MRI Image-to-Image Translation for Cross-Modality Image Registration and Segmentation

Qianye Yang[a,*], Nannan Li[a,*], Zixu Zhao[b,*], Xingyu Fan[c,*], Eric I-Chao Chang[d], Yan Xu[a,d,**]

[a] *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 EngineeringBeihang University, Beijing 100191, China*
[b] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*
[c] *Bioengineering College of Chongqing University, Chongqing 400044, China*
[d] *Microsoft Research, Beijing 100080, China*

Abstract

We develop a novel cross-modality generation framework that learns to generate predicted modalities from given modalities in MR images without real acquisition. Our proposed method performs image-to-image translation 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. 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 predicted modalities. Second, we propose an approach for MRI segmentation, translated multichannel segmentation (TMS), where given modalities, along with predicted modalities, are segmented by fully convolutional networks (FCN) in a multi-

[*] These four authors contribute equally to the study
[**] Corresponding author

Email addresses: QianyeYang@buaa.edu.cn (Qianye Yang), nannanli@buaa.edu.cn (Nannan Li), zixuzhao1218@gmail.com (Zixu Zhao), xingyu.fan02@gmail.com (Xingyu Fan), echang@microsoft.com (Eric I-Chao Chang), xuyan04@gmail.com (Yan Xu)

channel manner. Both 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 MRI datasets. We also observe encouraging results in cross-modality registration and segmentation on some widely adopted datasets. Overall, our work can serve as an auxiliary method in clinical diagnosis and be applied to various tasks in medical fields.

Keywords: Image-to-image, cross-modality, registration, segmentation, MRI

1. Introduction

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. joints in the brain; T2 images are utilized for locating inflammations and tumors; T2-Flair images present the location of lesions with water suppression.

However, there are three problems with MR images. (1) A series of scans of different modalities take a long time for real acquisition. (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 between one modality to another is hard to learn. As illustrated in Fig.1, there exist large differences among different modalities. The existing approaches cannot achieve satisfactory results for cross-modality synthesis. When dealing with the paired MRI data, the regression-based approach [4] will still lose some information of brain structures. Synthesizing a predicted modality from a given modality without real acquisitions, also known as cross-modality generation, is a nontrivial problem worthy of being studied. Take the transition from T1 (given modality) to T2 (target modality) as an example, $\widehat{T}2$ (predicted modality) can be generated through a cross-modality generation framework. Cross-modality generation tasks refer to transitions such as from T1 to T2, from T1 to T2-Flair, from T1 to T2-Flair, and vice versa.

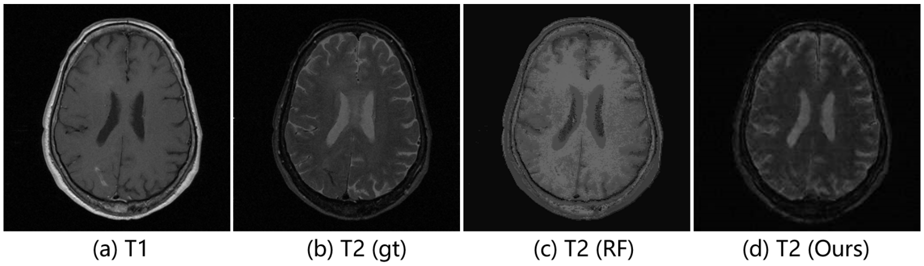

Figure 1: One example of two different modalities. (a), (b) show the difference between T1 and T2 images. (c), (d) show the cross-modality generation result of the regression-based approach based on random forests (RF) [4] and our method (transitions from T1 to T2). Note that the T2 image in (c) is relatively coarser and many contours are lost during the generation process, while accurate information of brain structures is preserved in (d).

Recently, image-to-image translation networks have provided a generic solution for image prediction problems in natural scenes, like mapping images to edges [5, 6], segments [7], semantic labels [8] (many to one), and mapping labels to realistic images (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) [9, 10, 11, 12] makes it possible to learn a loss adapting to the data and be applied to multiple translation tasks. Isola et al. [11] demonstrate that the conditional GAN (cGAN) is suitable for image-to-image translation tasks, where they condition on input images.

Previous work on image-to-image translation networks focuses on natural scenes [11, 13, 14, 15], however, such networks' effectiveness in providing a solution for translation tasks in medical scenes remains inconclusive. Motivated by [11], we introduce image-to-image translation networks to MRI cross-modality generation. 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 predicted 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 can serve as an auxiliary method in clinical diagnosis and also has great application potential, such as multimodal registration [16], segmentation [17], and virtual enhancement [18]. 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 [16, 19]. However, few of them adopt the cross-modality information from multiple modalities, especially from predicted modalities. Here, we propose a new method for cross-modality registration by adopting cross-modality information from our predicted modalities. In our method, inputting a given-modality image (e.g. T2 image) to our proposed framework yields a predicted modality (e.g. $\hat{T}1$ image). Both two modalities consist of our target space (T2 and $\hat{T}1$ images). The source images including T2 and T1 images are then registered to the identical modality in the target space with a registration algorithm (T2 is registered to T2, T1 is registered to $\hat{T}1$). The deformation generated in the registration process are finally combined in a weighted fusion process to propagate the atlas labels to the target space. 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.

The second application of our proposed framework is brain segmentation from MRI data, which also plays an important role in clinical auxiliary diagnosis. However, it is a difficult task owing to the artifacts and in-homogeneities introduced during the real image acquisition [20, 21]. To this point, we propose a novel approach for brain segmentation, called translated multichannel segmentation (TMS). In TMS, the predicted modality that generated in our proposed framework and its corresponding given modality are first input to separate channels and then segmented by fully convolutional networks (FCN) [8] for improvement of brain segmentation. The channels we mentioned are commonly used as image RGB channels in most neural networks. TMS is an effective method for brain segmentation by adding cross-modality information from predicted modalities. For instance, TMS can improve tumor segmentation performance by adding cross-modality information from predicted T2 modality into original T1 modality.

Contributions: (1) We introduce image-to-image translation networks for cross-modality MRI generation to synthesize predicted modalities from given modalities. Our proposed framework can cope with a great many MRI prediction tasks using the same objective and architecture. (2) Registration: We leverage our proposed framework to augment the target source with predicted modalities for atlas-based registration. Registering source images to target images and weighted fusion process enable us to make the most of cross-modality information without adding any extra data to our atlas source and target source. (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 predicted modality into separate channels, which allows us to adopt and fuse cross-modality information without using any extra data. (4) We demonstrate the universality of our framework method for cross-modality generation on five publicly available datasets. Experiments conducted on two sets of datasets also verify the effectiveness of two applications of our proposed framework. We finally observe competitive generation results of our proposed framework.

2. Related work

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

2.1. 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 [22], random forests [4], coupled dictionary learning [16], and convolutional neural network (CNN) [23], 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. [24]. 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 [25]. In [4], a regression-based approach is presented where a regression forest is trained using paired data from given modality to targeted modality. Later, the regression forest is utilized to regress target-modality patches from given modality patches. **(2) Deep learning based methods.** Nguyen et al. [23] 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. [26] 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 [26, 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 [4] as our baseline.

GANs in natural scenes: Recently, a Generative Adversarial Network (GAN) has been proposed by Goodfellow et al. [9]. 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 [10], image prediction [11], text-to-image translation [12] 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. [11] present a generalized solution to image-to-image translations in natural scenes. Our cross-modality image generation is inspired by [11] but we focus on medical images generation as opposed to natural scenes.

GANs in medical images: In spite of the success of existing approaches in natural scenes, there are few applications of GANs to medical images. Nie et al. [30] estimate CT images from MR images with a 3D 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 cross-modality image generation framework.

2.2. Image registration

A successful image registration application requires several components that are correctly combined, like the cost function and the transformation model. 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.

2.3. Image segmentation

A rich body of image segmentation algorithms exists in computer vision [37, 38, 8, 7]. We discuss two that are closely related to our work.

The Fully Convolutional Network (FCN) proposed by Long et al. [8] 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 [8], TMS adopts similar FCN architectures but focuses on fusing information of different modalities in a multichannel manner.

Xu et al. [7] propose an algorithm for gland instance segmentation, where the concept of multichannel learning is introduced. The proposed algorithm exploits features of edge, region, and location in a multichannel manner to generate instance segmentation. By contrast, TMS leverages features in predicted modalities to refine the segmentation performance of given modalities.

3. MRI Cross-Modality Image Generation

In this section, we mainly learn an end-to-end mapping from given-modality images to target-modality images. We introduce image-to-image translation networks to cross-modality generation. Here, cGANs are used to realize image-to-image translation networks. The flowchart of our algorithm is illustrated in Fig.2.

3.1. 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 (ground truth). 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 targeted-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 predicted-modality image $\widehat{y}$. Take the transition from T1 to T2 as an instance. Similar to a two-player min-max game, the training procedure of GAN 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 ground truth T2 (y) in order to puzzle discriminator D. On the other hand, D evolves to distinguish synthesized images $\hat{T}2$ ($\hat{y}$) generated

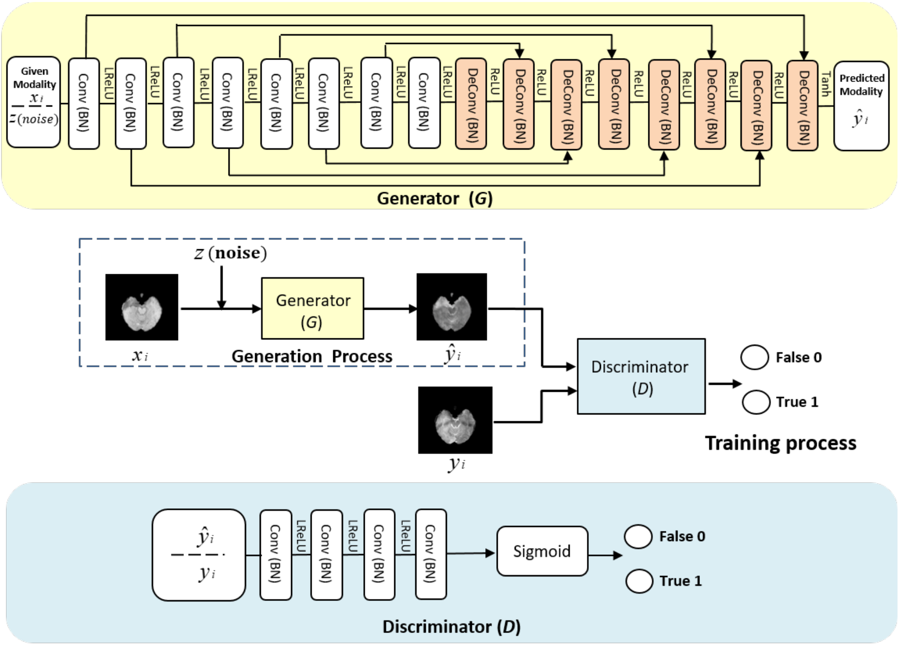

Figure 2: Overview of our framework for cross-modality generation. Notice that our training set is denoted as $S = \{(x_i, y_i), i = 1, 2, 3, \ldots, n\}$, where x_i and y_i refer to the ith input given-modality image and its corresponding targeted-modality image (ground truth). The training process 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 ground truth y_i. On the other hand, discriminator D evolves to distinguish between predicted-modality images $\hat{y}_i$ generated by G and the ground truth y_i. The output of D is 0 or 1, where 0 represents synthesized images and 1 represents the ground truth. In the generation process, predicted-modality images can be synthesized through the optimized G.

by G from the ground truth T2 (y). The overall objective function is defined:

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

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 ground truth. Existing algorithms [10] have found it favorable to combine traditional regularization terms with the objective function in GAN. An L1 loss, as described in [11], 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:

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

The overall objective function is then updated to:

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

Following [11], 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:

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

Here, the introduction of z leads to stochastic rather than deterministic outputs. We adopt dropout noise in several layers of G at both training and testing time. However, we discover that there is only minor stochasticity in the outputs although dropout noise is added to the inputs.

In addition, we also explore the effectiveness of other different loss functions. Generators with different loss functions are defined as follows: $cGAN$: Generator G together with an adversarial discriminator conditioned on the input; $L1$: Generator G with an L1 loss. It is essentially equivalent to a traditional CNN architecture with least absolute deviation; $cGAN + L1$: Generator G with both an L1 loss term and an adversarial discriminator conditioned on the input. We first try to apply the regression-based approach [4] by adopting a bagged ensemble of regression trees, which consists of 100 trees in total. The patch size is $3 \times 3 \times 3$ and thus we have 27-dimensional patch vectors. The results showed in the second column of Table 1 are unsatisfactory, which is expected since this algorithm is still not fine enough. We further explore whether the performance can be improved by adding three different loss functions to Generator G. Our proposed framework with different loss functions ($cGAN + L1, L1, cGAN$) outperforms the regression-based approach on three evaluation metrics (see the third, fourth, fifth column of Table 1).

Table 1: Generation performance on *BraTs2015* with different loss functions. "*RF*" indicates the regression-based approach using random forests (RF) [4]. "→" indicates the transition from given modality to predicted modality (the meaning of "→" is the same in the following tables); "↑" indicates a better performance if the value is higher, and "↓" indicates a better performance if the value is lower.

Transitions	*RF*			*cGAN + L1*			*L1*			*cGAN*		
	MAE ↓	PSNR ↑	MI ↑	MAE ↓	PSNR ↑	MI ↑	MAE ↓	PSNR ↑	MI ↑	MAE ↓	PSNR ↑	MI ↑
T1 → T2	45.758	17.990	0.690	**37.742**	**22.244**	0.845	38.511	22.072	**0.864**	38.909	21.935	0.828
T1 → T2-Flair	222.910	17.240	0.716	**28.991**	22.669	0.837	29.805	**23.232**	**0.894**	30.613	22.515	0.835

3.2. Network architecture

Our cross-modality generation framework is composed of two main submodels, **generator (G)** and **discriminator (D)**. It is similar to traditional GANs [9].

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 [39, 10, 40, 41, 42]. 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 [43]. These connections speed up information transmission since the bottleneck layer is ignored, and help to learn matching features for corresponding mirrored layers.

In Fig.2, we show the architecture of G. It has 8 convolutional layers, each of which contains a convolution, a Batch Normalization, and a leaky ReLu activation [44] (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 [44] (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 ground truth. 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 ground truth as true (label 1), then providing feedback to G. PixelGANs [11] have poor performance on spatial sharpness, and ImageGANs [11] with many parameters are hard to train. In contrast, PatchGANs [11] 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 [45, 46] 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. As shown in Fig.3, training with only the L1 loss gives the obscure predictions that lack some discernible details. Under same experimental setup, the results on the *BraTs2015* dataset are improved notably with the combination of the adversarial loss and L1 loss. Without direct control and guidance of discriminator D, our framework is heavily biased towards learning sophisticated brain structures.

As illustrated in Fig.2, D is configured with four layers of convolution-BatchNorm-ReLu (The slope of a ReLu is 0.2). 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.

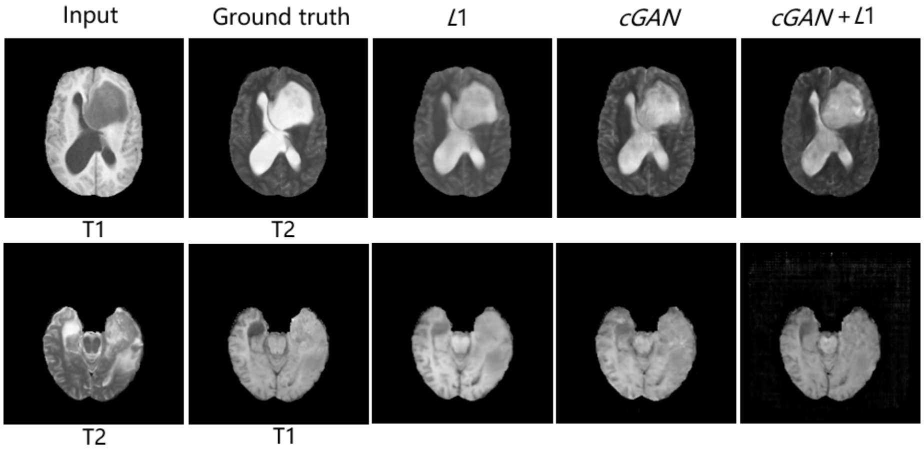

Figure 3: Samples of cross-modality generation results on *BraTs2015*. The right three columns shows results of our proposed framework with different loss functions ($L1, cGAN, cGAN+L1$).

4. 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.

4.1. Cross-Modality Registration

The first application of our cross-modality generation framework is to use the predicted modality for cross-modality image registration. Our method is inspired by an atlas-based registration, where the source image is registered to the target 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 target 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 target space can consist of T2 and $\hat{T}1$ images by using our cross-modality generation framework. The source space commonly consists of both T2 and T1 images from n subjects. (2) The second step is to register the source images

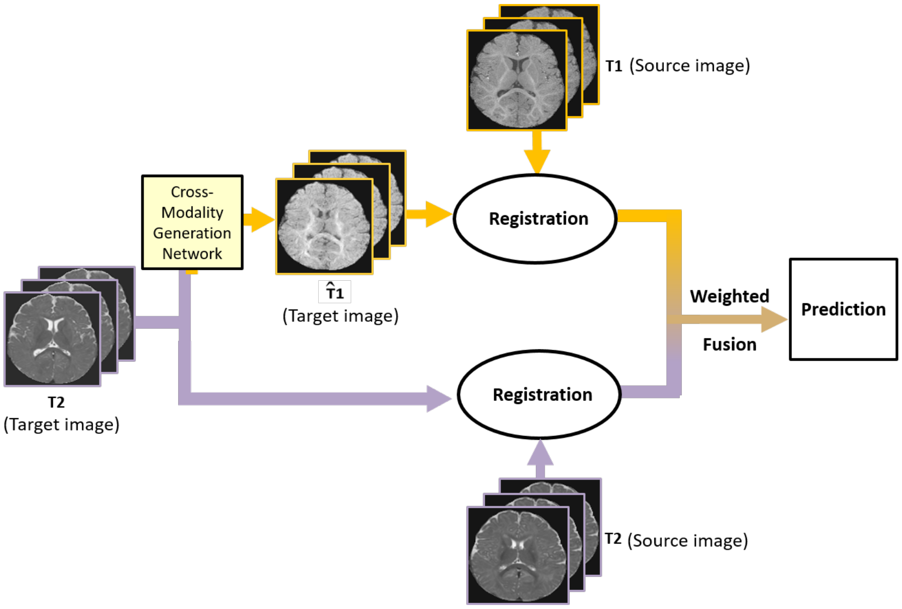

Figure 4: Flowchart of our approach for cross-modality registration. In the target space, inputting T2 images into our proposed framework yields $\hat{T}1$ images. T2 images from the source space are registered to T2 images from the target space. T1 images from the source space are registered to $\hat{T}1$ images from the target space. The corresponding deformations generated after registrations are combined in a weighted fusion process which can be used to predict the final results.

to the target images. For any target subject, we register all n source images to the target images and the deformation field that aligns the source image with the target image can be automatically computed with a registration algorithm. As illustrated in Fig.4, T2 images from the source space are registered to T2 images from the target space and T1 images from the source space are registered to $\hat{T}1$ images from the target 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.4). (4) Applying the deformations to the atlas labels can yield n predictions of the segmentation result. For any target subject, we compute the final results by averaging the n predictions of the target subject.

The registration algorithm mentioned in step (2) can be diverse. 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 [47]. 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.

4.2. 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 predicted-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 predicted-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 [8] 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 predicted-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:

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

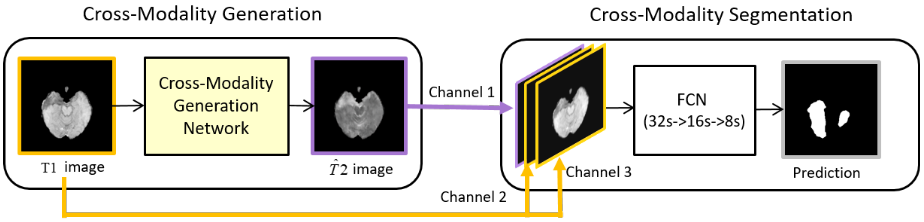
Cross-Modality Generation
Cross-Modality Segmentation
Cross-Modality Generation Network
T1 image
$\hat{T}2$ image
Channel 1
Channel 2
Channel 3
FCN (32s->16s->8s)
Prediction

Figure 5: Flowchart of our approach for cross-modality segmentation. First, we input a given-modality image to our cross-modality generation network to generate a predicted-modality image. For instance, given a T1 image, $\widehat{T}2$ images can be generated with our method. Second, two identical given-modality images and one corresponding predicted-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.

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.

5. Experiments and results

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. 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.

5.1. Implementation details

All our models are trained on NVIDIA Tesla K80 GPUs.

Generation: To obtain the appropriate loss weight in Equation 3, a five-fold cross-validation is conducted for hyper-parameters tuning. A weight of 100 is selected for λ. We train the models on a torch7 framework [48] using Adam optimizer [49] with 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 [50], instance normalization is effective at generation tasks by removing instance-specific information from the content image. Other parameters follow the reference [11]. All experiments use 70×70 PatchGANs.

Registration: A Windows release 2.1.0 version of ANTs [35] as well as its auxiliary registration tools are 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 [51]. For Elastix, we adopt the parameters in [52].

Segmentation: We implement standard FCN-8s on a publicly available MXNET toolbox [53]. 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 [8].

5.2. Cross-Modality Generation

Evaluation metrics. We report results on mean absolute error (MAE), peak signal-to-noise ratio (PSNR), mutual information (MI), and FCN-score.

MAE is defined as below:

$$MAE = \frac{1}{256 \times 256} \sum_{i=0}^{255} \sum_{j=0}^{255} \|\hat{y}(i,j) - y(i,j)\|, \tag{6}$$

where targeted-modality image $\hat{y}$ and predicted-modality image y both have a size of 256×256 pixels, and (i,j) indicates the location of pixels.

PSNR 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 [54]. 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 targeted-modality image y and predicted-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.

FCN-score is used to capture the joint statistics of data and evaluate synthesized images across the board. It includes accuracy and F-score. 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 F-score 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 [11] and adopt a pre-trained FCN to evaluate our experiment results. Pre-trained semantic classifiers are used to measure the discriminability of 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 train an FCN-8s model. Both T2 (test data/ground truth) and $\widehat{T}2$ images are subsequently segmented through the trained model. We score the segmentation (classification) accuracy of synthesized images against the ground truth. The gap of FCN-score between T2 images and $\widehat{T}2$ images quantitatively evaluates the quality of $\widehat{T}2$ 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: All 2D images are then resized to a resolution of 256×256 pixels, after which we have the 2D input images. 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 [55] 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 [56] contains multi-contrast MR images from 23 infants, including T1, T2 images and corresponding labels of Grey Matter (gm) and White Matter (wm). This dataset is randomly split into training and testing at a ratio of 4:1. 661 training images are trained for 800 epochs with batch size 1. 163 images are used for testing.

(3)*MRBrain13*: The MRBrain13 dataset [57] 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 [58] 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.

Table 2: Generation performance on four publicly available datasets evaluated by MAE, PSNR, and MI. Our models using image-to-image translation networks achieve better results than the regression-based approach using random forests (RF) [4].

Datasets	Transitions	*RF*			*cGAN* + *L*1			*L*1			*cGAN*		
		MAE ↓	PSNR ↑	MI ↑	MAE ↓	PSNR ↑	MI ↑	MAE ↓	PSNR ↑	MI ↑	MAE ↓	PSNR ↑	MI ↑
BraTs2015	T1 → T2	45.758	17.990	0.690	**37.742**	**22.244**	0.845	38.511	22.072	**0.864**	38.909	21.935	0.828
	T2 → T1	40.486	17.176	0.683	**32.973**	22.004	0.774	33.212	**22.913**	**0.818**	34.811	21.825	0.768
	T1 → T2-Flair	222.910	17.240	0.716	**28.991**	22.669	0.837	29.805	**23.232**	**0.894**	30.613	22.515	0.835
	T2 → T2-Flair	215.306	19.967	0.753	**25.172**	21.450	0.863	30.253	**22.147**	**0.933**	30.651	21.412	0.864
Iseg2017	T1 → T2	**40.649**	21.475	0.722	43.459	27.237	0.908	46.748	**28.529**	0.993	47.110	26.077	**1.157**
	T2 → T1	43.142	23.548	0.709	**42.311**	**24.100**	0.860	42.521	23.325	**0.880**	42.602	20.932	0.786
MRBrain13	T1 → T2-Flair	158.165	22.367	0.827	**45.426**	**27.861**	1.104	48.338	26.885	**1.252**	48.493	25.068	1.099
ADNI	PD → T2	135.625	24.860	0.790	**60.371**	27.338	1.276	60.891	**28.650**	**1.460**	62.285	26.044	1.089
	T2 → PD	80.500	23.666	1.055	**40.242**	30.728	1.401	40.445	**32.303**	**1.567**	43.245	29.425	1.311
RIRE	T1 → T2	122.592	16.893	0.537	**88.693**	28.763	0.618	100.075	**30.157**	**0.745**	100.282	27.198	0.542
	T2 → T1	222.016	19.688	0.546	**82.032**	23.946	0.927	87.883	**25.112**	**1.048**	89.156	22.688	0.865

Results. Generation performance on the five datasets are summarized in Table 2. Overall, our approach can effectively generate predicted-modality images from given-modality images and vice versa. We also present some samples of generation results to visualize the improvement of our approach over the regression-based method using RF [4] (see Fig.6). The images are intensity standardized before training and testing. Besides, we evaluate the segmentation results of our generated images on *BraTs2015* and *Iseg2017* to explore generation performance from another perspective (see Tables 3 and 4).

Fig.6 shows the qualitative results of various losses on five datasets. We have reasonable but blurry results with $L1$ alone. The $cGAN$ alone leads to improvements in visual performance but causes some artifacts in cross-modality MR image generation. Using $cGAN + L1$ terms achieves decent results and reduces artifacts. In contrast, the traditional method leads to rough and fuzzy

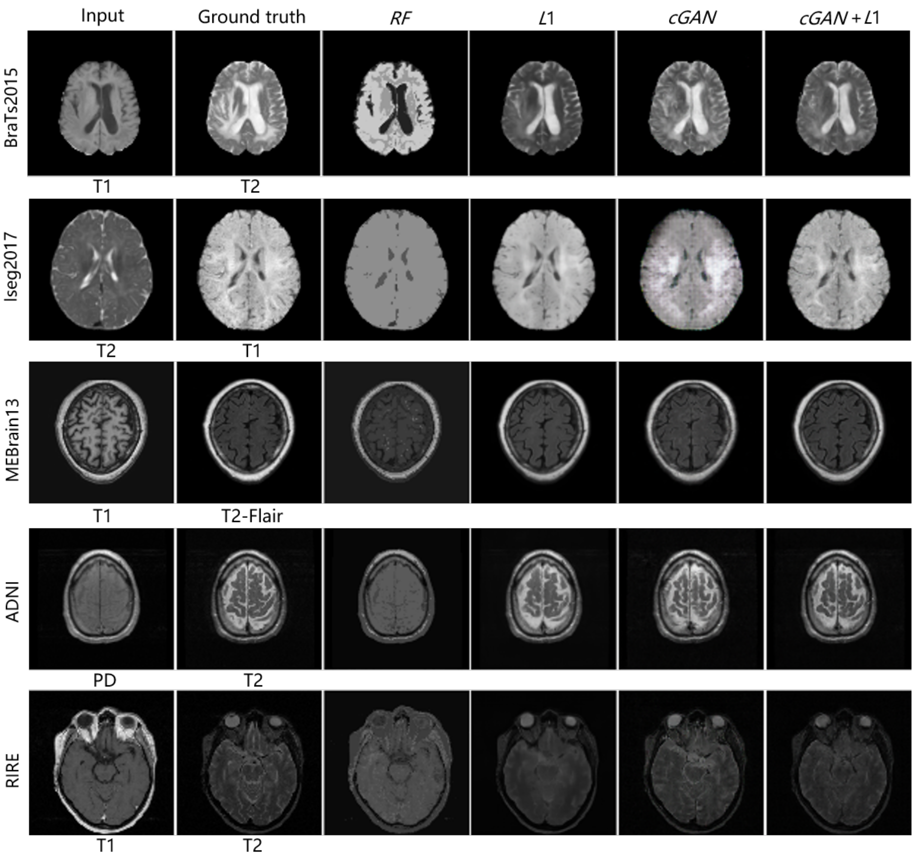

Figure 6: Samples of cross-modality generation results on five publicly available datasets including *BraTs2015* [55], *Iseg2017* [56], *MRBrain13* [57], *ADNI* [30], and *RIRE* [58]. Results are selected from top performing examples (relatively low MAE, high PSNR, and high MI collectively) with four approaches. The right four columns show results of the random forests (RF) [4] and our proposed framework with different loss functions ($L1, cGAN, cGAN + L1$).

results compared with image-to-image networks.

Table 2 quantitatively shows how using image-to-image translation networks allows us to achieve better generation results than the regression-based method using RF [4]. Note that different losses induce different quality of results, all of which outperform the traditional method on MAE, PSNR, and MI. In most cases, our approach with $cGAN + L1$ achieves the best results on MAE; $L1$ loss term contributes to superior performance on MI over other methods; both

Table 3: Segmentation results of generated images on *BraTs2015* evaluated by FCN-score. The gap between generated images and the ground truth can evaluate the generation performance of our method. Note that "gt" represents the ground truth; "all" represents mean accuracy of all pixels (the meanings of "gt" and "all" are the same in the following tables). We achieve close segmentation results between predicted-modality images and the ground truth.

Method	Accuracy		F-score
	all	tumor	tumor
T1 → T2	0.955	0.716	0.757
T2 (gt)	0.965	0.689	0.724
T2 → T1	0.958	0.663	0.762
T1 (gt)	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 (gt)	0.986	0.876	0.899

Table 4: Segmentation results of generated images on *Iseg2017* evaluated by FCN-score. Note that "gm" and "wm" indicate gray matter and white matter respectively. The minor gap between predicted-modality images and the ground truth shows decent generation performance of our framework.

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

$cGAN + L1$ and $L1$ lead to pretty good results on PSNR. Note that 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 [59]. Thus, it is reasonable that using L1 term contributes to superior results on MI.

We also quantify the generation results using FCN-score on *BraTs2015* and *Iseg2017* in Table 3 and Table 4. Our approach ($cGAN + L1$) is effective in generating realistic cross-modality MR images towards the ground truth. The

GAN-based objectives lead to high scores close to the ground truth.

Table 5: Running times under different loss functions. We compute running times by averaging multiple runs to remove other factors that affect running time speed. All images are the same size, thus their running times are nearly the same.

Method	$cGAN + L1$	$L1$	$cGAN$
Running time (s)	0.252	0.252	0.257

Running time. Comparisons of running times under different loss functions are summarized in Table 5. All three models with different loss functions are trained on the same training dataset and tested on the same testing dataset. The training process is carried on NVIDIA Tesla K80 GPUs.

5.3. Cross-Modality Registration

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

(1)*F-score*: The first metric is introduced to measure the overlap of ground truth segmentation labels. It is defined as $(2|H \cap G|)/(|H| + |G|)$ where G is the segmentation label of the target image and H is the segmentation prediction of the source image.

(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 target image.

Dataset. We preprocess the original MRI data from *Iseg2017* and *MR-Brain13* datasets with the following steps to make it applicable for our proposed framework. (1) We first shear the 3D image into a smaller cube, each side of which circumscribes the brain. (2) The brain cube is then resized to a size of $128 \times 128 \times 128$ voxels. (3) The last step is to slice the brain cubes from all the subjects into 2D data along 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 are treated as a pair source and target 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 source space and another 2 subjects in the target space ($n = 8$). For *MRBrain13* dataset, 4 subjects are selected for the source space while one subject in the target space ($n = 4$)

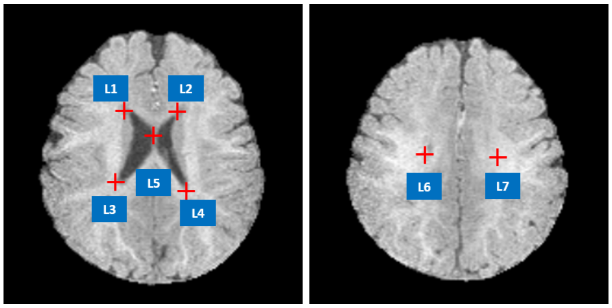

Figure 7: Illustration of the seven landmarks selected for cross-modality registration. L1: right lateral ventricle superior, L2: left lateral ventricle superior, L3: right lateral ventricle inferior, L4: left lateral ventricle inferior. L5: middle of the lateral ventricle, L6: right lateral ventricle posterior, L7: left lateral ventricle posterior.

Iseg2017 and *MRBrain13* datasets provide ground truth segmentation labels. Seven well-defined anatomic landmarks (see Fig.7) 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.

Results. Two sets of experiments are conducted to verify the effectiveness of our proposed method for cross-modality registration. In the first set of experiments, the source images and the target images are both T2 images. In the second set of experiments, the source images are T1 images while the target images are $\hat{T}1$ images. The deformations generated in each set of experiments are

Table 6: Registration results evaluated by Dist and F-score on *Iseg2017* and *MRBrain13*.

Datasets	Modalities	Structures	F-score		Dist	
			ANTs	Elastix	ANTs	Elastix
Iseg2017	T2	wm	0.508±0.008	0.473±0.006	2.105±0.006	2.836±0.014
		gm	0.635±0.015	0.592±0.012		
	$\widehat{T}1$	wm	0.503±0.004	0.469±0.005	1.884±0.011	2.792±0.008
		gm	0.622±0.014	0.580±0.012		
	$\mathbf{T2+\widehat{T}1}$	wm	**0.530±0.009**	**0.517±0.007**	**1.062±0.017**	**2.447±0.009**
		gm	**0.657±0.016**	**0.648±0.015**		
	T1	wm	0.529±0.008	0.514±0.014	1.136±0.009	2.469±0.012
		gm	0.650±0.016	0.639±0.018		
MRBrain13	T2-Flair	wm	0.431±0.025	0.412±0.010	3.417±0.031	3.642±0.023
		gm	0.494±0.026	0.463±0.023		
	$\widehat{T}1$	wm	0.468±0.032	0.508±0.012	3.159±0.016	3.216±0.014
		gm	0.508±0.024	0.487±0.018		
	$\mathbf{T2\text{-}Flair+\widehat{T}1}$	wm	**0.473±0.026**	**0.492±0.012**	**2.216±0.011**	**2.659±0.021**
		gm	**0.530±0.027**	**0.532±0.029**		
	T1	wm	0.484±0.038	0.534±0.009	2.524±0.022	2.961±0.019
		gm	0.517±0.025	0.510±0.018		

combined in a weighted fusion process to propagate the atlas labels to the target space, which yields the final predictions. Table 6 summarizes the registration results both in terms of Dist and F-score. The registration results of T1 images are considered as the upper bound of our cross-modality registration. We find that the registration performance of our predicted-modality images ($\widehat{T}1$ images) advances that of T2 images and T2-Flair images by achieving lower Dist and higher F-score, e.g. 0.622±0.014 for gray matter. This is reasonable as our predicted-modality images are realistic enough, as well as T1 image itself with high contrast for brain structure leads to lower registration errors.

We also introduce the cross-modality information from our $\widehat{T}1$ images into T2 images and T2-Flair images, of which the performance are denoted as T2+$\widehat{T}1$ and T2-Flair+$\widehat{T}1$ in the table 6. 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 observe statically significant improvements in registration accuracy: the Dist

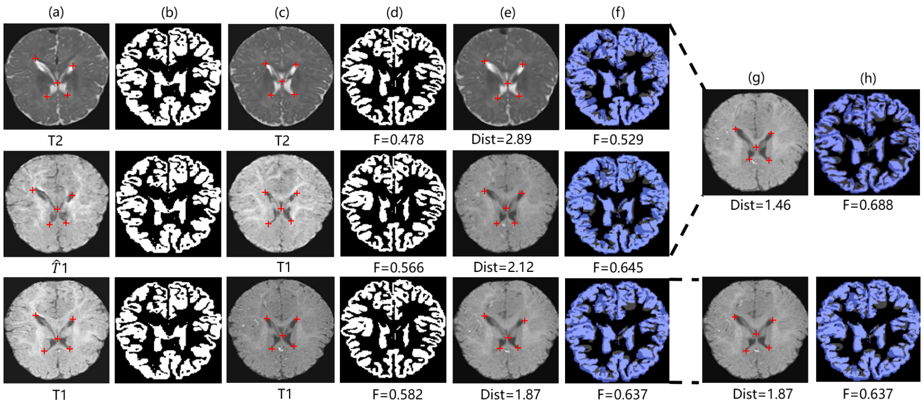

Figure 8: Samples of registration results of our method: (a) Target image, (b) Ground truth segmentation label of target image, (c) Source image, (d) Ground truth segmentation label of source image, (e) Warped image (source image warped by the best traditional registration algorithm (ANTs)), (f) Warped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

is greatly shorten (e.g. 1.062 vs 2.105 in *Iseg2017* dataset) and the F-score is notably increased (e.g. 0.532 vs 0.463 in *MRBrain13* dataset) compared to registrations without adding cross-modality information. In many cases, our method even advances the upper bound both in Dist and F-score. Fig.8 visualizes samples of the registration results of our methods.

5.4. Cross-Modality Segmentation

Evaluation metric. We report segmentation results on F-score (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 predicted 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.

(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. 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 matters and gray matters. To this point, we choose to add cross-modality information from T2 and T2-Flair images 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*.

Table 7: Tumor segmentation results of TMS on *Brats2015*. "T1+$\widehat{T}2$" and "T1+$\widehat{T}2$-$Flair$" indicate our approach (TMS) where inputs are both T1 and $\widehat{T}2$ images or T1 and $\widehat{T}2$-$Flair$ 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.

	F-score(tumor)	Δ
T1	0.760	-
T1+$\widehat{\mathbf{T}}$2	**0.808**	**6.32%**
T1+T2	0.857	-
T1+$\widehat{\mathbf{T}}$2-Flair	**0.819**	**7.89%**
T1+T2-Flair	0.892	-

As shown in Tables 7, cross-modality information from $\widehat{T}2$-$Flair$ and $\widehat{T}2$ images contributes improvements to tumor segmentation of T1 images (7.89% and 6.32% of tumors respectively). Likewise, Table 8 shows that cross-modality information from $\widehat{T}1$ images leads to improvements of wm and gm segmenta-

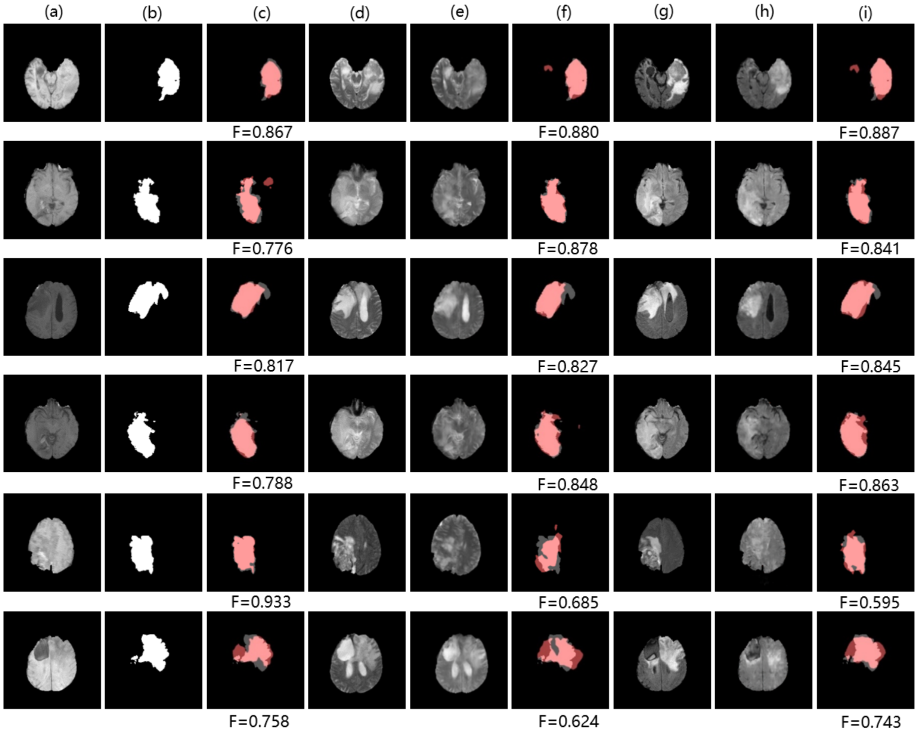

Figure 9: Samples of tumor segmentation results on *BraTs2015*: (a), (d), (e), (g), (h) denote T1 image, T2 image, $\widehat{T}2$ image, T2-Flair image, $\widehat{T}2$-*Flair* image. (b) denotes ground truth segmentation label of T1 image. (c), (f), (i) denote tumor segmentation results of T1 image using the FCN method, TMS (adding cross-modality information from $\widehat{T}2$ image), TMS (adding cross-modality information from $\widehat{T}2$-*Flair* image). Note that we have four decent samples in the first four rows and two abortive cases in the last two rows. Pink: true regions. Grey: missing regions. Dark red: false regions.

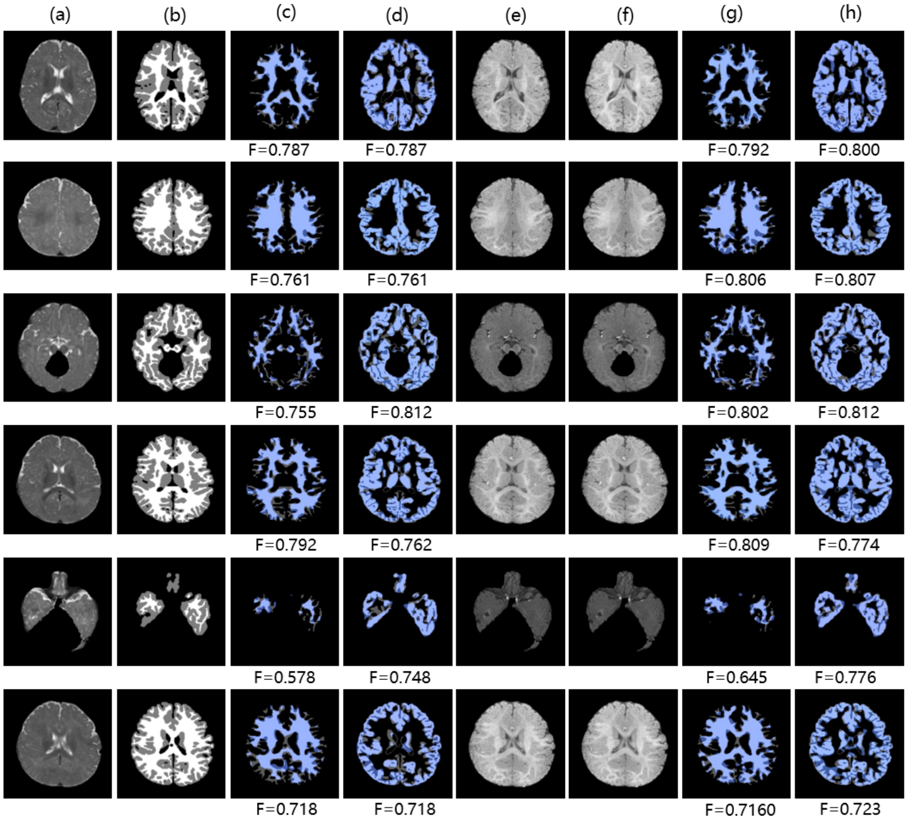

Figure 10: Samples of brain structure segmentation results on *Iseg2017*: (a), (e), (f) denote T2 image, T1 image, $\widehat{T}1$ image. (b) denotes ground truth segmentation label of T2 image. (c), (d) denote white matter and gray matter segmentation results of T2 image using the FCN method respectively. (g), (h) denote white matter and gray matter segmentation results of T2 image using TMS (adding cross-modality information from $\widehat{T}1$ image) respectively. Note that we have four decent samples in the first four rows and two abortive cases in the last two rows. Blue: true regions. Grey: missing regions. Dark blue: false regions.

Table 8: Brain structure segmentation results of TMS on *Iseg2017*. "T2+$\widehat{T}1$" indicates our method (TMS) where inputs are both T2 and $\widehat{T}1$ images. "T2" indicates the traditional FCN method where inputs are only T2 images. "T2+T1" indicates the upper bound.

	F-score(wm)	Δ	F-score(gm)	Δ
T2	0.649	-	0.767	-
T2+$\widehat{\mathbf{T}}$1	**0.669**	**3.08%**	**0.783**	**2.09%**
T2+T1	0.691	-	0.797	-

tion 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 which means that our predicted modalities are very close to real modalities. Overall, TMS outperforms the traditional FCN method when favorable cross-modality information is adopted. Fig.9 and Fig.10 visualize some samples of our segmentation results on *BraTs2015* and *Iseg2017* respectively.

5.5. Discussion

In most cases, cross-modality information in predicted-modality and original-modality images is conducive to observing the structures of tumors. As illustrated in Fig.11, there are also abortive cases where tiny structures may be mistaken, though we generally achieve excellent performance. In the yellow

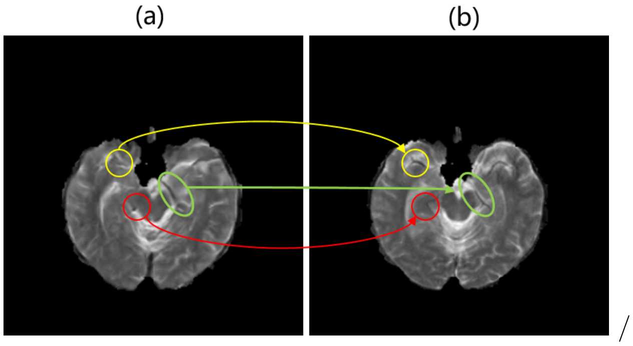

Figure 11: An abortive sample in our generation results:(a) $\hat{T}2$. (b) T2. Circles in $\hat{T}2$ indicate some misdescription of tiny structures. Different colourful circles indicate different problems.

box, the eyebrow-like structure is missing. The red box indicates a non-existent round structure which might be confounded with the vessel. In the green box, the learned structure seems to be discontinuous which might give rise to perplexity for radiologists to make a diagnosis. In the future, we will refine our algorithm to describe more tiny structures.

6. Conclusion

In this paper, we have developed a novel conditional-generative-adversarial-network-based framework for cross-modality translation that demonstrates competitive performance on cross-modality registration and segmentation. Our framework builds on top of the ideas of image-to-image translation networks. We also have proposed two new approaches for MR image registration and segmentation by adopting cross-modality information from predicted modality generated with our proposed framework. Our methods outperform the state-of-the-art 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. Our work is extensive and can be applied to a wide range of fields such as cross-modality translation from CT to MRI or from MRI to PET.

7. Acknowledgment

This work is supported by Microsoft Research under the eHealth program, the National Natural Science Foundation in China under Grant 81771910, the National Science and Technology Major Project of the Ministry of Science and Technology in China under Grant 2017YFC0110903, 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. The authors would like to thank all the dataset providers for making their databases publicly available.

References

References

[1] K. L. Tseng, Y. L. Lin, W. Hsu, C. Y. Huang, Joint sequence learning and cross-modality convolution for 3d biomedical segmentation, arXiv preprint arXiv:1704.07754.

[2] R. Rzedzian, B. Chapman, P. Mansfield, R. E. Coupland, M. Doyle, A. Chrispin, D. Guilfoyle, P. Small, Real-time nuclear magnetic resonance clinical imaging in paediatrics, Lancet 2 (8362) (1983) 1281–1282.

[3] J. Tsao, Ultrafast imaging: principles, pitfalls, solutions, and applications, Journal of Magnetic Resonance Imaging 32 (2) (2010) 252–266.

[4] A. Jog, S. Roy, A. Carass, J. L. Prince, Magnetic resonance image synthesis through patch regression, in: IEEE International Symposium on Biomedical Imaging, 2013, pp. 350–353.

[5] S. Xie, Z. Tu, Holistically-nested edge detection, ICCV (2015) 1–16.

[6] C. Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, AISTATS (2014) 562–570.

[7] Y. Xu, Y. Li, Y. Wang, M. Liu, Y. Fan, M. Lai, E. Chang, Gland instance segmentation using deep multichannel neural networks, TBME.

[8] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015, pp. 3431–3440.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.

[10] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: CVPR, 2016, pp. 2536–2544.

[11] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, CVPR.

[12] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, D. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, ICCV.

[13] Z. Tu, Learning generative models via discriminative approaches, in: CVPR, 2007, pp. 1–8.

[14] J. Lazarow, L. Jin, Z. Tu, Introspective neural networks for generative modeling, ICCV.

[15] L. Jin, J. Lazarow, Z. Tu, Introspective generative modeling: Decide discriminatively, arXiv preprint arXiv:1704.07820.

[16] S. Roy, A. Carass, J. Prince, Magnetic resonance image example based contrast synthesis., TMI 32 (12) (2013) 2348.

[17] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. V. Leemput, B. Fischl, Is synthesizing mri contrast useful for inter-modality analysis?, in: MICCAI, 2013, p. 631.

[18] R. Vemulapalli, H. V. Nguyen, S. K. Zhou, Unsupervised cross-modal synthesis of subject-specific scans, in: ICCV, 2016, pp. 630–638.

[19] I. J. Eugenio, S. M. Rory, L. K. Van, A unified framework for cross-modality multi-atlas segmentation of brain mri, Medical Image Analysis 17 (8) (2013) 1181.

[20] M. A. Balafar, A. R. Ramli, M. I. Saripan, S. Mashohor, Review of brain mri image segmentation methods, Artificial Intelligence Review 33 (3) (2010) 261–274.

[21] N. Sasirekha, K. R. Kashwan, Improved segmentation of mri brain images by denoising and contrast enhancement.

[22] W. T. Freeman, E. C. Pasztor, Learning low-level vision, International Journal of Computer Vision 40 (1) (2000) 25–47.

[23] H. V. Nguyen, K. Zhou, R. Vemulapalli, Cross-Domain Synthesis of Medical Images Using Efficient Location-Sensitive Deep Network, Springer International Publishing, 2015.

[24] M. I. Miller, G. E. Christensen, Y. Amit, U. Grenander, Mathematical textbook of deformable neuroanatomies., Proceedings of the National Academy of Sciences of the United States of America 90 (24) (1993) 11944.

[25] F. Rousseau, Brain Hallucination, Springer Berlin Heidelberg, 2008.

[26] R. Vemulapalli, H. V. Nguyen, S. K. Zhou, Unsupervised cross-modal synthesis of subject-specific scans, in: ICCV, 2015, pp. 630–638.

[27] Y. Huang, L. Shao, A. F. Frangi, Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding, CVPR.

[28] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, Computer Science.

[29] M. Mirza, S. Osindero, Conditional generative adversarial nets, Computer Science (2014) 2672–2680.

[30] D. Nie, R. Trullo, C. Petitjean, S. Ruan, D. Shen, Medical image synthesis with context-aware generative adversarial networks, arXiv preprint arXiv:1612.05362.

[31] J. M. Wolterink, T. Leiner, M. A. Viergever, I. Isgum, Generative adversarial networks for noise reduction in low-dose ct, TMI PP (99) (2017) 1–1.

[32] P. Viola, W. Wells, Alignment by maximization of mutual information, International Journal of Computer Vision 24 (2) (1997) 137–154.

[33] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, D. J. Hawkes, A comparison of similarity measures for use in 2-d-3-d medical image registration, IEEE Transactions on Medical Imaging 17 (4) (1998) 586–95.

[34] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, D. J. Hawkes, Nonrigid registration using free-form deformations: application to breast mr images, IEEE Transactions on Medical Imaging 18 (8) (1999) 712.

[35] Advanced normalization tools, http://stnava.github.io/ANTs/(2017(accessed Feb.2017)).

[36] S. Klein, M. Staring, K. Murphy, M. A. Viergever, J. P. Pluim, elastix: a toolbox for intensity-based medical image registration., IEEE Transactions on Medical Imaging 29 (1) (2010) 196.

[37] P. O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, CVPR (2015) 1713–1721.

[38] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, P.-A. Heng, Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks, TMI 35 (5) (2016) 1182–1195.

[39] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: ECCV, 2016, pp. 694–711.

[40] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, in: ECCV, 2016, pp. 318–335.

[41] D. Yoo, N. Kim, S. Park, A. S. Paek, I. S. Kweon, Pixel-level domain transfer, in: ECCV, 2016, pp. 517–532.

[42] Y. Zhou, T. L. Berg, Learning temporal transformations from time-lapse videos, in: ECCV, 2016, pp. 262–277.

[43] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015, pp. 234–241.

[44] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, ICML (2015) 448–456.

[45] S. Iizuka, E. Simo-Serra, H. Ishikawa, Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification, ACM TOG 35 (4) (2016) 110.

[46] G. Larsson, M. Maire, G. Shakhnarovich, Learning representations for automatic colorization, in: ECCV, 2016, pp. 577–593.

[47] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, R. V. Parsey, Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration, NeuroImage 46 (3) (2009) 786 – 802. doi:https://doi.org/10.1016/j.neuroimage.2008.12.037.
URL http://www.sciencedirect.com/science/article/pii/S1053811908012974

[48] R. Collobert, K. Kavukcuoglu, C. Farabet, Torch7: A matlab-like environment for machine learning, in: BigLearn, NIPS Workshop, no. EPFL-CONF-192376, 2011.

[49] D. Kingma, J. Ba, Adam: A method for stochastic optimization, Computer Science.

[50] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization.

[51] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, P. A. Yushkevich, Multi-atlas segmentation with joint label fusion, IEEE Transactions on Pattern Analysis & Machine Intelligence 35 (3) (2013) 611–623.

[52] X. Artaechevarria, A. Munoz-Barrutia, C. Ortiz-De-Solorzano, Combination strategies in multi-atlas image segmentation: application to brain mr data., IEEE Transactions on Medical Imaging 28 (8) (2009) 1266–1277.

[53] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang, Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems, Statistics.

[54] J. P. W. Pluim, J. B. A. Maintz, M. A. Viergever, Mutual-information-based registration of medical images: a survey, TMI 22 (8) (2003) 986–1004.

[55] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, The multimodal brain tumor image segmentation benchmark (brats), TMI 34 (10) (2015) 1993.

[56] L. Wang, Y. Gao, F. Shi, G. Li, J. H. Gilmore, W. Lin, D. Shen, LINKS: Learning-based multi-source IntegratioN frameworK for Segmentation of infant brain images, Vol. 108, 2015.

[57] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. D. Bresser, A. Alansary, M. D. Bruijne, A. Carass, A. El-Baz, Mrbrains challenge: Online evaluation framework for brain image segmentation in 3t mri scans, CIN 2015 (4-5) (2015) 1–16.

[58] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, M. C. Jr, R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, Comparison and evaluation of retrospective intermodality brain image registration techniques, J COMPUT ASSIST TOMO 21 (4) (1997) 554.

[59] A. B. L. Larsen, S. K. Snderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric (2015) 1558–1566.