall image contrast difference and blurring anatomic structure, respectively). The model trained with Institution-1 data (i.e., Uni-m1 and Uniz1) also showed overall image contrast difference compared with ground truth CE-MRI while the models trained with Institution-3 data showed tumor (Uni-m3) and normal vessel (Uni-z3) contrast enhancement failure.

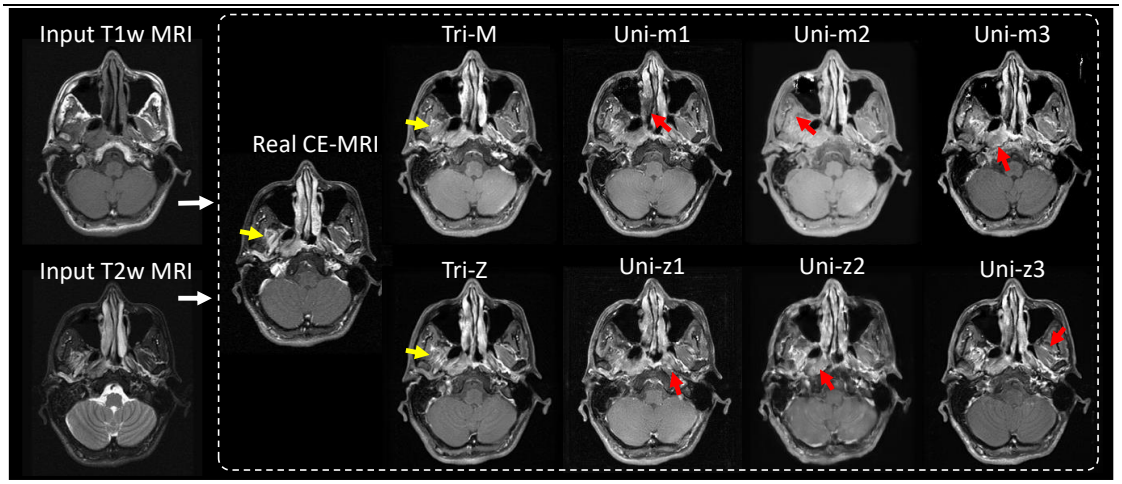

Figure 3-3. Illustration of GFCE-MRI generated from uni-institution and tri-institution models using Min-Max normalization and Z-Score normalization.

Both the two tri-institution models achieved promising generalizability to external data. The generated GFCE-MRI from both Tri-M and Tri-Z models achieved a better visual approximation of tumor contrast enhancement compared to uni-institution models. Compared with the Tri-M model, the Tri-Z model with Z-Score normalization obtained a better approximation of tumor surrounding structures (as indicated with yellow arrows).

3.5 Discussion

In radiotherapy, CE-MRI is commonly used for accurate tumor delineation, especially for the highly infiltrative NPC (Xiao, et al., 2022). However, GBCAs-associated safety issues have stimulated the medical community to eliminate the use of GBCAs. Recently, a worldwide interest has been promoted to synthesize the GFCE-MRI for providing a gadolinium-free alternative for precision tumor delineation (Bône et al., 2021; Chen et al., 2022; Gong et al., 2018a; Kleesiek et al., 2019a; Xiao, et al., 2022; Luo et al., 2021a;

Pasumarthi et al., 2021a; Xu et al., 2021a; Zhao et al., 2020a). Nevertheless, the model generalizability on external institution data remains unexplored and there is no standard multi-institutional MRI normalization method has been established. Herein, for the first time, we retrieved MRI data from seven institutions and investigated the model generalizability using different data normalizations for GFCE-MRI synthesis in NPC patients. In this discussion, we attempted to summarize our key findings, discuss the potential underlying mechanisms, and provide the research community with our perspectives in future directions.

The models trained with single-institution MRI data suffered from various degrees of performance drop on external MRI datasets. As shown in **Table 3-2** and **Table 3-3**, the quantitative results show that the uni-institution models performed well on internal testing datasets with lower MAE and higher PSNR but failed to generalize to external unseen data (i.e., with greater MAE and lower PSNR on external datasets). The visual comparisons (**Figure 3-3**) of synthetic GFCE-MRI among different models also showed that uni-institution models failed to predict the correct contrast enhancement, both in tumor and surrounding vessels. These results suggest that there exist significant MRI data bias across institutions, resulting in a phenomenon that performance of well-trained in-house models cannot generalize to external MRI datasets. The uni-instiution models obtained varied quantitative results on each individual external dataset (for example, the MAE ranges from 34.97 to 52.12 for Uni-m1), this may also be caused by the MRI data bias among external MRI datasets. These data bias may resulted from different MRI characteristics such as image contrast, resolution, texture, artifacts, etc. (as shown in **Figure 3-1**). In addition, the Uni-m2

model that normalized with Min-Max normalization obtained the best external results on Institution-7 dataset and worst results on Institution-6 dataset, while the Uni-z2 model (trained with the same source MRI) that normalized with Z-Score normalization obtained the best external results on Institution-5 and worst results on Institution-4, indicating that different normalization methods do influence the model generalizability. The possible reason might be that different normalization methods shorten the gap between the training dataset and the external dataset to different extent.

By involving diverse MRI data from multiple institutions, the overall external performance of Tri-M and Tri-Z model have been improved compared to uni-institution models, even with the same number of training samples (as shown in **Table 3-5**). This result indicates that involving diverse MRI data from multiple institutions is more capable of achieving a better external performance, possibly due to the view of the model has been enlarged. By training the model with diverse MRI data, the external testing data may have a higher chance to match the training data distribution, thus improving the external performance. However, the external performance improvement also vary depending on the specific normalization method used. As shown in **Table 3-2** and **Table 3-3**, though the external performance of Tri-M model obtained 8.65%, 14.51%, 4.91%， and 1.92% overall improvement in MAE, MSE, SSIM, and PSNR on the four external datasets, respectively, for each individual uni-institution model, the Tri-M model (normalized with Min-Max normalization) obtained comparable results to Uni-m1 and slightly worse results than Uni-m3, while the Tri-Z model (normalized with Z-Score normalization) achieved improved results compared to all uni-institution models, indicating that Z-Score normalization is capable of further improving the

GFCE-MRI model generalizability when training the model with multi-institutional MRI data. On the other hand, both Tri-M and Tri-Z did not obtain obvious performance degradation on the three intra-institution datasets, indicating that involving diverse MRI data from multiple institutions for model development is capable of maintain the intra-institution accuracy no matter what normalization method was used, though the two tri-institution models were trained with 1/3 number of samples from each individual institution.

Z-Score normalization outperformed Min-Max normalization in improving the model generalizability, for both uni-institution models and the tri-institution model. As shown in **Table 3-4** and **Table 3-5**, Z-Score normalization achieved 18.66%, 41.88%, 0.59% and 3.19% less external performance drop of MAE, MSE, SSIM and PSNR respectively than Min-Max normalization for uni-institution models. With Z-Score normalization, the tri-institution model Tri-Z also obtained additional 2.12%, 1.84% and 0.31% performance gain in MAE, MSE and PSNR than Tri-M. This is possibly due to Z-Score normalizes all the patients' mean and standard deviation to the same value (0 and 1, respectively), which minimized the distribution variations among all training patients and external testing patients (as shown in **Figure 3-2**), while Mix-Max normalization preserves the relationship (i.e., the intra-patient intensity ratio) among the original data intensities, which limited its contribution to narrowing the distribution gap across institutions.

In this study, we used percentage values instead of actual values to interpret the results obtained from different normalization methods. This is because the MRI

distributions across institutions are unidentical with different mean value and standard deviation, making the results incomparable. As demonstrated in (Lam, et al., 2022), the model trained with smaller mean intensity data obtained significantly better intra-institution quantitative results, even with the same number of training samples. Different normalization methods will further normalize the multi-institutional data to different distributions, making different normalization results uninterpretable. For example, the Uni-m2 model obtained better internal performance compared with Uni-z2 in MAE (24.45 v.s. 24.87), but the Uni-m3 model may not necessarily performed better than Uni-z3 since the distribution of the testing datasets are different after the two normalization methods. To quantitatively evaluate the results generated from two different normalization methods, we used percentage results (as shown in **Table 3-4** and **Table 3-5**) instead of the actual values to compare these two normalization results. For the multi-institutional setting, the Z-Score normalization may be a promising method for results interpretation compared to Min-Max normalization as the Min-Max normalization preserves the original data distribution across institutions, while the Z-Score normalization normalize the mean intensities and standard deviations of multi-institutional datasets to the same value and minimized the multi-institutional distribution diversity, making the normalized multi-institutional results comparable.

Our study has several limitations. Firstly, since our findings are based on MMgSN-Net (Xiao, et al., 2022), applicability of our results using other deep-learning models deserves future investigation. Secondly, this work takes into account the diversity of MRI images and signal intensities of MRI among institutions, as shown in **Figure 3-2**, after data normalization, small distribution variations also exist among

different institutional MRI, these variations may be caused by the image-based factors such as image texture, artifacts, and tumor size etc. As demonstrated in (Arega et al., 2021), MRI-specific data augmentation provides a potential solution to improve the model generalizability in aspect of training image, which will be considered in our future work to further improve the model generalizability.

3.6 Conclusion

In this study, we investigated the model generalizability for GFCE-MRI synthesis in NPC patients using data from seven institutions and explored potential model generalizability influence factors of diversity of training data and application of different normalization methods. Results of the present work showed that the tri-instituion models developed from multi-institutional MRI generally resulted in higher generalizability than the uni-institution models developed from single-institution datasets. Application of the Z-Score normalization was capable of improving the model generalizability and results interpretability in multi-institutional MRI setting, which outperformed Min-Max normalization.

4. Clinical evaluation of the GFCE-MRI in NPC radiotherapy

4.1 Abstract

Purpose: To investigate clinical efficacy of GFCE-MRI for gross-tumor-volume (GTV) delineation of NPC via a multi-institutional setting.

Methods and Materials: This study retrospectively retrieved T1w, T2w MRI, gadolinium-based CE-MRI and planning CT of 378 biopsy-proven NPC patients from three oncology centers. A MMgSN-Net was trained in 288 patients to leverage complementary features in T1w and T2w MRI for CE-MRI synthesis, which was validated independently in 90 patients. Two board-certified oncologists and one medical physicist participated in clinical evaluations in three aspects: image quality of GFCE-MRI, target volume delineation and treatment planning. Image quality of GFCE-MRI evaluation includes distinguishability between CE-MRI and GFCE-MRI, clarity of tumor-to-normal tissue interface, veracity of contrast enhancement in tumor invasion risk areas, and efficacy in primary tumor staging. Target volume delineation and treatment planning were manually performed by oncologists and the medical physicist, respectively. Paired two-tailed t-test with a significant level of 0.05 was performed to assess statistical difference of the results.

Results: The mean accuracy to distinguish GFCE-MRI from CE-MRI was 53.33%; no significant difference was observed in the clarity of tumor-to-normal tissue interface between GFCE-MRI and CE-MRI; for the veracity of contrast enhancement in tumor invasion risk areas and efficacy in primary tumor staging, a Jaccard Index (JI) of

76.04% and accuracy of 86.67% were obtained, respectively. The image quality evaluation suggests that the quality of GFCE-MRI is approximated to CE-MRI. The Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) of the GTVs that delineated from GFCE-MRI and CE-MRI were 0.762 (0.673-0.859) and 1.932mm (0.763mm-2.974mm) respectively, and the mean dosimetric difference of planning target volume (PTV) was less than 1%, which were clinically acceptable.

Conclusions: The GFCE-MRI is highly promising to replace the use of CE-MRI in tumor delineation of NPC patients.

4.2 Introduction

NPC is a highly infiltrative malignancy and is characterized by a distinct geographical distribution in East and Southeast Asia (Song et al., 2022). In 2020, 133,354 new cases and 80,008 deaths of NPC were recorded globally (World Cancer Research Fund International, 2020). At present, radiotherapy is the primary treatment modality for NPC due to its high radiosensitivity. As reported by Xu et al. (Xu et al., 2016), the 5-year survival rate for NPC patients achieved 66%-83% with radiotherapy alone. For early-stage NPC, the overall survival rate is greater than 90% (Xu et al., 2016). In radiotherapy treatment planning, accurate tumor delineation is the foremost prerequisite to achieve optimal tumor control and improve patient survival (Li et al., 2019). However, as a soft-tissue mass, NPC shows a high propensity to invade surrounding critical structures, such as neural systems and bony skull base, posing significant challenges for clinical oncologists to delineate the tumor volume accurately. To enhance tumor visibility for more precise tumor delineation, CE-MRI using GBCA is

widely used in clinical practice. It has been reported that approximately 45% of MRI scans performed in the United States involves the use of GBCA on a routine basis (Enterline, Jan 2021).

However, GBCA-based CE-MRI imaging is costly, time-consuming, resource-demanding, and can potentially increase the risk of toxicity for patients with impaired renal function (Lang et al., 2019; Rogosnitzky & Branch, 2016). Firstly, GBCA is desperately needed in medical practice. More than 450 million doses have been administrated since its introduction to the market, and more than 30 million doses of GBCA are consumed annually worldwide (Guo et al., 2018; Jakobsen et al., 2021). The cost for each CE-MRI scan ranges from HKD 9,200 to 19,980 in Hong Kong (Adventist Hospital, 2022; Gleneagles Hospital, 2022; Hong Kong Baptist Hospital, 2021; Union Hospital, 2021). For cancer patients receiving adaptive radiotherapy, repeated GBCA-based CE-MRI scans significantly increase patients' costs by a factor of 3 to 5. Then, due to the high number of patients requiring CE-MRI scans, the waiting time for cancer patients to receive radiotherapy treatment increases. As reported by Wildeman et al. (Wildeman et al., 2013), the average CE-MRI scan time for the head and neck region lasts 30 to 90 minutes for each patient, and the median waiting time between diagnosis and first radiotherapy treatment is 120 days (range 13-500 days) for NPC patients. Long waiting time may lead to significantly worse treatment outcomes as a result of the progression of cancer during the waiting (Wildeman et al., 2013). More importantly, GBCA-associated patient safety issues, such as acute adverse reactions, NSF and gadolinium deposition, have raised serious concerns in medical community (Nguyen et al., 2020a). All things considered, there is a pressing demand for developing an

alternative method to GBCA for cost-effective, time-efficient, and safe radiotherapy.

To address this issue, worldwide interests have been raised in applying deep learning to synthesize GFCE-MRI without using GBCA (Gong et al., 2018a; Kleesiek et al., 2019a) and great successes have been achieved. Preetha et al. (Preetha et al., 2021) proposed a deep convolutional neural network to synthesize GFCE-MRI for brain tumor response assessment from contrast-free T1w, T2w, and fluid-attenuated inversion recovery sequences; they found that synthetizing GFCE-MRI from contrast-free MRI is feasible, and there was no significant difference in treatment response compared to GBCA-based CE-MRI. Zhang et al. (Zhang et al., 2021) utilized GBCA-free T1 maps and cine imaging to synthesize cardiovascular GFCE-MRI through a modified conditional generative adversarial network; they demonstrated that the synthetic GFCE-MRI could achieve high agreement with GBCA-based CE-MRI in lesion distribution and quantification, while the GFCE-MRI achieved significantly better image quality than CE-MRI. Following these works, Li et al. (Li, et al., 2022) first applied the GFCE-MRI technique into the field of radiotherapy. They synthesized the GFCE-MRI for NPC tumor delineation from contrast-free T1w and T2w MRI and demonstrated that the synthetic GFCE-MRI has a high approximation to GBCA-based CE-MRI, especially for the visualization of tumor-to-muscle interface and intratumor texture, which is highly promising for tumor delineation. However, this work mostly focused on technical development of the synthetic network, more clinical evidence to demonstrate its clinical efficacy on tumor delineation is warranted. Clinical evaluation plays a pivotal role in demonstrating the performance of the new technology in real-world setting, which is essential prior to bench-to-bedside translation of the novel

GFCE-MRI technique.

In this study, we invited two board-certified clinical oncologists and one experienced medical physicist to conduct a series of clinical evaluations to investigate the clinical efficacy of GFCE-MRI in radiotherapy of multi-institutional NPC patients. Specially, the evaluations including image quality of GFCE-MRI (distinguishability between CE-MRI and GFCE-MRI, clarity of tumor-to-normal tissue interface, veracity of contrast enhancement in tumor invasion risk areas, and efficacy in primary tumor staging), target volume delineation and treatment planning. To the best of our knowledge, this is the first clinical evaluation study of GFCE-MRI in NPC radiotherapy using multi-institutional MRI data. This study would fill the current knowledge gap and provide the community with a clinical reference prior to clinical application of the novel GFCE-MRI technique in NPC radiotherapy.

4.3 Methods and materials

4.3.1 Patient data

Patient data was retrospectively collected from three oncology centers in Hong Kong. This dataset includes 378 biopsy-proven (stage I-IVb) NPC patients who received radiation treatment during 2012-2016. The three hospitals were labelled as Institution-1 (134 patients), Institution-2 (71 patients), and Institution-3 (173 patients), respectively. For each patient, T1w MRI, T2w MRI, gadolinium-based CE-MRI, and planning CT (with original organs at risk contours) were retrieved. MRI images were automatically registered since MRI images for each patient were scanned in the same position. The use of this dataset was approved by the Institutional Review Board of the

University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB) with reference number UW21-412, and the Research Ethics Committee (Kowloon Central/Kowloon East) with reference number KC/KE-18-0085/ER-1. Due to the retrospective nature of this study, patient consent was waived. In our dataset, the primary tumor stage (T-stage) for majority patients were stage-III, accounting for 56.06% of the whole dataset, while patients with other stages were 15.40%, 10.76%, and 17.78% for stage-I, stage-II, and stage-IV, respectively. In this multi-institution study, we only focused on the head and neck region where the primary tumor was located due to the limited anatomical region for training with deep learning techniques. For model development, 288 patients were used for model training and 90 patients were used for model testing. The details of patient characteristics and the number split for training and testing of each dataset were illustrated in **Table 4-1**. Prior to model training, MRI images were resampled to 256*224 by bilinear interpolation (Gribbon & Bailey, 2004) due to the inconsistent matrix sizes of the three datasets.

Table 4-1. Details of the multi-institutional patient characteristics. FS: field strength; TR: repetition time; TE: echo time; No.: Number; Avg: average.

Institution (vendor-FS)	Patient No. (train/test)	Avg. age	Sex	Modality	TR (ms)	TE (ms)	Contrast Density
Institution-1 (Siemens-1.5T)	134 (105/29)	56 ± 11	Male: 98 Female: 36	T1w	562 - 739	13 - 17	/
				T2w	7640	97	/
				CE-MRI	562 - 739	13 - 17	0.1mmol/kg
Institution-2 (Philips-3T)	71 (53/18)	49 ± 15	Male: 55 Female: 16	T1w	4.8 - 9.4	2.4 - 8.0	/
				T2w	3500 - 4900	50 - 80	/
				CE-MRI	4.8 - 9.4	2.4 - 8.0	0.1mmol/kg
Institution-3 (Siemens-3T)	173 (130/43)	57 ± 12	Male: 136 Female: 37	T1w	620	9.8	/
				T2w	2500	74	/
				CE-MRI	3.42	1.11	0.1mmol/kg

4.3.2 GFCE-MRI synthesis network

The MMgSN-Net was applied to learn the mapping from T1w MRI and T2w MRI to CE-MRI. The MMgSN-Net was a 2D network. The effectiveness of this network in GFCE-MRI synthesis for NPC patients has been demonstrated by a previous study (Li, Xiao, et al., 2022). T1w MRI and T2w MRI were used as input and corresponding CE-MRI was used as learning target. In this work, we obtained 12806 image pairs for model training and 3589 image pairs for testing. Different from the original study, which used single institutional data for model development and utilized min-max value of the whole dataset for data normalization, in this work, we used mean and standard deviation of each individual patient to normalize MRI intensities due to the heterogeneity of the MRI intensities across institutions.

4.3.3 Clinical evaluations

In this study, we attempted to conduct a series of clinical evaluations to investigate the efficacy of GFCE-MRI in assisting primary GTV delineation for NPC patients. The evaluation methods used in this study included image quality assessment of GFCE-MRI, target volume delineation, and treatment planning. Two board-certified clinical oncologists (D.Z. and Z.H. with 8 years' and 6 years' clinical experience, respectively) were invited to perform the GFCE-MRI quality assessment and target volume delineation, and one clinical physicist (Z.C. with 7 years' treatment planning experience) was invited to generate treatment plans using the GFCE-MRI based contours that were delineated by the participating oncologists. Considering the clinical burden of oncologists and the physicist, 30 patients (10 patients from each center) were randomly selected for clinical evaluations, including 15 real patients (5 patients each

center) and 15 corresponding synthetic patients (5 patients each center). All clinical evaluations were performed on an Eclipse workstation (V5.0.10411.00, Varian Medical Systems, USA) by the oncologists and physicist. The results were obtained under the consensus of the two oncologists.

4.3.4 Image quality of GFCE-MRI

To evaluate the general quality of synthetic GFCE-MRI against the real CE-MRI, we conducted four radiotherapy-related evaluations: distinguishability between CE-MRI and GFCE-MRI, clarity of tumor-to-normal tissue interface, veracity of contrast enhancement in tumor invasion risk areas, and efficacy in primary tumor staging. The GFCE-MRI and CE-MRI volumes were imported as individual patients to Eclipse system and randomly and blindly shown to oncologists for evaluation. The MRI volumes were shown in axial view, sagittal view and coronal view, and the oncologists can scroll through the slices to view adjacent images.

Distinguishability between CE-MRI and GFCE-MRI:

To evaluate the reality of GFCE-MRI, oncologists were invited to differentiate the synthetic patients from real patients. Different from the previous studies that utilized limited number (20-50 slices, axial view) of 2D image slices for reality evaluation (Kleesiek et al., 2019a; Li, Xiao, et al., 2022), we used 3D volumes in this study to help oncologists visualize the inter-slice adjacent information. The judgement results were recorded and the accuracy for each institution and the overall accuracy were calculated.

Clarity of tumor-to-normal tissue interface:

The clarity of tumor-normal tissue interface is critical for tumor delineation, which directly affects the final delineation outcomes. Oncologists were asked to use a 5-point Likert scale ranging from 1 (poor) to 5 (excellent) to evaluate the clarity of tumor-to-normal tissue interface. Paired two-tailed t-test (with a significance level of $p = 0.05$) was applied to analyse if the scores obtained from real patients and synthetic patients are significantly different.

Veracity of contrast enhancement in tumor invasion risk areas:

In addition to the critical tumor-normal tissue interface, the areas arounding the NPC tumor will also be considered during delineation. To better evaluate the veracity of contrast enhancement in GFCE-MRI, we selected 25 tumor invasion risk areas according to (Liang et al., 2009), including 13 high-risk areas and 12 medium-risk areas, and asked oncologists to determine whether these areas were at risk of being invaded according to the contrast-enhanced tumor regions. The 13 high-risk areas include: retropharyngeal space, parapharyngeal space, levator veli palatine muscle, prestyloid compartment, Tensor veli palatine muscle, poststyloid compartment, nasal cavity, pterygoid process, basis of sphenoid bone, petrous apex, prevertebral muscle, clivus, and foramen lacerum. The 12 medium-risk areas include foramen ovale, great wing of sphenoid bone, medial pterygoid muscle, oropharynx, cavernous sinus, sphenoidal sinus, pterygopalatine fossa, lateral pterygoid muscle, hypoglossal canal, foramen rotundum, ethmoid sinus, and jugular foramen. The areas considered at risk of invasion were recorded.

The JI (Fletcher & Islam, 2018) was utilized to quantitatively evaluate the

results of recorded risk areas from CE-MRI and GFCE-MRI. The JI could be calculated by:

$$JI = |R_{CE} \cap R_{VCE}|/|R_{CE} \cup R_{VCE}| \quad (4\text{-}1)$$

where R_{CE} and R_{VCE} represents the set of risk areas that recorded from CE-MRI and corresponding GFCE-MRI, respectively. JI measures similarity of two datasets, which ranges from 0% to 100%. Higher JI percentage indicates more similar of two risk areas.

Efficacy in primary tumor staging:

A critical radiotherapy-related application of CE-MRI is tumor staging, which plays a critical role in treatment planning and prognosis prediction (Lee et al., 2018). To assess the efficacy of GFCE-MRI in NPC tumor staging, oncologists were asked to determine the stage of the primary tumor shown in CE-MRI and GFCE-MRI. The staging results from CE-MRI were taken as the ground truth and the staging accuracy of GFCE-MRI was calculated.

4.3.5 Target volume delineation

GTV delineation is the foremost prerequisite for a successful radiotherapy treatment of NPC tumor, which demands excellent precision (Jager et al., 2015). An accurate tumor delineation improves local control and reduce toxicity to surrounding normal tissues, thus potentially improving patient survival (Jameson et al., 2014). To evaluate the feasibility of eliminating the use of GBCA by replacing CE-MRI with GFCE-MRI in tumor delineation, oncologists were asked to contour the primary GTV under the assistance of GFCE-MRI. For comparison, CE-MRI was also imported to Eclipse for

tumor delineation but assigned as a different patient, which were shown to oncologists in a random and blind manner. To mimic the real clinical setting, contrast-free T1w, T2w MRI and corresponding CT of each patient were imported into the Eclipse system since sometimes T1w and T2w MRI will also be referenced during tumor delineation, the delineated contours were mapped to corresponding CT for treatment planning. Due to both real patients and synthetic patients were involved in delineation, to erase the delineation memory of the same patient, we separated the patients to two datasets, each with the same number of patients, both two datasets with mixed real patients and synthetic patients without overlaps (i.e., the CE-MRI and GFCE-MRI from the same patient are not in the same dataset).When finished the first dataset delineation, there was a one-month interval before the delineation of the second dataset. After the delineation of all patients, the DSC (Balagopal et al., 2021) and HD (Yang et al., 2015) of the GTVs delineated from real patients and corresponding synthetic patients were calculated to evaluate the accuracy of delineated contours.

Dice similarity coefficient (DSC): DSC is a broadly used metric to compare the agreement between two segmentations (Chang et al., 2009). It measures the spatial overlap between two segmentations, which ranges from 0 (no spatial overlap) to 1 (complete overlap). The DSC can be expressed as:

$$DSC = 2 * |C_{CE} \cap C_{GFCE}|/(|C_{CE}| + |C_{GFCE}|) \qquad (4\text{-}2)$$

where C_{CE} and C_{GFCE} represent the contours delineated from real patients and synthetic patients, respectively.

Hausdorff distance (HD): Even though DSC is a well-accepted segmentation comparison metric, it is easily influenced by the size of contours. Small contours typically receive lower DSC than larger contours (Schreier et al., 2020). Therefore, HD was applied as a supplementary to make a more thorough comparison. HD is a metric to measure the maximum distance between two contours. Given two contours C_{CE} and C_{GFCE}, the HD could be calculated as:

$$HD = max(max_{x \in C_{CE}} \mathrm{d}(x, C_{GFCE}), max_{y \in C_{GFCE}} \mathrm{d}(y, C_{CE})) \qquad (4\text{-}3)$$

where d(x, C_{GFCE}) and d(y, C_{CE}) represent the distance from point x in contour C_{CE} to contour C_{GFCE} and the distance from point y in contour C_{GFCE} to contour C_{CE}.

4.3.6 Treatment planning

Measures such as DSC and HD sometimes do not reflect the clinical impact in actual radiotherapy treatment (Schreier et al., 2020). To measure the real clinical dose disagreement between contours delineated from CE-MRI and GFCE-MRI, the dosimetric differences between PTVs of real patients and synthetic patients were compared. The PTVs were delineated for each patient based on C_{CE} and C_{GFCE} by oncologists according to clinical guidance and their clinical experience. The delineated PTVs were labelled as P_{CE} for real patients and P_{GFCE} for synthetic patients, respectively. The PTV receiving 70Gy (PTV70), 66Gy (PTV66), and 60Gy (PTV60) were delineated by oncologists for each patient. A VMAT plan was generated by the physicist based on P_{GFCE} with prescription dose of 70Gy. Original organs at risk contours were transferred into the planning CT for dose limitation of normal organs.

The P_{CE} of each patient was also transferred to corresponding synthetic patient for dose distribution comparison. After treatment planning, the dose-volume histogram (DVH) was compared. Similar to (Kazemifar et al., 2019), the minimum dose delivered to 5% volume ($D_{5\%}$), minimum dose delivered to 95% volume ($D_{95\%}$), maximum dose (D_{max}) and the mean dose (D_{mean}) of P_{CE} and P_{GFCE} were calculated and compared for each patient. Paired two-tailed t-test was performed (with a significance level of $p = 0.05$) to analyze if there are significance difference in these metrics between P_{CE} and P_{GFCE}.

4.4 Results

4.4.1 Image quality of GFCE-MRI

Table 4-2 summarizes the results of the four GFCE-MRI quality evaluation metrics, including: (A) distinguishability between CE-MRI and GFCE-MRI; (B) clarity of tumor-to-normal tissue interface; (C) veracity of contrast enhancement in tumor invasion risk areas; and (D) efficacy in primary tumor staging.

(A) Distinguishability between CE-MRI and GFCE-MRI: The overall judgement accuracy for the MRI volumes was 53.33%, which is close to a random guess accuracy (i.e., 50%). For Institution-1, 2 (/5) real patients were judged as synthetic and 1(/5) synthetic patient was considered as real. For Institution-2, 2(/5) real patients were determined as synthetic and 4(/5) synthetic patients were determined as real. For Institution-3, 2(/5) real patients were judged as synthetic and 3(/5) synthetic patients were considered to be real. In total, 6(/15) real patients were judged as synthetic and 8(/15) synthetic patients were judged as real.

(B) Clarity of tumor-to-normal tissue interface: The overall clarity scores of tumor-to-normal tissue interface for real and synthetic patients were 3.67 with a median of 4 and 3.47 with a median of 4, respectively. No significant difference was observed between these two scores (p = 0.38). The average scores for real and synthetic patients were 3.6 and 3, 3.6 and 3.8, 3.8 and 3.6 for Institution-1, Institution-2, and Institution-3, respectively. 5(/15) real patients got a higher score than synthetic patients and 3(/15) synthetic patients obtained a higher score than real patients. The scores of the other 7 patient pairs were the same.

(C) Veracity of contrast enhancement in tumor invasion risk areas: The overall JI score between the recorded tumor invasion risk areas from CE-MRI and GFCE-MRI was 74.06%. The average JI obtained from Institution-1, Institution-2, and Institution-3 dataset were similar with a result of 71.54%, 74.78% and 75.85%, respectively. In total, 126 risk areas were recorded from the CE-MRI for all of the evaluation patients, while 10 (7.94%) false positive high risk invasion areas and 9 (7.14%) false negative high risk invasion areas were recorded from GFCE-MRI.

(D) Efficacy in primary tumor staging: A T-staging accuracy of 86.67% was obtained using GFCE-MRI. 13(/15) patient pairs obtained the same staging results. For the Institution-2 data, all synthetic patients observed the same stages as real patients. For the two T-stage disagreement patients, one synthetic patient was staged as phase IV while the corresponding real patient was staged as phase III, the other synthetic patient was staged as I while corresponding real patient was staged as phase III.

Table 4-2. GFCE-MRI image quality evaluation results in: (A) Distinguishability between CE-

MRI and GFCE-MRI; (B) Clarity of tumor-to-normal tissue interface; (C) Veracity of contrast enhancement in risk areas; and (D) T-staging.

	(A) Distinguishability between CE-MRI and GFCE-MRI			**(B) Clarity of tumor-to-normal tissue interface**					
	Institution-1	Institution-2	Institution-3	Institution-1		Institution-2		Institution-3	
	/	/	/	Real	Syn	Real	Syn	Real	Syn
Center average	70%	40%	50%	3.6	3	3.6	3.8	3.8	3.6
Overall average		53.33%			Real: 3.67			Syn: 3.47	

	(C) Veracity of contrast enhancement in risk areas			**(D) Efficacy in primary tumor staging**		
	Institution-1	Institution-2	Institution-3	Institution-1	Institution-2	Institution-3
Center average	71.54%	74.78%	75.85%	80%	100%	80%
Overall average		74.06%			86.67%	

4.4.2 Target volume delineation

The average DSC and HD between the C_{CE} and C_{GFCE} was 0.762 (0.673-0.859) with a median of 0.774 and 1.932mm (0.763mm-2.974mm) with a median of 1.913mm, respectively. For Institution-1, Institution-2, and Institution-3, the average DSC were 0.741, 0.794 and 0.751 respectively, while the average HD were 2.303mm, 1.456mm, and 2.037mm respectively. **Figure 4-1** illustrated the delineated primary GTV contours from an average patient with the DSC of 0.765 and HD of 1.938mm. The green contour shows the primary GTV that delineated form the synthetic patient, while the red contour was delineated from corresponding real GBCA-based patient.

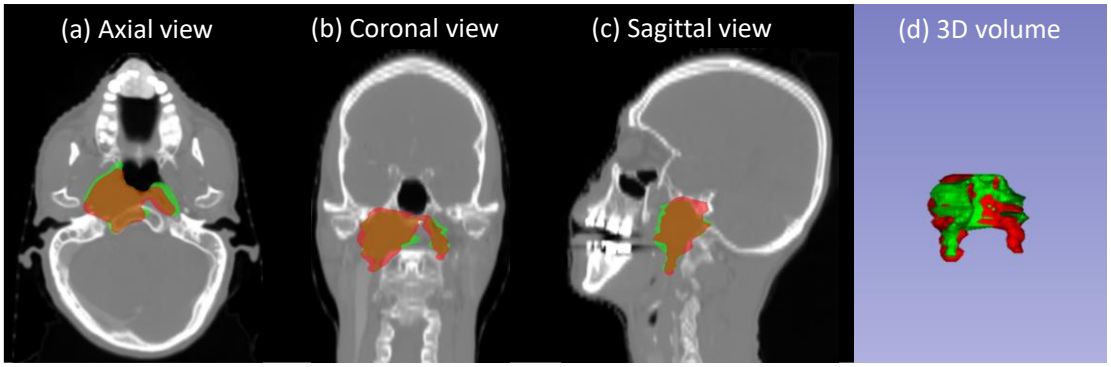

Figure 4-1. Illustration of the primary GTVs from a typical patient with an average DSC and HD. The green volume was delineated from the synthetic patient, while the red volume was delineated from the real GBCA-based patient.

4.4.3 Treatment planning

Table 4-3 illustrates the dosimetric differences between the P_{CE} and P_{GFCE}. The overall dose difference between P_{GFCE} and P_{CE} was less than 1% for all of the $D_{5\%}$, $D_{95\%}$, D_{max}, and D_{mean}. No significant difference was observed between P_{GFCE} and P_{CE} for these dose metrics (p > 0.05). Less than 0.5Gy dose difference were obtained for $D_{5\%}$ (0.046Gy), D_{max} (0.055Gy), and D_{mean} (0.398Gy). **Figure 4-2** shows the isodose lines (**Figure 4-2(A)**) and DVH (**Figure 4-2(B)**) of a typical patient. Red line and green line represent P_{GFCE} and P_{CE} respectively in both sub-figures. From **Figure 4-2(A)**, we observed that the P_{GFCE} and P_{CE} contours overlapped well, and the prescription 70Gy region (translucent red) covers well with both P_{GFCE} and P_{CE}. From **Figure 4-2(B)**, we see that the DVH of P_{GFCE} matches well with DVH of P_{CE}.

Table 4-3. The dose distribution differences between $\boldsymbol{P_{CE}}$ and $\boldsymbol{P_{GFCE}}$ with respect to $D_{5\%}$, $D_{95\%}$, D_{max}, and D_{mean}. NS: not significant. $\boldsymbol{P_{CE}}$: planning target volume from CE-MRI, $\boldsymbol{P_{GFCE}}$:

planning target volume from GFCE-MRI.

Metric	Mean dose differences (range)	*p*-value
$D_{5\%}$	0.043Gy (-0.287Gy ~0.230Gy)	NS (p=0.12)
$D_{95\%}$	0.960Gy (-0.115Gy ~2.581Gy)	NS (p=0.20)
D_{max}	-0.074Gy (-1.103Gy ~0.138Gy)	NS (p=0.48)
D_{mean}	0.271Gy (-0.184Gy ~0.879Gy)	NS (p=0.17)

Figure 4-2. (A) Dose distribution comparison of P_{GFCE} and P_{CE} from a single VMAT plan with prescription dose of 70Gy. The most inner red line and green line are P_{GFCE} and P_{CE}, respectively. (B) DVH plot with P_{GFCE} and P_{CE}, squares and triangles are based on P_{GFCE} and P_{CE}, respectively.

4.5 Discussion

The GBCA has been used for decades to improve the visibility of tumors and has been considered essential in GTV delineation, especially for the highly infiltrative NPC tumors. Since 2006, several GBCA-related safety issues have been reported (Flood et al., 2017; Kanda et al., 2015a; H. S. Thomsen, 2006). To provide a safe and high-quality clinical care, deep learning-based GFCE-MRI technics have been proposed in recent years, aiming at replacing the use of GBCA-based CE-MRI via utilizing the multiparametric information from contrast-free MRI sequences. However, most of these studies are technical contributions or feasibility studies in different anatomy structures using single institutional data. In this study, we conducted a series of GFCE-MRI-based clinical evaluations using multi-institutional MRI data to explore the clinical efficacy of using the GFCE-MRI for GTV delineation in NPC patients. The evaluations included overall image quality of synthetic GFCE-MRI, the performance of GFCE-MRI in target volume delineation, and dosimetric discrepancy between the dose generated from GFCE-MRI-based plans and real CE-MRI-based plans. To the best of our knowledge, this is the first clinical-oriented study of GFCE-MRI in radiotherapy. In this discussion, we sought to highlight our main findings, discuss the possible reasons, and provide potential future directions for the research community.

The evaluations for image quality of GFCE-MRI showed that the quality of synthetic GFCE-MRI is highly similar to the gadolinium-based CE-MRI, as shown in **Table 4-2**. Firstly, it is challenging for oncologists to distinguish the real CE-MRI from synthetic GFCE-MRI, which obtained a judgement accuracy of 53.33%. This result is slightly better than random guessing (i.e., with an accuracy of 50%), indicating that the generated GFCE-MRI has similar image quality compared to real CE-MRI. The

judgement accuracy in this study is similar to the reported results from the previous study (Li et al., 2022). However, the previous work used limited 2D images (50 single images, axial view) for evaluation, the inter-slice adjacent information was not considered, while we used the 3D volume in this study, meaning that the axial view, coronal view and sagittal view were also shown to the oncologists for evaluation. Secondly, for the critical delineation-related tumor-to-normal tissue interface, we observed no significant difference between the CE-MRI and GFCE-MRI, suggesting that the synthetic GFCE-MRI preserves the similar tumor-to-normal tissue interface clarity compared to CE-MRI. Thirdly, for the veracity evaluation of tumor invasion risk areas, the JI of risk areas obtained 74.06% between the CE-MRI and GFCE-MRI. In total, 126 risk areas were recorded from the CE-MRI for all of the evaluation patients, while we recorded 10 (7.94%) false positive high risk invasion areas and 9 (7.14%) false negative high risk invasion areas from GFCE-MRI, indicating there are still some mis-enhancement in tumor-surrounding risk areas. Lastly, we obtained a T-staging accuracy of 86.67% using GFCE-MRI. In our experiments, only two synthetic patients were mis-staged, indicating that most synthetic cases were with the similar quality in the aspect of tumor staging, while still having few cases with unsatisfied staging performance, which is potentially be improved by designing the deep learning model focus on tumor and surrounding regions learning. As reported by oncologists, the overall image quality of GFCE-MRI in NPC delineation is acceptable.

We obtained an average DSC of 0.762 and an average HD of 1.932mm between the GTV contours generated from real patients and synthetic patients, and we consider this GFCE-MRI-based results are acceptable. As this is the first study to investigate

clinical efficiency of GFCE-MRI in NPC delineation, these is no reference for comparison. However, there are some automatically NPC delineation studies. Tsuji et al. (Tsuji et al., 2010) applied a registration-based approach for adaptive GTV delineation and obtained a DSC of 0.69 between the contours generated from their approach and manually delineated contours, Yang et al. (Yang et al., 2015) proposed a multichannel auto-segmentation method to automatically segment the GTV of head and neck cancer, and they obtained a DSC of 0.75. Similarly, Guo et al. (Guo et al., 2019) developed a Dense-Net to automatically segment the GTV from head and neck cancer patients, they obtained a DSC of 0.73. Moreover, due to the complexity of nasopharynx involves multiple critical structures such as parotid glands and neural systems, which always challenges oncologists to consistently delineate the same target volume. As reported by Lu et al. (Lu et al., 2006), the DSC of interobserver variations in GTV delineation of head and neck patients is 0.75 for the same patient, which is similar to our GFCE-MRI-based results (0.762). We consider that directly comparing the DSC from different works is unsuitable since diverse datasets were used in different studies. However, it is reasonable to conclude that the GFCE-MRI resulted in a good agreement with ground truth GTV using these works as references.

Importantly, treatment planning is an essential evaluation for NPC radiotherapy, which directly demonstrates the dosimetric performance of GFCE-MRI-based contours. The dose distribution differences between P_{CE} and P_{GFCE} shows that the dose difference between these two PTVs was less than 1%, and we did not observe significant dose difference between these two target volumes, suggesting that the GTV contours delineated under assistant of GFCE-MRI is sufficient to generate clinically

equivalent treatment plans. In a study performed by Kazemifar et al. (Kazemifar et al., 2019), they generated synthetic CT from CE-MRI for treatment planning and clinically evaluated the synthetic CT on 14 patients. They found the dose error was also within 1% and demonstrated the effectiveness of their work.

There are several limitations of the current deep learning-based GFCE-MRI technique. Firstly, the GFCE-MRI generated from different institutions obtained variated clinical results, as shown in **Table 4-2(A)**, suggesting there exists large multi-institutional MRI heterogeneous, which could potentially influence wide applications of the GFCE-MRI technique. As such, methodologies to solve the multi-institutional data heterogeneous problem for GFCE-MRI synthesis will be an interesting area to be explored in the future. Secondly, as shown in the results of clarity of tumor-to-normal tissue interface and T-staging, there were still slight disagreements between the GFCE-MRI and real CE-MRI for the tumor and surrounding invasion risk areas, suggesting a potential to further improve the model performance, especially in regions of tumor and surrounding risk areas. We believe that this issue could be alleviated by including more MRI data and advanced deep learning architecture for model development. Then, a limitation of this study is that we did not conduct the clinical evaluations of metastasises such as neck lymph nodes and other metastatic anatomies due to the restriction of the diseased areas for model training. This could be a future study for using contrast-free MRI images of metastasises to synthesize the GFCE-MRI and evaluate its clinical efficacy.

4.6 Conclusion

In this study, we conducted a series of clinical evaluations to evaluation the potential clinical efficacy of GFCE-MRI in radiotherapy of NPC patients. Results showed that the GFCE-MRI has great potential to provide an alternative to GBCA-based CE-MRI for NPC delineation. The improvement of model generalization ability to multi-institutional MRI data and the model performance on tumor and surrounding risk areas are warranted in future study to generate more accurate multi-institutional GFCE-MRI.

5. Discussion

Gadolinium associated safety issues have raised worldwide concerns in recent years. To find an alternative to GBCA-based CE-MRI, in this study, we applied DL-assisted GFCE-MRI technique to the field of radiotherapy and successfully developed a MMgSN-Net to synthesize GFCE-MRI that tailored for NPC patients. Then, we investigated and improved the MMgSN-Net model generalizability using multi-institutional MRI data and patient-based data normalization, respectively. With the assistance of two radiation oncologists and a clinical physicist, we conducted a series of clinical evaluations to explore the clinical efficacy of the synthetic GFCE-MRI using multi-institutional MRI data. In this discussion, we attempt to highlight our key findings and limitations of the current research and provide our considerations for future research.

5.1 Current key findings and limitations

In this study, we for the first time developed a MMgSN-Net to synthesize the GFCE-

MRI that tailored for NPC patients. The major novelties of the MMgSN-Net are: (i) this is the first model to synthesize the GFCE-MRI for application of radiotherapy and the first study for NPC patients; (ii) the MMgSN-Net has the ability to extract the complementary information from input T1w and T2w MRI for GFCE-MRI synthesis; (iii) the quantitative results indicate the MMgSN-Net outperforms state-of-the-art U-Net, CycleGAN, and Hi-Net, and the Turing test results show the clinical oncologists were difficult to differentiate the GFCE-MRI from the real CE-MRI. Nevertheless, there are some limitations about the current work: (i) during the model development, MMgSN-Net was trained with small-sized NPC data from single institution. Synthesis failure may be observed for specific patients with unseen pattern; (ii) the MMgSN-Net is a supervised model. The performance of MMgSN-Net highly relies on the input-target alignment performance.

A generalizable GFCE-MRI model is highly needed for clinical practice, which enables the trained model could be directly used in external data. Training a generalizable GFCE-MRI model is challenging due to the highly heterogenous external MRI data. To investigate the MMgSN-Net model generalizability and explore potential solutions to improve the model generalizability, we utilized MRI from seven institutions to train and test different models with different normalization methods. According to our results, we found that using multi-institutional MRI for model training was helpful for improving the model generalizability. We also observed that Z-Score normalization makes multi-institutional results comparable and helps model generalizability improvement compared to the wildly used Min-Max normalization. The main limitations of this study are: (i) the patient samples of external testing datasets

were limited; and (ii) our findings were obtained from the proposed MMgSN-Net. More models could be involved to further validate our current observations.

Clinical evaluation of the GFCE-MRI is essential for bench-to-bedside translation of this technique. To assess the clinical efficacy of the synthetic GFCE-MRI, we conducted a series of clinical evaluations. Our results indicate that the GFCE-MRI is highly promising for NPC delineation. The dosimetric differences between synthetic patients and real patients were less than 1%. However, the main limitations of this study are: (i) the results were obtained from two oncologists from the same cancer center. Considering the large NPC delineation variations across oncologists, more oncologists from different cancer centers are warranted to obtain a more robust conclusion; (ii) the contrast enhancement accuracy of tumor surrounding risk regions still needs to be improved, this could be alleviated by improving the model performance by including more patient samples for model training or making the model focus on tumor and surrounding risk regions during training.

5.2 Future directions

There are still several aspects need to be explored in future research for more comprehensive analysis.

Firstly, though we have utilized multi-institutional data for model generalizability analysis, these institutions are all located in Hong Kong, and the number of patients were limited. The patients from different countries or regions may have bias in MRI characteristics. In addition, due to the limited number of patient

samples, we did not focus on specific patient group such as children and old populations. MRI characteristics such as image shape and tumor size may vary among sub-age groups. Herein, in terms of patient characteristics, a larger number of MRI data from different geographic patients and sub-age groups are warranted for more robust analysis.

Then, MRI data from more institutions should be involved to develop a generalization GFCE-MRI model in future study. Though the data of our work is retrieved from multiple institutions, which were generated from different patients and scanned with varies image protocols using different image machines, we did not focus on investigate the influence of scanning machine or imaging protocols to model generalizability due to the number of patients from each scanning machine and imaging protocol is limited. Currently we know that multi-institutional data is critical to validate or improve the model generalizability. Nevertheless, the heterogeneity of MRI, such as different MRI contrasts, resolutions, noises, or artifacts etc., are mostly caused by different imaging parameters and conditions of scanning machines in specific institutions, which makes a well-trained model cannot generalize to external MRI. So, the problem of multi-institutional data is a matter of the imaging parameters and scanning machine conditions for MRI data. Combining with the results we have observed in our current work, we consider that besides the patient characteristics, a key factor to improve the GFCE-MRI model generalizability is to involve the MRI data from various MRI scanners (E.g., the MRI machines from different manufactures with different field strength and model, etc.) with multiple scanning parameters (e.g., TE, TR, number of excitation (Nex), etc.) for model development. Additionally, there are two factors that may affect the model generalizability: the type of GBCA used during

CE-MRI imaging and the scanning start time and duration, which may affect the tumor contrast in generated CE-MRI. Due to the CE-MRI imaging is a dynamic process, the degree of contrast enhancement in CE-MRI for specific patients depends on the harmonization between scanning start time and patient metabolism velocity. The difference of contrast enhancement in CE-MRI may increase the MRI heterogeneity of CE-MRI, even with the same scanning conditions. Considering the patient privacy, collecting MRI data from multiple institutions, and combining them to develop a generalizable GFCE-MRI model is challenging. In recent years, federated learning has been proposed to protect patient privacy. Federated learning (Li et al., 2020) is a machine learning technique that enables integrating multi-institutional data for model training while without data sharing, which is potential for developing a GFCE-MRI model with higher accuracy and improved generalizability. Herein, we believe the application of federated learning in GFCE-MRI synthesis should be another interesting future direction to improve the GFCE-MRI model generalizability.

Finally, investigation of GFCE-MRI technique in radiotherapy of other cancer types than NPC could be another future research direction. Though GFCE-MRI technique in other cancer types such as brain cancer, liver cancer and breast cancer has been investigated, these studies were focus on non-radiotherapy applications such as disease diagnosis. In radiotherapy, the CE-MRI is used for different applications, such as target delineation and tumor staging, so the application of GFCE-MRI technique in radiotherapy has its own specialty and will face different challenges (E.g., radiotherapy pays more attention on the size of tumor and its boundary). Moreover, radiotherapy in different cancer types will face anatomy-specific challenges, such as the respiratory

motion of the liver in liver cancer radiotherapy. Feasibility studies of GFCE-MRI in different cancer types will be another interesting area to be explored in the future.

6. Conclusion

In this study, we for the first time developed a MMgSN-Net for synthesizing GFCE-MRI for radiotherapy of NPC patients. To apply the MMgSN-Net from bench to bedside, we further investigated and improved the generalizability of MMgSN-Net, and clinically evaluated the efficacy of the synthetic GFCE-MRI using multi-institutional data. To the best of our knowledge, this is the first work to apply DL to synthesize the GFCE-MRI for radiotherapy of NPC patients. The application GFCE-MRI technique to radiotherapy of other cancer types is warranted for future investigations.

7. References

Adoga, A. A., Kokong, D. D., Ma'an, N. D., Silas, O. A., Dauda, A. M., Yaro, J. P., Mugu, J. G., Mgbachi, C. J., & Yabak, C. J. J. S. A. j. o. c. (2018). The epidemiology, treatment, and determinants of outcome of primary head and neck cancers at the Jos University Teaching Hospital. *7*(03), 183-187.

Adventist Hospital. (2022, January 1, 2022). *Fees and Charges*. Adventist Hospital. Retrieved January 1 from https://www.hkah.org.hk/en/fees-and-charges/ancillary-services-fees/diagnostic-imaging-department

Amin, J., Sharif, M., Yasmin, M., & Fernandes, S. L. J. F. G. C. S. (2018). Big data analysis for brain tumor detection: Deep convolutional neural networks. *87*, 290-297.

Arega, T. W., Legrand, F., Bricq, S., & Meriaudeau, F. (2021). Using MRI-specific Data Augmentation to Enhance the Segmentation of Right Ventricle in Multi-disease, Multi-center and Multi-view Cardiac MRI. International Workshop on Statistical Atlases and Computational Models of the Heart,

ASCO. (2020). *Nasopharyngeal Cancer: Types of Treatment*. https://www.cancer.net/cancer-types/nasopharyngeal-cancer/types-treatment

Bahig, H., Koay, E., Barkati, M., Fuller, D. C., & Menard, C. (2019). Clinical Applications of MRI in Radiotherapy Planning. In *MRI for Radiotherapy* (pp. 55-70). Springer.

Balagopal, A., Nguyen, D., Morgan, H., Weng, Y., Dohopolski, M., Lin, M.-H., Barkousaraie, A. S., Gonzalez, Y., Garant, A., & Desai, N. (2021). A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. *Medical image analysis*, *72*, 102101.

Blanchard, P., Nguyen, F., Moya-Plana, A., Pignon, J., Even, C., Bidault, F., Temam, S., Ruffier, A., & Tao, Y. J. C. R. (2018). Nouveautés dans la prise en charge des carcinomes nasopharyngés. *22*(6-7), 492-495.

Bône, A., Ammari, S., Lamarque, J.-P., Elhaik, M., Chouzenoux, É., Nicolas, F., Robert, P., Balleyguier, C., Lassau, N., & Rohé, M.-M. (2021). Contrast-enhanced brain MRI synthesis with deep learning: key input modalities and asymptotic performance. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI),

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. J. C. a. c. j. f. c. (2018). Global cancer statistics 2018: GLOBOCAN estimates of

incidence and mortality worldwide for 36 cancers in 185 countries. *68*(6), 394-424.

Broome, D. R., Girguis, M. S., Baron, P. W., Cottrell, A. C., Kjellin, I., & Kirk, G. A. J. A. J. o. R. (2007). Gadodiamide-associated nephrogenic systemic fibrosis: why radiologists should be concerned. *188*(2), 586-592.

Buell, P. J. C. r. (1974). The effect of migration on the risk of nasopharyngeal cancer among Chinese. *34*(5), 1189-1191.

Calabrese, E., Rudie, J. D., Rauschecker, A. M., Villanueva-Meyer, J. E., & Cha, S. J. R. A. I. (2021). Feasibility of Simulated Postcontrast MRI of Glioblastomas and Lower Grade Gliomas Using 3D Fully Convolutional Neural Networks. e200276.

Cao, Y. (2011). The promise of dynamic contrast-enhanced imaging in radiation therapy. Seminars in radiation oncology,

Chang, E. T., Ye, W., Zeng, Y.-X., & Adami, H.-O. (2021). The evolving epidemiology of nasopharyngeal carcinoma. *Cancer Epidemiology, Biomarkers & Prevention, 30*(6), 1035-1047.

Chang, H.-H., Zhuang, A. H., Valentino, D. J., & Chu, W.-C. (2009). Performance measure characterization for evaluating neuroimage segmentation algorithms. *Neuroimage, 47*(1), 122-135.

Chen, C., Raymond, C., Speier, B., Jin, X., Cloughesy, T. F., Enzmann, D., Ellingson, B. M., & Arnold, C. W. J. a. p. a. (2021). Synthesizing MR Image Contrast Enhancement Using 3D High-resolution ConvNets.

Chen, C., Raymond, C., Speier, W., Jin, X., Cloughesy, T. F., Enzmann, D., Ellingson, B. M., & Arnold, C. W. (2022). Synthesizing MR image contrast enhancement using 3D high-resolution ConvNets. *IEEE Transactions on Biomedical Engineering*.

Chen, Y.-P., Chan, A. T., Le, Q.-T., Blanchard, P., Sun, Y., & Ma, J. J. T. L. (2019). Nasopharyngeal carcinoma. *394*(10192), 64-80.

Cheng, J., Dong, L., & Lapata, M. J. a. p. a. (2016). Long short-term memory-networks for machine reading.

Cheng, J. C.-H., Yuan, A., Chen, J.-H., Lu, Y.-C., Cho, K.-H., Wu, J.-K., Wu, C.-J., Chang, Y.-C., & Yang, P.-C. J. P. o. (2013). Early detection of Lewis lung carcinoma tumor control by irradiation using diffusion-weighted and dynamic contrast-enhanced MRI. *8*(5), e62762.

Chikui, T., Obara, M., Simonetti, A. W., Ohga, M., Koga, S., Kawano, S., Matsuo, Y.,

Kamintani, T., Shiraishi, T., & Kitamoto, E. J. I. j. o. d. (2012). The principal of dynamic contrast enhanced MRI, the method of pharmacokinetic analysis, and its application in the head and neck region. *2012*.

Chua, M. L., Wee, J. T., Hui, E. P., & Chan, A. T. J. T. L. (2016). Nasopharyngeal carcinoma. *387*(10022), 1012-1024.

Diop, A. D., Braidy, C., Habouchi, A., Niang, K., Gageanu, C., Boyer, L., & Chabrot, P. J. A. J. o. R. (2013). Unenhanced 3D turbo spin-echo MR angiography of lower limbs in peripheral arterial disease: a comparative study with gadolinium-enhanced MR angiography. *200*(5), 1145-1150.

Dou, Q., So, T. Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., Li, Z., Si, W., Lee, H. H., & Yu, K. (2021). Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ digital medicine*, *4*(1), 1-11.

Enterline, D. (Jan 2021). A Review of MR Contrast Agents: Why Gadolinium Matters Today. *Applied Radiology*.

Fei, N., Gao, Y., Lu, Z., & Xiang, T. (2021). Z-score normalization, hubness, and few-shot learning. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Fletcher, S., & Islam, M. Z. (2018). Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems*, *22*.

Flood, T. F., Stence, N. V., Maloney, J. A., & Mirsky, D. M. (2017). Pediatric brain: repeated exposure to linear gadolinium-based contrast material is associated with increased signal intensity at unenhanced T1-weighted MR imaging. *Radiology*, *282*(1), 222-228.

Frangi, A. F., Tsaftaris, S. A., & Prince, J. L. J. I. t. o. m. i. (2018). Simulation and synthesis in medical imaging. *37*(3), 673-679.

Fraum, T. J., Ludwig, D. R., Bashir, M. R., & Fowler, K. J. J. J. o. M. R. I. (2017). Gadolinium‐based contrast agents: a comprehensive risk assessment. *46*(2), 338-353.

Freedman, J. N., Collins, D. J., Gurney-Champion, O. J., McClelland, J. R., Nill, S., Oelfke, U., Leach, M. O., Wetscherek, A. J. R., & Oncology. (2018). Super-resolution T2-weighted 4D MRI for image guided radiotherapy. *129*(3), 486-493.

Fu, Z., Guo, X., Zhang, S., Zeng, H., Sun, K., Chen, W., & He, J. J. Z. z. l. z. z. (2018). Incidence and mortality of nasopharyngeal carcinoma in China, 2014. *40*(8), 566-571.

Gajera, V., Gupta, R., & Jana, P. K. (2016). An effective multi-objective task scheduling algorithm using min-max normalization in cloud computing. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT),

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Springer.

Gibby, W. A., Gibby, K. A., & Gibby, W. A. J. I. r. (2004). Comparison of Gd DTPA-BMA (Omniscan) versus Gd HP-DO3A (ProHance) retention in human bone tissue by inductively coupled plasma atomic emission spectroscopy. *39*(3), 138-142.

Gleneagles Hospital. (2022, January 1, 2022). *Radiology Services Fee and Charges*. Gleneagles Hospital. Retrieved January 1 from https://gleneagles.hk/fee-charges/radiology-services

Gong, E., Pauly, J. M., Wintermark, M., & Zaharchuk, G. (2018a). Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *Journal of magnetic resonance imaging*, *48*(2), 330-340.

Gong, E., Pauly, J. M., Wintermark, M., & Zaharchuk, G. J. J. o. m. r. i. (2018b). Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *48*(2), 330-340.

Granata, V., Fusco, R., Salati, S., Petrillo, A., Di Bernardo, E., Grassi, R., Palaia, R., Danti, G., La Porta, M., Cadossi, M. J. I. J. o. E. R., & Health, P. (2021). A Systematic Review about Imaging and Histopathological Findings for Detecting and Evaluating Electroporation Based Treatments Response. *18*(11), 5592.

Gribbon, K. T., & Bailey, D. G. (2004). A novel approach to real-time bilinear interpolation. Proceedings. DELTA 2004. Second IEEE international workshop on electronic design, test and applications,

Grobner, T., & Prischl, F. J. K. i. (2007). Gadolinium and nephrogenic systemic fibrosis. *72*(3), 260-264.

Grobner, T. J. N. D. T. (2006). Gadolinium–a specific trigger for the development of nephrogenic fibrosing dermopathy and nephrogenic systemic fibrosis? , *21*(4), 1104-1108.

Guo, B. J., Yang, Z. L., & Zhang, L. J. (2018). Gadolinium deposition in brain: current scientific evidence and future perspectives. *Frontiers in molecular neuroscience*, *11*, 335.

Guo, R., Mao, Y.-P., Tang, L.-L., Chen, L., Sun, Y., & Ma, J. J. T. B. j. o. r. (2019). The evolution of nasopharyngeal carcinoma staging. *92*(1102), 20190244.

Guo, Z., Guo, N., Gong, K., & Li, Q. (2019). Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Physics in Medicine & Biology*, *64*(20), 205015.

Han, X. (2017). MR‐based synthetic CT generation using a deep convolutional neural network method. *Medical physics*, *44*(4), 1408-1419.

Henke, L., Contreras, J., Green, O., Cai, B., Kim, H., Roach, M., Olsen, J., Fischer-Valuck, B., Mullen, D., & Kashani, R. J. C. O. (2018). Magnetic resonance image-guided radiotherapy (MRIgRT): A 4.5-year clinical experience. *30*(11), 720-727.

Hinton, G., Vinyals, O., & Dean, J. J. a. p. a. (2015). Distilling the knowledge in a neural network.

Ho, Y., Tsao, S.-W., Zeng, M., & Lui, V. W. Y. J. C. l. (2013). STAT3 as a therapeutic target for Epstein-Barr virus (EBV)–associated nasopharyngeal carcinoma. *330*(2), 141-149.

Holowka, S., Shroff, M., & Chavhan, G. B. (2019). Use and safety of gadolinium based contrast agents in pediatric MR imaging. *The Indian Journal of Pediatrics*, *86*(10), 961-966.

Hong Kong Baptist Hospital. (2021, July 1, 2021). *General Service Charges*. Hong Kong Baptist Hospital. Retrieved July 1 from https://www.hkbh.org.hk/fees-charges/general-services-charges/?lang=en

Hu, T., Itoh, H., Oda, M., Hayashi, Y., Lu, Z., Saiki, S., Hattori, N., Kamagata, K., Aoki, S., & Kumamaru, K. K. (2022). Enhancing Model Generalization for Substantia Nigra Segmentation Using a Test-time Normalization-Based Method. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Huynh, T., Gao, Y., Kang, J., Wang, L., Zhang, P., Lian, J., & Shen, D. J. I. t. o. m. i. (2015). Estimating CT image from MRI data using structured random forest and auto-context model. *35*(1), 174-183.

Hylton, N. J. J. C. O. (2006). Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *24*(20), 3293-3298.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Jager, E. A., Kasperts, N., Caldas-Magalhaes, J., Philippens, M. E., Pameijer, F. A., Terhaard, C. H., & Raaijmakers, C. P. (2015). GTV delineation in supraglottic laryngeal carcinoma: interobserver agreement of CT versus CT-MR delineation.

Radiation Oncology, *10*(1), 1-9.

Jakobsen, J. Å., Quattrocchi, C. C., Müller, F. H., Outteryck, O., Alcázar, A., Reith, W., Fraga, P., Panebianco, V., Sampedro, A., & Pietura, R. (2021). Patterns of use, effectiveness and safety of gadolinium contrast agents: a European prospective cross-sectional multicentre observational study. *BMC medical imaging*, *21*(1), 1-10.

Jameson, M. G., Kumar, S., Vinod, S. K., Metcalfe, P. E., & Holloway, L. C. (2014). Correlation of contouring variation with modeled outcome for conformal non-small cell lung cancer radiotherapy. *Radiotherapy and Oncology*, *112*(3), 332-336.

Jia, X., Ren, L., & Cai, J. (2020). Clinical implementation of AI technologies will require interpretable AI models. *Medical physics*, *47*(1), 1-4.

Just, N. J. B. j. o. c. (2014). Improving tumour heterogeneity MRI assessment with histograms. *111*(12), 2205-2213.

Kamran, S. C., Riaz, N., & Lee, N. J. S. O. C. (2015). Nasopharyngeal carcinoma. *24*(3), 547-561.

Kanda, T., Fukusato, T., Matsuda, M., Toyoda, K., Oba, H., Kotoku, J. i., Haruyama, T., Kitajima, K., & Furui, S. (2015a). Gadolinium-based contrast agent accumulates in the brain even in subjects without severe renal dysfunction: evaluation of autopsy brain specimens with inductively coupled plasma mass spectroscopy. *Radiology*, *276*(1), 228-232.

Kanda, T., Fukusato, T., Matsuda, M., Toyoda, K., Oba, H., Kotoku, J. i., Haruyama, T., Kitajima, K., & Furui, S. J. R. (2015b). Gadolinium-based contrast agent accumulates in the brain even in subjects without severe renal dysfunction: evaluation of autopsy brain specimens with inductively coupled plasma mass spectroscopy. *276*(1), 228-232.

Kanda, T., Ishii, K., Kawaguchi, H., Kitajima, K., & Takenaka, D. J. R. (2014). High signal intensity in the dentate nucleus and globus pallidus on unenhanced T1-weighted MR images: relationship with increasing cumulative dose of a gadolinium-based contrast material. *270*(3), 834-841.

Kazemifar, S., McGuire, S., Timmerman, R., Wardak, Z., Nguyen, D., Park, Y., Jiang, S., & Owrangi, A. (2019). MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. *Radiotherapy and Oncology*, *136*, 56-63.

Kim, E., Cho, H.-h., Ko, E., & Park, H. (2021). Generative Adversarial Network with Local Discriminator for Synthesizing Breast Contrast-Enhanced MRI. 2021 IEEE EMBS International Conference on Biomedical and Health Informatics

(BHI),

Kleesiek, J., Morshuis, J. N., Isensee, F., Deike-Hofmann, K., Paech, D., Kickingereder, P., Köthe, U., Rother, C., Forsting, M., & Wick, W. (2019a). Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study. *Investigative radiology*, *54*(10), 653-660.

Kleesiek, J., Morshuis, J. N., Isensee, F., Deike-Hofmann, K., Paech, D., Kickingereder, P., Köthe, U., Rother, C., Forsting, M., & Wick, W. J. I. r. (2019b). Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study. *54*(10), 653-660.

Lang, S. M., Alsaied, T., Moore, R. A., Rattan, M., Ryan, T. D., & Taylor, M. D. (2019). Conservative gadolinium administration to patients with Duchenne muscular dystrophy: decreasing exposure, cost, and time, without change in medical management. *The International Journal of Cardiovascular Imaging*, *35*(12), 2213-2219.

Lee, A. W., Ng, W., Chan, Y., Sze, H., Chan, C., Lam, T. J. R., & Oncology. (2012). The battle against nasopharyngeal cancer. *104*(3), 272-278.

Lee, A. W., Ng, W. T., Pan, J. J., Poh, S. S., Ahn, Y. C., AlHussain, H., Corry, J., Grau, C., Grégoire, V., & Harrington, K. J. (2018). International guideline for the delineation of the clinical target volumes (CTV) for nasopharyngeal carcinoma. *Radiotherapy and Oncology*, *126*(1), 25-36.

Lee, A. W., Zong, J. F., Pan, J. J., Choi, H. C., & Sze, H. C. (2019). Staging of Nasopharyngeal Carcinoma Based on the 8th Edition of the AJCC/UICC Staging System. In *Nasopharyngeal Carcinoma* (pp. 179-203). Elsevier.

Li, C., Sun, H., Liu, Z., Wang, M., Zheng, H., & Wang, S. (2019). Learning cross-modal deep representations for multi-modal MR image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Li, S., Xiao, J., He, L., Peng, X., & Yuan, X. (2019). The tumor target segmentation of nasopharyngeal cancer in CT images based on deep learning methods. *Technology in cancer research & treatment*, *18*, 1533033819884561.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, *37*(3), 50-60.

Li, W., Kazemifar, S., Bai, T., Nguyen, D., Weng, Y., Li, Y., Xia, J., Xiong, J., Xie, Y., Owrangi, A. J. B. P., & Express, E. (2021). Synthesizing CT images from MR images with deep learning: model generalization for different datasets through transfer learning. *7*(2), 025020.

Li, W., Lam, S., Li, T., Cheung, A. L.-Y., Xiao, H., Liu, C., Zhang, J., Teng, X., Zhi, S., & Ren, G. (2022). Multi-institutional Investigation of Model Generalizability for Virtual Contrast-Enhanced MRI Synthesis. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Li, W., Li, Y., Qin, W., Liang, X., Xu, J., Xiong, J., & Xie, Y. (2020). Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy. *Quantitative imaging in medicine and surgery*, *10*(6), 1223.

Li, W., Li, Y., Qin, W., Liang, X., Xu, J., Xiong, J., Xie, Y. J. Q. i. i. m., & surgery. (2020). Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy. *10*(6), 1223.

Li, W., Xiao, H., Li, T., Ren, G., Lam, S., Teng, X., Liu, C., Zhang, J., Lee, F. K.-h., & Au, K.-h. (2022). Virtual Contrast-Enhanced Magnetic Resonance Images Synthesis for Patients With Nasopharyngeal Carcinoma Using Multimodality-Guided Synergistic Neural Network. *International Journal of Radiation Oncology* Biology* Physics*, *112*(4), 1033-1044.

Li, W., Xiao, H., Li, T., Ren, G., Lam, S., Teng, X., Liu, C., Zhang, J., Lee, F. K.-h., & Au, K.-h. J. I. J. o. R. O. B. P. (2021). Virtual Contrast-enhanced Magnetic Resonance Images Synthesis for Patients with Nasopharyngeal Carcinoma using Multimodality-guided Synergistic Neural Network.

Liang, S.-B., Sun, Y., Liu, L.-Z., Chen, Y., Chen, L., Mao, Y.-P., Tang, L.-L., Tian, L., Lin, A.-H., & Liu, M.-Z. (2009). Extension of local disease in nasopharyngeal carcinoma detected by magnetic resonance imaging: improvement of clinical target volume delineation. *International Journal of Radiation Oncology* Biology* Physics*, *75*(3), 742-750.

Liang, X., Chen, L., Nguyen, D., Zhou, Z., Gu, X., Yang, M., Wang, J., Jiang, S. J. P. i. M., & Biology. (2019). Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. *64*(12), 125002.

Lin, M., Yu, X., Ouyang, H., Luo, D., & Zhou, C. J. S. r. (2015). Consistency of T2WI-FS/ASL fusion images in delineating the volume of nasopharyngeal carcinoma. *5*(1), 1-8.

Liu, Q., Dou, Q., Yu, L., & Heng, P. A. (2020a). MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE transactions on medical imaging*, *39*(9), 2713-2724.

Liu, Q., Dou, Q., Yu, L., & Heng, P. A. J. I. t. o. m. i. (2020b). MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data.

39(9), 2713-2724.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,

Long, M., Wang, J., Ding, G., Sun, J., & Yu, P. S. (2013). Transfer feature learning with joint distribution adaptation. Proceedings of the IEEE international conference on computer vision,

Lu, L., Cuttino, L., Barani, I., Song, S., Fatyga, M., Murphy, M., Keall, P., Siebers, J., & Williamson, J. (2006). SU - FF - J - 85: Inter - Observer Variation In The Planning Of Head/Neck Radiotherapy. *Medical physics*, *33*(6Part6), 2040-2040.

Luo, H., Zhang, T., Gong, N.-J., Tamir, J., Venkata, S. P., Xu, C., Duan, Y., Zhou, T., Zhou, F., & Zaharchuk, G. (2021a). Deep learning–based methods may minimize GBCA dosage in brain MRI. *European Radiology*, *31*(9), 6419-6428.

Luo, H., Zhang, T., Gong, N.-J., Tamir, J., Venkata, S. P., Xu, C., Duan, Y., Zhou, T., Zhou, F., & Zaharchuk, G. J. E. R. (2021b). Deep learning–based methods may minimize GBCA dosage in brain MRI. 1-10.

Mann, R. M., Kuhl, C. K., & Moy, L. J. J. o. M. R. I. (2019). Contrast - enhanced MRI for breast cancer screening. *50*(2), 377-390.

Marckmann, P., Skov, L., Rossen, K., Dupont, A., Damholt, M. B., Heaf, J. G., & Thomsen, H. S. J. J. o. t. A. S. o. N. (2006). Nephrogenic systemic fibrosis: suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging. *17*(9), 2359-2362.

Mathur, M., Jones, J. R., & Weinreb, J. C. J. R. (2020). Gadolinium deposition and nephrogenic systemic fibrosis: a radiologist's primer. *40*(1), 153-162.

Maximova, N., Gregori, M., Zennaro, F., Sonzogni, A., Simeone, R., & Zanon, D. J. R. (2016). Hepatic gadolinium deposition and reversibility after contrast agent–enhanced MR imaging of pediatric hematopoietic stem cell transplant recipients. *281*(2), 418-426.

McDermott, D. J. T. C. h. o. c. (2007). Artificial intelligence and consciousness. 117-150.

Nam, J.-m., McLaughlin, J. K., & Blot, W. J. J. J. J. o. t. N. C. I. (1992). Cigarette smoking, alcohol, and nasopharyngeal carcinoma: a case-control study among US whites. *84*(8), 619-622.

Nguyen, N. C., Molnar, T. T., Cummin, L. G., & Kanal, E. (2020a). Dentate nucleus signal intensity increases following repeated gadobenate dimeglumine

administrations: A retrospective analysis. *Radiology*, *296*(1), 122-130.

Nguyen, N. C., Molnar, T. T., Cummin, L. G., & Kanal, E. J. R. (2020b). Dentate nucleus signal intensity increases following repeated gadobenate dimeglumine administrations: A retrospective analysis. *296*(1), 122-130.

Nie, D., Cao, X., Gao, Y., Wang, L., & Shen, D. (2016). Estimating CT image from MRI data using 3D fully convolutional networks. In *Deep Learning and Data Labeling for Medical Applications* (pp. 170-178). Springer.

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017). Medical image synthesis with context-aware generative adversarial networks. International Conference on Medical Image Computing and Computer-Assisted Intervention,

O'Connor, J. P., Rose, C. J., Waterton, J. C., Carano, R. A., Parker, G. J., & Jackson, A. J. C. C. R. (2015). Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *21*(2), 249-257.

Observatory, T. G. C. (2020). *Nasopharynx*. World Health Organization Retrieved from https://gco.iarc.fr/today/data/factsheets/cancers/4-Nasopharynx-fact-sheet.pdf

Olchowy, C., Cebulski, K., Łasecki, M., Chaber, R., Olchowy, A., Kałwak, K., & Zaleska-Dorobisz, U. J. P. o. (2017). The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity-A systematic review. *12*(2), e0171704.

Parikh, A. P., Täckström, O., Das, D., & Uszkoreit, J. J. a. p. a. (2016). A decomposable attention model for natural language inference.

Pasumarthi, S., Tamir, J. I., Christensen, S., Zaharchuk, G., Zhang, T., & Gong, E. (2021a). A generic deep learning model for reduced gadolinium dose in contrast-enhanced brain MRI. *Magnetic Resonance in Medicine*, *86*(3), 1687-1700.

Pasumarthi, S., Tamir, J. I., Christensen, S., Zaharchuk, G., Zhang, T., & Gong, E. J. M. R. i. M. (2021b). A generic deep learning model for reduced gadolinium dose in contrast-enhanced brain MRI. *86*(3), 1687-1700.

Preetha, C. J., Meredig, H., Brugnara, G., Mahmutoglu, M. A., Foltyn, M., Isensee, F., Kessler, T., Pflüger, I., Schell, M., & Neuberger, U. (2021). Deep-learning-based synthesis of post-contrast T1-weighted MRI for tumour response assessment in neuro-oncology: a multicentre, retrospective cohort study. *The Lancet Digital Health*, *3*(12), e784-e794.

Raisch, D. W., Garg, V., Arabyat, R., Shen, X., Edwards, B. J., Miller, F. H., McKoy, J. M., Nardone, B., & West, D. P. J. E. o. o. d. s. (2014). Anaphylaxis associated

with gadolinium-based contrast agents: data from the Food and Drug Administration's Adverse Event Reporting System and review of case reports in the literature. *13*(1), 15-23.

Ren, G., Zhang, J., Li, T., Xiao, H., Cheung, L. Y., Ho, W. Y., Qin, J., & Cai, J. J. I. J. o. R. O. B. P. (2021). Deep learning-based computed tomography perfusion mapping (DL-CTPM) for pulmonary CT-to-perfusion translation.

Roberts, D. R., Chatterjee, A., Yazdani, M., Marebwa, B., Brown, T., Collins, H., Bolles, G., Jenrette, J. M., Nietert, P. J., & Zhu, X. (2016). Pediatric patients demonstrate progressive T1-weighted hyperintensity in the dentate nucleus following multiple doses of gadolinium-based contrast agent. *American Journal of Neuroradiology*, *37*(12), 2340-2347.

Roberts, D. R., Welsh, C. A., & Davis, W. C. (2017). Gadolinium deposition in the pediatric brain. *JAMA pediatrics*, *171*(12), 1229-1229.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., & Beer, L. J. N. M. I. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *3*(3), 199-217.

Rogosnitzky, M., & Branch, S. (2016). Gadolinium-based contrast agent toxicity: a review of known and proposed mechanisms. *Biometals*, *29*(3), 365-376.

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

Saba, T., Mohamed, A. S., El-Affendi, M., Amin, J., & Sharif, M. J. C. S. R. (2020). Brain tumor detection using fusion of hand crafted and deep learning features. *59*, 221-230.

Schmidt, M. A., Payne, G. S. J. P. i. M., & Biology. (2015). Radiotherapy planning using MRI. *60*(22), R323.

Schreier, J., Genghi, A., Laaksonen, H., Morgas, T., & Haas, B. (2020). Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT. *Radiotherapy and Oncology*, *145*, 1-6.

Semelka, R. C., Ramalho, J., Vakharia, A., AlObaidy, M., Burke, L. M., Jay, M., & Ramalho, M. J. M. r. i. (2016). Gadolinium deposition disease: initial description of a disease that has been around for a while. *34*(10), 1383-1390.

Shkolyar, E., Jia, X., Chang, T. C., Trivedi, D., Mach, K. E., Meng, M. Q.-H., Xing, L., & Liao, J. C. J. E. u. (2019). Augmented bladder tumor detection using deep

learning. *76*(6), 714-718.

Song, Y., Cheng, W., Li, H., & Liu, X. (2022). The global, regional, national burden of nasopharyngeal cancer and its attributable risk factors (1990–2019) and predictions to 2035. *Cancer Medicine*.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. J. C. a. c. j. f. c. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *71*(3), 209-249.

Thomsen, H. S. (2006). Nephrogenic systemic fibrosis: a serious late adverse reaction to gadodiamide. In (Vol. 16, pp. 2619-2621): Springer.

Thomsen, H. S. J. E. r. (2006). Nephrogenic systemic fibrosis: a serious late adverse reaction to gadodiamide. In (Vol. 16, pp. 2619-2621): Springer.

Tofts, P. S., Brix, G., Buckley, D. L., Evelhoch, J. L., Henderson, E., Knopp, M. V., Larsson, H. B., Lee, T. Y., Mayr, N. A., & Parker, G. J. J. J. o. M. R. I. A. O. J. o. t. I. S. f. M. R. i. M. (1999). Estimating kinetic parameters from dynamic contrast - enhanced T1 - weighted MRI of a diffusable tracer: standardized quantities and symbols. *10*(3), 223-232.

Tofts, P. S. J. s. (2010). T1-weighted DCE imaging concepts: modelling, acquisition and analysis. *500*(450), 400.

Torheim, T., Malinen, E., Kvaal, K., Lyng, H., Indahl, U. G., Andersen, E. K., & Futsaether, C. M. J. I. t. o. m. i. (2014). Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines. *33*(8), 1648-1656.

Tsuji, S. Y., Hwang, A., Weinberg, V., Yom, S. S., Quivey, J. M., & Xia, P. (2010). Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. *International Journal of Radiation Oncology* Biology* Physics*, *77*(3), 707-714.

Union Hospital. (2021). https://www.union.org/new/english/charges/charges_diagnostic.htm

Vajapeyam, S., Stamoulis, C., Ricci, K., Kieran, M., & Poussaint, T. Y. J. A. J. o. N. (2017). Automated processing of dynamic contrast-enhanced MRI: correlation of advanced pharmacokinetic metrics with tumor grade in pediatric brain tumors. *38*(1), 170-175.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems,

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, *312*, 135-153.

Wen, N., Cao, Y., & Cai, J. J. F. i. o. (2020). Magnetic Resonance Imaging for Radiation Therapy. *10*, 483.

Wildeman, M. A., Fles, R., Herdini, C., Indrasari, R. S., Vincent, A. D., Tjokronagoro, M., Stoker, S., Kurnianda, J., Karakullukcu, B., & Taroeno-Hariadi, K. W. (2013). Primary treatment results of nasopharyngeal carcinoma (NPC) in Yogyakarta, Indonesia. *PloS one*, *8*(5), e63706.

Wolleb, J., Sandkühler, R., Bieder, F., Barakovic, M., Hadjikhani, N., Papadopoulou, A., Yaldizli, Ö., Kuhle, J., Granziera, C., & Cattin, P. C. (2022). Learn to ignore: domain adaptation for multi-site MRI analysis. International Conference on Medical Image Computing and Computer-Assisted Intervention,

Wong, L. M., Ai, Q.-Y. H., Mo, F. K., Poon, D. M., & King, A. D. J. m. (2020). Non contrast-enhanced imaging as a replacement for contrast-enhanced imaging for MRI automatic delineation of nasopharyngeal carcinoma.

World Cancer Research Fund International. (2020). *Nasopharyngeal Cancer Statistics*. https://www.wcrf.org/cancer-trends/nasopharyngeal-cancer-statistics/

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. Proceedings of the IEEE international conference on computer vision,

Xing, L., Krupinski, E. A., & Cai, J. (2018). Artificial intelligence will soon change the landscape of medical physics research and practice. *Medical physics*, *45*(5), 1791-1793.

Xu, B.-Q., Tu, Z.-W., Tao, Y.-L., Liu, Z.-G., Li, X.-H., Yi, W., Jiang, C.-B., & Xia, Y.-F. (2016). Forty-six cases of nasopharyngeal carcinoma treated with 50 Gy radiotherapy plus hematoporphyrin derivative: 20 years of follow-up and outcomes from the Sun Yat-sen University Cancer Center. *Chinese Journal of Cancer*, *35*(1), 1-10.

Xu, B., Wang, N., Chen, T., & Li, M. J. a. p. a. (2015). Empirical evaluation of rectified activations in convolutional network.

Xu, C., Zhang, D., Chong, J., Chen, B., & Li, S. (2021a). Synthesis of gadolinium-enhanced liver tumors on nonenhanced liver MR images using pixel-level graph reinforcement learning. *Medical image analysis*, *69*, 101976.

Xu, C., Zhang, D., Chong, J., Chen, B., & Li, S. J. M. I. A. (2021b). Synthesis of gadolinium-enhanced liver tumors on nonenhanced liver MR images using pixel-level graph reinforcement learning. *69*, 101976.

Yang, J., Beadle, B. M., Garden, A. S., Schwartz, D. L., & Aristophanous, M. (2015). A multimodality segmentation framework for automatic target delineation in head and neck radiotherapy. *Medical physics*, *42*(9), 5310-5320.

Zack, G. W., Rogers, W. E., Latt, S. A. J. J. o. H., & Cytochemistry. (1977). Automatic measurement of sister chromatid exchange frequency. *25*(7), 741-753.

Zahra, M. A., Hollingsworth, K. G., Sala, E., Lomas, D. J., & Tan, L. T. J. T. l. o. (2007). Dynamic contrast-enhanced MRI as a predictor of tumour response to radiotherapy. *8*(1), 63-74.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. International conference on machine learning,

Zhang, L., Dai, J., Lu, H., He, Y., & Wang, G. (2018). A bi-directional message passing model for salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Zhang, Q., Burrage, M. K., Lukaschuk, E., Shanmuganathan, M., Popescu, I. A., Nikolaidou, C., Mills, R., Werys, K., Hann, E., & Barutcu, A. (2021). Toward replacing late gadolinium enhancement with artificial intelligence virtual native enhancement for gadolinium-free cardiovascular magnetic resonance tissue characterization in hypertrophic cardiomyopathy. *Circulation*, *144*(8), 589-599.

Zhao, J., Li, D., Kassam, Z., Howey, J., Chong, J., Chen, B., & Li, S. (2020a). Tripartite-GAN: synthesizing liver contrast-enhanced MRI to improve tumor detection. *Medical image analysis*, *63*, 101667.

Zhao, J., Li, D., Kassam, Z., Howey, J., Chong, J., Chen, B., & Li, S. J. M. i. a. (2020b). Tripartite-GAN: synthesizing liver contrast-enhanced MRI to improve tumor detection. *63*, 101667.

Zhao, P.-f., Qiao, P.-f., Niu, H., Wu, J., & Niu, G.-m. J. R. o. I. D. (2019). Application of dynamic contrast enhanced MRI in the diagnosis of brucellar spondylitis. *6*(2), 54-60.

Zhou, T., Fu, H., Chen, G., Shen, J., & Shao, L. J. I. t. o. m. i. (2020). Hi-net: hybrid-fusion network for multi-modal MR image synthesis.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision,