Original Article

Improving segmentation reliability of multi-scanner brain images using a generative adversarial network

Kai Niu[1]^, Xueyan Li[2,3]^, Li Zhang[4], Zhensong Yan[5], Wei Yu[5], Peipeng Liang[6], Yan Wang[7], Ching-Po Lin[8,9], Huimao Zhang[4]^, Chunjie Guo[4#]^, Kuncheng Li[10#]^, Tianyi Qian[5]

[1]Department of Otorhinolaryngology Head and Neck Surgery, the First Hospital of Jilin University, Changchun, China; [2]State Key Laboratory of Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, Changchun, China; [3]College of Electronic Science and Engineering, Peng Cheng Laboratory, Shenzhen, China; [4]Department of Radiology, the First Hospital of Jilin University, Changchun, China; [5]AI Lab, QuantMind, Beijing, China; [6]School of Psychology, Capital Normal University, Beijing, China; [7]Key Laboratory of Symbol Computation & Knowledge Engineering, Ministry of Education, College of Computer Science & Technology, Jilin University, Changchun, China; [8]Neurological Research Center, the First Hospital of Jilin University, Changchun, China; [9]Institute of Neuroscience, Yang-Ming University, Taipei 112, Taiwan, China; [10]Department of Radiology, Xuanwu Hospital, Capital Medical University, Beijing, China

Contributions: (I) Conception and design: C Guo, K Li, T Qian; (II) Administrative support: H Zhang, Y Wang, CP Lin; (III) Provision of study materials or patients: K Niu, L Zhang, P Liang, C Guo; (IV) Collection and assembly of data: K Niu, L Zhang, P Liang; (V) Data analysis and interpretation: X Li, Z Yan, W Yu, P Liang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Chunjie Guo. Department of Radiology, The First Hospital of Jilin University, 1 Xinmin St., Changchun 130021, Jilin, China. Email: guocj@jlu.edu.cn; Kuncheng Li. Department of Radiology, Xuanwu Hospital, Capital Medical University, 45 Changchun St., Beijing 100053, China. Email: cjr.likuncheng@vip.163.com.

Background: Magnetic resonance (MR) images generated by different scanners generally have inconsistent contrast properties, making it difficult to perform a combined quantitative analysis of images from a range of scanners. In this study, we aimed to develop an automatic brain image segmentation model to provide a more reliable analysis of MR images taken with different scanners.

Methods: The spatially localized atlas network tiles-27 (SLANT-27) deep learning model was used to train the automatic segmentation module, based on a multi-center dataset of 1,917 three-dimensional (3D) T1-weighted MR images. Subsequently, a framework called Qbrain, consisting of a new generative adversarial network (GAN) image transfer module and the SLANT-27 segmentation module, was developed. Another 3D T1-weighted MRI interscan dataset of 48 participants who were scanned in 3 MRI scanners (1.5T Siemens Avanto, 3T Siemens Trio Tim, and 3T Philips Ingenia) on the same day was used to train and test the Qbrain model. Volumetric T1-weighted images were processed with Qbrain, SLANT-27, and FreeSurfer (FS). The automatic segmentation reliability across the scanners was assessed using test-retest variability (TRV).

Results: The reproducibility of different segmentation methods across scanners showed a consistent trend in the greater reliability and robustness of QBrain compared to SLANT-27 which, in turn, showed greater reliability and robustness compared to FS. Furthermore, when the GAN image transfer module was added, the mean segmentation error of the TRV of the 3T Siemens *vs.* 1.5T Siemens, the 3T Philips *vs.* 1.5T Siemens, and the 3T Siemens *vs.* 3T Philips scanners was reduced by 1.57%, 2.01%, and 0.56%, respectively. In addition, the segmentation model improved intra-scanner variability (0.9–1.67%) compared with that of FS (2.47–4.32%).

^ ORCID: Kai Niu, 0000-0003-2040-6887; Xueyan Li, 0000-0002-9206-9480; Huimao Zhang, 0000-0003-0443-2831; Chunjie Guo, 0000-0002-5784-0479; Kuncheng Li, 0000-0002-1898-3979.

Conclusions: The newly developed QBrain method combined with GAN image transfer module and a SLANT-27 segmentation module was shown to improve the reliability of whole-brain automatic structural segmentation results across multiple scanners, thus representing a suitable alternative quantitative method of comparative brain tissue analysis for individual patients.

Keywords: Magnetic resonance imaging (MRI); brain, segmentation; deep learning; generative adversarial network (GAN)

Submitted Jun 22, 2021. Accepted for publication Oct 28, 2021.
doi: 10.21037/qims-21-653
View this article at: https://dx.doi.org/10.21037/qims-21-653

Introduction

Quantitative volumetric analysis of brain magnetic resonance imaging (MRI) is a widely used method applied to routine clinical practice in radiology departments and to research in the neurosciences. Brain volumetry is influenced by several technical factors, such as the MRI scanner field strength and imaging parameters defined by the scanning vendor (1,2), as well as several subject-related factors (3,4) and post-processing methods (5,6). Reliable and robust quantitative volumetric brain segmentation analysis is critical for individualized precision treatment of brain disorders, such as Alzheimer's disease (AD) and MS (6-8); however, it is unrealistic that every patient will be allocated to the same scanner in the clinical setting. This diversity of equipment makes it challenging to achieve reliable and robust quantitative volumetric brain segmentation analysis in individuals across different MRI field strengths and vendors.

Generally, automated segmentation methods can be classified into 3 categories: probabilistic-based, atlas-based, and deep learning (DL)-based segmentation. One of the most well-known probabilistic-based segmentation methods is FreeSurfer (FS) (9), which is renowned for its reliability and accuracy. It is used as an a priori method to identify anatomical structures and their relationship to corresponding landmarks. Atlas-based methods have also performed well in whole-brain segmentation. In general, these methods use non-rigid registrations to align a brain mask template, or many templates, with an input brain image and then assign template labels to the target brain region. Compared with probabilistic-based and atlas-based segmentation methods, DL-based methods can more accurately segment brain regions. The U-net (10) and V-net (11) models are 2 DL-based methods that have been designed for the accurate segmentation of medical images. The most basic DL-based method for whole-brain segmentation feeds the entire three-dimensional (3D) brain image into a 3D U-net or V-net model.

However, DL-based models usually require large amounts of data, and the manual labeling of training data is a laborious undertaking. To assemble data for the DL model, Roy *et al.* proposed using a large number of auxiliary labels to train 2D convolutional neural networks (CNNs), where the auxiliary labels could be generated by other automated tools, such as FS (12). In their study, the spatially localized atlas network tiles (SLANT) aligned multiple, traditional, spatially-distributed CNNs to learned contextual information for a fixed spatial location and more accurately generated a 3D whole-brain segmentation. It has also been shown that segmentation performance is best when the results from a network of 27 tiles of overlapped sub-spaces are combined; therefore, SLANT-27 (which divides the volume into 27 overlapped network tiles) was employed in in the current research. Compared to CNNs, the generative adversarial network (GAN) is a model generated for capturing data distribution in an adversarial way to produce more realistic images (13). There have been several attempts to generate brain images using MRI coupled with GAN. For example, Han *et al.* (14) produced synthetic multi-sequence brain MRI by GAN and Lei *et al.* (15) used GAN to realize a transformation between computerized tomography (CT) and MRI, demonstrating that GAN can decrease the variabilities caused by image modality or contrast. However, cross-scanner variabilities due to differences of field strength and MR image acquisition sequences have not been considered in previous studies.

In order to make comparable quantitative measurements of brain tissue in individual subjects, we propose an automatic whole-brain segmentation framework called Qbrain, which consists of the domain image transfer and a SLANT-27 segmentation module to reduce the variability effects of

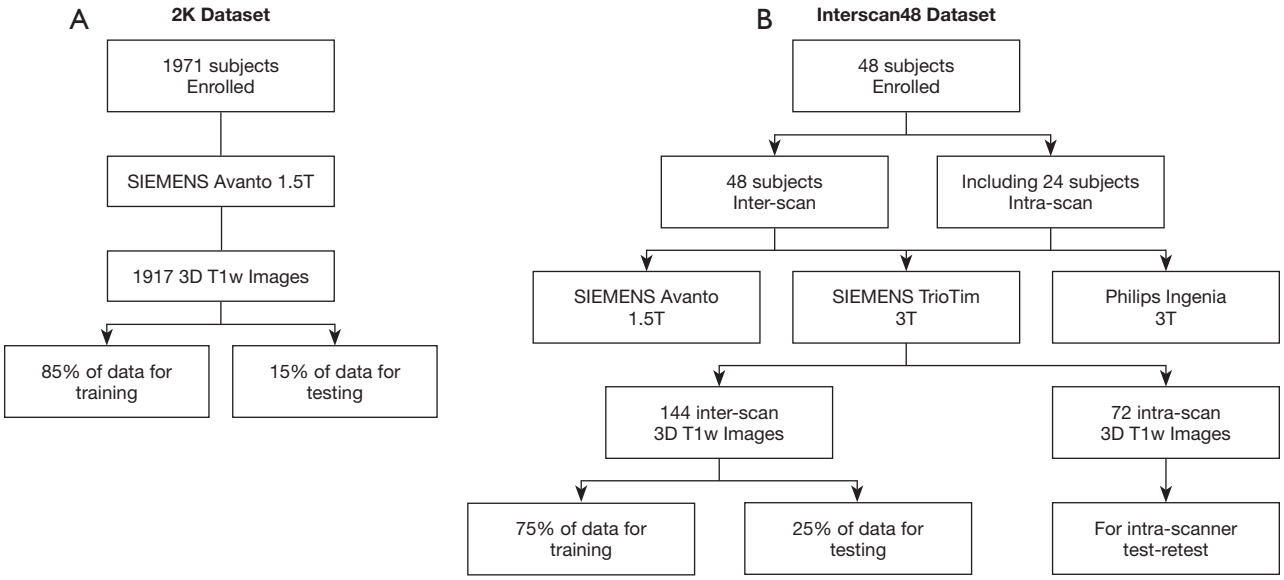

Figure 1 Experimental data sets: 2K dataset and Interscan48 dataset. T1w, T1-weighted.

field strength and vendor across different MRI scanners. We present the following article in accordance with the Materials Design Analysis Reporting (MDAR) checklist (available at https://dx.doi.org/10.21037/qims-21-653).

Methods

Participants

Two datasets were used in this study. The first dataset was used to train a segmentation model using the SLANT-27 (16) model. The second dataset was used to train and test the image domain transfer model, using GAN to support the SLANT-27 model. All the MRI data employed in this study were collected with the approval of the local ethics committee. The age of participants ranged from 10 to 78 years, and informed written consent was provided by each participant prior to data collection and analysis. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

The dataset for training the SLANT-27 automatic segmentation model consisted of 1,917 3D T1-weighted MRI scans of patients aged 18–78 years. The images were acquired using 1.5T Siemens Avanto MR scanners (Siemens Healthineers, Erlangen, Germany) and collected from multiple centers. High-resolution, T1-weighted images were acquired according to the parameters of the Alzheimer's Disease Neuroimaging Initiative (ADNI) project (voxel size 1.0×1.0×1.0 mm^3). For this study, the dataset was designated as 2K for simplicity. We randomly assigned 15% of the data in the 2K dataset as the testing set, and the remaining data were used as the training set. There were 60 labels for each scan, and the labels were established by clinical experts in the field. We used FS to generate 187 labels for each scan, and then the 187 labels were merged into 60 labels as the training mask. The experimental data set is illustrated in *Figure 1*.

The second dataset was collected to train and test the image-to-domain transfer GAN model. The dataset contained 48 participants from one center. For each subject, 3 clinical MRI scanners, including the 1.5T Siemens Avanto, 3T Siemens Trio Tim, and 3T Philips Ingenia (Philips Healthcare, Amsterdam, The Netherlands), were used to acquire 3D T1-weighted images on the same day. In addition, half of the participants were scanned twice with repositioning in-between, resulting in a total of 6 T1-weighted volumes per participant; thus, there were 216 T1-weighted scans in total, with 144 inter-scanner T1-weighted images in the dataset, thereafter referred to as Interscan48. Of the subjects in the Interscan48 dataset, 38 were used as the training set, and the remaining 12 participants who had both intra- and inter-scans were used as the testing set. The MRI exclusion criteria were contraindications to having an MRI scan, severe neurological disorders, or a history of serious head trauma or brain tumors (no participants were excluded). The MRI

Figure 2 The entire QBrain framework consisting of a GAN image transfer module and SLANT-27 segmentation module. GAN, generative adversarial network; SLANT, spatially localized atlas network tiles.

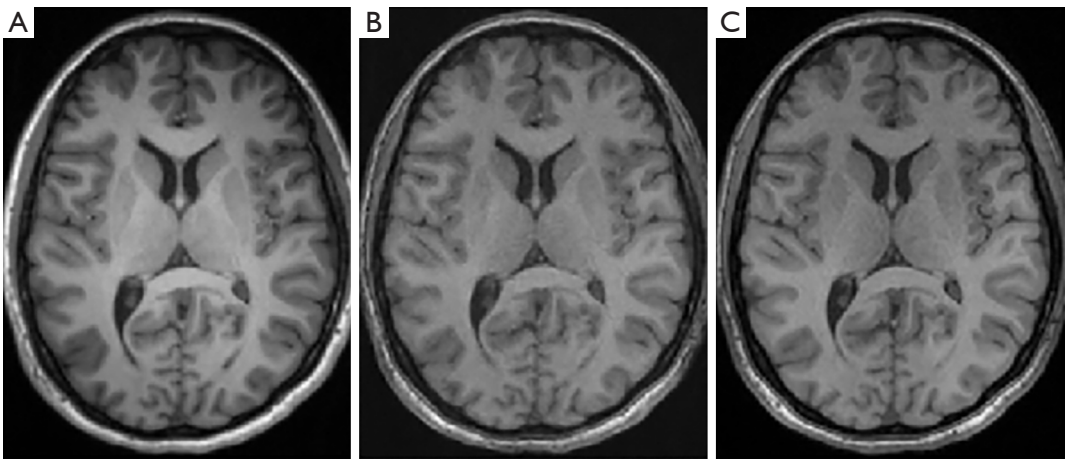

Figure 3 A comparison of the 3T image, the GAN converted image, and the 1.5T image. (A) A 3T MRI image; (B) the output image by the image domain transfer module, GAN; (C) the corresponding 1.5T image of the 3T image. GAN, generative adversarial network; MRI, magnetic resonance imaging.

acquisition parameters are detailed in Table S1.

Image processing

The entire QBrain framework is depicted in *Figure 2*. Preprocessing was performed for the input MR images, which included N4 bias field corrections and affine registrations. This was followed by the use of image domain transfer modules to transfer the input images from different scanners into the same scanner style, allowing for the following automatic segmentation process to be executed.

Image domain transfer

The image domain transfer model was used to transfer data to a unified target domain based on the GAN network. Namely, the model standardized the different field strengths (1.5T and 3T) or vendor scans to medium contrast images with fewer variations and thus improved image segmentation. A comparison between the 3T image, the GAN-converted image, and the 1.5T image is shown in *Figure 3*. In this section, we introduce the network architecture of the image domain transfer model and the objective function.

Network architecture

The architecture of our image domain transfer model, based on GAN, consisted of 2 parts, the generator (G), and the discriminator (D). The G adopted a 3D volume from the source domain (3T MRI) as input and generated a fake volume to imitate the target domain (1.5T MRI). Then, the D aimed to classify the inputs that were true volumes from the target domain. By adversarial training, the G was able to transfer data from the source domain to the target domain. The network architecture of the image domain transfer model is illustrated in *Figure 4*.

The G was a U-net-like structure with an encoder and a decoder. The encoder stage had 5 blocks; the first 4 blocks consisted of a convolution operation and a separate down-

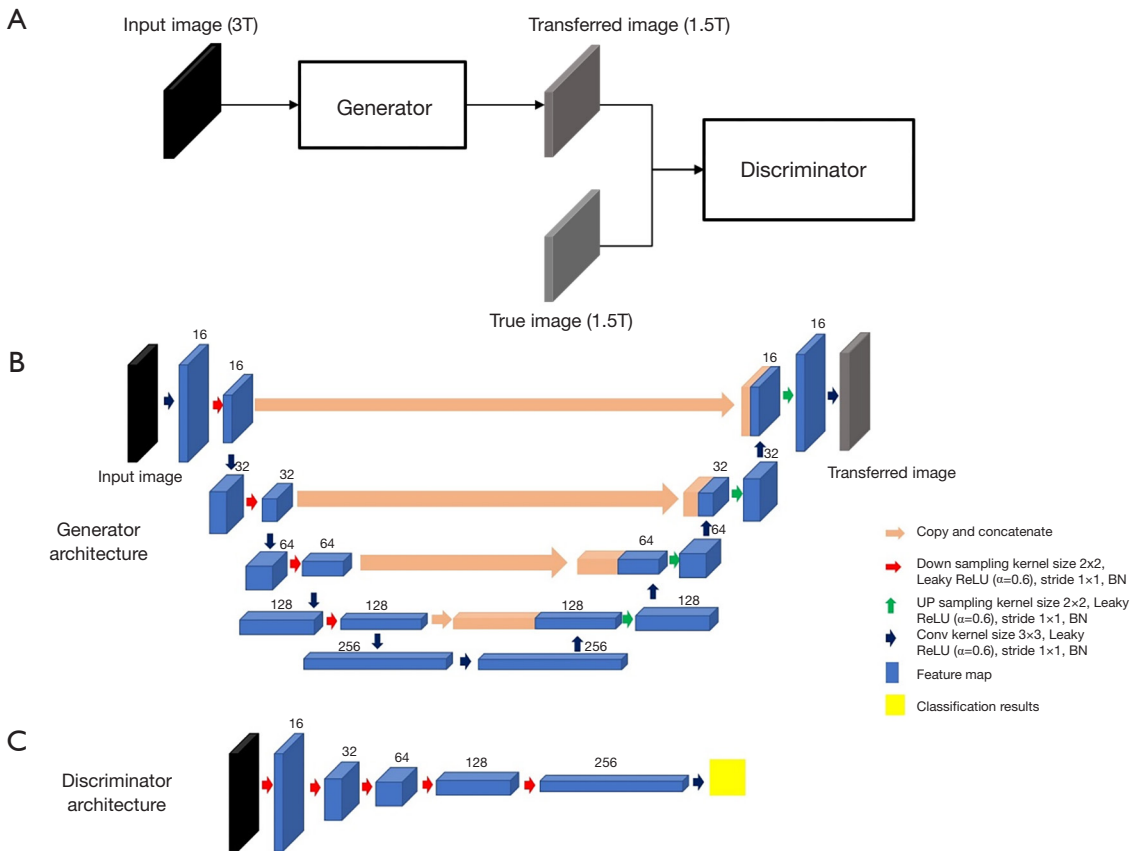

Figure 4 The network architecture of the image domain transfer model. (A) The basic architecture of a GAN used to transfer an image from the 3T scanner to an image from the 1.5T scanner; (B) the architecture of the generator, a U-net-like structure; (C) the architecture of the discriminator, a six-layer classifier. GAN, generative adversarial network.

sample operation. The convolution layer had kernel and stride sizes of 3×3 and 1, respectively. The activation was composed of the Leakey rectified linear unit (ReLU) with a ratio of 0.6 and a batch normalization (BN) that constituted the convolution operation.

The down-stream sampling consisted of the convolution with a stride size of 2. The last block was the only convolution operation performed. The numbers of feature maps in each convolution layer were 16, 16, 32, 32, 64, 64, 128, 128, and 256. The decoder also had 5 blocks, except for the last layer, with each layer consisting of an up-sampling and convolution operation. The convolution operation was only applied in the last layer. The numbers of feature maps in each convolution layer were 256, 128, 128, 64, 64, 32, 32, 16, and 1. The concatenate connection linked the corresponding down-sampling operation in the encoder with the up-sampling operation in the decoder.

The D was a traditional 6-layer classifier. The first 5 layers consisted of the down-sampling operation, and the last layer consisted of the convolution operation. The numbers of feature maps were 16, 32, 64, 128, 256, and 1. To improve the reliability of the GAN and accelerate the convergence model, we removed the BN layer in the first convolutional layer of the generator and the last convolutional layer of the discriminator, as proposed by Radford *et al.*, to use the deep convolutional (DC) GAN (17).

Objective function

The optimized function, L_{total}, of the above-described GAN consists of 4 parts: adversarial loss, L_{adv}, content loss, $L_{content}$, focal loss, L_{focal}, and feature loss, $L_{feature}$. This function can be described as:

$$L_{total} = L_{adv} + \lambda L_{content} + \gamma L_{focal} + \beta L_{feature} \quad [1]$$

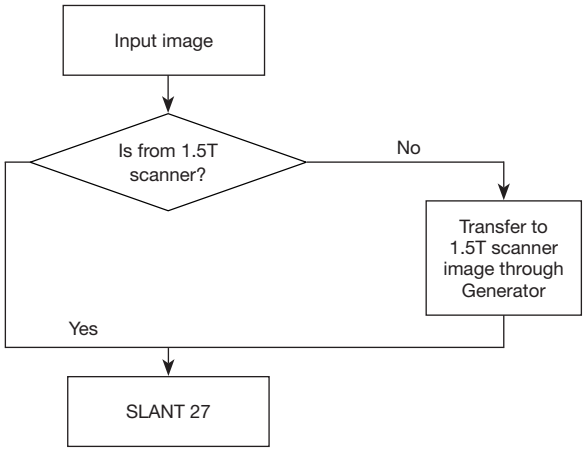

Figure 5 The flow chart to depict the automatic image segmentation steps. SLANT, spatially localized atlas network tiles.

where λ, γ, and β are the 3 regularization parameters used to strike the balance of the 4 terms.

Adversarial loss is defined as follows:

$$L_{adv} = \underset{D}{\operatorname{argmin}} \max_{G} \frac{1}{2} E_{I_t \sim p_{target}} \left[\left(D\left(I_t\right) - a \right)^2 \right] + E_{I_s \sim p_{source}} \left[\left(D\left(G\left(I_s\right)\right) - b \right)^2 \right] \quad [2]$$

where G (Is) represents the transferred images generated by the G. The D(It), D(G(Is)) represents the classification results from the D, and a and b denote the label assigned to the targeted and generated data, respectively. More specifically, adversarial loss in this form was the least square adversarial loss proposed in LSGAN (18). Compared with the log loss in the regular GAN, L_{adv} loss performed more stably and generated higher quality images.

The G can only guarantee that the generated image has the same distribution as the targeted image, it cannot guarantee the same structure; therefore, content loss was used to constrain the structure, as depicted in the following equation:

$$L_{content} = E_{I_f, I_t} \left\| I_f - I_t \right\|_F \quad [3]$$

where $\| \cdot \|$ represents the Frobenius norm and I_f represents the image generated by the GAN. Considering the data came from the source domain, and the target differed slightly in most areas, such as background and global structure, the areas with more detailed information were more likely to be considered. The focal loss in an object detection task, as described by Lin *et al.* (19), which led us to define the pixel-wise focal loss as:

$$L_{focal} = E_{I_f, I_t} \left\| W\left(I_f - I_t\right) \right\|_F \quad [4]$$

where W denotes the pixel-wise weight assigned to the remaining optimized pixel, depicted as:

$$W = \frac{error\ map}{total\ error} \quad [5]$$

where the error map denotes the pixel-wise error between the generated and target data, and the total error denotes a scalar calculated by the sum of the error maps. Coupled with L_{focal} optimization, our GAN methodology focused on detail resulting in higher quality images.

To further improve transfer ability, we made the synthesized data similar not only at the pixel level but also at the semantic level by applying feature loss, which was defined as:

$$L_{focal} = E_{I_f, I_t} \left\| \phi\left(\mathrm{I_f}\right) - \phi\left(\mathrm{I_t}\right) \right\|_F \quad [6]$$

where $\phi(\ .\ )$ represents feature maps extracted by the fifth layer of the discriminator, D, unlike the super-resolution (SR) GAN that uses the pre-trained visual geometry group (VGG) network (20) as a feature extractor. There were 2 reasons for selecting the D to replace the VGG network. Firstly, the pre-trained VGG network was trained on natural images that might not be able to adapt to medical imaging. Secondly, feature loss could also be used as a substitute for the original adversarial loss during training of the generator, according to Ouyang *et al.* (21).

Segmentation

The SLANT-27 (16) was used as the underlying segmentation method in QBrain. The whole-brain segmentation method SLANT-27 combines medical-image processing (registration, harmonization, and label fusion) with 3D network tiles. When a new T1-weighted MRI image arrived, the scan was divided into 27 overlapping tiles. For each tile, the 3D U-net was used to predict the tile label. Finally, the labels of all the tiles were fused into 1 and the inverse affine matrix was applied to transform the mask to the original space. The process of image segmentation is illustrated in *Figure 5*.

Table 1 The intra-scanner test-retest performance comparison of SLANT-27 and FS

Method	Mean (%)	Var (%)	Max (%)	Min (%)
SLANT-27	1.30	0.30	1.67	0.90
FS	3.06	0.48	4.32	2.47

SLANT, spatially localized atlas network tiles; FS, FreeSurfer; Var, variance; Max, maximum; Min, minimum.

Table 2 The inter-scanner test-retest performance comparison of SLANT-27, FS, and QBrain using 1.5T Siemens Avanto and 3T Siemens Trio Tim

Method	Mean (%)	Var (%)	Max (%)	Min (%)
SLANT-27	4.62	0.61	5.83	3.3
FS	6.68	0.92	8.19	5.37
QBrain	3.05	0.45	4.21	2.39

SLANT, spatially localized atlas network tiles; FS, FreeSurfer; Var, variance; Max, maximum; Min, minimum.

Table 3 The inter-scanner test-retest performance comparison of SLANT-27, FS, and QBrain using 1.5T Siemens Avanto and 3T Philips Ingenia

Method	Mean (%)	Var (%)	Max (%)	Min (%)
SLANT-27	3.97	0.38	4.81	3.36
FS	6.21	0.71	7.98	4.74
QBrain	1.96	0.25	2.60	1.63

SLANT, spatially localized atlas network tiles; FS, FreeSurfer; Var, variance; Max, maximum; Min, minimum.

Statistical analysis

The reproducibility of measuring the whole-brain structural volume using SLANT-27, FS, and QBrain in the Interscan48 test set was compared in section 3. The reproducibility was measured by test-retest variability (TRV), as follows:

$$TRV = \frac{|v_1 - v_2|}{v_1 + v_2} \times 200\% \quad [7]$$

where v_1 and v_2 represent the volumes of a brain region from the first (test) and second (retest) generated masks, respectively. A lower TRV value indicated better reproducibility compared with a higher TRV value. In test-retest statistics, the TRV of each brain tissue for each pair of volumes was computed, and then the mean, m, over these TRV values was calculated. Finally, the mean, variance (Var), maximum (Max), and minimum (Min) of m over all of the pairs were computed. The mean segmentation error was measured by subtracting the mean TRV of SLANT-27 from the mean TRV of Qbrain or FS and the paired t-test was used for the intra-scanner test-retest performance comparison of SLANT-27 and FS.

Results

Test-retest experiments of different field strengths and scanner vendors

The reproducibility of the whole-brain structural volume was measured by assessing:

(I) Intra-scanner reproducibility.

The intra-scanner test-retest segmentation performances of SLANT-27 and FS were first compared, as shown in *Table 1*. The results indicated that SLANT-27 was more reliable than FS for within-scanner reproducibility in the single center (mean TRV 1.30% *vs.* 3.06%, P<0.05).

(II) Inter-scanner reproducibility of different field strengths from the same vendor.

The 1.5T Avanto and the 3.0T Trio Tim MR scanners, both manufactured by Siemens, were the 2 scanners used for the inter-scanner test-retest performance comparison of SLANT-27, FS, and QBrain. These results are reported in *Table 2*. The SLANT-27 demonstrated better inter-scanner test-retest reproducibility than FS. Furthermore, compared to the SLANT-27 model, the mean segmentation error of TRV was reduced by 1.57% (from 4.62% to 3.05%) using the QBrain method on the 3T Siemens Trio Tim scanner.

(III) Inter-scanner reproducibility of different field strengths and vendors.

The inter-scanner TRV of SLANT-27, FS, and QBrain across different field strengths and vendors is shown in *Table 3*. Among the 3 methods, QBrain had the best performance with a TRV <2%. In addition, the mean segmentation error was reduced by 2.01% (from 3.97% to 1.96%) on the 3T Philips Ingenia scanner when the transfer module (GAN) was added.

(IV) Inter-scanner reproducibility of different vendors with the same field strength.

The inter-scanner TRV of SLANT-27, FS, and

QBrain across different vendors with the same field strengths is shown in *Table 4*. It was observed that the TRV value of QBrain was the lowest, at approximately 3% across both the Siemens and Philips vendors, and the mean segmentation error of TRV was reduced by 0.56% (from 3.52% to 2.96%) when the GAN module was added.

Table 4 The inter-scanner test-retest performance comparison of SLANT-27, FS, and QBrain using 3T Siemens Trio Tim and 3T Philips Ingenia

Method	Mean (%)	Var (%)	Max (%)	Min (%)
SLANT-27	3.52	0.46	4.25	2.70
FS	3.78	0.22	4.46	3.17
QBrain	2.96	0.24	3.27	2.56

SLANT, spatially localized atlas network tiles; FS, FreeSurfer; Var, variance; Max, maximum; Min, minimum.

Analysis by local regions

A TRV comparison between 60 different brain structures is shown in *Figure 6*. It can be seen that the overall level of TRV decreased significantly after the domain transfer module was applied to both the Trio Tim and Ingenia scanners. These results indicate that data from different domains can be transferred to similar distributions, which greatly improves the robustness of the segmentation model. Additionally, a comparison of the mean value of inter-scanner test-retest performance of Qbrain and SLANT27 is shown in *Figure 7*.

Discussion

In this study, we proposed a robust, automatic, whole-brain segmentation method, QBrain, which consists of domain transfer and segmentation modules. Domain transfer is used to transfer the data from different scanners to a

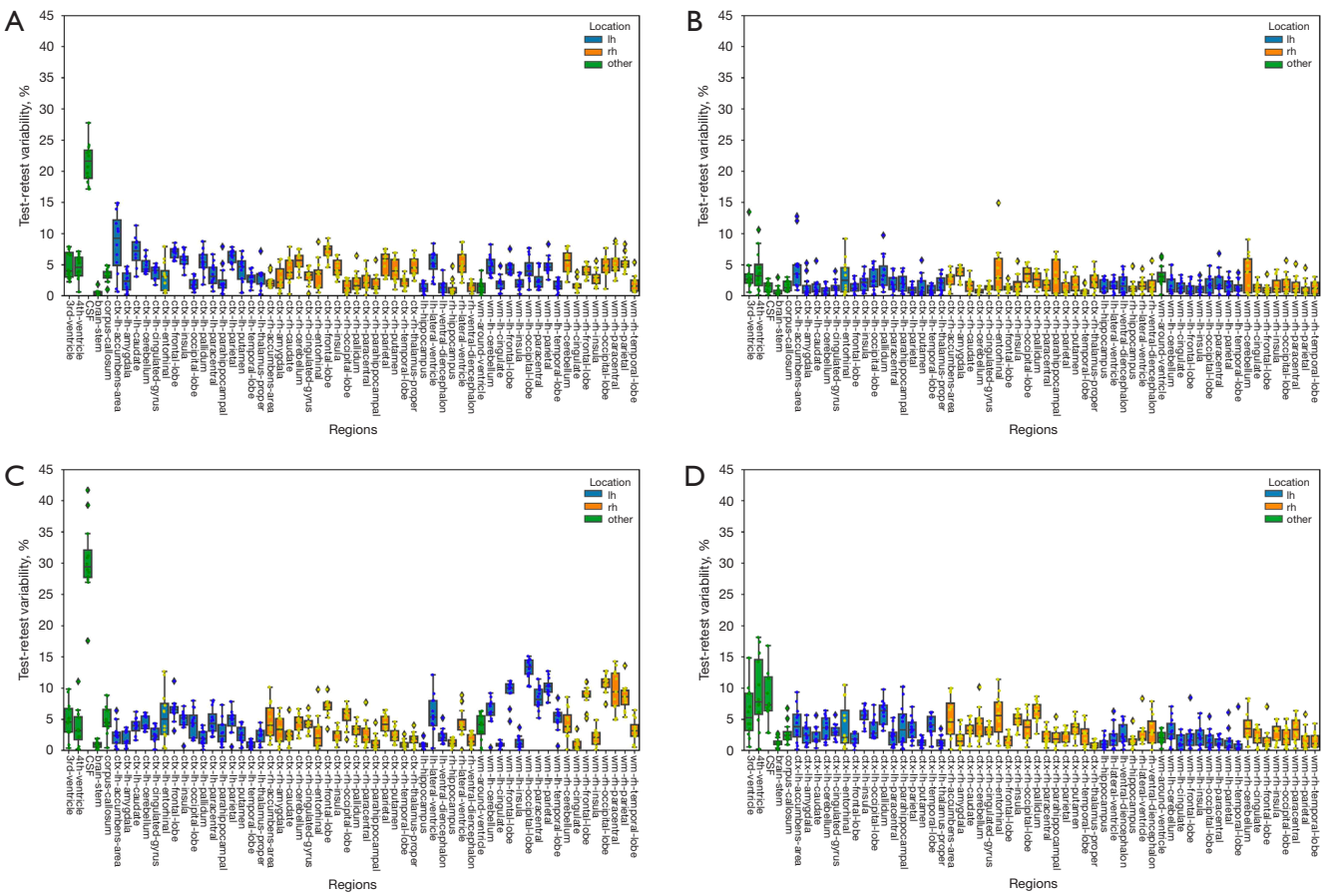

Figure 6 The flow chart to depict the automatic image segmentation steps. SLANT, spatially localized atlas network tiles.

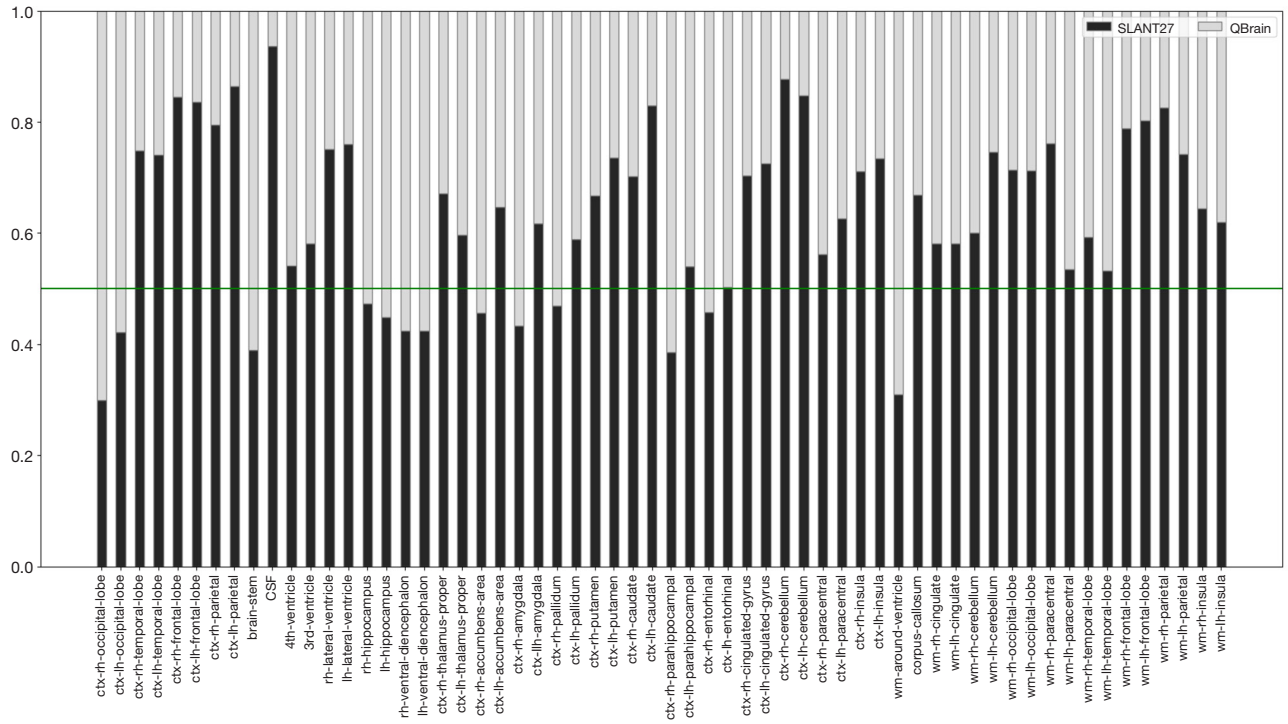

Figure 7 The comparison of the mean value of inter-scanner test-retest performance of Qbrain and SLANT27, based on data from 1.5T Siemens Avanto and 3T Philips Ingenia. SLANT, spatially localized atlas network tiles.

unified domain to improve the robustness of segmentation. Compared to the widely accepted and applied FS (9), QBrain obtained lower segmentation TRV values, both on intra-scanner and inter-scanner tests, which indicated that our method has better reliability and robustness. Additionally, QBrain improved the performance of SLANT-27 by image domain transfer. Compared with 3D CNN models (22,23), our method is less computationally intensive and easier to train. This is the first report of the TRV in structural volume across different field strengths and different scanners based on the GAN model.

Influence of functional loss

There are 4 loss functions employed in domain transfer modules, including L_{adv}, $L_{content}$, $L_{feature}$, and L_{focal}. L_{adv}, that are used to learn the distribution of 1.5T data by adversarial training between generators and discriminators, as well as to normalize the field strength of different scanners. The $L_{content}$ is a basic constraint for generating structural information at the pixel level and can also reduce the solution space and accelerate convergence of the modules. When only L_{adv} and $L_{content}$ were used, our generated results had similar signal distributions and structural features compared with the target data, and $L_{feature}$ was used to improve similarities at the feature level. Some detailed information, such as information on blood vessels, is sparsely distributed and cannot be made a focus because the above mentioned losses occur on a global scale. We, therefore, guided the model to focus on optimization of these regions by L_{focal}, which may show a reduced appearance of blood vessels.

Intra-scanner reproducibility

We found that repeatability was better with Qbrain compared to FS in whole-brain segmentation, as well as local regions of the brain. There were 2 reasons identified for intra-scanner errors in FS (9): (I) different images affect structural estimates, and (II) random noise and processing bias lead to small changes that arbitrarily accumulate and create large changes in a single patient. In contrast, once our model was trained, all parameters were fixed and thus avoided processing bias. Our method also learned high-level semantic information from deep convolution layers,

which resulted in intensity changes between 2 scans which had lesser impact on segmentation results. Therefore, our method was more reliable in the intra-scanner reliability test. Compared with previous results on the performance of Siemens scanners, the intra-scanner reliability of Qbrain (0.9–1.67%) was much smaller than that of FS (0–16.46%) and another work-in-progress package issued by Yan *et al.* (24) (0–9.89%).

Inter-scanner reproducibility

Generally, we found that the variability in volumetrics was lower on the same scanner than between scanners for both SLANT-27 and FS methods, and our current result of FS is in line with our previous research involving multiple sclerosis (MS) patients (6). The QBrain method, which consists of the domain image transfer module, has significantly improved inter-scanner variability compared with FS, especially between the 1.5T Siemens Avanto and the 3T Philips Ingenia scanners (1.96–6.21%). For whole-brain volume, the mean rate of atrophy in normal aging ranged from 0.3% to 0.7% per year (25), and reaches up to 1% to 4% per year in AD patients (26). This shows that our QBrain framework has an inter-scanner variability lower than that of the annual atrophy rate in AD patients, and it is expected that QBrain will be a feasible measurement for grouping AD data that are followed up within 1 to 2 years across different scanners.

As shown in *Tables 2,3*, our segmentation module based on SLANT-27 was only trained on the images collected using the 1.5T Siemens Avanto scanner across multiple centers. It was difficult to achieve superior reliability repeatedly on the 3T scanner due to the inconsistent image contrast. This is a common problem, with poor generalizations made by CNN-based methods. Various scanners have different hardware and acquisition sequences and thus lead to differences in images. In order to convert the image domain to a similar distribution, we used a GAN to transfer all of the 3T scanner data to the 1.5T scanner data. The mean segmentation error of TRV of different field strengths from the same vendors, different field strengths and vendors, and the same field strengths but different vendors was reduced by 1.57%, 2.01%, and 0.56%, respectively. The inter-scanner reliability of our system was significantly improved.

Influence of artifacts

There were large TRV values in the third brain ventricle, fourth brain ventricle, and cerebrospinal fluid (CSF) after domain transfer. A probable reason for the artifacts is that CSF flow and blood vessel pulses caused signal changes in these areas. Therefore, the artifacts could not be converted correctly by our transfer module.

Limitations and future work

Although we achieved markedly stable repeatability on intra-scanner and inter-scanner tests, our research still had several limitations. First, as discussed above, the presence of artifacts needs to be resolved. Second, in this study, only 3T MRI images collected from Siemens Trio Tim and Philips Ingenia scanners were converted by our method. In future, images from different 3T machine models and manufacturers need to be used to completely cover gaps in the other models. Third, our image domain transfer model was only trained on 34 paired 3T–1.5T images, and more paired data need to be collected to improve the performance of the presented method.

Conclusions

Herein, we have presented a new automatic segmentation method, QBrain, which includes a GAN network with new loss functions to transfer image and a SLANT-27 DL segmentation module to improve the reliability of within- and between-center brain image comparison. Our method could effectively minimize variability in multi-center and follow-up brain image segmentation studies.

Acknowledgments

We would like to thank the participants as well as the staff at the Department of Radiology at the First Hospital of Jilin University in China for making this study possible. We also thank Yanhua Wu from the Division of Clinical Research at the First Hospital of Jilin University who kindly provided statistical advice for this manuscript.

Funding: This study was supported by grants provided by the National Natural Science Foundation of China (62072212, 81600923), Natural Science Foundation of Jilin Province of China (20210101273JC), Foundation of Department of Finance of Jilin Province of China (2018SCZWSZX-026), Foundation of Health and Family Planning Commission of Jilin Province (2020J052), Bethune Project of Jilin University (2020B47), the National Key Research and Development Project of China (2020YFC2007302), Science

and Technology Achievement Transformation Fund of the First Hospital of Jilin University (JDYY2021-A0010) Beijing Municipal Science and Technology Project of Brain Cognition and Brain Medicine (Z171100000117001), and in part by the Jilin Province Development and Reform Commission Project under Grant 2020C019-2.

Footnote

Reporting Checklist: The authors have completed the MDAR checklist. Available at https://dx.doi.org/10.21037/qims-21-653

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/qims-21-653). TQ, ZY, and WY report that they are the employees of AI lab, QuantMind, which developed the Qbrain method used in this paper. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). All the MRI data employed in this study were collected with the approval of the local ethics committee. The age of participants ranged from 10 to 78 years, and informed written consent was provided by each participant prior to data collection and analysis.

References

1. Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage 2006;32:180-94.
2. Jovicich J, Czanner S, Han X, Salat D, van der Kouwe A, Quinn B, Pacheco J, Albert M, Killiany R, Blacker D, Maguire P, Rosas D, Makris N, Gollub R, Dale A, Dickerson BC, Fischl B. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage 2009;46:177-92.
3. Duning T, Kloska S, Steinsträter O, Kugel H, Heindel W, Knecht S. Dehydration confounds the assessment of brain atrophy. Neurology 2005;64:548-50.
4. Sormani MP, Arnold DL, De Stefano N. Treatment effect on brain atrophy correlates with treatment effect on disability in multiple sclerosis. Ann Neurol 2014;75:43-9.
5. Durand-Dubief F, Belaroussi B, Armspach JP, Dufour M, Roggerone S, Vukusic S, Hannoun S, Sappey-Marinier D, Confavreux C, Cotton F. Reliability of longitudinal brain volume loss measurements between 2 sites in patients with multiple sclerosis: comparison of 7 quantification techniques. AJNR Am J Neuroradiol 2012;33:1918-24.
6. Guo C, Ferreira D, Fink K, et al. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. Eur Radiol 2019;29:1355-64.
7. Guo S, Xiao B, Wu C. Identifying subtypes of mild cognitive impairment from healthy aging based on multiple cortical features combined with volumetric measurements of the hippocampal subfields. Quant Imaging Med Surg 2020;10:1477-89.
8. Zhang C, Kong M, Wei H, Zhang H, Ma G, Ba M. The effect of ApoE ε 4 on clinical and structural MRI markers in prodromal Alzheimer's disease. Quant Imaging Med Surg 2020;10:464-74.
9. Fischl B. FreeSurfer. Neuroimage 2012;62:774-81.
10. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Springer: International Conference on Medical image computing and computer-assisted intervention, 2015.
11. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. IEEE: 2016 Fourth International Conference on 3D Vision (3DV), 2016.
12. Roy AG, Conjeti S, Sheet D, Katouzian A, Navab N, Wachinger C. Error corrective boosting for learning fully convolutional networks with limited data. Springer: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017.

13. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks. Advances in Neural Information Processing Systems 2014;3:2672-80.
14. Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, Furukawa Y, Mauri G, Nakayama H. GAN-based synthetic brain MR image generation. IEEE: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018.
15. Lei Y, Harms J, Wang T, Liu Y, Shu HK, Jani AB, Curran WJ, Mao H, Liu T, Yang X. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med Phys 2019;46:3565-81.
16. Huo Y, Xu Z, Xiong Y, Aboud K, Parvathaneni P, Bao S, Bermudez C, Resnick SM, Cutting LE, Landman BA. 3D whole brain segmentation using spatially localized atlas network tiles. NeuroImage 2019;194:105-19.
17. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434 2015.
18. Mao X, Li Q, Xie H, Lau RY, Wang Z, Smolley SP. On the effectiveness of least squares generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 2019;41:2947-60.
19. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision; 2017.
20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014.
21. Ouyang J, Chen KT, Gong E, Pauly J, Zaharchuk G. Ultra-low-dose PET reconstruction using generative adversarial network with feature matching and task-specific perceptual loss. Med Phys 2019;46:3555-64.
22. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. Springer: International conference on information processing in medical imaging, 2017.
23. Rajchl M, Pawlowski N, Rueckert D, Matthews PM, Glocker B. Neuronet: fast and robust reproduction of multiple brain image segmentation pipelines. arXiv preprint arXiv:180604224 2018.
24. Yan S, Qian T, Maréchal B, Kober T, Zhang X, Zhu J, Lei J, Li M, Jin Z. Test-retest variability of brain morphometry analysis: an investigation of sequence and coil effects. Ann Transl Med 2020;8:12.
25. Scahill RI, Frost C, Jenkins R, Whitwell JL, Rossor MN, Fox NC. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch Neurol 2003;60:989-94.
26. Fox NC, Schott JM. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. Lancet 2004;363:392-4.

Cite this article as: Niu K, Li X, Zhang L, Yan Z, Yu W, Liang P, Wang Y, Lin CP, Zhang H, Guo C, Li K, Qian T. Improving segmentation reliability of multi-scanner brain images using a generative adversarial network. Quant Imaging Med Surg 2022;12(3):1775-1786. doi: 10.21037/qims-21-653

Supplementary

Table S1 MRI acquisition parameters

	MR1	MR2	MR3
Manufacturer	Siemens	Philips	Siemens
Model name	Avanto	Ingenia	Trio Tim
Station name	MRC25494	3FCD991	MRC35363
System version	syngo MR B17	R6.0.531	syngo MR B15
Field strength (T)	1.5	3	3
Head coil channels	8	16	8
3D T1w Images			
Voxel size, mm3	1.0×1.0×1.0	1.0×1.0×1.0	1.0×1.0×1.0
Number of slices	176	192	176
Repetition time (ms)	1900	7.07	1900
Echo time (ms)	3.37	3.19	2.96
Flip angle (°)	15	7	9

MRI, magnetic resonance imaging; T1WI, T1-weighted images.

https://dx.doi.org/10.21037/qims-21-653