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

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, 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 two popular normalization approaches. A multimodality-guided synergistic neural network (MMgSN-Net) was 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. 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 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 improved model generalizability.**

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

I. INTRODUCTION

Nasopharyngeal carcinoma (NPC), 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 [1]. Radiotherapy (RT) is currently the mainstay treatment modality for NPC, which achieved 66%-83% 5-year survival rate for NPC patients with RT alone [2]. Precise tumor delineation is the most critical prerequisite for a successful RT treatment. 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 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

---

This research was partly supported by research grants of Innovation and Technology Fund (ITS/080/19) of the Innovation and Technology Commission, Project of Strategic Importance Fund (P0035421) of The Hong Kong Polytechnic University, Shenzhen-Hong Kong-Macau ST Program (Category C) (SGDX20201103095002019) and Shenzhen Basic Research Program (JCYJ20210324130209023) of Shenzhen Science and Technology Innovation Committee.

Wen Li, Saikit Lam, 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. (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.

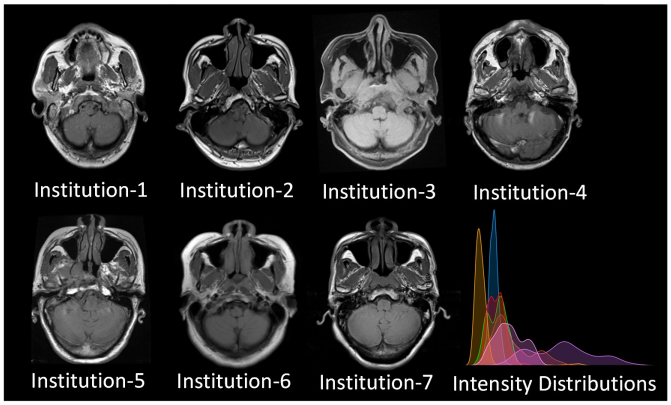


Fig. 1. Illustration of heterogeneity of multi-institutional MRI data.

deviations, and intensity ranges (Fig. 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.) [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 [19]. 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 [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. 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 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.

II. METHODS AND MATERIALS

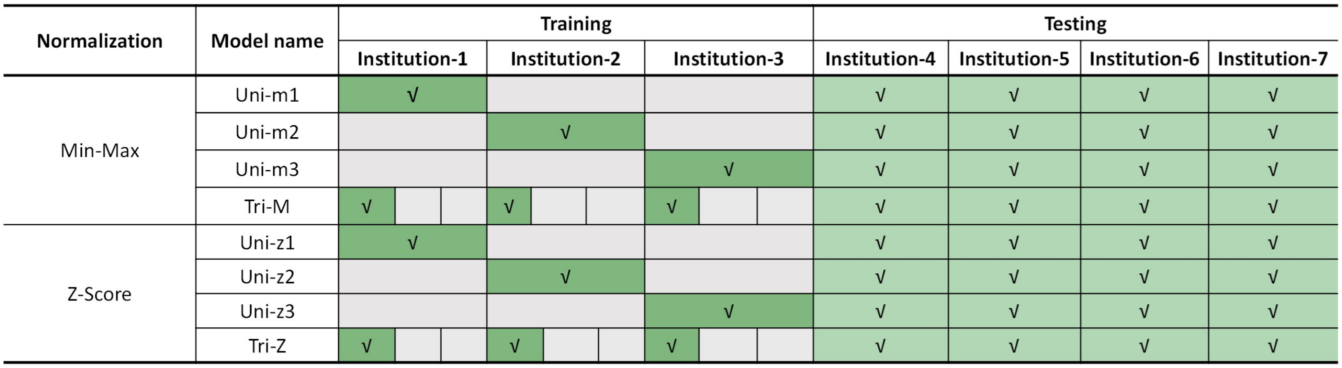
A. Patient Data

A total of 256 NPC patients from seven medical institutions were retrospectively collected in this study. For fair comparisons, 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. 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 [23]. For Institution-1, Institution-2, and Institution-3, the 71 patients were randomly divided into 53 and 18 for model training and validation, respectively.

B. 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. Fig. 2 illustrated the overall study design.

1) Neural Network: The multimodality-guided synergistic neural network (MMgSN-Net) was used as the base network in this study. The MMgSN-Net is a 2D deep learning algorithm [15], 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 Fig. 3. 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 synergistic guidance system (SGS) in synthesis network for complementary feature selection and fusion. In the decoder of synthesis network, the fused features



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

Fig. 2. The Overall Study Design.

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 CE-MRI, thus encouraging the synthesis network to generate more realistic GFCE-MRI.

2) Data Normalization: Data normalization plays a pivotal role in model development [24]. It minimizes feature bias by transforming the features into a common space so that larger numeric feature values cannot dominate smaller numeric feature values [25]. Currently different data normalization methods are applied in medical image translation tasks. The most popular two normalization methods are Min-Max (also called scaling) [26] and Z-Score [27]. 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 [18]. 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

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

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

Where x represent 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} 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 Fig. 4, 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.

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

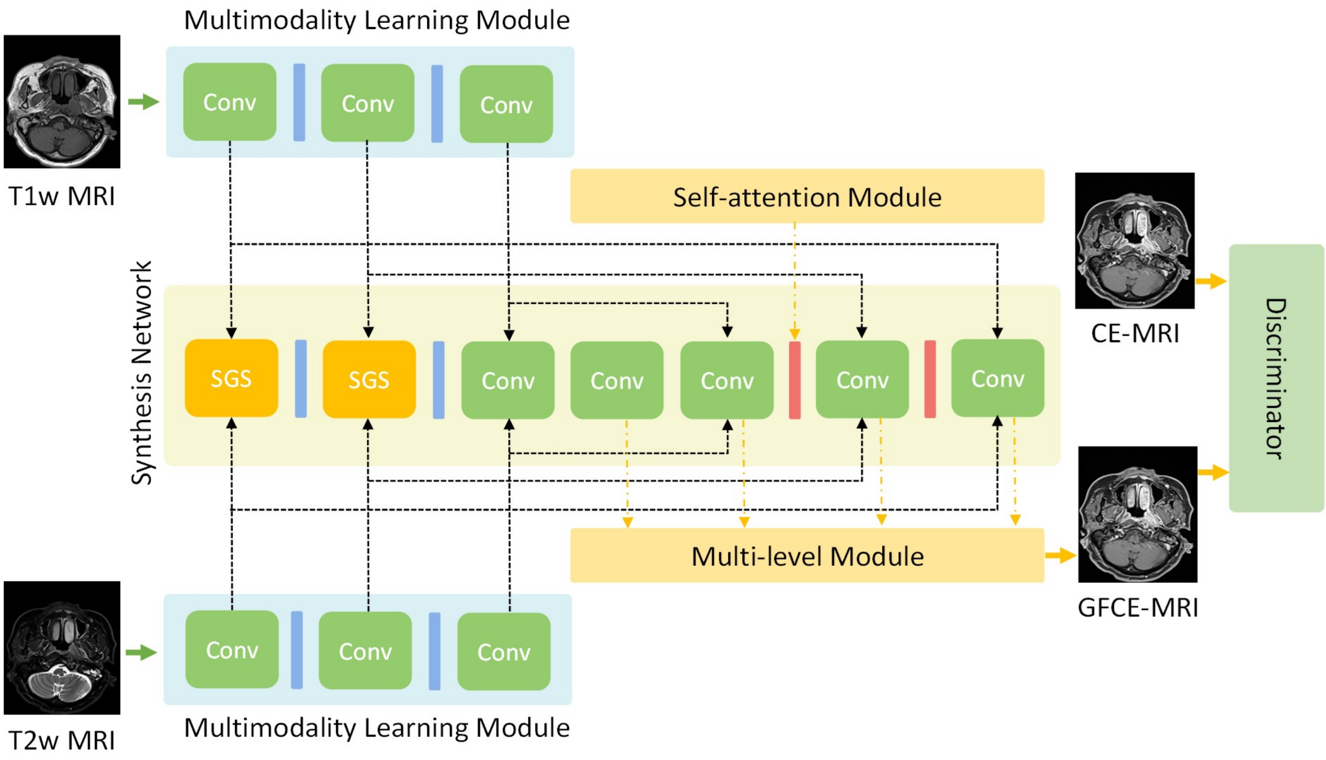
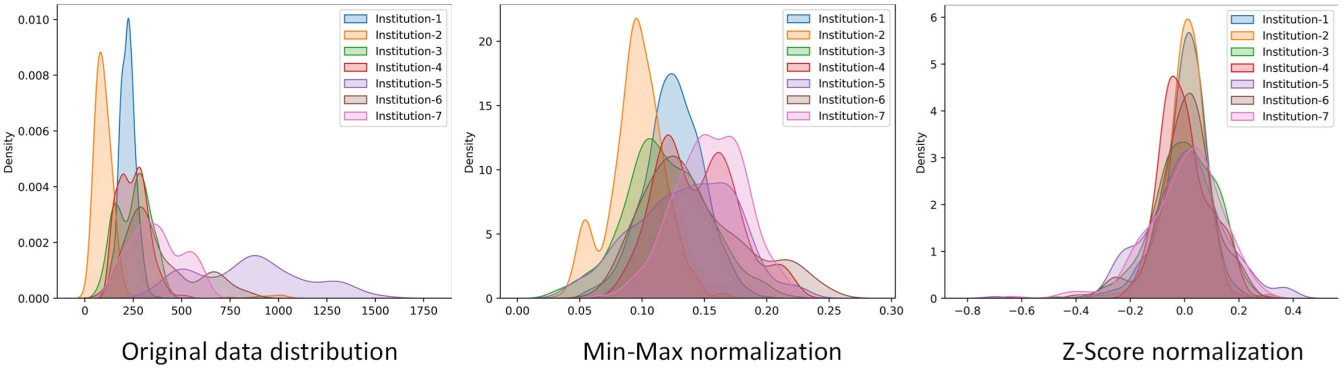


Fig. 3. The architecture of MMgSN-Net. It is a two-inputs network consisting of five key components: multimodality learning module, synthesis network, self-attention module, multi-level module, and a discriminator. T1-weighted MRI and T2-weight MRI were used as inputs, gadolinium-based contrast-enhanced MRI was used as the learning target. SGS, synergistic guidance system; Conv, convolutional layers.



Original data distribution

Min-Max normalization

Z-Score normalization

Fig. 4. 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.

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.

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], [28], [29]. 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.

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

$$PSNR = 20 \cdot \lg \frac{max(y_x) \cdot \sqrt{n}}{\|y_i - f(x_i)\|_2}. \tag{4}$$

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].

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.

III. RESULTS

A. 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 I and Table II 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 40.73 ± 7.65 for Uni-m1, 24.45 ± 3.67 to 51.51 ± 9.29 for Uni-m2, 25.56 ± 6.92 to 39.72 ± 9.12 for Uni-m3, and from 23.03 ± 3.18 to 36.53 ± 6.5 for Uni-z1, 24.87 ± 4.64 to 37.59 ± 7.35 for Uni-z2, 26.84 ± 6.17 to 33.63 ± 6.41 for Uni-z3, respectively. The percentage of uni-institution models' external performance degradation were shown in Table III. The average performance drop for MAE were 68.49% and 44.42% 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 110.67% drop) for Min-Max normalization and Uni-z1 (with 58.62% 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.

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 IV, the overall external performance obtained 7.34% improvement for Tri-M model and 9.66% improvement for Tri-Z model in MAE and 1.57% improvement for Tri-M model and 2.36% improvement for Tri-Z model in PSNR.

TABLE I
INTERNAL AND EXTERNAL QUANTITATIVE RESULTS USING MIN-MAX NORMALIZATION

Model	Testing	MAE ± SD ($\times 10^3$)	PSNR ± SD
Uni-m1	***Institution-1***	***25.39 ± 3.59***	***33.45 ± 1.38***
	Institution-4	52.12 ± 10.89	27.65 ± 1.72
	Institution-5	35.03 ± 6.56	30.47 ± 1.24
	Institution-6	34.97 ± 4.02	31.65 ± 0.67
	Institution-7	40.80 ± 9.12	29.35 ± 1.51
	Overall	***40.73 ± 7.65***	***29.78 ± 1.29***
Uni-m2	***Institution-2***	***24.45 ± 3.67***	***32.17 ± 0.89***
	Institution-4	50.26 ± 7.11	27.50 ± 0.95
	Institution-5	51.76 ± 6.28	27.83 ± 1.02
	Institution-6	58.74 ± 19.93	27.05 ± 2.13
	Institution-7	45.27 ± 3.83	28.41 ± 0.73
	Overall	***51.51 ± 9.29***	***27.70 ± 1.21***
Uni-m3	***Institution-3***	***25.56 ± 6.92***	***31.30 ± 1.72***
	Institution-4	44.53 ± 7.63	28.51 ± 1.32
	Institution-5	35.67 ± 5.09	30.09 ± 0.78
	Institution-6	45.36 ± 15.96	29.41 ± 2.08
	Institution-7	33.30 ± 7.81	30.69 ± 1.48
	Overall	***39.72 ± 9.12***	***29.68 ± 1.42***
Tri-M	Institution-1	26.27 ± 4.01	33.06 ± 1.30
	Institution-2	26.27 ± 4.19	31.74 ± 0.86
	Institution-3	28.91 ± 6.38	31.45 ± 2.05
	Overall	***27.15 ± 4.86***	***32.08 ± 1.40***
	Institution-4	41.82 ± 7.82	28.97 ± 1.20
	Institution-5	41.55 ± 9.04	29.19 ± 1.51
	Institution-6	46.12 ± 13.55	29.29 ± 1.84
	Institution-7	33.53 ± 8.21	30.57 ± 1.56
	Overall	***40.76 ± 9.66***	***29.51 ± 1.53***

TABLE II
INTERNAL AND EXTERNAL QUANTITATIVE RESULTS USING Z-SCORE NORMALIZATION

Model	Testing	MAE ± SD ($\times 10^3$)	PSNR ± SD
Uni-z1	***Institution-1***	***23.03 ± 3.18***	***34.21 ± 1.58***
	Institution-4	43.10 ± 5.91	28.96 ± 1.20
	Institution-5	32.74 ± 6.27	31.03 ± 1.16
	Institution-6	32.07 ± 5.05	32.36 ± 1.07
	Institution-7	38.22 ± 8.77	29.84 ± 1.42
	Overall	***36.53 ± 6.5***	***30.55 ± 1.21***
Uni-z2	***Institution-2***	***24.87 ± 4.64***	***32.28 ± 1.10***
	Institution-4	48.47 ± 7.30	27.62 ± 1.22
	Institution-5	31.35 ± 7.52	31.33 ± 1.51
	Institution-6	33.27 ± 5.23	31.68 ± 1.14
	Institution-7	37.27 ± 9.36	29.76 ± 1.57
	Overall	***37.59 ± 7.35***	***30.10 ± 1.36***
Uni-z3	***Institution-3***	***26.84 ± 6.17***	***31.97 ± 2.09***
	Institution-4	38.30 ± 5.53	29.50 ± 1.21
	Institution-5	31.92 ± 7.32	31.06 ± 1.42
	Institution-6	30.78 ± 4.70	32.52 ± 1.08
	Institution-7	33.51 ± 8.08	30.95 ± 1.50
	Overall	***33.63 ± 6.41***	***31.01 ± 1.30***
Tri-Z	Institution-1	23.71 ± 3.12	33.72 ± 1.43
	Institution-2	25.74 ± 4.80	32.01 ± 1.10
	Institution-3	27.36 ± 6.80	31.87 ± 2.23
	Overall	***25.60 ± 4.91***	***32.53 ± 1.59***
	Institution-4	37.20 ± 5.14	29.72 ± 1.21
	Institution-5	29.94 ± 6.43	31.69 ± 1.25
	Institution-6	29.60 ± 4.94	32.78 ± 1.12
	Institution-7	33.04 ± 8.38	30.87 ± 1.57
	Overall	***32.45 ± 6.22***	***31.27 ± 1.04***

3) Influence of normalization methods to model generalizability: The quantitative results from Table III and Table IV indicate that Z-Score normalization outperformed the Min-Max normalization on external datasets, with less average performance drop for uni-institution models (44.42% v.s. 68.49% for MAE and 6.82% v.s. 10.46% for PSNR, respectively) and more average improvement for tri-institution models (9.66% v.s. 7.34% for MAE and 2.36% v.s. 1.57% for PSNR). Moreover, as shown in Table I and Table II, though the overall external performance of Tri-M outperformed Uni-m2 but with comparable external performance with Uni-m1 and slightly

TABLE III
EXTERNAL PERFORMANCE DROP OF UNI-INSTITUTION MODELS

Min-Max			Z-Score		
Model	MAE	PSNR	Model	MAE	PSNR
Uni-m1	60.42%	12.32%	Uni-z1	58.62%	10.70%
Uni-m2	110.67%	13.89%	Uni-z2	51.15%	6.75%
Uni-m3	34.37%	5.18%	Uni-z3	25.30%	3.00%
Overall	***68.49%***	***10.46%***	***Overall***	***44.42%***	***6.82%***

TABLE IV
EXTERNAL PERFORMANCE IMPROVEMENT OF TRI-INSTITUTION MODELS

Model	MAE	PSNR
Tri-M	7.34%	1.57%
Tri-Z	9.66%	2.36%

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.

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. 5. 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 Uni-z1) 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.

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).

IV. DISCUSSION

In radiotherapy, 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 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 I and Table II, 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 (Fig. 5) 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 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 Fig. 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 IV). 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 I and Table II, though the external performance of Tri-M model obtained 7.34% overall improvement in MAE and 1.57% improvement in PSNR on the four external datasets, 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

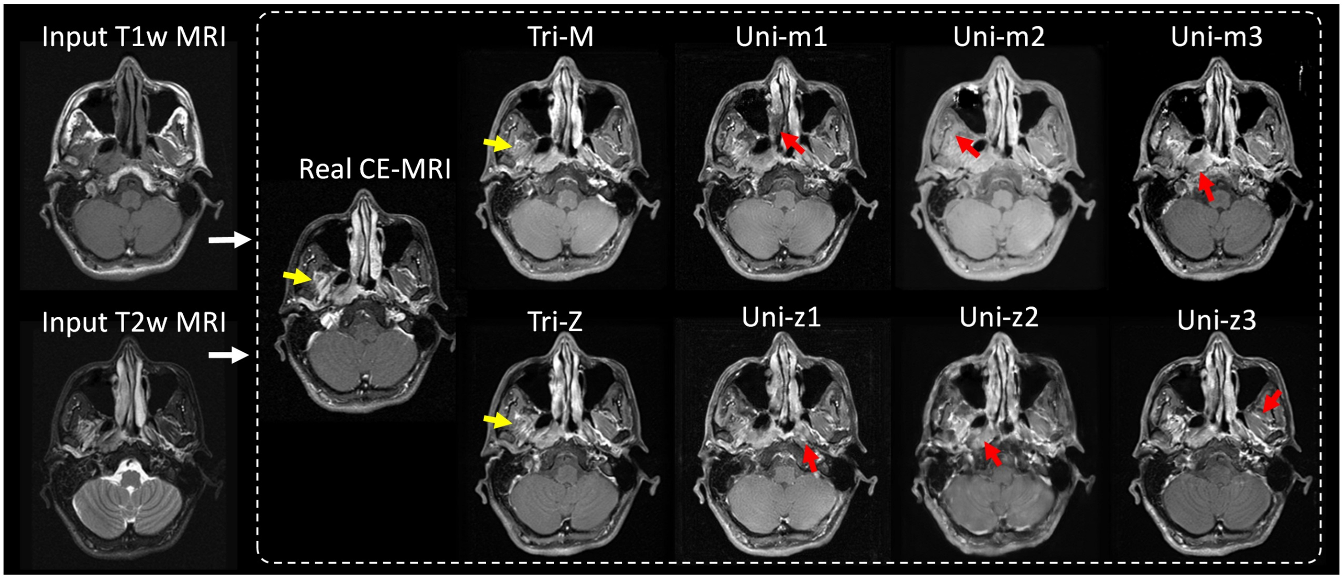


Fig. 5. Illustration of GFCE-MRI generated from uni-institution and tri-institution models using Min-Max normalization and Z-Score normalization.

(normalized with Z-Score normalization) achieved improved results compared to all uni-institution models, indicting that Z-Score normalization is capable of futher 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 III and Table IV, Z-Score normalization achieved 24.07% and 3.64% less external performance drop of MAE and PSNR respectively than Min-Max normalization for uni-institution models. With Z-Score normalization, the tri-institution model Tri-Z also obtained additional 2.32% and 0.79% performance gain in MAE 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 Fig. 4), 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 [21], 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-m3 model obtained better internal performance compared with Uni-z3 in MAE (25.56 v.s. 26.84), 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 III) 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 [15], 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 Fig. 4, 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 [30], 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.

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 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.

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. Bône, S. Ammari, J.-P. Lamarque, M. Elhaik, 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, 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, Conference Proceedings, pp. 765–773.

[22] J. Wolleb, R. Sandkühler, F. Bieder, M. Barakovic, N. Hadjikhani, A. Papadopoulou, 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, Conference Proceedings, pp. 725–735.

[23] 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, Conference Proceedings, pp. 126–131.

[24] 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, Conference Proceedings, pp. 736–744.

[25] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015, vol. 72.

[26] 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, Conference Proceedings, pp. 812–816.

[27] 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*, Conference Proceedings, pp. 142–151.

[28] 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.

[29] X. Han, "Mr-based synthetic ct generation using a deep convolutional neural network method," *Medical physics*, vol. 44, no. 4, pp. 1408–1419, 2017.

[30] 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, Conference Proceedings, pp. 250–258.