SCIENTIFIC REPORTS
nature research

OPEN

MRI Cross-Modality Image-to-Image Translation

Qianye Yang[1,5], Nannan Li[1,4,5], Zixu Zhao[1,5], Xingyu Fan[2,5], Eric I-Chao Chang[3] & Yan Xu[1,3*]

We present a cross-modality generation framework that learns to generate translated modalities from given modalities in MR images. Our proposed method performs Image Modality Translation (abbreviated as IMT) by means of a deep learning model that leverages conditional generative adversarial networks (cGANs). Our framework jointly exploits the low-level features (pixel-wise information) and high-level representations (e.g. brain tumors, brain structure like gray matter, etc.) between cross modalities which are important for resolving the challenging complexity in brain structures. Our framework can serve as an auxiliary method in medical use and has great application potential. Based on our proposed framework, we first propose a method for cross-modality registration by fusing the deformation fields to adopt the cross-modality information from translated modalities. Second, we propose an approach for MRI segmentation, translated multichannel segmentation (TMS), where given modalities, along with translated modalities, are segmented by fully convolutional networks (FCN) in a multichannel manner. Both of these two methods successfully adopt the cross-modality information to improve the performance without adding any extra data. Experiments demonstrate that our proposed framework advances the state-of-the-art on five brain MRI datasets. We also observe encouraging results in cross-modality registration and segmentation on some widely adopted brain datasets. Overall, our work can serve as an auxiliary method in medical use and be applied to various tasks in medical fields.

Magnetic Resonance Imaging (MRI) has become prominent among various medical imaging techniques due to its safety and information abundance. They are broadly applied to clinical treatment for diagnostic and therapeutic purposes. There are different modalities in MR images, each of which captures certain characteristics of the underlying anatomy. All these modalities differ in contrast and function. Three modalities of MR images are commonly referenced for clinical diagnosis: T1 (spin-lattice relaxation), T2 (spin-spin relaxation), and T2-Flair (fluid attenuation inversion recovery)[1]. T1 images are favorable for observing structures, e.g. gray matter and white matter in the brain; T2 images are utilized for locating tumors; T2-Flair images present the location of lesions with water suppression. Each modality provides a unique view of intrinsic MR parameters. Examples of these three modalities are shown in Fig. 1. Taking full consideration of all these modalities is conducive to MR image analysis and diagnosis.

However, the existence of complete multi-modality MR images is limited by the following factors: (1) There is a certain probability of failure during the scanning process. (2) Motion artifacts are produced along with MR images. These artifacts are attributed to the difficulty of staying still for patients during scanning (e.g. pediatric population[2]), or motion-sensitive applications such as diffusion imaging[3]. (3) The mapping from one modality to another is hard to learn. Each of modality captures different characteristics of the underlying anatomy, and the relationship between any two modalities is highly non-linear. Owing to differences in the image characteristics across modalities, existing approaches cannot achieve satisfactory results for cross-modality synthesis as mentioned in[4]. For example, when dealing with the paired MRI data, the regression-based approach[5] even lose some information of brain structures. Synthesizing a translated modality from a given modality without real acquisitions, also known as cross-modality generation, is a nontrivial problem worth of studying. Take the transition from T1 (given modality) to T2 (target modality) as an example, $\hat{T}2$ (translated modality) can be generated through a cross-modality generation framework. In this paper,$\hat{}$denotes translated modalities. Cross-modality

[1]State Key Laboratory of Software Development Environment and Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education and Research Institute of Beihang University in Shenzhen, Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, 100191, China. [2]Bioengineering College of Chongqing University, Chongqing, 400044, China. [3]Microsoft Research Asia, Beijing, 100080, China. [4]Ping An Technology (Shenzhen) Co., Ltd., Shanghai, 200030, China. [5]These authors contributed equally: Qianye Yang, Nannan Li, Zixu Zhao and Xingyu Fan. *email: xuyan04@gmail.com

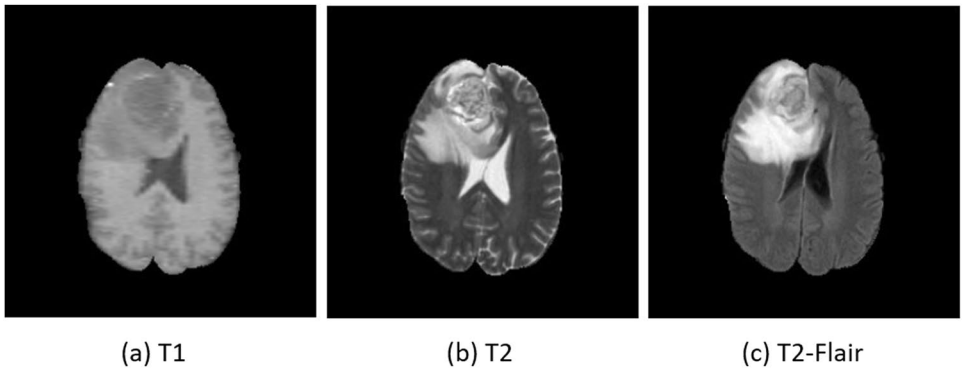

Figure 1. Examples of three different modalities: (**a**) T1, (**b**) T2, and (**c**) T2-Flair.

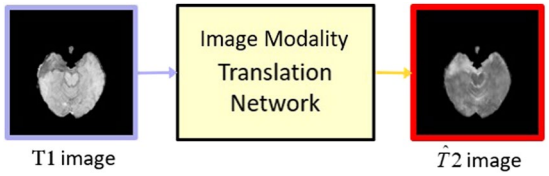

Figure 2. Overview of our IMT network. It learns to generate translated modality images ($\hat{T}2$) from given modality images (T1). The red box indicates our translated images.

generation tasks refer to transitions such as from T1 to T2, from T1 to T2-Flair, from T2 to T2-Flair, and vice versa.

Recently, image-to-image translation networks have provided a generic solution for image prediction problems in natural scenes, like mapping images to edges[6,7], segments[8], semantic labels[9] (many to one), and mapping labels to realistic images[10] (one to many). It requires an automatic learning process for loss functions to make the output indistinguishable from reality. The recently proposed Generative Adversarial Network (GAN)[11–14] makes it possible to learn the distribution of the input data and be applied to multiple translation tasks. Isola *et al.*[13] demonstrate that the conditional GAN (cGAN) is suitable for image-to-image translation tasks.

Previous work on image-to-image translation networks focuses on natural scenes[13,15–17]. Motivated by[13], we introduce Image Modality Translation networks (IMT) to brain MRI cross-modality generation (see Fig. 2). Unlike some classic regression-based approaches that leverage an L1 loss to capture the low-level information, we adopt cGANs to capture high-level information and an L1 loss to ensure low-level information at the same time, which allows us to recover more details from the given modality and reduce the noise generated along with the translated modality.

In this paper, we mainly focus on developing a cross-modality generation framework which provides us with novel approaches of cross-modality registration and segmentation. Our proposed cross-modality generation framework has great application potential, such as multimodal registration[18], segmentation[19], and virtual enhancement[4]. Among all these applications, we choose cross-modality registration and segmentation as two examples to illustrate the effectiveness of our cross-modality generation framework.

The first application of our proposed framework is cross-modality image registration which is necessary for medical image processing and analysis. With regard to brain registration, accurate alignment of the brain structures such as hippocampus, gray matter, and white matter are crucial for monitoring brain disease like Alzheimer Disease (AD). The accurate delineation of brain structures in MR images can provide neuroscientists with volumetric and structural information on the structures, which has been already achieved by existing atlas-based registrations[18,20]. However, few of them adopt the cross-modality information from multiple modalities, especially from translated modalities.

Here, we propose a new method for cross-modality registration by adopting cross-modality information from our translated modalities. The flowchart is illustrated in Fig. 3. In our method, inputting a given-modality image (e.g. T2 image) to our proposed framework yields a translated modality (e.g. $\hat{T}1$ image). Both two modalities compose our fixed images space (T2 and $\hat{T}1$ images). The moving images including T2 and T1 images are then registered to the identical modality in the fixed images space with a registration algorithm. Specifically, T2 (moving) is registered to T2 (fixed), T1 (moving) is registered to $\hat{T}1$ (fixed). The deformation generated in the registration process are finally combined in a weighted fusion process and then propagate the moving images labels to the fixed images space. It is feasible since the introduction of translated modality provides us with richer anatomical information in comparison with only one modality is given, leading to more precise registration results. Our method is applicable to dealing with cross-modality registration problems by making the most of cross-modality information without adding any extra data at the same time.

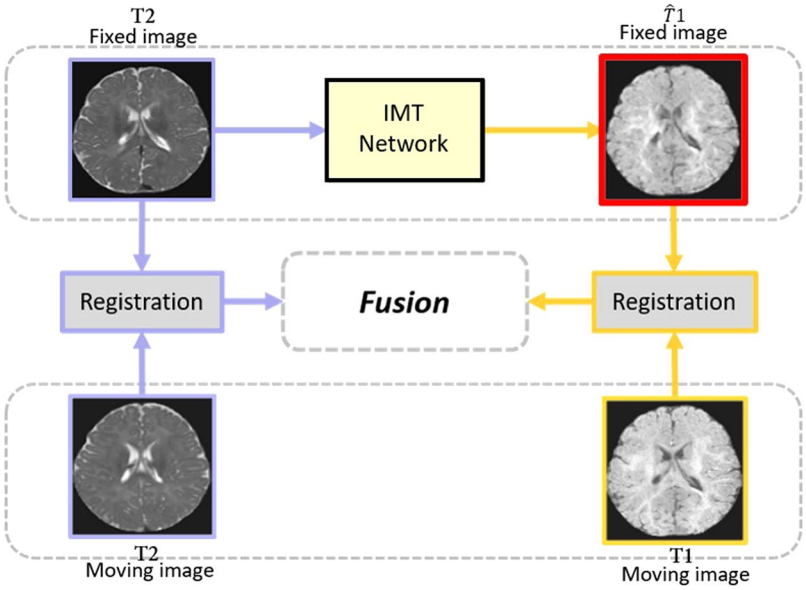

Figure 3. Overview of our approach for cross-modality registration. Inputting a given-modality image (T2) to IMT framework yields a translated modality ($\hat{T}1$). Then T2 (moving) is registered to T2 (fixed), T1 (moving) is registered to $\hat{T}1$ (fixed). The deformation generated in the registration process are finally combined in a weighted fusion process, obtaining our final registration result. The red box indicates our translated images.

The second application of our proposed framework is brain segmentation for MRI data. However, it is a difficult task owing to the artifacts and in-homogeneities introduced during the real image acquisition[21,22]. To this point, we propose a novel approach for brain segmentation, called translated multichannel segmentation (TMS). In TMS, the translated modality and its corresponding given modality are fed into fully convolutional networks (FCN)[9] for brain segmentation. Here, we fine tune Imagenet-FCN model using our MRI images. (Other well-performing segmentation models such as U-Net and dilated CNNs might as well be selected.) Thus we follow its original three-channel network, inputting one translated modality and two given modality images to serve as three channels. TMS is an effective method for brain segmentation by adding cross-modality information from translated modalities since different MRI modalities have unique tissue contrast profiles and therefore provide complementary information that could be of use to the segmentation process. For instance, TMS can improve tumor segmentation performance by adding cross-modality information from translated T2 modality into original T1 modality.

Contributions: (1) We introduce the end-to-end Image Modality Translation (IMT) network for cross-modality MRI generation to synthesize translated modalities from given modalities. A comprehensive comparison is provided with five datasets representing real-world clinical applications, each has its unique characteristics in data size, patient cohort and disease status.The results show that our IMT framework can cope with a variety of brain MRI modality translation tasks using the same objective and architecture.

(2) Registration: We proposed a registration method which is able to leverage our IMT framework to augment the fixed images space with translated modalities for atlas-based registration. Registering moving images to fixed images and weighted fusion process enable us to make the most of cross-modality information without adding any extra data.

(3) Segmentation: Our proposed approach, translated multichannel segmentation (TMS), performs cross-modality image segmentation by means of FCNs. We input two identical given modalities and one corresponding translated modality into separate channels, which allows us to adopt and fuse cross-modality information and improve the segmentation performance without using any extra data.

Related Work

In this section, we mainly focus on methods related to cross-modality image generation, its corresponding registration and segmentation.

Image generation. Related work on image generation can be broadly divided into three categories: cross-modality synthesis, GANs in natural scenes, and GANs in medical images.

Cross-modality synthesis. In order to synthesize one modality from another, a rich body of algorithms have been proposed using non-parametric methods like nearest neighbor (NN) search[23], random forests[5], coupled dictionary learning[18], and convolutional neural network (CNN)[24], etc. They can be broadly categorized into two classes: **(1) Traditional methods**. One of the classical approaches is an atlas-based method proposed by Miller *et al.*[25]. The atlas contains pairs of images with different tissue contrasts co-registered and sampled on the same voxel locations in space. An example-based approach is proposed to pick several NNs with similar properties from low-resolution images to generate high-resolution brain MR images using a Markov random field[26]. In[5],

a regression-based approach is presented where a regression forest is trained using paired data from a given modality to a target modality. Later, the regression forest is utilized to regress target-modality patches from given modality patches. **(2) Deep learning based methods**. Nguyen *et al.*[24] present a location-sensitive deep network (LSDN) to incorporate spatial location and image intensity feature in a principled manner for cross-modality generation. Vemulapalli *et al.*[4] propose a general unsupervised cross-modal medical image synthesis approach that works without paired training data. Huang *et al.*[27] attempt to jointly solve the super-resolution and cross-modality generation problems in 3D medical imaging using weakly-supervised joint convolutional sparse coding.

Our image generation task is essentially similar to these issues. We mainly focus on developing a novel and simple framework for cross-modality image generation and we choose paired MRI data as our case rather than unpaired data to improve the performance. To this point, we try to develop a 2D framework for cross-modality generation tasks according to 2D MRI principle. The deep learning based methods[4,27] are not perfectly suitable for our case on the premise of our paired data and MRI principle. We thus select the regression-based approach[5] as our baseline.

GANs in natural scenes. Recently, a Generative Adversarial Network (GAN) has been proposed by Goodfellow *et al.*[11]. They adopt the concept of a min-max optimization game and provide a thread to image generation in unsupervised representation learning settings. To conquer the immanent hardness of convergence, Radford *et al.*[28] present a deep convolutional Generative Adversarial Network (DCGAN). However, there is no control of image synthesis owing to the unsupervised nature of unconditional GANs. Mirza *et al.*[29] incorporate additional information to guide the process of image synthesis. It shows great stability refinement of the model and descriptive ability augmentation of the generator. Various GAN-family applications have come out along with the development of GANs, such as image inpainting[12], image prediction[13], text-to-image translation[14] and so on. Whereas, all of these models are designed separately for specific applications due to their intrinsic disparities. To this point, Isola *et al.*[13] present a generalized solution to image-to-image translations in natural scenes. Our cross-modality image generation is inspired by[13] but we focus on medical images generation as opposed to natural scenes.

GANs in medical images. Except of the success of existing approaches in natural scenes, there are several applications of GANs to medical images as well. Nie *et al.*[30] estimate CT images from MR images with a Context-Aware GAN model. Wolterink *et al.*[31] demonstrate that GANs are applicable to transforming low-dose CT into routine-dose CT images. However, all these methods are designed for specific rather than general applications. Loss functions need to be modified when it comes to multi-modality transitions. Thus, a general-purpose strategy for medical modality transitions is of great significance. Fortunately, this is achieved by our cross-modality image generation framework. The previous version of our manuscript is uploaded to Arxiv in early 2018.

Image registration. A successful image registration application requires several components that are correctly combined, like the cost function and transformation model. The cost function, also called similarity metrics, measures how well two images are matched after transformation. It is selected with regards to the types of objects to be registered. As for cross-modality registration, commonly adopted cost functions are mutual information (MI)[32] and cross-correlation (CC)[33]. Transformation models are determined according to the complexity of deformations that need to be recovered. Some common parametric transformation models (such as rigid, affine, and B-Splines transformation) are enough to recover the underlying deformations[34].

Several image registration toolkits such as ANTs[35] and Elastix[36] have been developed to facilitate research reproduction. These toolkits have effectively combined commonly adopted cost functions and parametric transformation models. They can estimate the optimal transformation parameters or deformation fields based on an iterative framework. In this work, we choose ANTs and Elastix to realize our cross-modality registration. More registration algorithms can be applied to our method.

Image segmentation. A rich body of image segmentation algorithms exists in computer vision[8,9,37,38]. We discuss two that are closely related to our work.

The Fully Convolutional Network (FCN) proposed by Long *et al.*[9] is a semantic segmentation algorithm. It is an end-to-end and pixel-to-pixel learning system which can predict dense outputs from arbitrary-sized inputs. Inspired by[9], TMS adopts similar FCN architectures but focuses on fusing information of different modalities in a multichannel manner.

Xu *et al.*[8] propose an algorithm for gland instance segmentation, which adopts the idea of multichannel learning. The proposed algorithm exploits features of edge, region, and location in a multichannel manner to generate instance segmentation. By contrast, TMS leverages features in translated modalities to refine the segmentation performance of given modalities.

Methods

In this section, we mainly learn an end-to-end mapping from given-modality images to target-modality images. We introduce Image Modality Translation (IMT) networks to cross-modality generation. Here, cGANs are used to realize IMT networks. The flowchart of our algorithm is illustrated in Fig. 4.

Training. We denote our training set as $S = \{(x_i, y_i), i = 1, 2, 3, \ldots, n\}$, where x_i refers to the ith input given-modality image, and y_i indicates the corresponding target-modality image. We subsequently drop the subscript i for simplicity, since we consider each image holistically and independently. Our goal is to learn a mapping from given-modality images $\{x_i\}_{i=1}^{n} \in X$ to target-modality images $\{y_i\}_{i=1}^{n} \in Y$. Thus, given an input image x and a random noise vector z, our method can synthesize the corresponding translated-modality image $\hat{y}$. Take the transition from T1 to T2 as an instance. Similar to a two-player min-max game, the training procedure of GAN

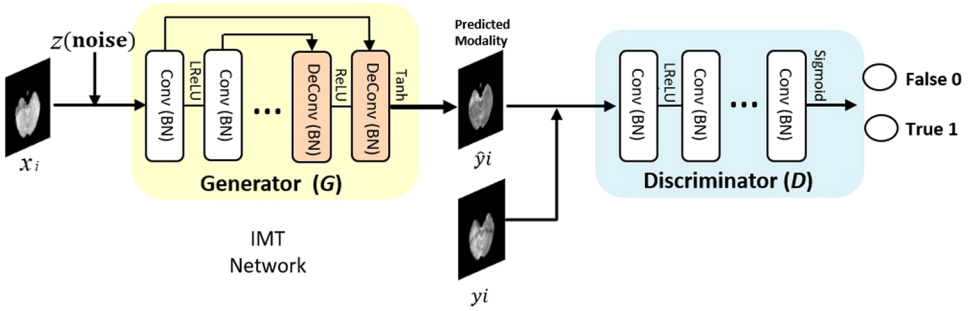

Figure 4. Overview of our end-to-end IMT network for cross-modality generation. Notice that our training set is denoted as $S = \{(x_i, y_i), i = 1, 2, 3, \ldots, n\}$, where x_i and y_i refer to the ith input given-modality image and its corresponding target-modality image. The training process involves two aspects. On the one hand, given an input image x_i and a random noise vector z, generator G aims to produce indistinguishable images $\hat{y}_i$ from the real images y_i. On the other hand, discriminator D evolves to distinguish between translated-modality images $\hat{y}_i$ generated by G and the real images y_i. The output of D is 0 or 1, where 0 represents synthesized images and 1 represents the real data. In the generation process, translated-modality images can be synthesized through the optimized G.

mainly involves two aspects: On one hand, given an input image T1 (x), generator G produces a realistic image $\hat{T}2$ ($\hat{y}$) towards the real data T2 (y) in order to puzzle discriminator D. On the other hand, D evolves to distinguish synthesized images $\hat{T}2$ ($\hat{y}$) generated by G from the real data T2 (y). The overall objective function is defined:

$$\begin{aligned}\mathscr{L}_{cGAN}(G, D) &= \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x, y)]+ \\ &\mathbb{E}_{x\sim p_{data}(x),zp_z(z)}[\log(1 - D(x, G(x, z))],\end{aligned} \tag{1}$$

where $p_{data}(x)$ and $p_{data}(z)$ refer to the distributions over data x and z, respectively. G is not only required to output realistic images to fool D, but also to produce high-quality images close to the real data. Existing algorithms[12] have found it favorable to combine traditional regularization terms with the objective function in GAN. An L1 loss, as described in[13,39], usually guarantees the correctness of low-level features and encourages less blurring than an L2 loss. Thus, an L1 loss term is adopted into the objective function in our method. The L1 loss term is defined as follows:

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)}[\|y - G(x, z)\|_1]. \tag{2}$$

The overall objective function is then updated to:

$$\mathscr{L} = \mathscr{L}_{cGAN}(G, D) + \lambda\mathscr{L}_{L1}(G), \tag{3}$$

where λ is a hyper-parameter specified manually to balance the adversarial loss and L1 loss. The appropriate weight of λ is based on the cross-validation of training data. A value of 100 is eventually selected for λ.

Following[13], the optimization is an iterative training process with two steps: (1) fix parameters of G and optimize D; (2) fix parameters of D and optimize G. The overall objective function can be formulated as follows:

$$G^* = arg\ \min_G \max_D \mathscr{L}_{cGAN}(G, D) + \lambda\mathscr{L}_{L1}(G). \tag{4}$$

Here, the introduction of z enables it to match any distribution rather than just a delta function. As[40] described, dropout can also be interpreted as a way of regularizing a neural network by adding noise to its hidden units. Thus we replace the noise vector z with several dropout layers in G to achieve the same effect.

In addition, we also explore the effectiveness of each component in our objective function. Generators with different loss functions are defined as follows: *cGAN*: Generator G together with an adversarial discriminator conditioned on the input; *L*1: Generator G with an L1 loss. It is essentially equivalent to a traditional CNN architecture with least absolute deviation; *cGAN* + *L*1: Generator G with both an L1 loss term and an adversarial discriminator conditioned on the input.

Network architecture. Our cross-modality generation framework is composed of two main submodels, generator (G) and discriminator (D). It is similar to traditional GANs[11].

Generator. Although appearances of input and output images are different, their underlying structures are the same. Shared information (e.g. identical structures) needs to be transformed in the generative network. In this case, encoder-decoder networks with an equal number of down-sampling layers and up-sampling layers are proposed as one effective generative network[12,41–44]. However, it is a time-consuming process when all mutual information between input and output images (such as structures, edges and so on) flows through the entire network layer by layer. Besides, the network efficiency is limited due to the presence of a bottleneck layer which restricts information flow. Thus, skip connections are added between mirrored layers in the encoder-decoder network,

following the "U-Net" shape in[45]. These connections speed up information transmission since the bottleneck layer is ignored, and help to learn matching features for corresponding mirrored layers.

The architecture of G has 8 convolutional layers, each of which contains a convolution, a Batch Normalization, and a leaky ReLu activation[46] (a slope of 0.2) with numbers of filters at 64, 128, 256, 512, 512, 512, 512, and 512 respectively. Following them are 8 deconvolutional stages, each of which includes a deconvolution, a Batch Normalization, and an unleaky ReLu[46] (a slope of 0.2) with numbers of filters at 512, 1024, 1024, 1024, 1024, 512, 256, and 128 respectively. It ends with a tanh activation function.

Discriminator. GANs can generate images that are not only visually realistic but also quantitatively comparable to the real images. Therefore, an adversarial discriminator architecture is employed to confine the learning process of G. D identifies those generated outputs of G as false (label 0) and the real data as true (label 1), then providing feedback to G. PixelGANs[13] have poor performance on spatial sharpness, and ImageGANs[13] with many parameters are hard to train. In contrast, PatchGANs[13] enable sharp outputs with fewer parameters and less running time since PatchGANs have no constraints on the size of each patch. We thus adopt a PatchGAN classifier as our discriminator architecture. Unlike previous formulations[47,48] that regard the output space as unstructured, our discriminator penalizes structures at the scale of image patches. In this way, high-level information can be captured under the restriction of D, and low-level information can be ensured by an L1 term.

The architecture of D contains four stages of convolution-BatchNorm-ReLu with the kernel size of (4,4). The numbers of filters are 64, 128, 256, and 512 for convolutional layers. Lastly, a sigmoid function is used to output the confidence probability that the input data comes from real MR images rather than generated images.

Application

In this section, we choose cross-modality registration and segmentation from multiple applications as two examples to verify the effectiveness of our proposed framework. Details of our approaches and algorithms are discussed in the following subsections.

Cross-modality registration. The first application of our cross-modality generation framework is to use the translated modality for cross-modality image registration. Our method is inspired by an atlas-based registration, where the moving image is registered to the fixed image with a non-linear registration algorithm. Images after registration are called the warped images. Our method contains four steps: (1) We first build our fixed images space with only one modality images being given. We use T1 and T2 images as one example to illustrate our method. Given T2 images, our fixed images space can consist of T2 and $\hat{T}1$ images by using our cross-modality generation framework. The moving images space commonly consists of both T2 and T1 images from n subjects. (2) The second step is to register the moving images to the fixed images, constructing n corresponding atlases. Since multiple atlases encompass richer anatomical variability than a single atlas, we used multi-atlas-based rather than single-atlas-based registration approach. For any fixed subject, we register all n moving images to the fixed images and the deformation field that aligns the moving image with the fixed image can be automatically computed with a registration algorithm. As illustrated in Fig. 3, T2 images from the moving images space are registered to T2 images from the fixed images space and T1 images from the moving images space are registered to $\hat{T}1$ images from the fixed images space. (3) The deformations generated in (2) are combined in a weighted fusion process, where the cross-modality information can be adopted. We fuse the deformations generated from T2 registrations with deformations generated from $\hat{T}1$ registrations (see Fig. 3). (4) Applying the deformations to the atlas segmentation labels can yield n registered segmentation labels of fixed images. For any fixed subject, we obtain the final registration results by averaging the n registered labels of the fixed subject.

Among multiple registration algorithms, we select ANTs[35] and Elastix[36] to realize our method. Three stages of cross-modality registration are adopted via ANTs. The first two stages are modeled by rigid and affine transforms with mutual information. In the last stage, we use SyN with local cross-correlation, which is demonstrated to work well with cross-modality scenarios without normalizing the intensities[49]. For Elastix, affine and B-splines transforms are used to model the nonlinear deformations of the atlases. Mutual information is adopted as the cost function.

Cross-modality segmentation. We propose a new approach for MR image segmentation based on cross-modality images, namely translated multichannel segmentation (TMS). The main focus of TMS is the introduction of the translated-modality images obtained in our proposed framework, which enriches the cross-modality information without any extra data. TMS inputs two identical given-modality images and one corresponding translated-modality image into three separate channels which are conventionally used for RGB images. Three input images are then fed into FCN networks for improving segmentation results of given-modality images. Here, we employ the standard FCN-8s[9] as the CNN architecture of our segmentation framework because it can fuse multi-level information by combining feature maps of the final layer and last two pooling layers. Fig. 5 depicts the flowchart of our segmentation approach.

We denote our training dataset as $S = \{(x_i, \hat{y}_i, l_i),\ i = 1, 2, 3, \ldots, n\}$, where x_i refers to the ith given-modality image, $\hat{y}_i$ indicates the ith corresponding translated-modality image obtained in our proposed framework, and l_i represents the corresponding segmentation label. We denote the parameters of the FCN architecture as θ and the model is trained to seek optimal parameters θ^*. During testing, given an input image x, the segmentation output $\hat{l}$ is defined as below:

$$P(\hat{l} = k|x;\ \theta^*) = s_k(h(x,\ \theta^*)), \tag{5}$$

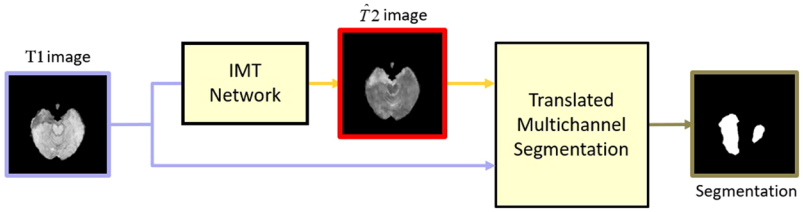

Figure 5. Flowchart of our approach for cross-modality segmentation. First, we input a given-modality image to our IMT network to generate a translated-modality image. For instance, given a T1 image, $\hat{T}2$ images can be generated with our method. Second, two identical given-modality images and one corresponding translated-modality image are fed to channels 1, 2, and 3 and segmented by FCN networks. Under the standard FCN-32s, standard FCN-16s, and standard FCN-8s settings, we output our segmentation results. The red box indicates our translated images.

where k denotes the total number of classes, $h(\cdot)$ denotes the feature map of the hidden layer, $s(\cdot)$ refers to the softmax function and s_k indicates the output of the kth class.

Experiments

In this section, we demonstrate the generalizability of our framework for MR image generation and apply it to cross-modality registration and segmentation. We first conduct a large number of experiments on five publicly available datasets for MR image generation (*BraTs2015, Iseg2017, MRBrain13, ADNI, RIRE*). Then we choose *Iseg2017* and *MRBrain13* for cross-modality registration. We finally choose *BraTs2015* and *Iseg2017* for cross-modality segmentation. Among these five MRI datasets, the *BraTs2015*, *Iseg2017*, and *MRBrain13* datasets provide ground truth segmentation labels.

Implementation details. All our models are trained on NVIDIA Tesla K80 GPUs. Our code will be publicly released upon acceptance.

Generation. We train the models on a torch7 framework[50] using Adam optimizer[51] with a momentum term $\beta 1 = 0.5$. The learning rate is set to 0.0002. The *batchsize* is set to 1 because our approach can be regarded as "instance normalization" when *batchsize* $= 1$ due to the use of batch normalization. As demonstrated in[52], instance normalization is effective at generation tasks by removing instance-specific information from the content image. Other parameters follow the reference[13]. All experiments use 70×70 PatchGANs.

Registration. A Windows release 2.1.0 version of ANTs[35] is used in our experiments. As for the Elastix[36], a Windows 64 bit release 4.8 version is adopted. All the registration experiments are run in a Microsoft High-Performance Computing cluster with 2 Quad-core Xeon 2.43 GHz CPU for each compute node. We choose the parameters by cross-validation. For ANTs, we use the parameters in[53]. For Elastix, we adopt the parameters in[54].

Segmentation. We implement standard FCN-8s on the MXNET toolbox[55]. A pre-trained VGG-16 model, a trained FCN-32s model, and a trained FCN-16s model are used for initialization of FCN-32s, FCN-16s, and FCN-8s respectively. The learning rate is set to 0.0001, with a momentum of 0.99 and a weight decay of 0.0005. Other parameters are set to the defaults in[9].

Cross-modality generation. *Evaluation metrics.* We report results on mean absolute error (MAE), peak signal-to-noise ratio (PSNR), mutual information (MI), Structural Similarity Index (SSIM) and FCN-score.

We follow the definition of MAE in[56]:

$$MAE = \frac{1}{m \times n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|\hat{y}(i, j) - y(i, j)\|, \tag{6}$$

where target-modality image y and translated-modality image $\hat{y}$ both have a size of $m \times n$ pixels, and (i, j) indicates the location of pixels.

PSNR[57] is defined as below:

$$PSNR = 10\log 10\frac{MAX^2}{MSE}, \tag{7}$$

where MAX is the maximum pixel value of two images and MSE is the mean square error between two images.

MI is used as a cross-modality similarity measure[58]. It is robust to variations in modalities and calculated as:

$$I(y; \hat{y}) = \sum_{m\in y}\sum_{n\in \hat{y}} p(m, n)\log\left(\frac{p(m, n)}{p(m)p(n)}\right), \tag{8}$$

where m, n are the intensities in target-modality image y and translated-modality image $\hat{y}$ respectively. $p(m, n)$ is the joint probability density of y and $\hat{y}$, while $p(m)$ and $p(n)$ are marginal densities.

SSIM[59] is defined as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{9}$$

where μ_x and μ_y denote the mean values of original and distorted images. σ_x and σ_y denote the standard deviation of original and distorted images, and σ_{xy} is the covariance of both images.

FCN-score is used to capture the joint statistics of data and evaluate synthesized images across the board. It includes accuracy and Dice. On one hand, accuracy consists of the mean accuracy of all pixels (denoted as "all" in the tables) and per-class accuracy (such as mean accuracy of tumors, gray matter, white matter, etc.). On the other hand, the Dice is defined as follows: $(2|H \cap G|)/(|H| + |G|)$ where G is the ground truth map and H is the prediction map.

Here, we follow the definitions of FCN-score in[13] and adopt a pre-trained FCN to evaluate our experiment results. The semantic segmentation task in essence is to label each pixel with its enclosing object or region class. Pre-trained semantic classifiers are used to measure the discriminability of the synthesized images as a fake-metric. If synthesized images are plausible, classifiers pre-trained on real images would classify synthesized images correctly as well. Take the transition from T1 to T2 for instance. T2 images (training data) are utilized to fine tune an FCN-8s model. Both T2 (test data/real data) and $\hat{T}2$ (synthesized data) images are subsequently segmented through the well-trained model. We score the segmentation (classification) accuracy of synthesized images against the real images. The gap of FCN-score between T2 images and $\hat{T}2$ images quantitatively evaluates the quality of $\hat{T}2$ images.

Datasets. The data preprocessing mainly contains three steps. (1) Label Generation: Labels of necrosis, edema, non-enhancing tumor, and enhancing tumor are merged into one label, collectively referred to as tumors. Labels of Grey Matter (gm) and White Matter (wm) remain the same. Thus, three types of labels are used for training: tumors, gm, and wm. (2) Dimension Reduction: We slice the original volumetric MRI data along the z-axis because our network currently only supports 2D input images. For example, the 3D data from BraTs2015 datasets, with a size of $240 \times 240 \times 155$ voxels (respectively representing the pixels of x-, y-, z-direction), is sliced to 2D data (155×220, 155 slices and 220 subjects). (3) Image Resizing and Scaling: All 2D images are then resized to a resolution of 256×256 pixels, after which we generate the 2D input images. Then the input images are scaled to [0.0, 1.0] and normalized with mean value of 0.5 and standard deviation of 0.5. So, all the input data are normalized in range $[-1.0, 1.0]$. Note that different modalities of the same subject from five brain MRI datasets that we choose are almost voxel-wise spatially aligned. We do not choose to coregister the data in our datasets since this is beyond the scope of our discussion. We respectively illustrate five publicly available datasets used for cross-modality MRI generation.

(1)*BraTs2015*: The BraTs2015 dataset[60] contains multi-contrast MR images from 220 subjects with high-grade glioma, including T1, T2, T2-Flair images and corresponding labels of tumors. We randomly select 176 subjects for training and the rest for testing. 1924 training images are trained for 600 epochs with batch size 1. 451 images are used for testing.

(2)*Iseg2017*: The Iseg2017 dataset[61] contains multi-contrast MR images from 23 infants, including T1, T2 images and corresponding labels of Grey Matter (gm) and White Matter (wm). We randomly select 18 subjects for training and remaining 5 subjects for testing. 661 training images are trained for 800 epochs with batch size 1. 163 images from the 5 subjects are used for testing.

(3)*MRBrain13*: The MRBrain13 dataset[62] contains multi-contrast MR images from 20 subjects, including T1 and T2-Flair images. We randomly choose 16 subjects for training and the remaining 4 for testing. 704 training images are trained for 1200 epochs with batch size 1. 176 images are used for testing.

(4)*ADNI*: The ADNI dataset[30] contains T2 and PD images (proton density images, tissues with a higher concentration or density of protons produce the strongest signals and appear the brightest on the image) from 50 subjects. 40 subjects are randomly selected for training and the remaining 10 for testing. 1795 training images are trained for 400 epochs with batch size 1. 455 images are used for testing.

(5)*RIRE*: The RIRE dataset[63] includes T1 and T2 images collected from 19 subjects. We randomly choose 16 subjects as for training and the rest for testing. 477 training images are trained for 800 epochs with batch size 1. 156 images are used for testing.

In this study, we have adopted datasets which represent typical training data sizes in medical imaging problems. For example, the RIRE and MRBrain13 datasets. Whether these datasets are sufficient remains an open question. In theory, the more training data, the better performance. However, being sufficient is usually an application-dependent measure. The generator in our framework used a modified U-Net architecture. In the original U-Net paper[45], it was trained on ISBI cell tracking challenge datasets of "PhC-U373" and "DIC-HeLa", which contain 35 and 20 images separately for training with partial annotation. This means that although the size of training set might not enough for the network to reach the best performance, it is still possible for it to learn useful features, satisfying application needs. For testing purposes, on one hand, compare to the image translation experiment[13] using CMP Facades dataset (train images: 400, test images: 100) and the ADNI dataset for MRI to CT translation (train + test subjects: 16) in[30], all of the datasets used in our paper contain more images than those. On the other hand, we included p-values of the t-test to show the statistical significance for the experiments of image generation and registration. These indicate that the test data size is sufficient to support the conclusions in this study.

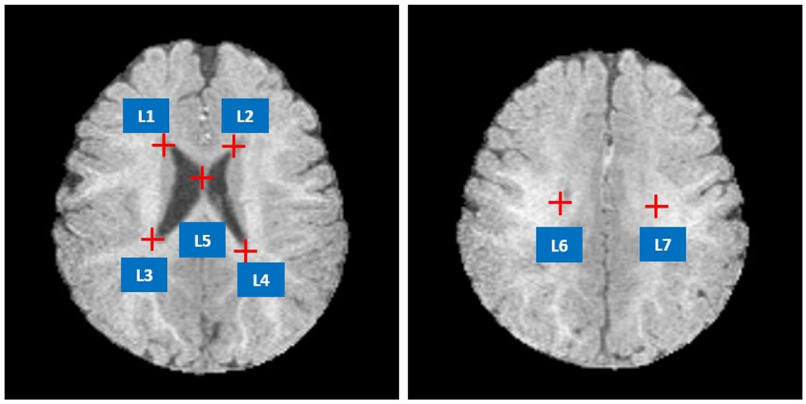

Figure 6. Illustration of the seven landmarks selected for cross-modality registration. L1: right lateral ventricle superior, L2: left lateral ventricle superior, L3: right lateral ventricle inferior, L4: left lateral ventricle inferior. L5: middle of the lateral ventricle, L6: right lateral ventricle posterior, L7: left lateral ventricle posterior.

Datasets	Transitions	RF				CA-GAN				IMT											
										cGAN + L1				cGAN				L1			
		MAE ↓	PSNR ↑	MI ↑	SSIM ↑	MAE ↓	PSNR ↑	MI ↑	SSIM ↑	MAE ↓	PSNR ↑	MI ↑	SSIM ↑	MAE ↓	PSNR ↑	MI ↑	SSIM ↑	MAE ↓	PSNR ↑	MI ↑	SSIM ↑
BraTs2015	T1 → T2	**6.025**	**24.717**	0.617	**0.910**	11.947	19.738	0.787	0.826	8.292	22.560	0.862	0.866	10.692	20.301	0.788	0.575	8.654	22.517	**0.901**	0.880
	T2 → T1	**7.921**	**23.385**	0.589	**0.893**	16.587	17.462	0.661	0.723	9.937	22.518	0.777	0.854	15.430	18.507	0.673	0.723	10.457	22.374	**0.818**	**0.896**
	T1 → T2-Flair	**8.176**	**23.222**	0.609	**0.873**	13.999	19.157	0.722	0.756	**7.934**	**22.687**	0.833	0.837	11.671	19.969	0.749	0.797	8.462	22.642	**0.879**	0.857
	T2 → T2-Flair	**7.318**	**23.138**	0.610	**0.875**	12.658	18.848	0.756	0.749	8.858	21.664	0.848	0.836	10.469	20.656	0.817	0.823	8.950	21.791	**0.928**	0.860
Iseg2017	T1 → T2	3.955	28.028	0.803	0.902	12.175	21.992	0.804	0.690	**3.309**	**29.979**	0.931	0.887	8.028	22.860	0.782	0.748	3.860	28.874	**0.993**	**0.913**
	T2 → T1	11.466	22.342	0.788	**0.808**	17.151	18.401	0.789	0.662	**9.586**	**23.610**	**0.868**	0.745	17.311	18.121	0.777	0.620	10.591	**23.325**	**0.880**	0.754
MRBrain13	T1 → T2-Flair	7.609	24.780	1.123	0.863	13.643	19.503	0.805	0.782	**6.064**	**26.495**	1.066	0.823	9.906	22.616	1.009	0.785	6.505	**26.299**	**1.185**	**0.881**
ADNI	PD → T2	9.485	24.006	1.452	0.819	16.575	19.008	0.674	0.728	6.757	26.477	1.266	0.812	7.211	26.330	1.184	0.779	**4.898**	**29.089**	**1.484**	**0.891**
	T2 → PD	5.856	29.118	1.515	**0.880**	17.648	18.715	0.659	0.713	**4.590**	**31.014**	1.381	0.856	5.336	29.032	1.282	0.820	5.055	30.614	**1.536**	**0.881**
RIRE	T1 → T2	38.047	12.862	0.694	0.501	18.625	18.248	**0.724**	0.749	**5.250**	**28.994**	0.636	0.736	13.690	21.038	0.513	0.506	9.105	**28.951**	**0.698**	**0.760**
	T2 → T1	17.022	19.811	0.944	0.622	23.374	16.029	0.650	0.728	**9.035**	**24.043**	0.916	0.692	13.964	20.450	0.737	0.538	9.105	**24.003**	**0.969**	**0.741**

Table 1. Generation performance on five publicly available datasets evaluated by MAE, PSNR, MI, and SSIM. The bold entries in this table indicate the algorithm which gets the best performance in each task. The standard for choosing the best algorithm is to have statistical significance over the other algorithms (p-value < 0.05). If an algorithm gets the best evaluation metrics but has no statistical significance over the others (p-value > 0.05), all of them will be regarded as the best algorithms. The result show that our IMT approach outperforms both Random Forest (RF) based method[5] and Context-Aware GAN (CA-GAN)[30] method on most datasets.

Cross-modality registration. *Evaluation metric.* We use the two evaluation metrics for cross-modality registration, namely Dice and Distance Between Corresponding Landmarks (Dist).

(1)*Dice*: The first metric is introduced to measure the overlap of ground truth segmentation labels and registered segmentation labels. It is defined as $(2|H \cap G|)/(|H| + |G|)$ where G is the ground truth segmentation label of the fixed image and H is the registered segmentation label of the fixed image. Since image registration involves identification of a transformation to fit a fixed image to a moving image. The success of the registration process is vital for correct interpretation of many medical image-processing applications, including multi-atlas segmentation. A higher Dice, which measures the overlap of propagated segmentation labels through deformation and the ground truth labels, indicates a more accurate registration.

(2)*Distance Between Corresponding Landmarks (Dist)*: The second metric is adopted to measure the capacity of algorithms to register the brain structures. The registration error on a pair of images is defined as the average Euclidean distance between a landmark in the warped image and its corresponding landmark in the fixed image. To compute the Euclidean distance, all 2D-slices after registration are stacked into 3D images.

Dataset. We preprocess the original MRI data from *Iseg2017* and *MRBrain13* datasets. We didn't change the size of the images from the MRBrain13 dataset. The preprocessing is only to slice the 3D images from all the subjects into 2D images along the z-axis. The Iseg2017 dataset contains MR images of infant brains have much smaller fields-of-view, it was preprocessed with the following steps for considerations in computational efficiency. (1) We first crop the 3D image into a smaller cube, each side of which circumscribes the brain. (2) The brain cubes are

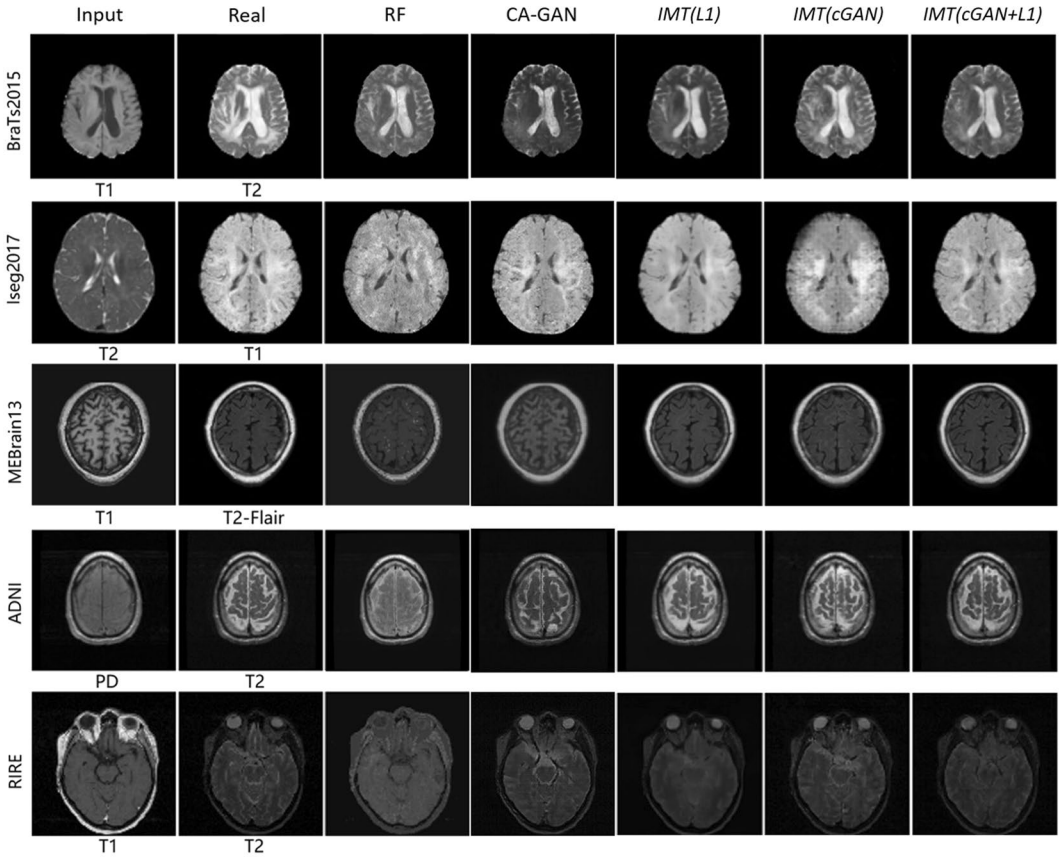

Figure 7. Samples of cross-modality generation results on five publicly available datasets including *BraTs2015*[60], *Iseg2017*[61], *MRBrain13*[62], *ADNI*[30], and *RIRE*[63]. Results are selected from top performing examples (relatively low MAE, high PSNR, high MI, and high PSNR collectively) with four approaches. The right five columns show results of the random-forests-based method (RF)[5], the Context-Aware GAN (CA-GAN)[30] and IMT framework with different loss functions ($L1$, $cGAN$, $cGAN + L1$).

	Accuracy		Dice
Method	**all**	**tumor**	**tumor**
T1 → T2	0.955	0.716	0.757
T2 (real)	0.965	0.689	0.724
T2 → T1	0.958	0.663	0.762
T1 (real)	0.972	0.750	0.787
T1 → T2-Flair	0.945	0.729	0.767
T2 → T2-Flair	0.966	0.816	0.830
T2-Flair (real)	0.986	0.876	0.899

Table 2. Segmentation results of IMT images on *BraTs2015* evaluated by FCN-score. The gap between translated images and the real images can evaluate the generation performance of our method. Note that "all" represents mean accuracy of all pixels (the meanings of "all" are the same in the following tables). We achieve close segmentation results between translated-modality images and target-modality images.

resized to a size of 128 × 128 × 128 voxels, without significant down-sampling or information lost. (3) The last step is to slice the brain cubes from all the subjects into 2D data along the z-axis (128 × 128, 128 slices).

After preprocessing, the brain slices with the same depth value from different subjects are spatially aligned. During the training phase, a pair of brain slices from two different subjects with the same depth value is treated as a pair moving and fixed images. In order to conduct five-fold cross-validation for our experiments, the value of n (numbers of atlases) is selected differently in each dataset. For *Iseg2017* dataset, we choose 8 subjects in the moving images space and another 2 subjects in the fixed images space ($n = 8$). For *MRBrain13* dataset, 4 subjects are selected for the moving images space while one subject in the fixed images space ($n = 4$).

Method	Accuracy			Dice	
	all	gm	wm	gm	wm
T1 → T2	0.892	0.827	0.506	0.777	0.573
T2 (real)	0.920	0.829	0.610	0.794	0.646
T2 → T1	0.882	0.722	0.513	0.743	0.569
T1 (real)	0.938	0.811	0.663	0.797	0.665

Table 3. Segmentation results of IMT translated images on *Iseg2017* evaluated by FCN-score. Note that "gm" and "wm" indicate gray matter and white matter respectively. The minor gap between translated-modality images and the target-modality images shows decent generation performance of our framework.

Datasets	Modalities	Structures	Dice		Dist	
			ANTs	Elastix	ANTs	Elastix
Iseg2017	T2	wm	0.508	0.475	2.105	2.836
		gm	0.635	0.591		
	$\hat{T}1$	wm	0.503	0.469	1.884	2.792
		gm	0.622	0.580		
	T2 + $\hat{T}1$	wm	**0.530**	**0.519**	**1.062**	**2.447**
		gm	**0.657**	**0.648**		
	T1	wm	0.529	0.500	1.136	2.469
		gm	0.650	0.607		
	$\hat{T}2$	wm	0.495	0.457	2.376	3.292
		gm	0.617	0.573		
	T1 + $\hat{T}2$	wm	**0.538**	**0.527**	**1.097**	**2.116**
		gm	**0.664**	**0.650**		
	T1 + T2	wm	0.540	0.528	1.013	2.109
		gm	0.666	0.651		
MRBrain13	T2-Flair	wm	0.431	0.412	3.417	3.642
		gm	0.494	0.463		
	$\hat{T}1$	wm	0.468	0.508	3.159	3.216
		gm	0.508	0.487		
	T2-Flair + $\hat{T}1$	wm	**0.473**	**0.492**	**2.216**	**2.659**
		gm	**0.530**	**0.532**		
	T1	wm	0.484	0.534	2.524	2.961
		gm	0.517	0.510		
	$\hat{T}2$-*Flair*	wm	0.431	0.410	3.568	3.726
		gm	0.497	0.458		
	T1 + $\hat{T}2$ -*Flair*	wm	**0.486**	**0.505**	**2.113**	**2.556**
		gm	**0.534**	**0.540**		
	T2-Flair + T1	wm	0.486	0.503	2.098	2.508
		gm	0.534	0.539		

Table 4. Registration results evaluated by Dist and Dice on *Iseg2017* and *MRBrain13*. The bold entries indicate the experiments which used the combination of the real and the translated images in another modality generated by the real images.

Iseg2017 and *MRBrain13* datasets provide ground truth segmentation labels. Seven well-defined anatomic landmarks (see Fig. 6) that are distributed in the lateral ventricle are manually annotated by three doctors. We consider the average coordinates from three doctors as the ground truth positions of the landmarks.

Cross-modality segmentation. *Evaluation metric.* We report segmentation results on Dice (higher is better).

Dataset. The original training set is divided into *PartA* and *PartB* at the ratio of 1:1 based on the subjects. The original test set maintains the same (denoted as *PartC*). *PartA* is used to train the generator. *PartB* is then used to infer the translated modality. *PartB* is then used to train the segmentation model, which is tested on *PartC*.

(1)*Brats2015*: The original *Brats2015* dataset contains 1924 images (*PartA*: 945, *PartB*: 979) for training and 451 images (*PartC*) for testing. After preprocessing, 979 images are trained for 400 epochs and 451 images are used for testing.

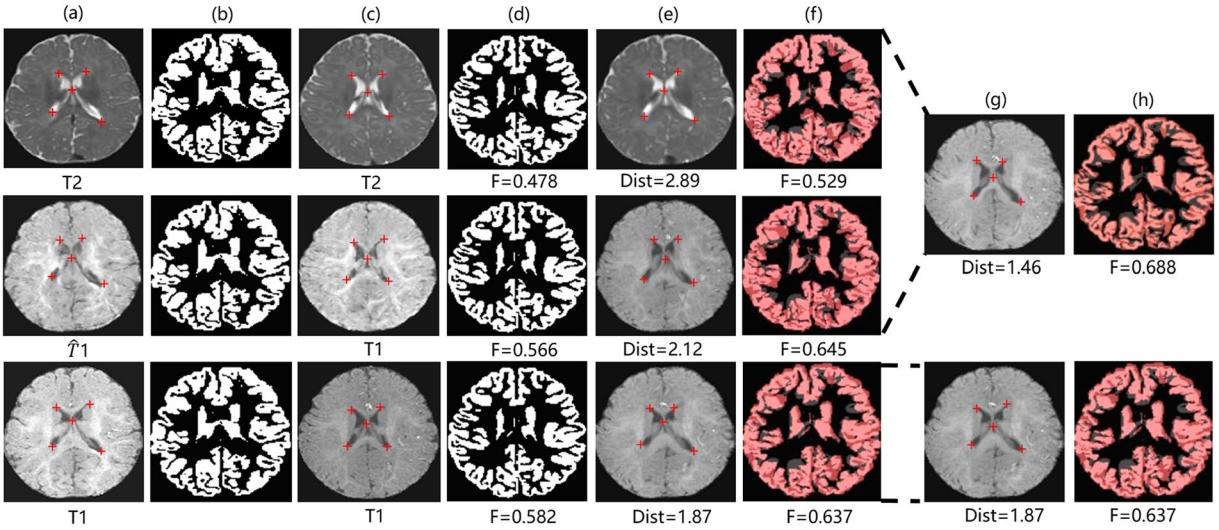

Figure 8. Samples of registration results of our method: (**a**) Fixed image, (**b**) Ground truth segmentation label of fixed image, (**c**) Moving image, (**d**) Ground truth segmentation label of moving image, (**e**) Warped image (moving image warped by the best traditional registration algorithm (ANTs), (**f**) Warped ground truth segmentation label of moving image, (**g**) Fused image, (**h**) Segmentation prediction of fused image. The pink, dark red, grey areas in (**f**) denote true regions, false regions, and missing regions respectively. The red crosses denote landmarks in the fixed and moving images.

(2)*Iseg2017*: The original *Iseg2017* dataset contains 661 images (*PartA*: 328, *PartB*:333) for training and 163 images (*PartC*) for testing. After preprocessing, 333 images are trained for 800 epochs and 163 images remain for testing.

Results

Cross-modality generation. Generation performance with different methods on the five datasets are summarized in Table 1. It quantitatively shows how using IMT network allows us to achieve better generation results than the regression-based method using RF[5] and the latest proposed Context-Aware GAN method from[30] on most datasets evaluated by MAE, PSNR, MI, and SSIM. However, there are also some cases where the RF method surpasses our IMT network on the *BraTs2015* dataset (images with tumors). It is explicable since the RF method incorporates additional context features, taking full advantages of structural information and thus leading to comparable generation results on images with tumors. The method of CA-GAN utilized the spatial information of the images as well, but the results show that a series of metrics have not been improved. Although a fair comparison is difficult between a 2D and 3D networks directly, according to[64], the results provides a preliminary evidence that including 3D spatial information may not necessarily improve the predicative performance for the applications of interest. For example, the authors designed a network which could incorporate 3D spatial information by taking one or more Transrectal Ultrasound slices neighboring each slice to be segmented as input. However, it did not improve the segmentation performance in most of their experiment results. In addition, our task is more difficult compare to[30] since the MR image is superior in the detail of the image while the CT image has relatively low soft tissue contrast. Considering the structure complexity of the CA-GAN and the increased task difficulty, we believe it is the widely-observed difficulties in training generative adversarial networks, as reported in[11], which diminished the potential benefit in using full 3D spatial information.

Note that different losses induce different quality of generated images. In most cases, our IMT network with $cGAN + L1$ achieves the best results on MAE and PSNR; $L1$ loss term contributes to superior performance on MI over other methods. MI focuses more attention on the matching of pixel-wise intensities and ignores structural information in the images. Meanwhile, the L1 loss term ensures pixel-wise information rather than the properties of human visual perception[65]. Thus, it is reasonable that using L1 term contributes to superior results on MI.

Figure 7 shows the qualitative results of cross-modality image generation using different approaches on five datasets. We have reasonable but blurry results using IMT network with $L1$ alone. The IMT network with $cGAN$ alone leads to improvements in visual performance but causes some artifacts in cross-modality MR image generation. Using $cGAN + L1$ terms obtains sharp and realistic images and reduces artifacts. In contrast, the RF method and Context-Aware GAN lead to rough and fuzzy results compared with IMT networks.

We also quantify the generation results using FCN-score on *BraTs2015* and *Iseg2017* in Table 2 and Table 3. Our approach ($cGAN + L1$) is effective in generating realistic cross-modality MR images towards the real images. The cGAN-based objectives lead to high scores close to the real images.

To validate the perceptual realism of our generated images, two more experiments are conducted. One is conducted by three radiologists. The other is done by five well-trained medical students. For the first experiment, we randomly select 1100 pairs of images, each of which consists of an image generated by our framework and its corresponding real image. On each trial, three radiologists are respectively asked to select which one is fake in the

Datasets	Modalities	Structures	Dice	Dist
Iseg2017	T2	wm	0.823	0.475
		gm	0.859	
	$\hat{T}1$	wm	0.882	0.183
		gm	0.910	
	T2 + $\hat{T}1$	wm	**0.883**	**0.190**
		gm	**0.857**	
	T1	wm	0.868	0.179
		gm	0.898	
	$\hat{T}2$	wm	0.807	0.218
		gm	0.846	
	T1 + $\hat{T}2$	wm	**0.868**	**0.186**
		gm	**0.898**	
	T1 + T2	wm	0.868	0.184
		gm	0.898	
MRBrain13	T2-Flair	wm	0.976	0.182
		gm	0.976	
	$\hat{T}1$	wm	0.966	0.181
		gm	0.968	
	T2-Flair + $\hat{T}1$	wm	**0.971**	**0.180**
		gm	**0.974**	
	T1	wm	0.976	0.179
		gm	0.981	
	$\hat{T}2$-Flair	wm	0.985	0.180
		gm	0.983	
	T1 + $\hat{T}2$-Flair	wm	**0.985**	**0.179**
		gm	**0.985**	
	T2-Flair + T1	wm	0.978	0.178
		gm	0.982	

Table 5. Results of our additional registration experiments evaluated by Dist and Dice on *Iseg2017* and *MRBrain13* implemented by ANTS. The bold entries indicate the experiments which used the combination of the real and the translated images in another modality generated by the real images.

	Dice(tumor)	Δ
T1	0.760	—
T1 + $\hat{T}2$	**0.808**	**6.32%**
T1 + T2	0.857	—
T1 + $\hat{T}2$-Flair	**0.819**	**7.89%**
T1 + T2-Flair	0.892	—

Table 6. Tumor segmentation results of TMS on *Brats2015*. "T1 + $\hat{T}2$" and "T1 + $\hat{T}2$-*Flair*" in bold font indicate our approach (TMS) where inputs are both T1 and $\hat{T}2$ images or T1 and $\hat{T}2$-*Flair* images. "T1" indicates the traditional FCN method where inputs are only T1 images. "T1 + T2" and "T1 + T2-Flair" indicate the upper bound. Δ indicates the increment between TMS and the the traditional FCN method.

	Dice(wm)	Δ	Dice(gm)	Δ
T2	0.649	—	0.767	—
T2 + $\hat{T}1$	**0.669**	**3.08%**	**0.783**	**2.09%**
T2 + T1	0.691	—	0.797	—

Table 7. Brain structure segmentation results of TMS on *Iseg2017*. "T2 + $\hat{T}1$" in bold font indicates our method (TMS) where inputs are both T2 and $\hat{T}1$ images. "T2" indicates the traditional FCN method where inputs are only T2 images. "T2 + T1" indicates the upper bound.

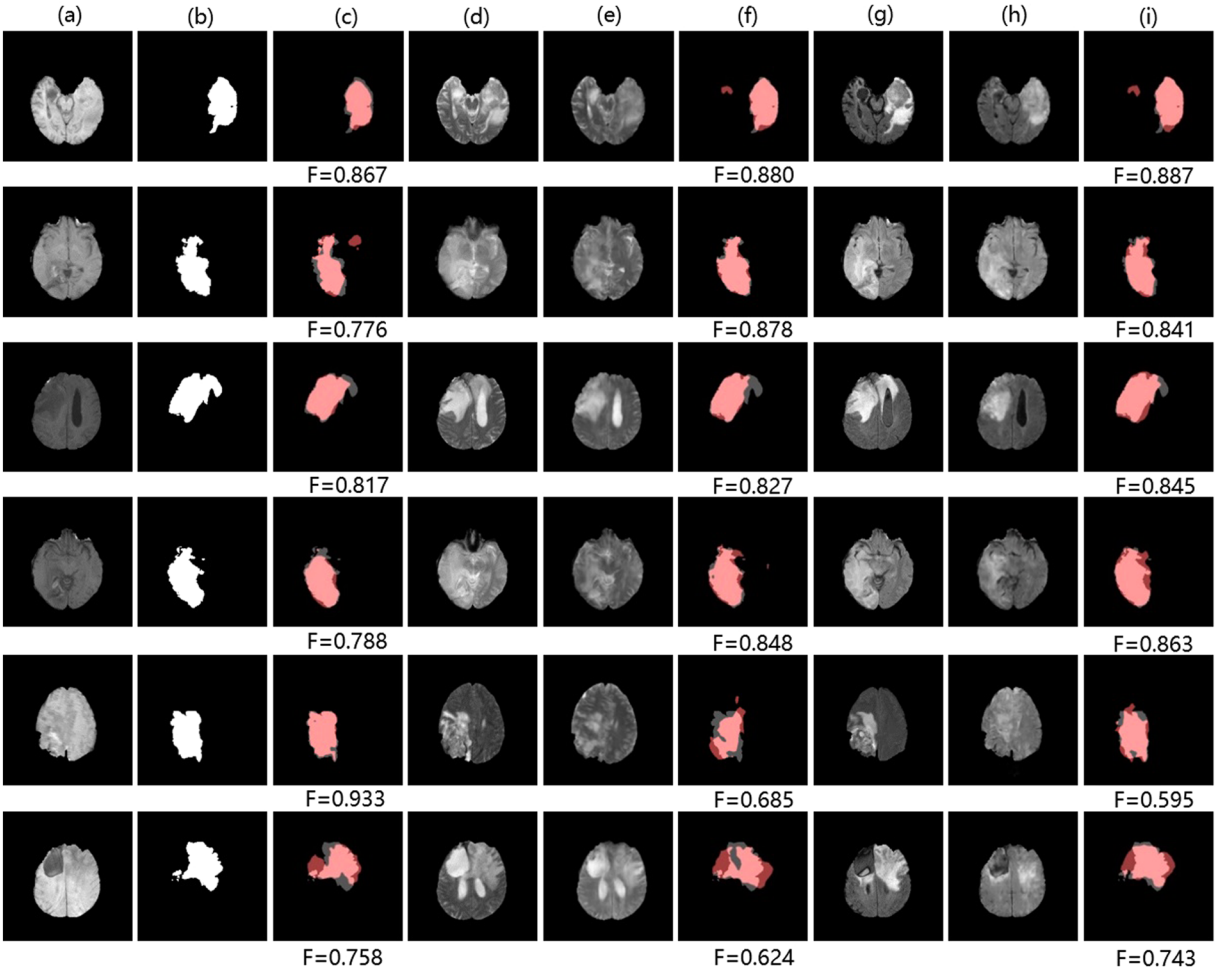

Figure 9. Samples of tumor segmentation results on *BraTs2015*: (**a,d,e,g,h**) denote T1 image, T2 image, $\hat{T}2$ image, T2-Flair image, $\hat{T}2$-*Flair* image. (**b**) Denotes ground truth segmentation label of T1 image. (**c,f,i**) Denote tumor segmentation results of T1 image using the FCN method, TMS (adding cross-modality information from $\hat{T}2$ image), TMS (adding cross-modality information from $\hat{T}2$-*Flair* image). Pink: true regions. Grey: missing regions. Dark red: false regions.

image pair. The first 100 trials are practice after which they are given feedback. The following 1000 trials are the main experiment where no feedback are given. The average performance of the three radiologists quantitatively evaluates the perceptual realism of our approach. For the second experiment, the experimental setting is perfectly identical. Results indicate that our generated images fooled radiologists on 25% trials and fooled students on 27.6% trials.

Cross-modality registration. Our experiments not only include registration with real data, but also with translated images ($\hat{T}1$ and $\hat{T}2$ images for *Iseg2017* dataset, $\hat{T}1$ and $\hat{T}2$-*Flair* images for *MRBrain13* dataset). The deformations generated in each set of experiments are combined in a weighted fusion process, yielding the final registration deformation. In order to compute the Euclidean distance of those corresponding landmarks between warped images and fixed images, all 2D-slices are then stacked into 3D images. Besides, we also employ the fused deformation to segmentation labels of moving images, obtaining registered segmentation results of fixed images. Table 4 summarizes the registration results both in terms of Dist and Dice. We introduce the cross-modality information from our $\hat{T}1$ images into T2 images and T2-Flair images, of which the performance are denoted as "T2 + $\hat{T}1$" and "T2-Flair + $\hat{T}1$". Likewise, "T1 + $\hat{T}2$" and "T1 + $\hat{T}2$-*Flair*" indicate performance of registrations with cross-modality information from our $\hat{T}2$-*Flair* images added into T1 images. We also show the upper bounds of registrations with translated images, which are denoted as "T1 + T2" and "T2-Flair + T1". The weights for the combination are determined through five-fold cross-validation. The optimal weights of 0.92 and 0.69 are selected for $\hat{T}1$ images in terms of white matter and gray matter on *Iseg2017* and 0.99 and 0.82 are selected on *MRBrain13*.

After the weighted fusion process, we find that registrations with translated images show better performance than those with real data by achieving higher Dice, e.g. 0.657 (T2 + $\hat{T}1$) vs. 0.635 (T2) and 0.534 (T1 + $\hat{T}2$-*Flair*) vs. 0.517 (T1), both got p-value < 0.0001 on t-test. We also observe that the Dist is greatly shortening (e.g. 2.216 (T2-Flair + $\hat{T}1$) vs. 3.417(T2-Flair), p-value < 0.0001 on t-test) compared to registrations without adding

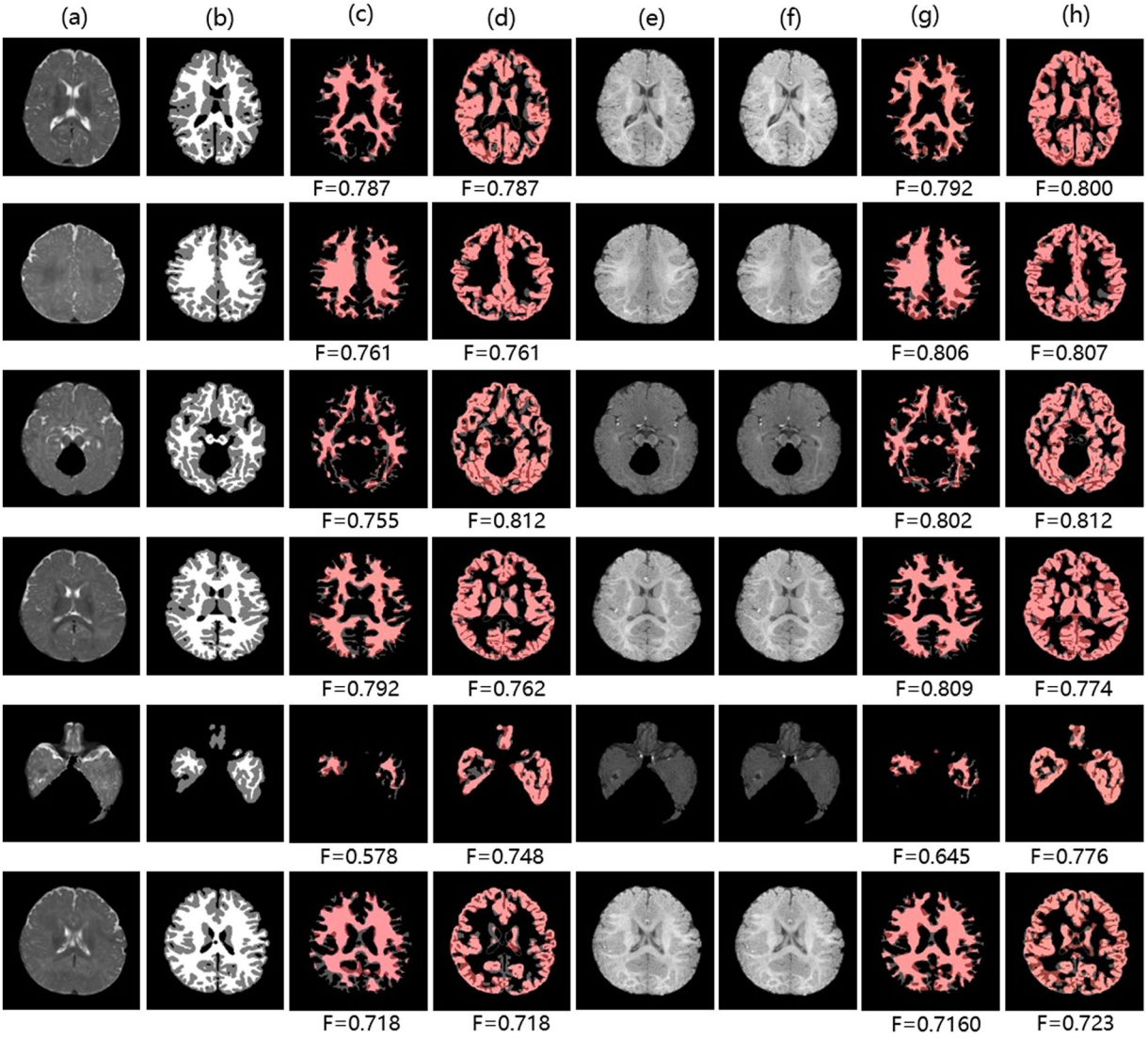

Figure 10. Samples of brain structure segmentation results on *Iseg2017*: (**a,e,f**) denote T2 image, T1 image, $\hat{T}1$ image. (**b**) Denotes ground truth segmentation label of T2 image. (**c,d**) Denote white matter and gray matter segmentation results of T2 image using the FCN method respectively. (**g,h**) Denote white matter and gray matter segmentation results of T2 image using TMS (adding cross-modality information from $\hat{T}1$ image) respectively. Pink: true regions. Grey: missing regions. Dark red: false regions.

cross-modality information. In many cases, our method even advances the upper bound both in Dist and Dice. These results are reasonable because our translated images are realistic enough, as well as the real data itself with high contrast for brain structure leads to lower registration errors. Figure 8 visualizes samples of the registration results of our methods. More details can be found there.

To demonstrate the effectiveness of our cross-modality registration approach with translated images, we propose an additional experiment by employing a known transformation to the moving images to generate transformed images that can be used as our "fixed". This allows us to directly estimate the benefit of adding translated modalities to the registration process when finding the known transformation during the registration step. Take T1 and T2 images as one example. The T1 and T2 images from the moving images space are first rotated a certain degree. Here we rotate them by 30 degrees. The $\hat{T}1$ images generated from our framework are also rotated 30 degrees. All these rotated images are used as our "fixed". T2 (moving) images are registered to rotated T2 (fixed) images and T1 (moving) images are registered to rotated $\hat{T}1$ (fixed) images. The following fusion processes are the same as our stated method. Table 5 shows the results of our additional experiments.

Cross-modality segmentation. Our experiments focus on two types of MRI brain segmentation: tumor segmentation and brain structure segmentation. Among all MRI modalities, some modalities are conducive to locating tumors (e.g. T2 and T2-Flair) and some are utilized for observing brain structures (e.g. T1) like white matter and gray matter. To this point, we choose to add cross-modality information from T2 and T2-Flair images

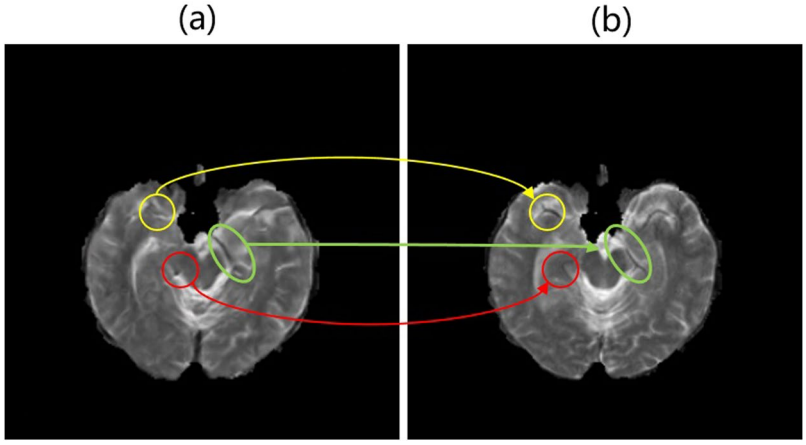

Figure 11. An abortive sample in our generation results: (**a**) $\hat{T}2$. (**b**) T2. Circles in $\hat{T}2$ indicate some misdescription of tiny structures. Circles in different colors indicate different problems.

into T1 images for tumor segmentation and add cross-modality information from T1 images into T2 images for brain structure segmentation. Experiments of tumor segmentation are conducted on *Brats2015* and experiments of brain structure segmentation are conducted on *Iseg2017*.

As shown in Tables 6, cross-modality information from $\hat{T}2$-*Flair* and $\hat{T}2$ images contributes improvements to tumor segmentation of T1 images (7.89% and 6.32% of tumors respectively). Likewise, Table 7 shows that cross-modality information from $\hat{T}1$ images leads to improvements of wm and gm segmentation of T2 images (3.08% of wm and 2.09% of gm). We also add cross-modality information from real modalities to make an upper bound. We observe a minor gap between results of TMS and the upper bound though our translated modalities are very close to real modalities. It is explicable by the presence of abnormal tissue anatomy (eg. tumors) and the cortex in MR images. The tumors are diffuse and even a small difference in the overlap can cause a low value for the Dice. In addition, in some finer cortex regions (unlike large homogeneous gray matter and white matter), our approach may produce some relatively coarse images, leading to a lower Dice. Moreover, our approach aims to help incorporate extra cross-modality information for more accurate segmentation rather than replace the real images. Overall, TMS outperforms the traditional FCN method when favorable cross-modality information is adopted. Figures 9 and 10 visualize some samples of our segmentation results on *BraTs2015* and *Iseg2017* respectively.

Discussion

We have described a new approach for cross-modality MR image generation using IMT network. Experimental results in section Experiments have highlighted the capability of our proposed approach to handle complex cross-modality generation tasks. The rationales are as follows. First, the cGAN rather than GAN network is essentially conceived of as a supervised network. It not only pursues realistic looking images, but also penalizes the mismatch between input and output so as to produce grounded enough real images. Second, the L1 term, which introduces pixel-wise regularization constraints into our generation task, guarantees the quantifications of low-level textures. Besides, we also described registration and segmentation applications of generated images. Both given-modality images and generated translated-modality images are used together to provide enough contrast information to differentiate different tissues and tumors, contributing to improvements for MR images registration and segmentation.

Although our approach generally achieves excellent performance, we recognize that in some cases our generated images are still not as good as real images at tiny structures. As illustrated in Fig. 11, there are also abortive cases where tiny structures may be mistaken. In the yellow circle, the eyebrow-like structure is missing. The red circle indicates a non-existent round structure which might be confounded with the vessel. In the green circle, the learned structure seems to be discontinuous which might give rise to perplexity for radiologists to make a diagnosis. In the future, we will improve our algorithm to describe more tiny structures.

Conclusion

In this paper, we develop a conditional generative adversarial network based framework for cross-modality translation and provided a comprehensive comparison with five datasets representing real-world clinical applications, each has its unique characteristics in data size, patient cohort and disease status. Important algorithmic options such as different loss functions were compared with traditional non-deep-learning machine learning methods, also with each other. We also have reported the performance in using the proposed methods for the difference in downstream tasks, registration and segmentation, arguably representing a more clinically relevant value from the proposed methodology. Our methods lead to comparable results in cross-modality generation, registration and segmentation on widely adopted MRI datasets without adding any extra data on the premise of only one modality image being given. In the future, we will make efforts in cross-modality translation tasks beyond MRI, such as from MRI to PET.

Received: 9 May 2019; Accepted: 12 February 2020;
Published online: 28 February 2020

References

1. Tseng, K. L., Lin, Y. L., Hsu, W. & Huang, C. Y. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In *CVPR, 2017*, 3739–3746 (2017).
2. Rzedzian, R. *et al*. Real-time nuclear magnetic resonance clinical imaging in paediatrics. *Lancet* **2**, 1281–1282 (1983).
3. Tsao, J. Ultrafast imaging: principles, pitfalls, solutions, and applications. *J. Magn. Reson. Imag.* **32**, 252–266 (2010).
4. Vemulapalli, R., Nguyen, H. V. and Zhou, S. K. Unsupervised cross-modal synthesis of subject-specific scans. In *ICCV, 2016*, 630–638 (2016).
5. Jog, A., Roy, S., Carass, A. & Prince, J. L. Magnetic resonance image synthesis through patch regression. In *Proc. IEEE Int. Symp. Biomed. Imaging*, 350–353 (2013).
6. Xie, S. & Tu, Z. Holistically-nested edge detection. *ICCV, 2015* 1–16 (2015).
7. Lee, C. Y., Xie, S., Gallagher, P., Zhang, Z. & Tu, Z. Deeply-supervised nets. *Artif. Intell*. 562–570 (2014).
8. Xu, Y. *et al*. Gland instance segmentation using deep multichannel neural networks. *IEEE Trans. Biomed. Eng.* **64**, 2901–2912 (2017).
9. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *CVPR, 2015*, 3431–3440 (2015).
10. Zhang, R., Isola, P. & Efros, A. A. Colorful image colorization. In *ECCV*, 649–666 (Springer, 2016).
11. Goodfellow, I. *et al*. Generative adversarial nets. In *NIPS, 2014*, 2672–2680 (2014).
12. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A. Context encoders: Feature learning by inpainting. In *CVPR, 2016*, 2536–2544 (2016).
13. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *CVPR, 2017*, 5967–5976 (2017).
14. Zhang, H. *et al*. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV, 2017*, 5907-5915 (2017).
15. Tu, Z. Learning generative models via discriminative approaches. In *CVPR, 2007*, 1–8 (2007).
16. Lazarow, J., Jin, L. & Tu, Z. Introspective neural networks for generative modeling. *ICCV, 2017* 5907–5915 (2017).
17. Lazarow, J., Jin, L. & Tu, Z. Introspective neural networks for generative modeling. In *CVPR, 2017*, 2774–2783 (2017).
18. Roy, S., Carass, A. & Prince, J. Magnetic resonance image example based contrast synthesis. *IEEE Trans. Med. Imaging* **32**, 2348–2363 (2013).
19. Iglesias, J. E. *et al*. Is synthesizing mri contrast useful for inter-modality analysis? In *MICCAI, 2013*, 631–638 (2013).
20. Eugenio, I. J., Rory, S. M. & Van, L. K. A unified framework for cross-modality multi-atlas segmentation of brain mri. *Med. Image Anal.* **17**, 1181–1191 (2013).
21. Balafar, M. A., Ramli, A. R., Saripan, M. I. & Mashohor, S. Review of brain mri image segmentation methods. *Artif. Intell. Rev.* **33**, 261–274 (2010).
22. Sasirekha, N. & Kashwan, K. Improved segmentation of mri brain images by denoising and contrast enhancement. *Indian J. Sci. Technol.* **8**, 1–7 (2015).
23. Freeman, W. T. & Pasztor, E. C. Learning low-level vision. *Int. J. Comput. Vision* **40**, 25–47 (2000).
24. Van Nguyen, H., Zhou, K. & Vemulapalli, R. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In *MICCAI, 2015*, 677–684 (2015).
25. Miller, M. I., Christensen, G. E., Amit, Y. & Grenander, U. Mathematical textbook of deformable neuroanatomies. *Proc. Acad. Nat. Sci. Phila* **90**, 11944–11948 (1993).
26. Rousseau, F. Brain hallucination. In *ECCV, 2008*, 497–508 (2008).
27. Huang, Y., Shao, L. & Frangi, A. F. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In *CVPR, 2017*, 5787–5796 (2017).
28. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
29. Mirza, M. & Osindero, S. Conditional generative adversarial nets. In *ICLR, 2014*, 2672–2680 (2014).
30. Nie, D., Trullo, R., Petitjean, C., Ruan, S. & Shen, D. Medical image synthesis with context-aware generative adversarial networks. In *MICCAI, 2017*, 417-425 (2017).
31. Wolterink, J. M., Leiner, T., Viergever, M. A. & Isgum, I. Generative adversarial networks for noise reduction in low-dose ct. *IEEE Trans. Med. Imaging* **36**, 2536–2545 (2017).
32. Viola, P. & Wells, W. Alignment by maximization of mutual information. *Int. J. Comput. Vision* **24**, 137–154 (1997).
33. Penney, G. P. *et al*. A comparison of similarity measures for use in 2-d-3-d medical image registration. *IEEE Trans. Med. Imaging* **17**, 586–595 (1998).
34. Rueckert, D. *et al*. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Trans. Med. Imaging* **18**, 712–721 (1999).
35. Avants, B. B., Tustison, N. & Song, G. Advanced normalization tools (ants). *Insight J* **2**, 1–35 (2009).
36. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29**, 196–205 (2010).
37. Pinheiro, P. O. & Collobert, R. From image-level to pixel-level labeling with convolutional networks. *CVPR, 2015* 1713–1721 (2015).
38. Dou, Q. *et al*. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE Trans. Med. Imaging* **35**, 1182–1195 (2016).
39. You, C. *et al*. Structurally-sensitive multi-scale deep neural network for low-dose ct denoising. *IEEE Access* **6**, 41839–41855 (2018).
40. Srivastava, N. Improving neural networks with dropout. *UofT* **182**, 566 (2013).
41. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *ECCV, 2016*, 694–711 (2016).
42. Wang, X. & Gupta, A. Generative image modeling using style and structure adversarial networks. In *ECCV, 2016*, 318–335 (2016).
43. Yoo, D., Kim, N., Park, S., Paek, A. S. and Kweon, I. S. Pixel-level domain transfer. In *ECCV, 2016*, 517–532 (2016).
44. Zhou, Y. & Berg, T. L. Learning temporal transformations from time-lapse videos. In *ECCV, 2016*, 262–277 (2016).
45. Ronneberger, O., Fischer, P. & Brox, T. Convolutional networks for biomedical image segmentation. In *MICCAI, 2015*, 234–241 (2015).
46. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML, 2015*, 448-456 (2015).
47. Iizuka, S., Simo-Serra, E. & Ishikawa, H. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph* **35**, 110–119 (2016).
48. Larsson, G., Maire, M. & Shakhnarovich, G. Learning representations for automatic colorization. In *ECCV, 2016*, 577–593 (2016).
49. Boltcheva, D., Yvinec, M. & Boissonnat, J.-D. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *NeuroImage* **46**, 786–802 (2009).
50. Collobert, R., Kavukcuoglu, K. & Farabet, C. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop, 2011*, 192376–192381 (2011).

51. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
52. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).
53. Wang, H. *et al.* Multi-atlas segmentation with joint label fusion. *IEEE Trans. Pattern Anal. Mach. Intel.* **35**, 611–623 (2013).
54. Artaechevarria, X. & Munoz-Barrutia, A. &Ortiz-De-Solorzano, C. Combination strategies in multi-atlas image segmentation: application to brain mr data. *IEEE Trans. Med. Imaging* **28**, 1266–1277 (2009).
55. Chen, T. *et al.* Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274* (2015).
56. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn Res.* **12**, 2825–2830 (2011).
57. Hore, A. & Ziou, D. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2366–2369 (2010).
58. Pluim, J. P. W., Maintz, J. B. A. & Viergever, M. A. Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Imaging* **22**, 986–1004 (2003).
59. Wang, Z. & Bovik, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Process. Mag.* **26**, 98–117 (2009).
60. Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2015).
61. Wang, L. *et al.* Links: Learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage* **108**, 160–172 (2015).
62. Mendrik, A. M. *et al.* Mrbrains challenge: Online evaluation framework for brain image segmentation in 3t mri scans. *Comput. Intel. Neurosc.***2015**, 1–16 (2015).
63. West, J. *et al. Comparison and evaluation of retrospective intermodality brain image registration techniques* **21**, 554–566 (1997).
64. Ghavami, N. *et al. Integration of spatial information in convolutional neural networks for automatic segmentation of intraoperative transrectal ultrasound images. Journal of Medical Imaging* **6**, 011003 (2018).
65. Larsen, A. B. L., Sønderby, S. K., Larochelle, H. & Winther, O. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300* 1558–1566 (2015).

Acknowledgements

This work is supported by the National Science and Technology Major Project of the Ministry of Science and Technology in China under Grant 2017YFC0110903, Microsoft Research under the eHealth program, the National Natural Science Foundation in China under Grant 81771910, the Beijing Natural Science Foundation in China under Grant 4152033, the Technology and Innovation Commission of Shenzhen in China under Grant shenfagai2016-627, Beijing Young Talent Project in China, the Fundamental Research Funds for the Central Universities of China under Grant SKLSDE-2017ZX-08 from the State Key Laboratory of Software Development Environment in Beihang University in China, the 111 Project in China under Grant B13003.

Author contributions

Q.Y., N.L., Z.Z. and X.F. built the algorithm. Q.Y. conducted experiments on image generation and segmentation. N.L., Z.Z. and X.F. collected the datasets, conducted the experiments on image generation and registration. E.C. provided the guidance on mathematics. Y.X. provided the guidance on the overall plan of this projects. All authors contributed to manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.X.