Model Generalizability Investigation for GFCE-MRI Synthesis in NPC Radiotherapy Using Multi-institutional Patient-based Data Normalization

Wen Li, Saikit Lam, Yinghui Wang, Chenyang Liu, Tian Li, Jens Kleesiek, Andy Lai-Yin Cheung, Ying Sun, Francis Kar-ho Lee, Kwok-hung Au, Victor Ho-fun Lee, and Jing Cai

***Abstract*— Recently, deep learning has been demonstrated to be feasible in eliminating the use of gadolinium-based contrast agents (GBCAs) through synthesizing gadolinium-free contrast-enhanced MRI (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 five popular normalization approaches. Three state-of-the-art neural networks were applied to map from T1-weighted and T2-weighted MRI to contrast-enhanced MRI (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. The five normalization methods were applied to normalize the training and testing data 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 mean absolute error (MAE) and peak signal-to-noise ratio (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 the best model generalizability improvement.**

***Index Terms*— Contrast enhanced MRI, data normalization, nasopharyngeal carcinoma**

This research was partly supported by research grants of Innovation and Technology Fund-Mainland-Hong Kong Joint Funding Scheme (ITF-MHKJFS) (MHP/005/20), Innovation and Technology Fund (ITS/080/19) of the Innovation and Technology Commission, Project of Strategic Importance Fund (P0035421) and Projects of RISA (P0043001) of The Hong Kong Polytechnic University, and Shenzhen Basic Research Program (JCYJ20210324130209023) of Shenzhen Science and Technology Innovation Committee.

Wen Li, Saikit Lam, Yinghui Wang, Chenyang Liu, Tian Li, Andy Lai-Yin Cheung, and Jing Cai are are with the Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China.

Jing Cai is with the Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China, and also with The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518000, Shenzhen, China (E-mail: jing.cai@polyu.edu.hk).

Jens Kleesiek is with the Institute for AI in Medicine (IKIM), University Hospital Essen, 45131 Essen, Germany.

Ying Sun is with the Department of Radiation Oncology, Sun Yat-sen University Cancer Center, Guangzhou, China.

Francis Kar-ho Lee, Kwok-hung Au are with the Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong SAR, China.

Victor Ho-fun Lee is with the Department of Clinical Oncology, The University of Hong Kong, Hong Kong SAR, China.

I. Introduction

Nasopharyngeal carcinoma (NPC), a highly aggressive epithelial carcinoma originating in the mucosal lining of the nasopharynx, has long been a notorious malignancy in the population of East and Southeast Asia [1]. Radiotherapy (RT) is currently the mainstay treatment modality for NPC, which achieved 66%-83% 5-year survival rate with RT alone [2]. Precise tumor delineation is the most critical prerequisite for a successful RT treatment, therefore, contrast-enhanced MRI (CE-MRI), using gadolinium-based contrast agents (GBCAs), has become an indispensable part in accurate NPC tumor delineation [3] in routine RT treatment planning practice. Nevertheless, emerging evidence has shown that nephrogenic systemic fibrosis (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 [4]. Further evidence has shown that gadolinium accumulation in the dentate nucleus and globus pallidus has been observed in paediatric patients [5]. Apart from this, gadolinium deposition was also observed in patients with normal renal function [6]. 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 [4]. 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 gadolinium-free contrast-enhanced MRI (GFCE-MRI), which serves similar purposes as the CE-MRI, through deep learning approaches [7]–[15]. 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 [16]–[18], which largely limits the wide application of the 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 chal-

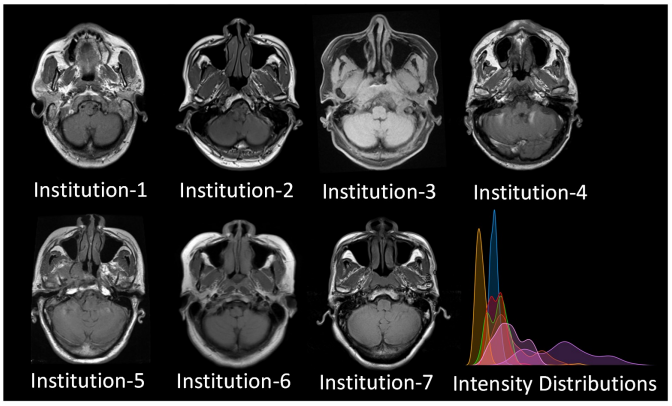

Fig. 1. Visualization of MRI data heterogeneity among the 7 institutions, in aspects of intensity distribution.

lenges: 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 (Fig. 1). These challenges have raised a growing concern of model generalizability developed using deep learning algorithms, which strongly relies on the assumption that the training data and testing data are independent and identically distributed (i.i.d.) [19]. 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 on external datasets [19]. To tackle this, one of the potential remedies to improve model generalizability is to integrate multi-institutional MRI images during model training to enlarge the view of deep learning models [20], [21], 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 [19], [22], 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. Various normalization methods have been applied to normalize data, which can be broadly categorized into linear normalization and nonlinear normalization techniques. The linear normalization linearly maps the MRI pixel intensities from original space to a target space, typically with a far smaller data range, such as Min-Max normalization [23], Z-Score normalization [24], and Decimal normalization [25], etc., while the nonlinear normalization maps the MRI pixel intensities in a nonlinear manner, such as Sigmoid normalization [26] and Tanh normalization [27]. Nevertheless, related research in multi-institutional setting that contain various real-world distributions of MRI data is severely scarce in the body of literature.

We hypothesize that minimizing the distribution variations between training and external testing MRI data by applying various data normalization techniques would improve the model generalizability. This approach is deemed practical as it waives the requirements of model architecture refinement and model training. In this study, we included MRI data from seven medical 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 delineating the authentic morphology of NPC tumors for precise radiation delivery. Hence, the success of this study may not only provide the medical community with valuable 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.

II. Methods and Materials

A. Patient Data

A total of 256 NPC patients from seven medical institutions were retrospectively collected in this study. For fair comparisons, identical amount of patients (n=71) were retrieved from Institution-1, Institution-2, and Institution-3, respectively for uni-institution and tri-institution model development. Institution-4 (n=18), Institution-5 (n=9), Institution-6 (n=9), and Institution-7 (n=7) were adopted as external datasets for evaluating generalizability of the developed models. T1-weighted (T1w) MRI, T2-weighted (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/HA HKW 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 [28]. For each of the Institution-1, Institution-2, and Institution-3, all the patients (n=71) were randomly divided into training (n=53) and internal evaluation (n=18), respectively.

B. Study Design

The overall idea of this study was firstly to apply 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. Three state-of-the-art (SOTA) neural networks were used to assess the consistency of the results. The separately and jointly trained models were referred to as uni-institution models and tri-institution models, respectively. In total, 60 models were trained. Fig. 2 illustrated the overall study design.

1) Neural Networks:

MMgSN-Net: The MMgSN-Net is a 2D deep learning algorithm [15], which consists of five key modules: multi-modality learning module, synthesis network, self-attention module, multi-level module, and a discriminator. The structure of the MMgSN-Net is illustrated in Fig. 3a. 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 subsequently transferred to the synergistic guidance system (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 size of NPC tumor can be remarkably different between patients, some patients (especially advanced T-stage patients) exhibit bulky tumor phenotypes, which may exist across different regions in image. Without application of the self-attention module, it may be technically difficult for the algorithms to effectively capture anatomical information of such large-sized tumor (e.g., morphology of the infiltrative tumor) due to the limited size of convolutional kernels. Hence, the self-attention module was added to capture anatomical information of large-sized NPC tumor, therefore enabling the MMgSN-Net to preserve the morphology of large-sized anatomic structures. NPC tumors are also highly infiltrative and polymorphic in shape, thus the contrast enhancement of tumor edge is critical for discriminating the tumor from surrounding normal tissues. Several studies [29]–[31] have shown that integrating features from multiple deep layers can improve the performance in image segmentation and, more remarkably, in tumor edge detection. Therefore, the multi-level module was used to aggregate the multi-level features for edge detection of NPC tumor. The discriminator was utilized to distinguish the synthetic GFCE-MRI from ground-truth CE-MRI, thus encouraging the synthesis network to generate more realistic GFCE-MRI.

Hi-Net: Hi-Net [32] (illustrated in Fig. 3b) was proposed for multi-modal MRI synthesis, with the aim of learning a mapping from multi-modal source images (i.e., existing modalities) to a target image modality. Hi-Net consists of four main components: a modality-specific network, a multi-modal fusion network, a multi-model synthesis network, and a discriminator. The modality-specific network was utilized to learn representations for each individual modality, while the fusion network was employed to learn the common latent representation of multi-modal data. The multi-modal synthesis network was designed to densely combine the latent representation with hierarchical features from each modality, acting as a generator to synthesize the target images. Additionally, a layer-wise multi-modal fusion strategy was presented to effectively exploit the correlations among multiple modalities. For this purpose, a Mixed Fusion Block (MFB) was proposed to adaptively weight different fusion strategies, including element-wise summation, product, and maximization.

ResViT: ResViT [33] (illustrated in Fig. 3c) is a novel generative adversarial approach used for MRI synthesis. Unlike traditional convolutional neural networks, ResViT combined the contextual sensitivity of vision transformers with the precision of convolution operators and the realism of adversarial learning. The ResViT generator featured a central bottleneck comprising novel Aggregated Residual Transformer (ART) blocks that synergistically combined residual convolutional and transformer modules to capture diverse representations. Additionally, a channel compression module was used to distill task-relevant information. To mitigate computational burden, ART blocks employed a weight sharing strategy. ResViT also introduced a unified implementation that eliminated the need to rebuild separate synthesis models for varying source-target modality configurations.

2) Data Normalization: Data normalization plays a pivotal role in model development [34]. It minimizes feature bias by transforming the features into a common space so that larger numeric feature values cannot dominate smaller numeric feature values [35]. Currently different data normalization methods have been applied in medical image translation tasks, such as Min-Max (also called scaling) [23] and Z-Score [24], Decimal [25], Sigmoid [26], and Tanh [27] etc. These normalization methods have also been applied to different objects prior to training, i.e., dataset-based, patient-based,and single-image based normalization. In natural image tasks, most studies adopted 2D networks, which typically used the statistical values of each single image or the entire dataset for data normalization [18]. For medical images, however, image and dataset-based normalization may not be 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 the 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, five patient-based normalization methods (Min-Max, Z-Score, Decimal, Sigmoid, Tanh) were applied to mitigate the data distribution variations among training datasets and external unseen datasets using the statistical values of each patient. Subsequently, impacts of different data normalization techniques on model generalizability were evaluated. The five normalization methods can be mathematically described as

$$x_{min\_max} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{1}$$

$$x_{z\_score} = \frac{x - \mu_x}{\delta_x}. \tag{2}$$

$$x_{decimal} = \frac{x}{10^j}, j = log_{10}^{max(x)}. \tag{3}$$

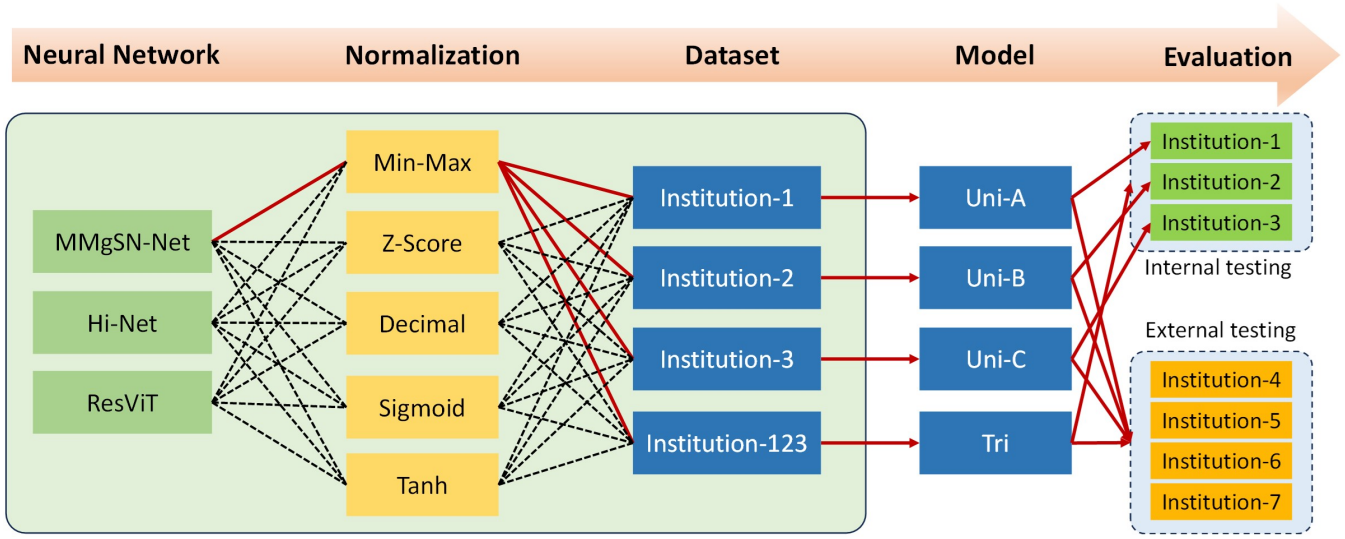

Fig. 2. Overall Study Design. To investigate the model generalizability, three networks, five normalization methods, as well as four training datasets were used to train different models. The solid red lines show the three uni-institution models and one tri-institution model that trained using MMgSN-Net and normalized with Min-Max normalization. Each model was evaluated using both the internal and four external datasets.

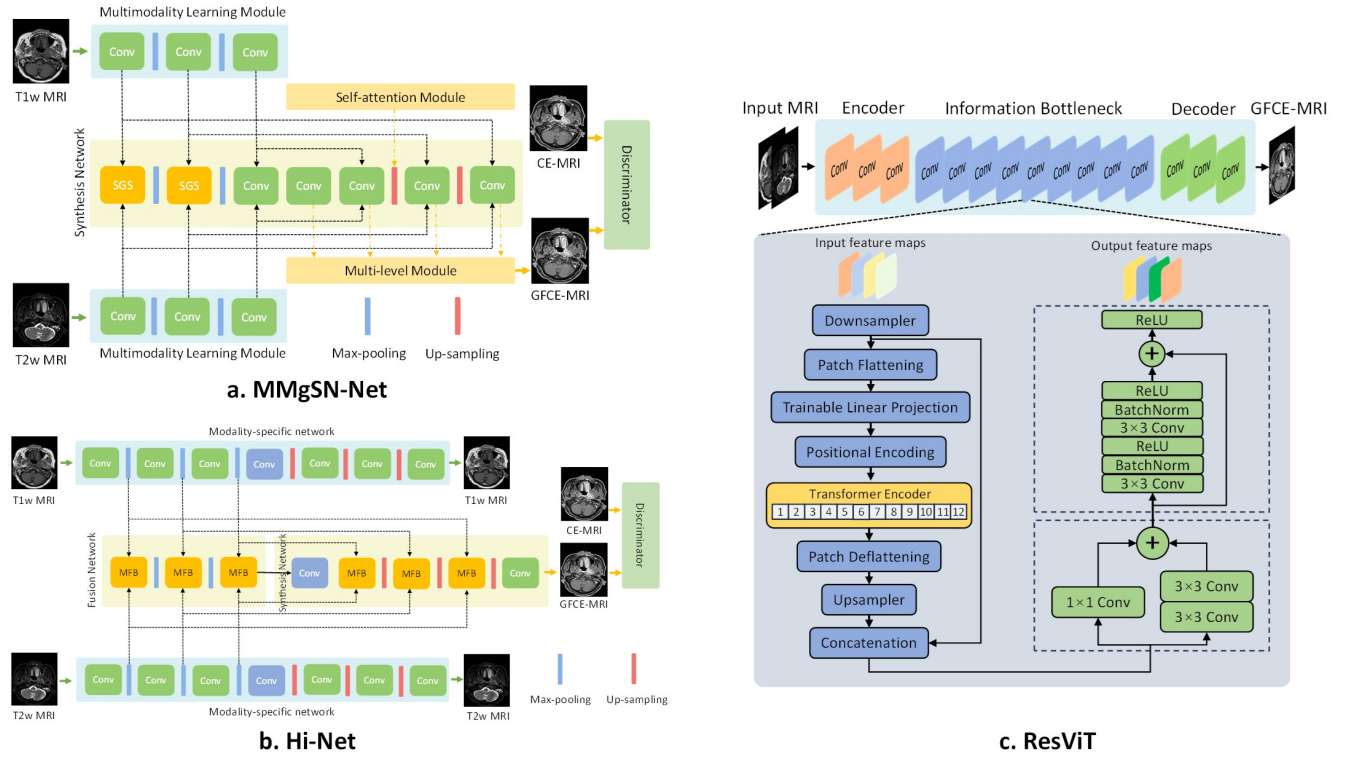

Fig. 3. Architecture of the three studied neural networks (a: MMgSN-Net; b: Hi-Net; c: ResViT). T1-weighted MRI and T2-weight MRI were used as inputs, while the gadolinium-based contrast-enhanced MRI was used as the learning target. SGS, synergistic guidance system; Conv, convolutional layers; MFB, mixed fusion block.

$$x_{sigmoid} = \frac{1}{1 + e^{-(\frac{x - \mu_x}{\delta_x})}}. \quad (4)$$

$$x_{tanh} = \frac{1}{2}\{tanh(0.01(\frac{x - \mu_x}{\delta_x}) + 1\}. \quad (5)$$

Where x represents the intensities of each patient volume, while x_{min}, x_{max}, μ_x, and δ_x are minimum value, maximum value, mean value and standard deviation of the patient. x_{min_max}, x_{z_score}, $x_{decimal}$, $x_{sigmoid}$, and x_{tanh} represent the corresponding values of patient data after Min-Max, Z-Score, Decimal, Sigmoid, and Tanh normalization methods, respectively.

The Min-Max normalization rescales the intensity range to [0, 1] and preserves the relationship among the original data distributions due to its linear transformation nature. The Z-

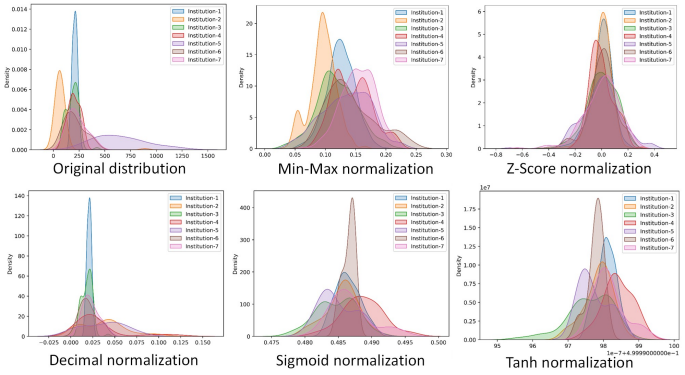

Fig. 4. Variations of data distribution among the 7 institutions, before and after application of various patient-based normalization methods.

Score method normalizes the mean value and standardization of the patient to 0 and 1 respectively, which enables the comparison of two datasets with different distributions. Similar to Mix-Max normalization, the Decimal normalization also rescales the pixel intensity range to [0, 1], which is achieved by shifting the decimal point of the pixel intensities, and the amount of shifting depends mainly on the maximum pixel value of the patient. Unlike Min-Max, Z-Score, and Decimal normalization methods, the Sigmoid and Tanh techniques are two nonlinear normalization methods. Sigmoid is typically used when data is not evenly distributed around its mean. Outliers that lie far from the mean are exponentially squashed, resulting in a more balanced distribution of pixel intensities [36]. The Tanh method was introduced by Hample and has been considered as an efficient normalization technique. It is not as sensitive to outliers as other normalization methods, making it a useful choice for datasets with extreme values [27]. As shown in Fig. 4, prior to data normalization, severe inter-institutional distribution discrepancy exists. The distribution discrepancy was mitigated after application of data normalization, especially after the Z-Score normalization method.

3) Comparison Models: To investigate how significant is the external performance degradation for the GFCE-MRI models that were trained with single-institution MRI data, we first trained three uni-institution models using data from Institution-1, Institution-2, and Institution-3 separately for each of the studied normalization methods. A total of 53 patients were used for training of each uni-institution model and 18 patients were used for internal testing to evaluate the internal performance. Each of the five normalization methods were separately applied prior to model training. The three uni-institution models were labeled as Uni-"xy", where "x" represents the first lowercase letter of the normalization method, and "y" represents the institution number of the training dataset. For example, Uni-z1 refers to the uni-institution model trained with institution-1 dataset and normalized using Z-Score normalization. The generalizability of these models was evaluated using four external datasets (i.e., Institution-4 to Institution-7).

To investigate how significant is the external performance improvement for models that were trained with diversified multi-institution MRI data, we trained the GFCE-MRI model using jointed with data from Institution-1 to Institution-3. Considering that the number of training samples may influence assessment of the tri-institution model as it could be practically difficult to analyze whether the model generalizability improvement is caused by adoption of a diverse dataset or by increasement of the training samples. Therefore, 18 patients were randonly assigned 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. Each of the five normalization methods was also applied to develop the tri-institution models prior to training. The tri-institution models with different normalization methods were labeled as Tri-"X", where "X" represents the first uppercase letter of the normalization method. The four datasets from Institution-4 to Institution-7 were used for external testing to evaluate the model generalizability.

C. Evaluations

1) Quantitative Evaluation: To quantitatively evaluate the performance of uni- and tri-institution models, mean absolute error (MAE) and peak signal-to-noise ratio (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 [15], [37], [38]. Smaller MAE and larger PSNR values indicate better quantitative results. Prior to quantitative evaluation, we rescaled the intensities of the CE-MRI and the predicted GFCE-MRI data to [0, 1] for computing the underlying percentage differences for multi-institutional data comparison, taking into account the intensity distribution variations. Paired two-tailed t-test (significance level, p=0.05) was performed to analyze if there was statistically significant difference between results from different models.

$$MAE = \frac{\sum_{i=1}^{N} |y_i - f(x_i)|}{n}. \tag{6}$$

$$PSNR = 20 \cdot \lg \frac{max(y_x) \cdot \sqrt{n}}{\|y_i - f(x_i)\|_2}. \tag{7}$$

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 intensities of the CE-MRI and GFCE-MRI data to [0, 1].

2) Qualitative Evaluation: To visually assess the performance of the developed models on external datasets, the trained uni- and tri-institution models were directly applied to the external datasets without fine-tuning. Prior to results inference, the five patient-based normalization methods were applied to uni-institution models and tri-institution models for comparisons of results between external datasets. The input T1w, T2w MRI and ground-truth CE-MRI were shown alongside the GFCE-MRI generated from different models.

III. Results

A. Quantitative Results

The quantitative results of the 60 uni-institution and tri-institution models are illustrated in Fig. 5. MAE values beyond 100 on external datasets and those exceed 80 on the internal datasets were truncated for the sake of highlighting more useful information. The first and the second rows in Fig. 5 show the external and internal results, respectively, for MMgSN-Net (left), ResViT (middle) and Hi-Net (right). Each bar represents the result from a model that was trained with a specific network and normalization method. For example, the four red bars from left to right in Fig. 5(a) represent the results from the MMgSN-Net model, which was trained with data from Institution-1, Institution-2, Institution-3, and Institution-123 using Z-Score normalization, and tested directly on the four external datasets. All numeric quantitative results are supplemented in Sup. Table I.

1) Generalizability of single-institution models: All uni-institution models suffered from severe performance drop on external MRI datasets across all normalization methods, despite the use of neural networks. Table I presents the percentage performance drop of synthetic GFCE-MRI for the uni-institution models on the four external datasets, as compared to the internal results. As MAE and PSNR had similar trends, MAE was used here as an indicator to illustrate the results. Among the five normalization methods, the Decimal normalization obtained the greatest performance drop, with percentage MAE drop ranging from 76.63% to 287.18% across the three neural networks. The Tanh model obtained a relative lower external performance drop (15.94% using ResViT and 19.81% using Hi-Net), while the internal results of the uni-institution models using Tanh normalization are dissatisfactory (as shown in Fig. 5 (e) and (f)), with a MAE of 37.59 ± 4.68, 43.75 ± 11.08, 36.84 ± 5.67 for uni-t1, uni-t2, and uni-t3 respectively using ResViT, and 24.95 ± 2.72, 128.49 ± 11.24, 78.2 ± 11.33 for uni-t1, uni-t2, and uni-t3 respectively using Hi-Net, please refer to Sup. Table I for more details. For a comparison, Z-Score normalization achieved the best internal results, with a MAE of 23.87 ± 3.39, 24.98 ± 4.78, and 26.99 ± 6.24 for uni-z1, uni-z2, and uni-z3 respectively using ResViT, and 23.23 ± 2.97, 27.47 ± 4.5, and 28.29 ± 6.19 for uni-z1, uni-z2, and uni-z3 respectively using Hi-Net, please refer to Sup. Table I for more details. As shown in Fig. 5(a)-(c), the Z-Score normalization also achieved the best external performance on all (no significant difference with Uni-t1 using Hi-Net) uni-institution models across the three neural networks with the lowest MAE between GFCE-MRI and real CE-MRI, suggesting that the model trained with Z-Score normalization could attain superior generalizability when trained using single-institution MRI data.

2) Generalizability of tri-institution models: The model generalizability was improved when training the model with more diverse MRI data for majority of the all comparing normalization methods except Decimal normalization on MMgSN-Net and ResViT, as shown in Table II. Besides the Decimal normalization, the overall external performance obtained 7.34%-16.21%, 10.82%-21.41%, and 5.37%-25.01% improvement for MMgSN-Net, ResViT, and Hi-Net respectively across different normalization methods, suggesting that increasing the diversity of training data helps improving the model generalizability.

Apart from the external performance improvement of tri-institution models, it was also observed that tri-institution models obtained comparable internal results when compared with its single-institution counterparts, as shown in Fig. 5(d)-(f). The bars in Fig. 5(d)-(f) with the same color show the models trained using the same normalization method but with different datasets. For example, the Uni-z1, Uni-z2 and Uni-z3 using MMgSN-Net obtained a MAE of 23.03 ± 3.18, 24.87 ± 4.64 and 26.84 ± 6.17 respectively on their respective intra-institution testing set, while Tri-Z obtained an average MAE of 25.60 ± 4.91 on all of the three testing datasets. It should be noted that unlike the uni-institution models, the internal testing results of tri-institution models are average of the three internal testing datasets (Institution-1, 2, and 3) instead of a single testing dataset, since the data from all three institutions (1/3 of each) were involved in the tri-institution model training. As shown in Fig. 5(d)-(f), the average internal results of tri-institution models does not indicate a remarkable performance degradation when compared with the three uni-institution models (indicated by black dashed lines). In contrast, the uni-institution models suffered from substantial performance drops when tested on MRI data from other institutions, highlighting the advantages of incorporating a more diverse MRI data for model development.

3) Influence of normalization methods to model generalizability: The quantitative results from Fig. 5 indicate that Z-Score normalization outperformed all comparing normalization methods on external datasets across three neural networks, with the lowest MAE of 32.45 ± 6.22, 32.53 ± 7.27 and 33.43 ± 7.26 for MMgSN-Net, ResViT and Hi-Net, respectively. The Tanh normalization is network-dependent. The models trained with the Tanh normalization using MMgSN-Net and ResViT outperformed Decimal, Sigmoid and Min-Max, but obtained inferior performance on Uni-t2, Uni-t3 and Tri-T when trained using Hi-Net. This may be attributed to the compression of effective image information to a small pixel intensity scale (1e-7) after application of the Tanh normalization, as illustrated in Fig. 4. This compression renders the effective information unrecognizable by Hi-Net. Among the five normalization methods, Decimal normalization obtained the worst external results across the three neural networks, with the highest MAE of 358.05 ± 210.14 for Uni-d2 that was trained using Hi-Net. This may partly be explained by the difference of pixel range of T1w MRI, T2w MRI, and CE-MRI for some patients in our datasets. For example, the maximum values of T1w MRI and CE-MRI for one patient in institution-2 are 813 and 403 respectively (less than 1000), while the maximum value of T2w MRI is 1105 (larger than 1000), after Decimal normalization, the range of T1w MRI and CE-MRI becomes [0-0.813] and [0-0.403] respectively, while the range for T2w MRI becomes [0-0.1105], which may in turn interfere the model training procedure, suggesting that Decimal normalization may not be suitable for our GFCE-MRI synthesis task or other MRI-related task considering the diverse MRI pixel intensity ranges. In this study, different normalization methods produced varying degrees of impact

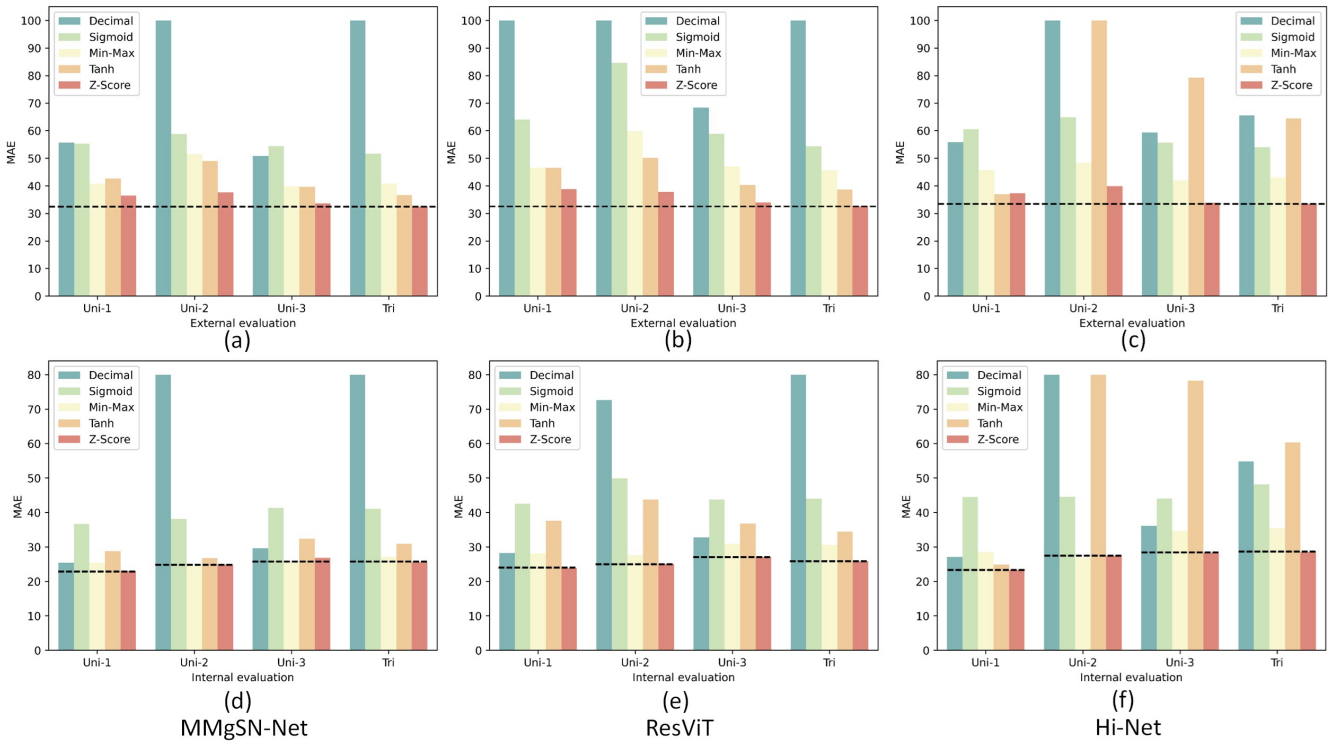

Fig. 5. Quantitative results of the 60 models on the external test datasets (a-c) and internal test datasets (d-f). Each bar represents the MAE result of a model that was trained with a specific network (left-most column: MMgSN-Net; middle column: ResViT; right-most column: Hi-Net) and normalization method.

TABLE I
EXTERNAL PERFORMANCE DROP OF UNI-INSTITUTION MODELS, FROM TOP TO BOTTOM ARE RESULTS FROM MMGSN-NET, RESVIT AND HI-NET, RESPECTIVELY

Min-Max			Z-Score			Decimal			Sigmoid			Tanh		
Model	MAE	PSNR	Model	MAE	PSNR	Model	MAE	PSNR	Model	MAE	PSNR	Model	MAE	PSNR
Uni-m1	60.42%	12.32%	Uni-z1	58.62%	10.70%	Uni-d1	119.11%	16.58%	Uni-s1	50.85%	10.56%	Uni-t1	48.23%	8.48%
Uni-m2	110.67%	13.89%	Uni-z2	51.15%	6.75%	Uni-d2	211.55%	38.83%	Uni-s2	54.09%	8.57%	Uni-t2	83.03%	6.34%
Uni-m3	34.37%	5.18%	Uni-z3	25.30%	3.00%	Uni-d3	71.52%	9.42%	Uni-s3	31.65%	4.39%	Uni-t3	22.55%	2.13%
Overall	***68.49%***	***10.46%***	***Overall***	***44.42%***	***6.82%***	***Overall***	***134.06%***	***21.61%***	***Overall***	***45.53%***	***7.84%***	***Overall***	***51.27%***	***5.65%***
Uni-m1	65.50%	11.08%	Uni-z1	62.51%	11.11%	Uni-d1	460.31%	20.09%	Uni-s1	50.45%	9.92%	Uni-t1	23.89%	5.54%
Uni-m2	116.84%	13.37%	Uni-z2	51.44%	5.91%	Uni-d2	292.45%	41.36%	Uni-s2	69.56%	9.31%	Uni-t2	14.63%	2.37%
Uni-m3	52.22%	6.85%	Uni-z3	25.97%	0.99%	Uni-d3	108.79%	11.06%	Uni-s3	34.60%	4.85%	Uni-t3	9.31%	1.12%
Overall	***78.19%***	***10.43%***	***Overall***	***46.64%***	***6.00%***	***Overall***	***287.18%***	***24.17%***	***Overall***	***51.54%***	***8.03%***	***Overall***	***15.94%***	***3.01%***
Uni-m1	60.34%	11.01%	Uni-z1	60.44%	11.97%	Uni-d1	106.05%	15.06%	Uni-s1	36.16%	8.17%	Uni-t1	47.82%	3.79%
Uni-m2	80.02%	10.22%	Uni-z2	45.14%	6.19%	Uni-d2	59.60%	32.21%	Uni-s2	45.47%	7.37%	Uni-t2	10.24%	-0.09%
Uni-m3	21.28%	3.71%	Uni-z3	19.69%	2.67%	Uni-d3	64.24%	9.03%	Uni-s3	26.20%	4.81%	Uni-t3	1.36%	0.50%
Overall	***53.88%***	***8.31%***	***Overall***	***41.76%***	***6.94%***	***Overall***	***76.63%***	***18.77%***	***Overall***	***35.94%***	***6.78%***	***Overall***	***19.81%***	***1.40%***

TABLE II
EXTERNAL PERFORMANCE IMPROVEMENT OF TRI-INSTITUTION MODELS, FROM TOP TO BOTTOM ARE RESULTS FROM MMGSN-NET, RESVIT AND HI-NET, RESPECTIVELY

Min-Max			Z-Score			Decimal			Sigmoid			Tanh		
Model	MAE	PSNR	Model	MAE	PSNR	Model	MAE	PSNR	Model	MAE	PSNR	Model	MAE	PSNR
Tri-M	7.34%	1.57%	Tri-Z	9.66%	2.36%	Tri-D	-9.01%	-1.54%	Tri-S	7.89%	2.28%	Tri-T	16.21%	6.17%
Tri-M	10.82%	1.56%	Tri-Z	11.78%	2.40%	Tri-D	-0.92%	-6.49%	Tri-S	21.41%	4.57%	Tri-T	15.23%	3.51%
Tri-M	5.37%	1.48%	Tri-Z	9.65%	2.99%	Tri-D	58.44%	17.13%	Tri-S	10.48%	3.21%	Tri-T	25.01%	13.76%

on model external performance, indicating that normalization methods exerted tremendous influence on the model generalizability, even when the models were trained with the same MRI data. The consistent results of the three neural networks demonstrated that the multi-institution MRI data normalized with Z-Score normalization achieved an improved model generalizability, which outperformed other comparing normalization methods.

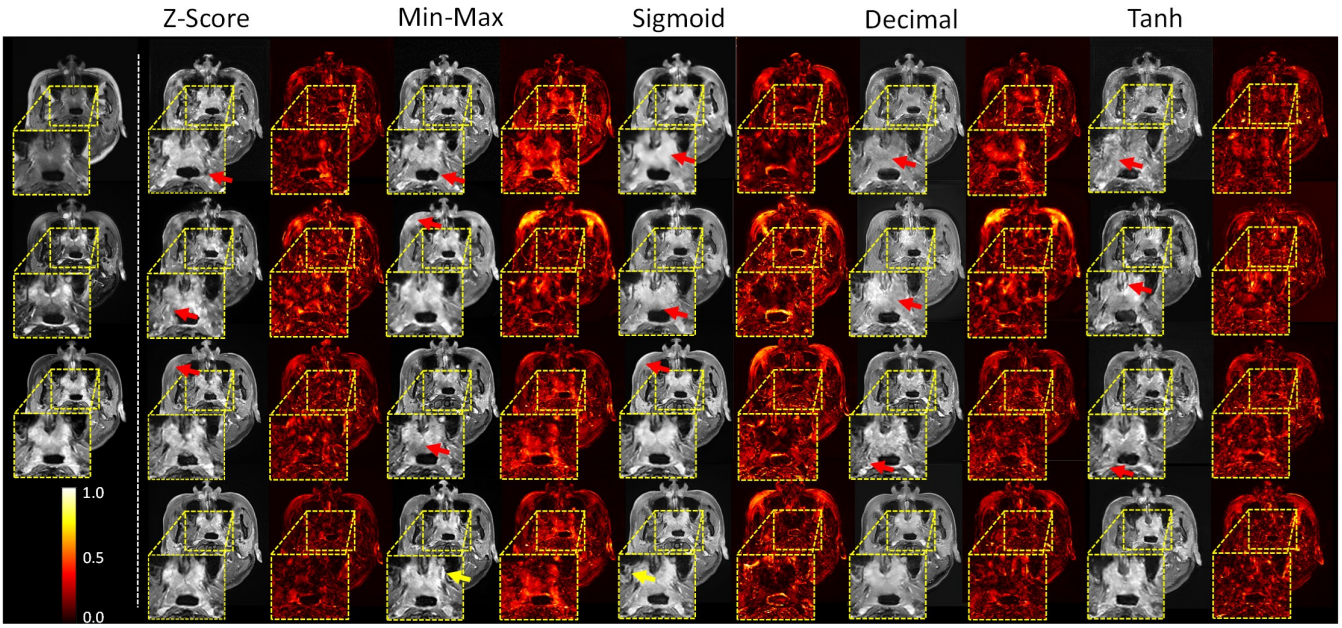

Fig. 6. Illustration of GFCE-MRI generated from uni-institution and tri-institution models using different normalization methods.

B. 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 Fig. 6. The first column from top to bottom are input T1w MRI, T2w MRI and real CE-MRI respectively. Other columns from top to bottom are synthetic GFCE-MRI from three uni-institution models and the tri-institution model using the five normalization methods, respectively. The red images are difference maps between synthetic GFCE-MRI and real CE-MRI. The NPC tumors were zoomed in using yellow boxes for better visualization of region of interest. As shown in Fig. 6, the GFCE-MRI generated from the model that was trained using multi-institution MRI data and normalized with Z-Score normalization outperformed other models, resulting in a better approximation of the real CE-MRI reflected by a darker appearance of the difference map for both tumor region and the entire image.

The uni-institution models suffered from varying degrees of performance degradation on external MRI data with diverse contrast enhancement failures in tumor and tumor surrounding areas (indicated with red arrows), especially in the models trained with Institution-1 and Institution-2 data (with overall image contrast difference and blurring anatomic structure). The five tri-institution models obtained improved generalizability on external data, which achieved an improved visual approximation of tumor contrast enhancement compared to their uni-institution counterparts. Compared with Tri-Z model that was normalized with Z-Score normalization, Tri-M model and Tri-S model that were normalized with Min-Max normalization and Sigmoid normalization, respectively, obtained an inferior enhancement of tumor surrounding structures (as indicated with yellow arrows in Fig. 6).

IV. Discussion

In RT, CE-MRI is commonly used for accurate tumor delineation, especially for the highly infiltrative NPC [15]. 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 [7]–[15]. 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 five different data normalization techniques 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 Fig. 5 and Table I, the quantitative results show that the uni-institution models performed well on internal testing datasets with lower MAE while they failed to generalize to external unseen data (i.e., with greater MAE and lower PSNR on external datasets). The visual comparisons (Fig. 6) of synthetic GFCE-MRI among different models also underscored that the uni-institution models failed to predict the correct contrast enhancement, both in tumor and surrounding structures. These results suggest that there exist significant MRI data bias across institutions, resulting in a phenomenon that performance of well-trained in-house models fail to generalize to external MRI datasets. As shown in Table I, uni-instiution models trained with different intra-institution dataset obtained varied performance drop on the same external dataset (e.g., the percentage MAE drop of

MMgSN-Net normalized with Min-Max normalization ranged from 34.37% to 110.67%), this may also be caused by the MRI data bias among the three training datasets. These data biases may result from variations in MRI data characteristics, such as image contrast, resolution, texture, artifacts, etc., as well as differences in imaging scanners and scanning parameters across different institutions. (as shown in Fig. 1). In addition, with the same training data and network structure (e.g., Fig. 5(a)),the external performance of the models trained with different normalization methods are were markedly different (with the MAE values ranging from 36.53 ± 6.5 to 55.72 ± 22.91), indicating that normalization methods do influence the model generalizability. A possible reason might be that different normalization methods mitigate the variations 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-institution models have been improved compared to uni-institution models, even with the same number of training samples (as shown in Table II). This result indicates that involving diverse MRI from multiple institutions is more capable of achieving a better model generalizability, possibly due to the enlarged view of the model. 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. On the other hand, the tri-institution models did not obtain obvious performance degradation on the three internal testing datasets (as demonstrated in section *Results A. 2)*), indicating that involving diverse MRI data from multiple institutions for model development is also capable of maintaining the intra-institution accuracy, though the five tri-institution models were trained with only 1/3 number of samples from each individual institution.

Z-Score normalization outperformed other comparison normalization methods in improving the model generalizability, for both uni-institution models and the tri-institution model. As shown in Fig. 5, Z-Score normalization achieved the best performance, compared to other normalization methods, with the lowest MAE of 32.45 ± 6.22, 32.53 ± 7.27 and 33.43 ± 7.26 for MMgSN-Net, ResViT and Hi-Net respectively. This is possibly attributed to the fact that the Z-Score method normalizes all the patients' mean and standard deviation to the same value (0 and 1, respectively), which effectively minimized the distribution variations among all training patients and external testing patients (as shown in Fig. 4). Decimal normalization and Tanh normalization may not suitable for our GFCE-MRI synthesis task due to the diversity of original pixel intensity ranges across patients and the messy pixel ranges after Decimal normalization. It is worth noting that the Tanh normalization is network-dependent with unstable results using Hi-Net, this is possibly due to the small effective pixel intensity scale after application of Tanh normalization (as shown in Fig. 4 and equation (5)). For other two normalization methods, Min-Max outperformed Sigmoid in both internal and external evaluation, with lower MAE and higher PSNR for both uni-institution models and tri-institution models, as shown in Fig. 5.

Intriguingly, the three studied neural networks demonstrated the consistency of our findings. As shown in Fig. 5 (a-c), the external performance of Z-Score normalization consistently outperformed other comparing normalization methods on all the three studied networks (MMgSN-Net, Hi-Net, and ResViT). Apart from this, Table I also demonstrated that all the uni-institution models of the three studied networks suffered from tremendous performance drop on external datasets, in terms of MAE and PSNR. By contrast, the model trained with multi-institution MRI data produced improved performance on the external dataset across the three studied networks, irrespective of the normalization methods used (as shown in Table II, here the Decimal normalization is not considered due to the inappropriate application scenario).

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 not identical with different mean values and standard deviations, making the results not comparable. As demonstrated in [21], the model trained with data of smaller mean intensity data lead to significantly better intra-institution quantitative results, even with the same number of training samples. Application of different normalization methods will further normalize the multi-institutional data to different distributions, making the normalized results uninterpretable. To quantitatively evaluate the results generated from different normalization methods, we used normalized values (to [0-1]) to compute the percentage difference and supplemented with percentage external drop and improvement results (as shown in Table I and Table II) instead of the absolute values for the sake of comparing the model performance. For the multi-institutional setting, the Z-Score normalization may be a promising method for results interpretation compared to other normalizations. For instance, Min-Max and Decimal preserves the original data distribution across institutions, while the Z-Score method normalizes 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, as the scope of this work was centered on the GFCE-MRI synthesis for NPC patients, applicability of our findings in other tasks deserves future investigation considering the scarcity of the multi-institutional data. Secondly, this work takes into account the diversity of MRI pixel intensities across institutions, as shown in Fig. 4, after application of data normalization, small distribution variations still exist among different institutional data, these variations may be caused by the image-based factors such as image texture, artifacts, and tumor size etc. As demonstrated in [39], MRI-specific data augmentation provides a promising solution to further enhance the model generalizability in aspect of training image, which will be considered in future work to further improve the model generalizability.

V. 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 influencing 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 data generally resulted in higher generalizability on external unseen data 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 a multi-institutional MRI setting, which outperformed other comparing normalization methods.

REFERENCES

[1] E. T. Chang, W. Ye, Y.-X. Zeng, and H.-O. Adami, "The evolving epidemiology of nasopharyngeal carcinoma," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 30, no. 6, pp. 1035–1047, 2021.

[2] B.-Q. Xu, Z.-W. Tu, Y.-L. Tao, Z.-G. Liu, X.-H. Li, W. Yi, C.-B. Jiang, and Y.-F. Xia, "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*, vol. 35, no. 1, pp. 1–10, 2016.

[3] A. W. Lee, W. T. Ng, J. J. Pan, S. S. Poh, Y. C. Ahn, H. AlHussain, J. Corry, C. Grau, V. Grégoire, and K. J. Harrington, "International guideline for the delineation of the clinical target volumes (ctv) for nasopharyngeal carcinoma," *Radiotherapy and Oncology*, vol. 126, no. 1, pp. 25–36, 2018.

[4] S. Holowka, M. Shroff, and G. B. Chavhan, "Use and safety of gadolinium based contrast agents in pediatric mr imaging," *The Indian Journal of Pediatrics*, vol. 86, no. 10, pp. 961–966, 2019.

[5] D. R. Roberts, C. A. Welsh, and W. C. Davis, "Gadolinium deposition in the pediatric brain," *JAMA pediatrics*, vol. 171, no. 12, pp. 1229–1229, 2017.

[6] D. R. Roberts, A. Chatterjee, M. Yazdani, B. Marebwa, T. Brown, H. Collins, G. Bolles, J. M. Jenrette, P. J. Nietert, and X. Zhu, "Pediatric patients demonstrate progressive t1-weighted hyperintensity in the dentate nucleus following multiple doses of gadolinium-based contrast agent," *American Journal of Neuroradiology*, vol. 37, no. 12, pp. 2340–2347, 2016.

[7] J. Kleesiek, J. N. Morshuis, F. Isensee, K. Deike-Hofmann, D. Paech, P. Kickingereder, U. Köthe, C. Rother, M. Forsting, and W. Wick, "Can virtual contrast enhancement in brain mri replace gadolinium?: a feasibility study," *Investigative radiology*, vol. 54, no. 10, pp. 653–660, 2019.

[8] A. Bone, S. Ammari, J.-P. Lamarque, M. Elhaik, E. Chouzenoux, F. Nicolas, P. Robert, C. Balleyguier, N. Lassau, and M.-M. Rohé, "Contrast-enhanced brain mri synthesis with deep learning: key input modalities and asymptotic performance," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, Conference Proceedings, pp. 1159–1163.

[9] E. Gong, J. M. Pauly, M. Wintermark, and G. Zaharchuk, "Deep learning enables reduced gadolinium dose for contrast-enhanced brain mri," *Journal of magnetic resonance imaging*, vol. 48, no. 2, pp. 330–340, 2018.

[10] H. Luo, T. Zhang, N.-J. Gong, J. Tamir, S. P. Venkata, C. Xu, Y. Duan, T. Zhou, F. Zhou, and G. Zaharchuk, "Deep learning–based methods may minimize gbca dosage in brain mri," *European Radiology*, vol. 31, no. 9, pp. 6419–6428, 2021.

[11] S. Pasumarthi, J. I. Tamir, S. Christensen, G. Zaharchuk, T. Zhang, and E. Gong, "A generic deep learning model for reduced gadolinium dose in contrast-enhanced brain mri," *Magnetic Resonance in Medicine*, vol. 86, no. 3, pp. 1687–1700, 2021.

[12] C. Xu, D. Zhang, J. Chong, B. Chen, and S. Li, "Synthesis of gadolinium-enhanced liver tumors on nonenhanced liver mr images using pixel-level graph reinforcement learning," *Medical Image Analysis*, vol. 69, p. 101976, 2021.

[13] C. Chen, C. Raymond, W. Speier, X. Jin, T. F. Cloughesy, D. Enzmann, B. M. Ellingson, and C. W. Arnold, "Synthesizing mr image contrast enhancement using 3d high-resolution convnets," *IEEE Transactions on Biomedical Engineering*, 2022.

[14] J. Zhao, D. Li, Z. Kassam, J. Howey, J. Chong, B. Chen, and S. Li, "Tripartite-gan: synthesizing liver contrast-enhanced mri to improve tumor detection," *Medical image analysis*, vol. 63, p. 101667, 2020.

[15] W. Li, H. Xiao, T. Li, G. Ren, S. Lam, X. Teng, C. Liu, J. Zhang, F. K.-h. Lee, and K.-h. Au, "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*, vol. 112, no. 4, pp. 1033–1044, 2022.

[16] L. Xing, E. A. Krupinski, and J. Cai, "Artificial intelligence will soon change the landscape of medical physics research and practice," *Medical physics*, vol. 45, no. 5, pp. 1791–1793, 2018.

[17] X. Jia, L. Ren, and J. Cai, "Clinical implementation of ai technologies will require interpretable ai models," *Medical physics*, vol. 47, no. 1, pp. 1–4, 2020.

[18] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.

[19] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[20] Q. Dou, T. Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H. H. Lee, and K. Yu, "Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–11, 2021.

[21] W. Li, S. Lam, T. Li, A. L.-Y. Cheung, H. Xiao, C. Liu, J. Zhang, X. Teng, S. Zhi, and G. Ren, "Multi-institutional investigation of model generalizability for virtual contrast-enhanced mri synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, Conference Proceedings, pp. 765–773.

[22] J. Wolleb, R. Sandkühler, F. Bieder, M. Barakovic, N. Hadjikhani, A. Papadopoulou, O. Yaldizli, J. Kuhle, C. Granziera, and P. C. Cattin, "Learn to ignore: domain adaptation for multi-site mri analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, Conference Proceedings, pp. 725–735.

[23] V. Gajera, R. Gupta, and P. K. Jana, "An effective multi-objective task scheduling algorithm using min-max normalization in cloud computing," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016, Conference Proceedings, pp. 812–816.

[24] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, Conference Proceedings, pp. 142–151.

[25] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[26] B. D. Fulcher and A. Fornito, "A transcriptional signature of hub connectivity in the mouse connectome," *Proceedings of the National Academy of Sciences*, vol. 113, no. 5, pp. 1435–1440, 2016.

[27] S. Bhanja and A. Das, "Impact of data normalization on deep neural network for time series forecasting," *arXiv preprint arXiv:1812.05519*, 2018.

[28] K. T. Gribbon and D. G. Bailey, "A novel approach to real-time bilinear interpolation," in *Proceedings. DELTA 2004. Second IEEE international workshop on electronic design, test and applications*. IEEE, 2004, Conference Proceedings, pp. 126–131.

[29] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[31] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1741–1750.

[32] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-net: hybrid-fusion network for multi-modal mr image synthesis," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2772–2781, 2020.

[33] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: Residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[34] T. Hu, H. Itoh, M. Oda, Y. Hayashi, Z. Lu, S. Saiki, N. Hattori, K. Kamagata, S. Aoki, and K. K. Kumamaru, "Enhancing model generalization for substantia nigra segmentation using a test-time normalization-based method," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, Conference Proceedings, pp. 736–744.

[35] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015, vol. 72.
[36] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020.
[37] W. Li, Y. Li, W. Qin, X. Liang, J. Xu, J. Xiong, and Y. Xie, "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*, vol. 10, no. 6, p. 1223, 2020.
[38] X. Han, "Mr-based synthetic ct generation using a deep convolutional neural network method," *Medical physics*, vol. 44, no. 4, pp. 1408–1419, 2017.
[39] T. W. Arega, F. Legrand, S. Bricq, and F. Meriaudeau, "Using mri-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center and multi-view cardiac mri," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2021, Conference Proceedings, pp. 250–258.