

A Universal Intensity Standardization Method Based on a Many-to-One Weak-Paired Cycle Generative Adversarial Network for Magnetic Resonance Images

Yuan Gao^{ID}, Yingchao Liu, Yuanyuan Wang^{ID}, *Senior Member, IEEE*, Zhifeng Shi, and Jinhua Yu^{ID}, *Member, IEEE*

Abstract—In magnetic resonance imaging (MRI), different imaging settings lead to various intensity distributions for a specific imaging object, which brings huge diversity to data-driven medical applications. To standardize the intensity distribution of magnetic resonance (MR) images from multiple centers and multiple machines using one model, a cycle generative adversarial network (CycleGAN)-based framework is proposed. It utilizes a unified forward generative adversarial network (GAN) path and multiple independent backward GAN paths to transform images in different groups into a single reference one. To preserve image details and prevent resolution loss, two jump connections are applied in the CycleGAN generators. A weak-pair strategy is designed to fully utilize the prior knowledge of the organ structure and promote the performance of the GANs. The experiments were conducted on a T2-FLAIR image database with 8192 slices from 489 patients. The database was obtained from four hospitals and five MRI scanners and was divided into nine groups with different imaging parameters. Compared with the representative algorithms, the peak signal-to-noise ratio, the histogram correlation, and the structural similarity were increased by 3.7%, 5.1%, and 0.1% on average, respectively; the gradient magnitude similarity deviation, the mean square error, and the average disparity were reduced by 19.0%, 15.7%, and

9.9% on average, respectively. Experiments also showed the robustness of the proposed model with a different training set configuration and effectiveness of the proposed framework over the original CycleGAN. Therefore, the MR images with different imaging settings could be efficiently standardized by the proposed method, which would benefit various data-driven applications.

Index Terms—Cycle generative adversarial network, intensity standardization, magnetic resonance imaging, many-to-one, weak-pair strategy.

I. INTRODUCTION

MAGNETIC resonance imaging (MRI) is currently one of the most important medical imaging methods. With the rapid development of techniques such as big data, deep learning and Radiomics [1]–[3], many studies have been conducted to reveal the underlying diagnosis and treatment information from MR images, such as noninvasive disease diagnosis [4], [5], differential diagnosis [6], [7], treatment prescription selection [8], [9], prognosis prediction [10], [11] and therapeutic efficacy evaluation [12]. The ongoing data-driven studies tend to use magnetic resonance (MR) images obtained from multi-center, multi-device with multi-parameter to establish models with reasonable robustness and universality, which place higher demands on the quality, especially the uniformity of image dataset.

However, image settings such as the applied magnetic field, the sequence of radio frequency pulses, the image reconstruction algorithm and so on have strong impacts on the MR signal intensity. The intensity distribution of images obtained from different scanners or under different protocols, therefore, tends to be different. As a result, for image data driven studies that rely heavily on various image features, the effectiveness would be eventually degenerated because the image feature differences caused by imaging parameters may exceed feature differences caused by the pathological characteristic itself. Therefore, to ensure the uniformity of the dataset, in practice, some of the studies use MR images acquired using a fixed protocol from one scanner, which will decrease the effectiveness of the radiomics model when the test sets are acquired using different conditions. Hence, the generalization

Manuscript received December 3, 2018; revised January 12, 2019; accepted January 16, 2019. Date of publication January 24, 2019; date of current version August 30, 2019. This work was supported in part by the National Basic Research Program of China under Grant 2015CB755500, in part by the National Key Research and Development Program of China under Grant 2016YFC0106105, and in part by the Taishan Scholars Program under Grant TSQN20161070. (Yuan Gao and Yingchao Liu contributed equally to this work.) (Corresponding authors: Yuanyuan Wang; Jinhua Yu.)

Y. Gao, Y. Wang, and J. Yu are with the Department of Electronic Engineering, Fudan University, Shanghai 200433, China, and also with the Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Fudan University, Shanghai 200433, China (e-mail: 15110720016@fudan.edu.cn; yywang@fudan.edu.cn; jhyu@fudan.edu.cn).

Y. Liu is with the Department of Neurosurgery, Provincial Hospital Affiliated to Shandong University, Jinan 250021, China (e-mail: 13805311573@126.com).

Z. Shi is with the Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai 200040, China (e-mail: 13917793493@139.com).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMI.2019.2894692

performance of the model is reduced considerably. Therefore, it is highly necessary to standardize different MR images from multiple centers in the preprocessing step so that all images can be aggregated to form a large and consistent dataset for the following computer aided diagnostic processes.

For the sake of eliminating the intensity difference among different MR image groups, researchers have proposed many methods. These methods can be divided into two major categories: the global histogram-matching method and the joint histogram registration method.

The global histogram-matching methods utilize different intensity characteristics as histogram landmark points to establish the transforming function with the form of piecewise linear functions. Nyul *et al.* [13] proposed transforming functions that use the minimum and maximum overall percentiles of histograms and several percentile markers of a “foreground” histogram. Collewet *et al.* [14] introduced three sets of landmarks, including the maximum intensity of images, the mean intensity of images, and the $\pm 3\sigma$ of the normal distribution of the image intensity. Madabhushi and Udupa [15] brought forward the intensity landmarks that were not affected by the diseased or abnormal tissues by finding the largest fuzzily connected homogenous region. Sun *et al.* [16] employed the average intensity μ , the maximum decile and the minimum decile of the intensity as landmarks and stretched the histograms of the target images according to the landmarks. De Nunzio *et al.* [17] separately standardized the intensities of different brain tissues by using different linear functions and formed a full histogram transformation by joining the different transformations together with spline smoothness. However, these methods are based upon a hypothesis that the MR image intensity relationship of tissues is constant between the target image group and the reference one, but this generally could not be fully satisfied.

The other method category is the joint histogram registration method. Jager *et al.* [18] first retrieved the joint probability density functions (PDFs) from multimodality histograms. Jager and Hornegger [19] then estimated the intensity mapping with a registration between the target and reference PDFs. Dzyubachyk *et al.* [20] applied the original joint histogram registration method to the whole-body MRI scan. Superior intensity standardization was obtained by comprehensively using the transforming relationships among different parts. Robitaille *et al.* [21] combined the global histogram matching method with the joint histogram registration method by using the characteristic points in the joint histograms as the landmarks for global matching. Our group proposed a method based on the nonrigid intensity transformation, which was applied to the weighted subregion intensity distribution instead of point intensity vector distribution, to enhance the stability and reduce the impact of registration error [22], [23]. In this group of methods, multimodality MR images are registered to a standard space and the pixel intensities obtained from each modality are used to form features in the intensity domain, which constitute a multidimensional point cloud used for the intensity mapping. However, this kind of method requires high multimodality MR image registration quality and hence the estimation accuracy is strongly influenced by the MR image registration.

These two groups of methods mentioned above focus on the intensity transformation between two fixed imaging settings. If the intensity standardization needs to be done for images coming from multiple centers, multiple transforming models need to be established. Meanwhile, these methods require features extracted from paired target-reference MR images obtained from the same person in a short interval, which not only seriously limits the number of training data but also makes it almost impossible to conduct standardization models for multicenter studies. Furthermore, these methods are not able to process new images that are not from any MR image group in the training data, which also limit their usability.

Therefore, in the paper, we propose a universal MR image standardization method in order to standardize all target MR images with the same modality. This method is based on the cycle generative adversarial network (CycleGAN) [24]. An extended framework is applied with a unified forward generator/discriminator pair and multiple independent backward generator/discriminator pairs for building the transformation model using multiple groups of images. Each generator, in particular, has two jump connection paths established between the downsampling path and upsampling one in order to provide more high-resolution details to the contracting layers. The weak-pair strategy is applied, which makes the images sent to the discriminator mainly contain differences in intensity distribution while other features such as structure and position that are not to be adjusted are as similar as possible. That ensures the discriminator’s concentration.

Meanwhile, in order to prove the effectiveness of the proposed algorithm in clinical data mining, we try to standardize MR images acquired from different MR scanners and attempt to differentiate high grade glioma (HGG) from lower grade glioma (LGG). The diagnostic accuracy by using MR images before and after intensity standardization is compared.

The paper is organized as follows. Section II expatiates on the proposed method including network architecture, loss function and training process in detail. The dataset as well as preprocessing steps containing weak-pair input strategy and data augmentation are shown in section III. Section IV describes the experiments consists of the impact of network structure optimization, data input strategy and training dataset on the model performance, the comparison with other methods as well as the effect of the proposed method as an image preprocessing step on radiomics-based differential diagnosis. Section IV also demonstrates the results. Section V raises the discussion. The conclusion is made in section VI.

II. METHODS

A. Formulation

Assume that the training set contains N groups of T2-FLAIR image slices with M_n images in each, where $n \in \{1, 2, \dots, N\}$. Since the images in one group share the same imaging parameters, they could be regarded as images in one domain X_n . Therefore, the images can be represented as $\{x_{n,m}\}_{m=1}^{M_n}$ $n=1 \dots N$, where $x_{n,m} \in X_n$. The distribution of the data is $x_n \sim p_{data}(x_n)$. Taking the N_{th} group as a reference, the goal is to figure out the universal

transformation $G_{forward} : X_1 \sim X_N \rightarrow X_N$ ensuring that every $G_{forward}(x_{n,m})$ belongs to X_N while keeping its own image feature. To validate whether the image-specific feature is preserved during the transformation, N reverse transformations $G_{backward n} : X_N \rightarrow X_n$ are established. Then, we apply $N + 1$ adversarial discriminators $D_{forward}$ and $D_{backward n}$. $D_{forward}$ is used to distinguish between reference images $\{x_N\}$ and forward-transformed target ones $\{G_{forward}(x_n)\}$, while $D_{backward n}$ is used to identify the original target images $\{x_n\}$ and backward-transformed reference ones $\{G_{backward n}(x_N)\}$. Thus, the objective of the transformation problem includes the *adversarial losses* and the *cycle consistency losses*, which will be described in section C, D and E.

B. Network Architecture

The modules of the network are illustrated in Fig. 1. To obtain a unified universal transformation that can transform any T2-FLAIR image slice into the reference image domain, a universal forward generative network $G_{forward}$ and a universal forward discriminative network $D_{forward}$ are utilized. Since the training images are from N different domains, to ensure the information of the original target image is preserved after the forward transformation, the images have to be recovered to N different domains in the process of the cycle consistency loss calculation. Therefore, N backward generative networks $G_{backward n}$ and N backward discriminative networks $D_{backward n}$ are applied to evaluate the information loss during the transformation. The training processes of generators and discriminators are alternated, so that the data generated by the generator is getting closer to the corresponding reference one. Every generative network consists of a convolutional down-sampling path, a residual convolution path and a convolutional upsampling path, as illustrated in Fig. 2. The convolutional down-sampling path contains a 7×7 convolution with $stride = 1$ and two 3×3 down-sampling convolutions with $stride = 2$, each followed by a rectified linear unit (ReLU) [25] as an activation function in order to increase the nonlinear characteristics of the model and solve the problem of slow convergence of neural network learning caused by the disappearance of the gradient. The filter number is 64, 128 and 256, respectively. The reflection padding is also applied. The residual convolution path contains nine successive residual blocks [26]. Each contains two 3×3 convolutional layers, and the filter number is fixed to 256. The convolutional upsampling path contains two 3×3 upsampling transposed convolutions with $stride = 2$ and a 7×7 transposed convolution with $stride = 1$, each followed by a rectified linear unit (ReLU). In particular, the image-to-image transformation should focus on the image's intensity adjustment so that the information loss caused by resolution changes should be avoided. For this purpose, two jump-connection paths are established [27]. The first one feeds the output of the ReLU layer after the 7×7 convolution to the input of the 7×7 transposed convolution. The other one feeds the output of the ReLU layer after the first 3×3 convolution to the input of the output's 3×3 transposed convolution. Therefore,

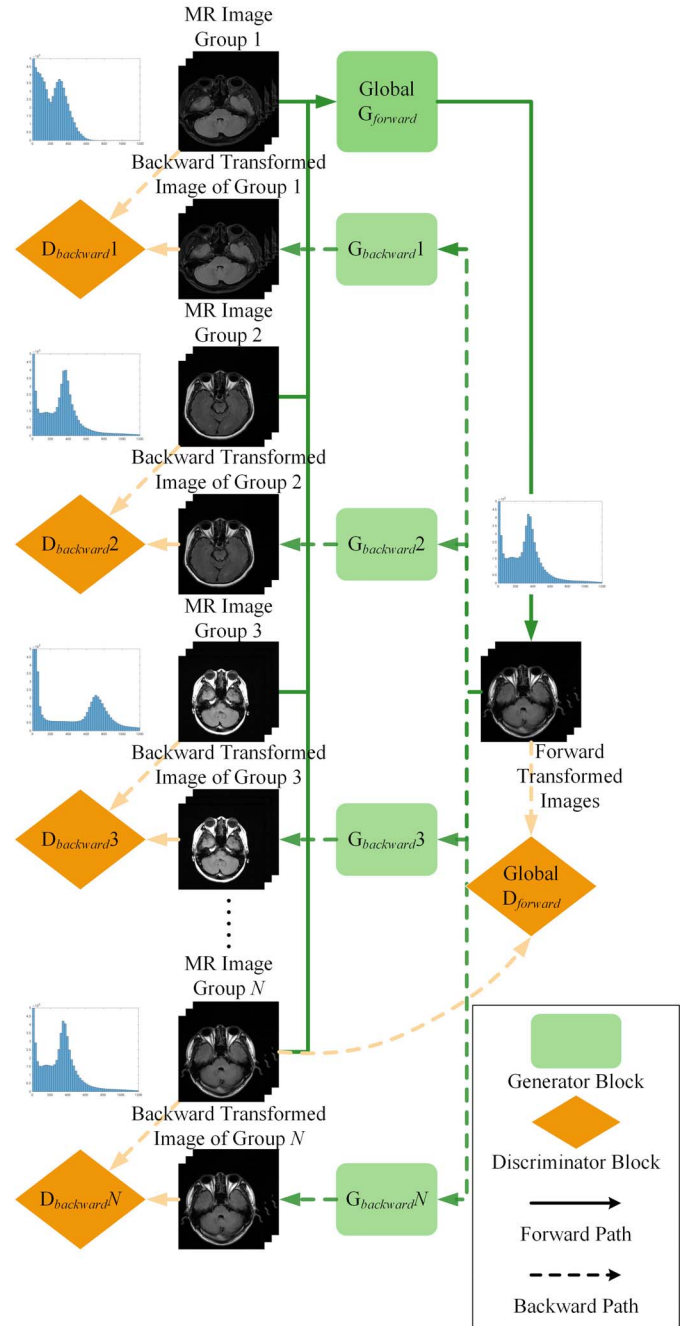


Fig. 1. The modules of the proposed networks.

the dimensions of the feature-map inputs for the transposed convolution layer are 256, 256 and 128, respectively. The discriminative networks use the classical PatchGAN [28] with 70×70 receptive field that consists of three successive 4×4 down-sampling convolutions with $stride = 2$ followed by a leaky rectified linear unit (LeakyReLU) [29] as the activation function and two convolutions with $stride = 1$ followed by a LeakyReLU. The LeakyReLU is an optimization of ReLU which reduces the number of inactive neurons and retains the information of the negative axis. The filter number is 64, 128, 256, 512 and 1 respectively. Therefore, a single layer 30×30 discriminative feature map is produced, which is then synthesized to obtain the final discriminative result.

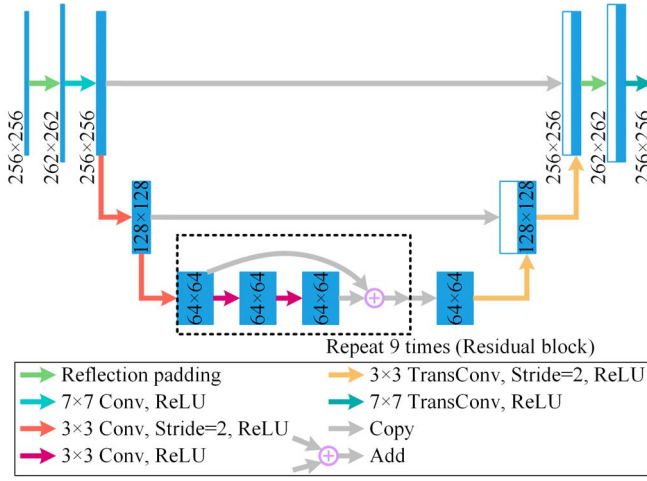


Fig. 2. The structure of the applied generative networks.

C. Adversarial Loss

The adversarial losses are used to make the generated images $\{G(x_n)\}$ approximate the distribution defined by the reference images and belong to the domain Y . The loss manifests as the classification capability of the discriminator D . Using the adversarial loss in least squares generative adversarial networks (LSGAN) [30], the loss function is expressed as

$$\begin{aligned} \mathcal{L}_{LSGAN}(G, D, X_n, Y) \\ = \mathbb{E}_{y \sim p_{data}(y)} [(D_Y(y) - 1)^2] \\ + \mathbb{E}_{x_n \sim p_{data}(x_n)} [D_Y(G(x_n))^2]. \end{aligned} \quad (1)$$

For GANs [31], the generator G aims to generate the images $G(x)$ that seems to be from domain Y , and the discriminator D aims to distinguish between the generated images and the real ones. Therefore, G tries to minimize the loss function, and D tries to maximize it. The optimization objective of the forward adversarial loss is $\min_{G_{forward}} \max_{D_{forward}} \mathcal{L}_{LSGAN}(G_{forward}, D_{forward}, X_n, X_N)$, and that of the backward adversarial loss is $\min_{G_{backward n}} \max_{D_{backward n}} \mathcal{L}_{LSGAN}(G_{backward n}, D_{backward n}, X_N, X_n)$.

D. Cycle Consistency Loss

If the output of the generator G is fixed to a particular image in the domain Y , the adversarial loss will not make sense since $D_Y(G(x_n)) = D_Y(y_i)$ where $y_i \in Y$. Thus, the cycle consistency loss is applied. The principle of the cycle consistency loss is straightforward. Once the image-specific feature is preserved, the standardized image could be transformed into the original image using a backward generator $G_{backward} : Y \rightarrow X$. Specifically, that is to say $G_{backward n}(G_{forward}(x_{n,m})) \approx x_{n,m}$ and $G_{forward}(G_{backward n}(x_{N,m})) \approx x_{N,m}$. By applying the L_1

norm, the loss function is expressed as

$$\begin{aligned} \mathcal{L}_{cycle}(G_{forward}, G_{backward n}) \\ = \mathcal{E}_{x_n \sim p_{data}(x_n)} [\|G_{backward n}(G_{forward}(x_n)) - x_n\|_1] \\ + \mathcal{E}_{x_N \sim p_{data}(x_N)} [\|G_{forward}(G_{backward n}(x_N)) - x_N\|_1]. \end{aligned} \quad (2)$$

E. Entire Loss Function

The entire loss function for the group n is defined as

$$\begin{aligned} \mathcal{L}(G_{forward}, G_{backward n}, D_{forward}, D_{backward n}) \\ = \mathcal{L}_{LSGAN}(G_{forward}, D_{forward}, X_n, X_N) \\ + \mathcal{L}_{LSGAN}(G_{backward n}, D_{backward n}, X_N, X_n) \\ + \mathcal{L}_{cycle}(G_{forward}, G_{backward n}). \end{aligned} \quad (3)$$

The optimization objective is to obtain the optimal $G_{forward}$ and the N optimal $G_{backward n}$ as defined as

$$\begin{aligned} G_{forward}^*, G_{backward n}^* \\ = \arg \min_{G_{forward}, G_{backward n}} \max_{D_{forward}, D_{backward n}} \\ \mathcal{L}(G_{forward}, G_{backward n}, D_{forward}, D_{backward n}). \end{aligned} \quad (4)$$

F. Training Process Overview

The training process of the whole framework could be described as follows:

- Randomly select an image $x_{n,m}$ without replacement in the WHOLE target dataset X .
- Randomly select one out of 150 MR images, $x_{N,\tilde{m}}$, in the reference dataset x_N closest to $x_{n,m}$ using the weak-pair data input strategy introduced in III.B.
- Generate $G_{forward}(x_{n,m})$, $G_{backward n}(G_{forward}(x_{n,m}))$, $G_{backward n}(x_{N,\tilde{m}})$ and $G_{forward}(G_{backward n}(x_{N,\tilde{m}}))$ using the unified forward generator $G_{forward}$ and the n_{th} backward generator $G_{backward n}$.
- Compute $\mathcal{L}(G_{forward}, G_{backward n}, D_{forward}, D_{backward n})$ and the gradient of $G_{forward}$ and $G_{backward n}$ using (3).
- Adjust the parameters of $G_{forward}$ and $G_{backward n}$ with the gradients computed in step (d).
- Compute $\mathcal{L}_{LSGAN}(G_{forward}, D_{forward}, X_n, X_N)$ and the gradient of $D_{forward}$ using (1).
- Adjust the parameters of $D_{forward}$ with the gradients computed in step (f).
- Compute $\mathcal{L}_{LSGAN}(G_{backward n}, D_{backward n}, X_N, X_n)$ and the gradient of $D_{backward n}$ using (1).
- Adjust the parameters of $D_{backward n}$ with the gradients computed in (h).
- Return to step (a) until the training process finishes.

After the training process, the $G_{forward}$ is extracted as a standardizer to transform any images into the reference domain.

TABLE I
THE DETAILS OF PATIENT IMAGING ACQUISITION PROTOCOLS FOR THE 9 GROUPS OF DIFFERENT IMAGING PARAMETERS FROM 5 MRI SCANNERS LOCATED AT THE 4 HOSPITALS

Site	Instrument	Image number	Patient number	Slice thickness (mm)	Pulse repetition time/Echo time/Inverse time (ms/ms/ms)	Imaging frequency (MHz)	Pixel bandwidth (Hz/pixel)	Receive coil	Pixel spacing (mm)
Department of Neurosurgery, Huashan Hospital, Fudan University	Siemens Magnetom Verio 3.0T MRI scanner	1904	119	6	8000/102/2370	123.1678	287	12-channel head coil	0.4492
		1472	92	6	9000/102/2500	123.2407	287	12-channel head coil	0.4492
		464	29	6	8500/102/2438.8	123.2463	287	12-channel head coil	0.4492
	GE Discovery MR750 3.0T MRI scanner	1196	80	6	8800/152.3/2100	127.7692	195.312	8-channel head coil	0.4688
		656	41	4	8525/141.9/2100	127.7706	195.312	8-channel head coil	0.4688
Department of Neurosurgery/Neuro-oncology, Sun Yat-sen University Cancer Center	Siemens Magnetom Trio Tim 3.0T MRI scanner	924	46	5	8500/90/2439.2	123.2587	289	12-channel head coil	0.6875
		548	30	5	8500/91/2439.2	123.2622	287	12-channel head coil	0.4297
Department of Neurosurgery, Huadong Hospital, Fudan University	Siemens Magnetom Verio 3.0T MRI scanner	644	32	5.5	9000/83/2500	123.2001	201	12-channel head coil	0.8984
Department of Neurosurgery, Shandong Provincial Hospital	Siemens Magnetom Skyra 3.0T MRI scanner	384	20	5	9000/128/2500	123.2292	285	32-channel head coil	0.8984

TABLE II
THE DETAILS OF VOLUNTEER IMAGING ACQUISITION PROTOCOLS

Instrument	Pulse sequence	Slice thickness (mm)	Pulse repetition time/Echo time/Inverse time (ms/ms/ms)	Pixel bandwidth (Hz/pixel)	Receive coil	Imaging frequency (MHz)	Pixel spacing (mm)
Siemens Magnetom Verio 3.0T MRI scanner	Axial T2W BLADE spin echo	6	3500/95/NA	287	12-channel head coil	123.1678	0.8984
	T1W FLAIR spin echo /inversion recovery	6	2000/17/857	287	12-channel head coil	123.1678	0.8984
	T2W FLAIR spin echo /inversion recovery	6	8000/102/2370	287	12-channel head coil	123.1678	0.4492
GE Discovery MR750 3.0T MRI scanner	Axial T2 PROPELLER spin echo	6	4581.9478/96.128/NA	195.312	8-channel head coil	127.7705	0.4688
	T1 FLAIR spin echo /inversion recovery	6	1872.34/21.864/720	195.312	8-channel head coil	127.7705	0.4688
	T2 FLAIR spin echo /inversion recovery	6	8525/145.684/2100	195.312	8-channel head coil	127.7705	0.4688

NA: Not applicable

III. DATA AND PREPROCESSING

A. MR Image Dataset in the Study

This study was approved by the ethics committees of the four participating hospitals in China and informed consent was obtained from every patient and volunteer.

There were 489 patients who participated in our study. A total of 8192 T2 FLAIR brain MR images were acquired from five MRI scanners located at the four hospitals. The training data was divided into nine groups according to different imaging parameters. The details of the patient imaging acquisition protocols are shown in Table I.

To independently test the proposed method, a test dataset was acquired from ten healthy volunteers from the Department of Radiology at Huashan Hospital at Fudan University.

To meet the data requirements for all algorithms used for comparison, T2, T2-FLAIR and T1-FLAIR brain MR images were successively obtained with a Siemens Magnetom Verio 3.0T MRI scanner and a GE Discovery MR750 3.0T MRI scanner. Therefore, 60 total image sets were obtained from these ten volunteers. The details of volunteer imaging acquisition protocols are shown in Table II.

B. The Weak-Pair Data Input Strategy

Fig. 3 shows a sample of an MRI scan with six different brain layers. Since the morphological structures of MR images significantly differ from one MRI layer to another, when all target images and reference ones are completely randomly paired like Fig. 4, the main difference between two image groups will be the morphological structure instead of the

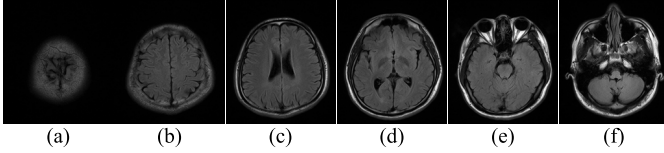


Fig. 3. An example demonstrating one set of 6 different brain layers from a single MRI scan.

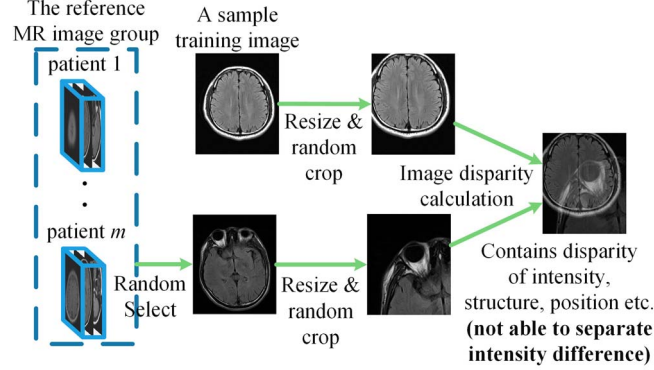


Fig. 4. All the target and reference images completely randomly paired.

intensity distribution, which may make the generative model mainly focus on morphological changes, which is undesirable. Therefore, we introduce a weak-pair approach to increase the structural similarity between two randomly paired MR images. First, in order to obtain the position of the MR image, a registration with a certain brain atlas is applied. According to [32], we apply the SPM12 tool to combine the registration and tissue classification into a single circular generative model to get a robust registration result, which is expressed as the axial-sagittal-coronal coordinate values of every pixel in the MR image. Furthermore, the layer position is estimated by averaging the axial coordinate value in the 1/3 center area of the corresponding image layer. Finally, in every epoch, one out of 150 MR images, $x_{N,\bar{m}}$, in the reference dataset x_N closest to $x_{n,m}$ is randomly selected to be paired for each $x_{n,m}$. The procedure of the proposed strategy is shown in Fig. 5.

C. Data Augmentation

Since the MR images used for training are all axial brain scans, in order to contain the entire brain, there will be a black background around the MR images. This part of the black border will cause the transformation model to be less sensitive to edge pixels. To improve the model stability and prevent overfitting, data augmentation is applied before each epoch. For every $x_{n,m}$, a random number $rand_{n,m}$ is obtained according to a Gaussian distribution with $\mu = 256$ and $\sigma = 64$. Then, the target image size is calculated as

$$size_{n,m} = \text{round}(256 + \text{absolute}(256 - rand_{n,m})), \quad (5)$$

where $\text{absolute}(\cdot)$ denotes the operation to calculate the absolute value and $\text{round}(\cdot)$ obtains the integer value around the variable. $x_{n,m}$ and its corresponding image $x_{N,\bar{m}}$ are both scaled to $size_{n,m} \times size_{n,m}$ with bilinear interpolation. Two images are later synchronously randomly cropped to 256×256 . Finally, such images are synchronously randomly flipped, and the image pair is generated.

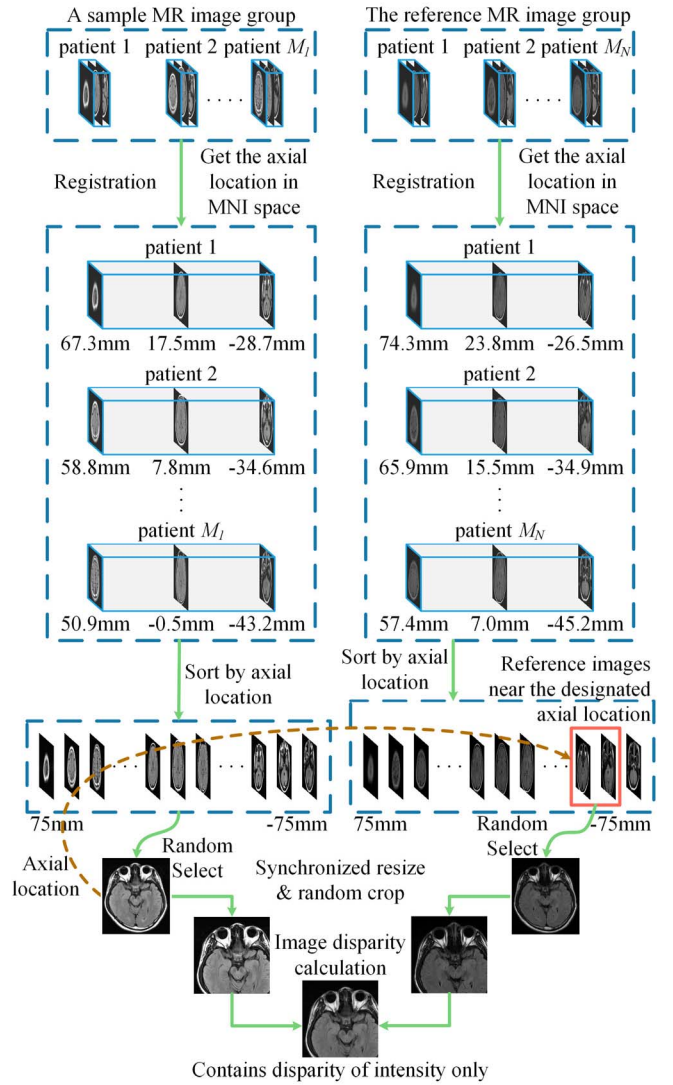


Fig. 5. The procedure of the proposed weak-pair data input strategy.

IV. EXPERIMENTS AND RESULTS

In this section, we first evaluate the effectiveness of the proposed method. Then, the comparison with different methods is performed. The peak signal-to-noise ratio (PSNR), the histogram correlation, the structural similarity index (SSIM), the gradient magnitude similarity deviation (GMSD) [33], the mean square error (MSE) and the average disparity are the applied quantitative criteria for the performance evaluation of the methods. The PSNR is defined as

$$PSNR = 10 \log_{10} \left(\frac{MAX_{Intensity}^2}{MSE} \right), \quad (6)$$

where $MAX_{Intensity}$ is the maximum value of the intensity in the MR images, and MSE is the mean squared error. The histogram correlation is defined as

Histogram Correlation

$$= 1 - \sqrt{1 - \frac{\sum \sqrt{Cnt_{standard} \cdot Cnt_{ref}}}{\sqrt{\sum Cnt_{standard} \times \sum Cnt_{ref}}}}, \quad (7)$$

where the $Cnt_{standard}$ is a vector containing the count of every bin in the histogram of the standardized image, and Cnt_{ref} is

TABLE III
THE EXPERIMENTAL RESULTS FOR THE STUDY OF THE FRAMEWORK OPTIMIZATION IN THE PROPOSED METHOD

Evaluation criteria	Original images	The proposed method	Direct Resnet structure	Without weak-pair strategy
PSNR	53.75	65.13	64.53	64.39
Histogram Correlation	0.2724	0.8865	0.8033	0.7917
SSIM	0.9829	0.9989	0.9986	0.9986
GMSD	0.1328	0.1027	0.1051	0.1040
MSE	18.50×10^3	1.353×10^3	1.523×10^3	1.602×10^3
Average Disparity	30.80	1.093	1.064	1.303

a vector containing the count of every bin in the histogram of the reference image. The SSIM is defined as

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (8)$$

where x and y are the reference and standardized image, correspondingly. μ_x and μ_y are the mean value of x and y . σ_x and σ_y are the standard deviation of x and y . $C_1 = 10^{-4}$ and $C_2 = 9 \times 10^{-4}$. The GMSD is defined as

$$GMSD = \sigma \left(\frac{2\mathbf{m}_r(i)\mathbf{m}_d(i) + c}{\mathbf{m}_r^2(i) + \mathbf{m}_d^2(i) + c} \right), \quad (9)$$

where $\mathbf{m}_r(i) = \sqrt{(\mathbf{r} * \mathbf{h}_x)^2(i) + (\mathbf{r} * \mathbf{h}_y)^2(i)}$ and $\mathbf{m}_d(i) = \sqrt{(\mathbf{d} * \mathbf{h}_x)^2(i) + (\mathbf{d} * \mathbf{h}_y)^2(i)}$ are the gradient magnitude of reference image $\mathbf{r}$ and standardized image $\mathbf{d}$ at location i . σ represents standard deviation pooling and $*$ denotes convolution, c is a positive constant for keeping numerical stability. The MSE is defined as

$$MSE = \frac{\sum (S_i - R_i)^2}{Count(R_i)}, \quad (10)$$

where S_i means the pixel value of a point in the standardized image, R_i is the pixel value of the corresponding point in the reference image, and $Count(R_i)$ denotes the amount of points. The average disparity is defined as

$$Average\ Disparity = 100 \text{mean} \left(\frac{|I_{standard} - I_{ref}|}{I_{ref}} \right), \quad (11)$$

where $I_{standard}$ is the standardized image, and I_{ref} is the reference image. The average disparity is the mean value of the pixel's relative error. Larger values of the PSNR, the histogram correlation and the SSIM as well as smaller values of the GMSD, the MSE and the average disparity denote the better performance.

The framework optimizations including the network structure and data input strategy in the proposed method are first explored. To evaluate the effectiveness of such optimizations, the performance comparisons between the frameworks with/without such optimizations are conducted. The training set is the patient image slices from nine groups and tests are applied with the volunteer dataset. The five quantitative criteria mentioned above are then gained.

Since the data and its represented data distribution are important for the image standardization, we further evaluate the impact of the dataset for the proposed method. For the

nine groups of patient images used for the model training, each time we select five groups and use them as the training set in this series of experiments. This process is repeated four times. Test A contained all kinds of training images. Test B, C and D did not contain images with lower, middle and higher average intensity respectively. The corresponding experimental results for the proposed methods with the tests applied using the volunteer dataset are obtained.

Meanwhile, we compare the proposed method with two major types of intensity standardization methods including the histogram-matching method proposed by Sun *et al.* and the joint histogram registration method previously proposed by our group. In Sun's method, the high-intensity region (HIR) and low-intensity region (LIR) are the values at the maximum and minimum deciles, respectively. In the joint histogram registration method, the number of reference points on one modality is set to 31, and the number of layers in the b -spline registration is set to seven. Both visual observation and quantitative comparison result are demonstrated.

In order to demonstrate the significance of the proposed method as a preprocessing step for multi-center data in the data mining process, we apply a set of reference MR images to train a differential diagnosis model and apply a set of target MR images and their standardized images to test it.

The model is used to distinguish gliomas' WHO grade (HGG: WHO IV; LGG: WHO II and III) [34], which is a classical clinical problem in glioma diagnosis. MRI images of 66 cases with glioma (34 HGG and 32 LGG) were from reference MR images and considered as the training set, and 28 (17 HGG and 11 LGG) were from target MR images and regarded as the test set. A total of 555 radiomics-based features are firstly extracted using a self-adaptive feature extraction method from the glioma region of the training cases. Then, the key feature selection and support vector machine (SVM) parameters optimization are proceeded simultaneously using the minimum redundancy maximum relevance (mRMR) based genetic algorithm (GA). The SVM model is finally obtained according to the predefined parameters and key features. After that, the MR images in the test set and their standardized images are used to test the SVM differential diagnosis model, respectively. Comparison of classification accuracy is shown.

A. The Performance Improvement Brought by Optimization in the Framework

These comparisons are shown in Table III. The performance by using the original Resnet-based structure without the jump connections as the generator in the proposed method is first revealed. Most performance indicators are improved with the

TABLE IV
THE EXPERIMENTAL RESULTS FOR THE STUDY OF THE TRAINING DATA USED IN THE PROPOSED METHOD

Evaluation criteria	Original images	The proposed method with nine MR image groups in a training set	Five of nine MR image groups in a training set			
			Test A	Test B	Test C	Test D
PSNR	53.75	65.13	64.96	64.52	63.78	64.45
Histogram Correlation	0.2724	0.8865	0.8700	0.8599	0.8360	0.8509
SSIM	0.9829	0.9989	0.9992	0.9991	0.9988	0.9991
GMSD	0.1328	0.1027	0.1032	0.1038	0.1038	0.1035
MSE	18.50×10^3	1.353×10^3	1.378×10^3	1.439×10^3	1.571×10^3	1.465×10^3
Average Disparity	30.80	1.093	1.770	1.903	2.585	2.782

proposed method. The PSNR, the histogram correlation and the SSIM are increased by 0.6 dB, 0.0832 and 0.0003 respectively, the GMSD and the MSE are reduced by 0.0024 and 170 respectively, and the change in the average disparity is not significant.

The performance by not applying the weak-pair strategy in the proposed method is then demonstrated. All performance indicators are improved with the application of the weak-pair strategy. The PSNR, the histogram correlation and the SSIM are increased by 0.74 dB, 0.0948 and 0.0003, respectively, and the GMSD, the MSE and the average disparity are reduced by 0.0013, 249 and 0.21, respectively.

Meanwhile, from Table III, it is also noted that the proposed intensity standardization method demonstrates a significantly improved performance based on all quantitative criteria compared with the original image data.

B. The Impact of Datasets on Model Performance

The results for the study of the training data used in the proposed method are shown in Table IV. It can be seen that the performance is decreased due to the significant reduction of the training data. After further examination, we found that if the diversity of the training images could be guaranteed, the degradation would not be significant. The proposed method can well represent the distribution, even with the limited training data. This may demonstrate the superiority of the proposed method.

C. Comparison With Other Methods

Fig. 6 demonstrates a typical set of the target, the reference and the standardized results of different methods. The target images and reference ones differ severely in intensity, which makes it hard for the computer-aided diagnosis system to perform properly before intensity standardization. Using visual observation, a slight intensity corruption is found in the standardization result with Sun's method. Meanwhile, the average intensity is slightly higher. Specifically, the blue region in Fig. 6 (d3) shows a contrast enhancement between the gray matter and the white matter, which may be attributed to the intensity correspondence deviation caused by the absence of the intensity relationship consistency. The standardization results with the joint histogram registration method are visually good in the brain regions, as with the proposed method. However, for the skull, muscle and skin areas, only the proposed method visually achieves a relatively good result. Table V provides the quantitative results of the comparison

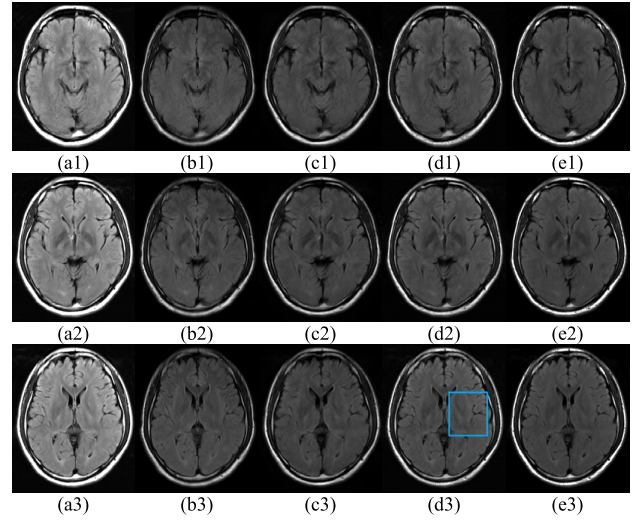


Fig. 6. The test on three continuous example slice aligned in MNI space from a volunteer. (a1)-(a3) MR images from GE scanner used as the original target, (b1)-(b3) MR images from Siemens scanner used as the original reference, (c1)-(c3) standardized results (transformation from the target to the reference) using the proposed method, (d1)-(d3) standardized results using the histogram matching method, (e1)-(e3) standardized results using the joint histogram registration method. The blue region in (d3) shows a contrast enhancement. (e1)-(e3) have slightly worse result in skull, muscle and skin areas.

for the different methods. It can be clearly seen that the proposed method achieves the state-of-the-art results, and the improvement is significant.

For the operation time of the standardization process, the histogram matching method is preferably fast since it takes 0.5867 second for an image slice. The joint histogram registration method consumes 2.1651 seconds for a slice mainly due to the complicated computation of the *b*-spline interpolation. For the proposed method, because convolutions are parallel and very suitable for implementation with graphic processing units (GPUs), the time cost is only 0.2868 second per slice, which is acceptable as a preprocessing step in the MR image analysis workflow.

D. The Radiomics-Based Differential Diagnosis Performance Improvement Brought by the Proposed Method

The accuracy of differential diagnosis between HGG and LGG with original images from target dataset is 71.4%, and that on the standardized images increases to 82.1%. Therefore, we believe that the proposed method can

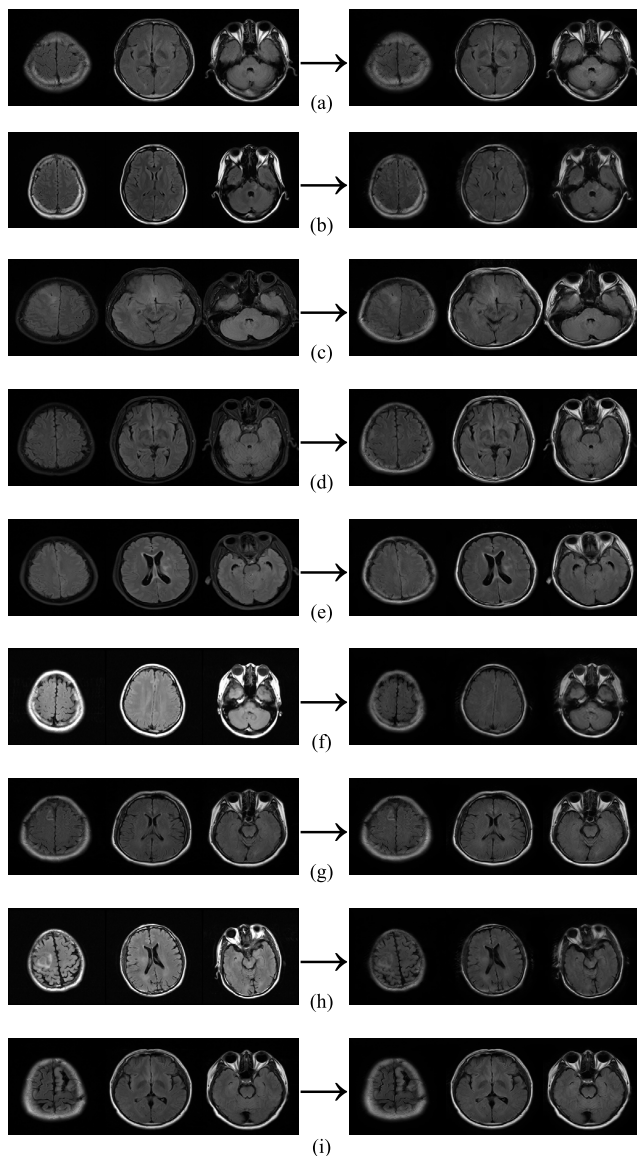


Fig. 7. The typical results for the 9 image groups in which the left is the target images and the right is the standardized results. (a) Huashan Hospital, Siemens Magnetom Verio scanner with TR=8000 ms, TE=102 ms, TI=2370 ms. (b) Shandong Provincial Hospital, Siemens Magnetom Skyra scanner with TR=9000 ms, TE=128 ms, TI=2500 ms. (c) Sun Yat-sen University Cancer Center, Siemens Magnetom Trio Tim scanner with TR=8500 ms, TE=90 ms, TI=2439.2 ms. (d) Sun Yat-sen University Cancer Center, Siemens Magnetom Trio Tim scanner with TR=8500 ms, TE=91 ms, TI=2439.2 ms. (e) Huadong Hospital, Siemens Magnetom Verio scanner with TR=9000 ms, TE=83 ms, TI=2500 ms. (f) Huashan Hospital, GE Discovery MR750 scanner with TR=8800 ms, TE=152.3 ms, TI=2100 ms. (g) Huashan Hospital, Siemens Magnetom Verio scanner with TR=9000 ms, TE=102 ms, TI=2500 ms. (h) Huashan Hospital, GE Discovery MR750 scanner with TR=8525 ms, TE=141.9 ms, TI=2100 ms. (i) Huashan Hospital, Siemens Magnetom Verio scanner with TR=8500 ms, TE=102 ms, TI=2438.8 ms.

effectively improve the accuracy of differential diagnosis methods based on modern techniques such as big data, deep learning and Radiomics under multi-center, multi-machine and multi-parameter conditions.

V. DISCUSSION

The purpose of this study is to establish a universal model for the intensity standardization of MR images.

The proposed method has many advantages compared with the previous methods.

First, when the MR images used to train the proposed framework are obtained from enough wide variety of sources thus the coverage is high throughout the MR image domain, thanks to the high-dimensional feature representation from the applied deep networks for the MR images instead of the low-dimensional feature representation of the conventional methods, the proposed method is able to extract features resulting from differences in tissue and structure over a relatively wide MR image domain, rather than differences between different machines or different parameters. Therefore, the proposed method may produce a universal transformation model for such a MR image domain. As a result, the method could be directly used for completely unknown testing images, which makes this model the first universal method that does not require the statistical information of the particular image domain to which the testing images belong. This allows the standardization of any MR image over a wide range that is not limited to the images defined by the training image domains. It is no doubt that, as shown in Table IV, when the MR images in the training set are from limited sources, the performance would be diminished because the model could not acquire characteristics of the entire MR image domain. However, the performance decrease is not so obvious and the results are still acceptable.

Second, it can be seen from Table III that the proposed method with the weak-pair strategy achieves better results compared with the one using the random pair strategy. It is because the application of the weak-pair data input strategy can make the discriminator concentrate on the intensity deviation rather than other differences, which results in more effective convergence. Meanwhile, since the model employs weak-paired MR images rather than precisely paired ones, the training sets of different machines or parameters can be acquired from different patients instead of using more than one machine to scan the same patients within short intervals. Moreover, the proposed method does not even require accurate registration results between target data and reference data, which is needed for the previous methods. Therefore, it is easy to form a training set with a large amount of data from multiple sources, which allows us to obtain a robust standardization model with strong generalization capability. Such an advantage makes it possible to standardize the multi-center or even cross-regional MR images in order to establish a large dataset that facilitates the development of various MR image-based clinical analysis and computer-aided diagnostic systems.

As for the network structure, Table III also shows that the proposed method which applies two jump connections between corresponding layers of the downsampling path and upsampling one achieves better results compared with the one using the direct Resnet structure. Such improvements are mainly attributed to the preservation of high-resolution features in the input images by directly combine them with up-sampled convolution result before the last trans-convolutional layer. The effectiveness of the proposed two jump connections is thus demonstrated.

TABLE V
THE COMPARISON RESULTS OF THE DIFFERENT METHODS

Evaluation criteria	Original images	The proposed method	Histogram-matching method	Joint histogram registration method
PSNR	53.75	65.13	61.39	64.27
Histogram Correlation	0.2724	0.8865	0.8049	0.8825
SSIM	0.9829	0.9989	0.9968	0.9988
GMSD	0.1328	0.1027	0.1291	0.1245
MSE	18.50×10^3	1.353×10^3	1.652×10^3	1.559×10^3
Average Disparity	30.80	1.093	13.54	8.467

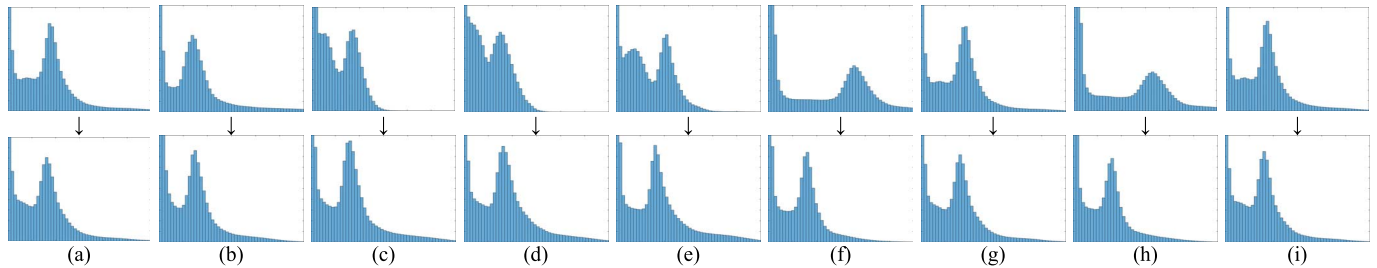


Fig. 8. The histograms of the standardized images and the reference images of the 9 image groups for which the top is the reference results and the bottom is the standardized results. The groups of the figure correspond to the groups of Fig. 7.

To explore more details of the experimental results, Fig. 7 shows the typical results of the proposed method that was directly performed on the training data of nine image groups. It can be seen that the transformation largely retains the original image information. Meanwhile, the intensity levels of the standardized images are consistent across the different image groups. Fig. 8 shows the histograms of the standardized images and the reference ones. It can be seen that the histograms of the standardized results for different groups of MR images are close to the histograms of the reference ones.

Meanwhile, the experimental results also illustrate the value of the proposed method in medical big data mining. With standardization of MR images from different MR scanners, the image difference caused by different MR scanners which may exceed the image difference caused by the pathological features to be discriminated is reduced or even eliminated. In this case, a large amount of standardized data from multiple sources can be used for training various computer-aided diagnosis models based on MR images. Therefore, with the proposed standardization method as an image preprocessing step, the result models can focus on the difference caused by the pathological features and can also be used in a wider range.

VI. CONCLUSION

In this paper, a universal MR image standardization method is proposed. It utilizes an optimized cycle generative adversarial network to standardize images from different image groups. Two jump connections are proposed in each generative network to prevent the transformation model from losing its image resolution. Meanwhile, a weak-pair approach is proposed to increase the structural similarity between two randomly paired MR images and focus the transformation on the intensity transformation rather than the structural modification. The experimental results demonstrate the state-of-the-art performance of the proposed method. This promising method may

vigorously promote the potential establishment of large MR image datasets, which might contribute to the development of medical big data processing and further improve the reliability and generalization capabilities of computer aided diagnosis.

REFERENCES

- [1] R. Lambin *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012.
- [2] V. Kumar *et al.*, "Radiomics: The process and the challenges," *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1234–1248, Nov. 2012.
- [3] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, Jun. 2014, Art. no. 4006.
- [4] J. Yu *et al.*, "Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma," *Eur. Radiol.*, vol. 27, no. 8, pp. 3509–3522, Aug. 2017.
- [5] L. Chen *et al.*, "Multi-label nonlinear matrix completion with transductive multi-task feature selection for joint MGMT and IDH1 status prediction of patient with high-grade gliomas," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1775–1787, Aug. 2018.
- [6] S. Kloppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, Mar. 2008.
- [7] A. Raj, A. Kuceyeski, and M. Weiner, "A network diffusion model of disease progression in dementia," *Neuron*, vol. 73, no. 6, pp. 1204–1215, Mar. 2012.
- [8] R. Shiradkar *et al.*, "Radiomics based targeted radiotherapy planning (Rad-TRaP): A computational framework for prostate cancer treatment planning with MRI," *Radiat. Oncol.*, vol. 11, Nov. 2016, Art. no. 148.
- [9] S. Rathore, H. Akbari, and J. Doshi, "Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: Implications for personalized radiotherapy planning," *J. Med. Imag.*, vol. 5, no. 2, Mar. 2018, Art. no. 021219.
- [10] B. Zhang *et al.*, "Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma," *Clin. Cancer Res.*, vol. 23, no. 15, pp. 4259–4269, Aug. 2017.
- [11] M. Ingrisch *et al.*, "Radiomic analysis reveals prognostic information in T1-weighted baseline magnetic resonance imaging in patients with glioblastoma," *Invest. Radiol.*, vol. 52, no. 6, pp. 360–366, Jun. 2017.
- [12] J. Bottcher *et al.*, "Response to neoadjuvant treatment of invasive ductal breast carcinomas including outcome evaluation: MRI analysis by an automatic CAD system in comparison to visual evaluation," *Acta Oncol.*, vol. 53, no. 6, pp. 759–768, Jun. 2014.

- [13] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.
- [14] G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification," *Magn. Reson. Imag.*, vol. 22, no. 1, pp. 81–91, Jan. 2004.
- [15] A. Madabhushi and J. K. Udupa, "New methods of MR image intensity standardization via generalized scale," *Med. Phys.*, vol. 33, no. 9, pp. 3426–3434, Sep. 2006.
- [16] X. Sun *et al.*, "Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions," *Biomed. Eng. OnLine*, vol. 14, Jul. 2015, Art. no. 73.
- [17] G. De Nunzio, R. Cataldo, and A. Carlà, "Robust intensity standardization in brain magnetic resonance images," *J. Digit. Imag.*, vol. 28, no. 6, pp. 727–737, Dec. 2015.
- [18] F. Jager *et al.*, "A new method for MRI intensity standardization with application to lesion detection in the brain," in *Proc. 11th Int. Fall Workshop Vis., Modeling, Vis.*, Aachen, Germany, 2006, pp. 269–276.
- [19] F. Jager and J. Hornegger, "Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 28, no. 1, pp. 137–150, Jan. 2009.
- [20] O. Dzyubachyk *et al.*, "Inter-station intensity standardization for whole-body MR data," *Magn. Reson. Med.*, vol. 77, no. 1, pp. 422–433, Jan. 2017.
- [21] N. Robitaille *et al.*, "Tissue-based MRI intensity standardization: Application to multicentric datasets," *Int. J. Biomed. Imag.*, vol. 2012, May 2012, Art. no. 347120.
- [22] Y. Gao *et al.*, "ND point cloud registration for intensity normalization on magnetic resonance images," in *Proc. VipIMAGE*, Porto, Portugal, 2017, pp. 121–130.
- [23] Y. Gao *et al.*, "Optimised MRI intensity standardisation based on multi-dimensional sub-regional point cloud registration," *Comput. Methods Biomech. Biomed. Eng., Imag. Vis.*, to be published, doi: [10.1080/21681163.2018.1511477](https://doi.org/10.1080/21681163.2018.1511477).
- [24] J. Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 2242–2251.
- [25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 807–814.
- [26] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. 29th IEEE Con. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
- [27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Munich, Germany, 2015, pp. 234–241.
- [28] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 702–716.
- [29] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, 2013, pp. 1–3.
- [30] X. Mao *et al.*, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Jul. 2017, pp. 2813–2821.
- [31] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 28th Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2672–2680.
- [32] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.
- [33] W. Xue *et al.*, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
- [34] D. N. Louis *et al.*, "The 2016 World Health Organization classification of tumors of the central nervous system: A summary," *Acta Neuropathol.*, vol. 131, no. 6, pp. 803–820, Jun. 2016.