2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Disentangled Representations based Unsupervised Deformable Framework for Cross-modality Image Registration

Jiong Wu[1,*] and Shuang Zhou[2]

Abstract— **Cross-modality magnetic resonance image (MRI) registration is a fundamental step in various MRI analysis tasks. However, it remains challenging due to the domain shift between different modalities. In this paper, we proposed a fully unsupervised deformable framework for cross-modality image registration through image disentangling. To be specific, MRIs of both modalities are decomposed into a shared domain-invariant content space and domain-specific style spaces via a multimodal unsupervised image-to-image translation approach. An unsupervised deformable network is then built based on the assumption that intrinsic information in the content space is preserved among different modalities. In addition, we proposed a novel loss function consists of two metrics, with one defined in the original image space and the other in the content space. Validation experiments were performed on two datasets. Compared to two conventional state-of-the-art cross-modality registration methods, the proposed framework shows a superior registration performance.**

Clinical relevance—**This work can serve as an auxiliary tool for cross-modality registration in clinical practice.**

I. INTRODUCTION

Different modality magnetic resonance images (MRIs) show specific tissue features in different spatial domains. The fusion of complementary information from different modalities will improve the performance of various MRI analysis tasks such as brain segmentation and disease progression analysis [1]. However, in clinical practice, different modality images are generally produced by scanning the patients using scanners at different times with some anatomical changes. Hence, there is a strong clinical need to develop a cross-modality image registration method for accurate information fusion thus accurately analysis and interpretation.

Due to the quite different intensity profiles across modalities, cross-modality image registration remains challenging. To tackle this problem, the majority of traditional approaches rely on information theories. These methods generally utilize information theory measures such as mutual information (MI) [2], normalized mutual information (NMI) [3] to calculate the misalignment between images. However, unlike other measures such as mean squared difference (MSD) and cross-correlation (CC), information theory measures often ignore local anatomical details during the registration process [1]. Some other traditional approaches convert the multiple modalities into a new one or one of the existing modalities to solve the cross-modality registration problem. For instance, Roy et al. [4] use an expectation-maximization framework to convert T1-weighted (T1w) MRIs into CT images before registration. Wachinger et al. [5] convert the multiple modality images into the images which only contained structural information before registration via patch entropy calculation and manifold learning. Although these methods obtained some promising results, losing some anatomical information may reduce the registration accuracy.

To fully utilize anatomical features of original images from different modalities, image synthesis-based approaches are proposed. These methods firstly synthesize proxy images of missing modalities and then perform the registration in a multi-modality manner. For instance, Iglesias et al. [6] use a K-nearest neighbor patch-based method to synthesize T1w MRIs for registration. Chen et al. [7] propose a cross-modality registration method that creates proxy images based on a trained regression forest. With the advent of the generative adversarial network (GAN) [8], lots of registration methods utilizing GANs to create the proxy missing modality images. Examples include adopting CycleGAN to synthesize multi-modality atlases to improve the accuracy of conventional registration algorithm [1], using a conditional GAN to realize multi-modality images translation for cross-modality image registration [9], and utilizing CycleGAN to generate CT image from MRI following an inverse-consistent network for MRI-CT registration [10].

Different from the aforementioned approaches, Qin et al. [11] propose an unsupervised learning-based method by directly embedding the registration network into a multimodal image-to-image translation framework (MUNIT) [12]. Although this method alleviates the issue of some dissimilar image generation in CycleGAN based methods, deformation fields generated from latent content domain introduces inconsistencies on the local level. In this paper, we propose a fully unsupervised deformable framework for cross-modality image registration via disentangled representations. A representations disentangling model (RDM) is introduced and learned to drive a deformable registration network for learning the mappings between T1w and T2w MRIs. In addition, we propose a novel loss function consisting of an image consistency metric defined in the original image space and a content consistency metric defined in the content space.

II. METHOD

Given two real-valued functions x and y from different modalities $\mathcal{X}$ and $\mathcal{Y}$ defined on the background space $\Omega \in \mathbb{R}^{H\times W}$, they respectively represent a 2D grayscale moving image and a 2D grayscale target image. Our goal

*This work was supported by the Scientific Research Project of Hunan University of Arts and Science (20ZD01)

*Correspondence at wujiong@huas.edu.cn

[1] School of Computer and Electrical Engineering, Hunan University of Arts and Science, Hunan, China.

[2] Furong College, Hunan University of Arts and Science, Hunan, China.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

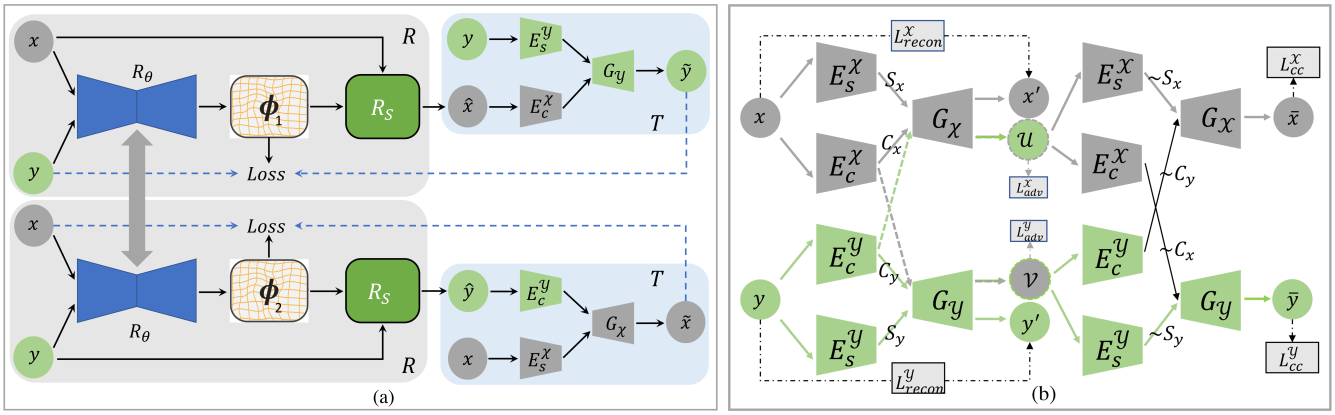

Fig. 1. (a) Overview of the proposed disentangled representations based unsupervised deformable cross-modality registration framework. It consists of two sub-frameworks, with the top one aligns image x to y and the bottom one aligns y to x. Each sub-framework consists of a registration network R and an image-to-image translation network T. All components of T come from the representation disentangling model (RDM). (b) Overview architecture of the RDM.

is to find an optimal mapping ϕ such that $x \circ \phi$ is well aligned to y. We achieve this by constructing a disentangled representations based unsupervised deformable framework and converting the cross-modality registration into intra-modality registration. The proposed framework, as shown in Fig. 1 (a), consists of two sub-framworks with the top one aligns image x to y and the bottom one aligns y to x. Each sub-framework consists of two components: (i) a registration network R and (ii) an image-to-image translation network T. Parameters of R are shared in these two networks and trained simultaneously. All components of T come from a pre-trained RDM, the architecture of RDM is depicted in Fig. 1 (b).

A. Registration Network

Registration network R is composed of a fully convolutional neural network R_θ and a re-sampling layer R_S. R_θ takes x and y as inputs, and outputs deformation fields $\phi_1 = R_\theta(x, y)$ or $\phi_2 = R_\theta(y, x)$. The fields are $H \times W$ matrices of 2D vectors which indicate the directions and displacements of pixels in the moving images. The architecture of R_θ has an encoder-decoder structure with skip connections. Specifically, each layer of R_θ adopts 2D convolutions followed by Leaky ReLU activations. To fully learn the relationship between image pairs and transformation fields, small kernels of sizes 3×3 and 2×2 are used alternatively.

In terms of the re-sampling layer R_S, taking the top registration network for example, we compute the location $\phi(\mathbf{q})$ of each pixel $\mathbf{q}$ in moving image x. The intensity value of $x(\phi(\mathbf{q}))$ is calculated using the neighboring pixels of it. Let $\delta(\phi(\mathbf{q}))$ be a set of the neighboring pixels, the intensity value of $x\left(\phi(\mathbf{q})\right)$ can be computed as

$$x(\phi(\mathbf{q})) = \sum_{\mathbf{u}\in\delta(\phi(\mathbf{q}))} x(\mathbf{u}) \prod_{d\in\{X,Y\}} \left(1 - |\phi^d(\mathbf{u}) - x^d|\right), \quad (1)$$

where X and Y denote two directions of the coordinate system.

B. Image-to-image Translation Network

As showed in Fig. 1 (a), the image-to-image translation network T is connected to the registration network and used to translate the transformed image $\hat{x}$ or $\hat{y}$ into $\tilde{y}$ or $\tilde{x}$. The translation processes can be calculated as

$$\tilde{x} = G_{\mathcal{X}}\left(E_c^{\mathcal{Y}}(\hat{y}), E_s^{\mathcal{X}}(x)\right), \tilde{y} = G_{\mathcal{Y}}\left(E_c^{\mathcal{X}}(\hat{x}), E_s^{\mathcal{Y}}(y)\right), \quad (2)$$

where $E_s^{\mathcal{X}}$ and $E_s^{\mathcal{Y}}$ denote style encoders, $E_c^{\mathcal{X}}$ and $E_c^{\mathcal{Y}}$ denote content encoders, $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ denote generators. All of the encoders and generators are pre-trained using RDM, as shown in Fig. 1 (b), based on the assumption that images of different modalities can be disentangled into content codes in a domain-invariant content space and style codes in different domain-specific style spaces, the images of different modalities can be translated by swapping the style codes.

To train the RDM, loss function is defined as a weighted sum of in-domain reconstruction loss $\mathcal{L}_{recon}$, cross-domain translation loss $\mathcal{L}_{adv}$, latent space reconstruction loss $\mathcal{L}_{lat}$ and cross-cycle consistency loss $\mathcal{L}_{cc}$, i.e.,

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{recon} + \lambda_{lat}\mathcal{L}_{lat} + \lambda_{cc}\mathcal{L}_{cc} + \lambda_{adv}\mathcal{L}_{adv}, \quad (3)$$

where $\mathcal{L}_{recon} = \mathcal{L}_{recon}^{\mathcal{X}} + \mathcal{L}_{recon}^{\mathcal{Y}}$, $\mathcal{L}_{lat} = \mathcal{L}_{lat}^{C_{\mathcal{X}}} + \mathcal{L}_{lat}^{S_{\mathcal{X}}} + \mathcal{L}_{lat}^{C_{\mathcal{Y}}} + \mathcal{L}_{lat}^{S_{\mathcal{Y}}}$ and $\mathcal{L}_{adv} = \mathcal{L}_{adv}^{\mathcal{X}\to\mathcal{Y}} + \mathcal{L}_{adv}^{\mathcal{Y}\to\mathcal{X}}$. $\mathcal{L}_{recon}^{\mathcal{X}}$ and $\mathcal{L}_{recon}^{\mathcal{Y}}$ are used to evaluate the dissimilarity between reconstructed images and original images, and respectively defined as

$$\mathcal{L}_{recon}^{\mathcal{X}} = \mathbb{E}_{x\sim\mathcal{X}} \left\| G_{\mathcal{X}}\left(E_c^{\mathcal{X}}(x), E_s^{\mathcal{X}}(x)\right) - x \right\|_1, \quad (4)$$

and

$$\mathcal{L}_{recon}^{\mathcal{Y}} = \mathbb{E}_{y\sim\mathcal{Y}} \left\| G_{\mathcal{Y}}\left(E_c^{\mathcal{Y}}(y), E_s^{\mathcal{Y}}(y)\right) - y \right\|_1. \quad (5)$$

$\mathcal{L}_{lat}^{C_{\mathcal{X}}}$ and $\mathcal{L}_{lat}^{S_{\mathcal{Y}}}$ are respectively calculated as

$$\mathcal{L}_{lat}^{C_{\mathcal{X}}} = \left\| E_c^{\mathcal{Y}}\left(G_{\mathcal{Y}}(C_x, S_y)\right) - C_x \right\|_1, \quad (6)$$

and

$$\mathcal{L}_{lat}^{S_{\mathcal{Y}}} = \left\| E_S^{\mathcal{Y}}\left(G_{\mathcal{Y}}(C_x, S_y)\right) - S_y \right\|_1, \quad (7)$$

where $C_x = E_c^{\mathcal{X}}(x)$ denotes the content codes and $S_y = E_s^{\mathcal{Y}}(y)$ denotes the style codes.

To better preserve the content information, a cross-cycle consistency loss $\mathcal{L}_{cc}$ is introduced with the following form,

$$\begin{aligned}\mathcal{L}_{cc} = \mathcal{L}_{cc}^{\mathcal{X}} + \mathcal{L}_{cc}^{\mathcal{Y}} =& \mathbb{E}_{x\sim\mathcal{X}} \left\| G_{\mathcal{X}}\left(E_c^{\mathcal{Y}}\left(v\right), E_s^{\mathcal{X}}\left(u\right)\right) - x \right\|_1 + \\ & \mathbb{E}_{y\sim\mathcal{Y}} \left\| G_{\mathcal{Y}}\left(E_c^{\mathcal{X}}\left(u\right), E_s^{\mathcal{Y}}\left(v\right)\right) - y \right\|_1,\end{aligned} \tag{8}$$

where $u = G_{\mathcal{X}}(C_y, S_x)$ and $v = G_{\mathcal{Y}}(C_x, S_y)$. The adversarial losses $\mathcal{L}_{adv}^{\mathcal{X}\to\mathcal{Y}}$ and $\mathcal{L}_{adv}^{\mathcal{Y}\to\mathcal{X}}$ are adopted to match the distributions of translated images and the distributions in the target domains. For instance, the loss $\mathcal{L}_{adv}^{\mathcal{X}\to\mathcal{Y}}$ is calculated as

$$\begin{aligned}\mathcal{L}_{adv}^{\mathcal{X}\to\mathcal{Y}} =& \mathbb{E}_{C_{\mathcal{X}}\sim p(C_{\mathcal{X}}), S_{\mathcal{X}}\sim p(S_{\mathcal{X}})}\left[log\left(1 - D_{\mathcal{Y}}(x_{\mathcal{X}\to\mathcal{Y}})\right)\right] + \\ & \mathbb{E}_{y\sim\mathcal{Y}}\left[log\left(D_{\mathcal{Y}}(y)\right)\right],\end{aligned} \tag{9}$$

where $p(C_x)$ and $p(S_x)$ respectively denote the distribution of C_y and the distribution of S_x.

C. Training Losses

After training the RDM, the encoders and generators including $E_c^{\mathcal{X}}$, $E_s^{\mathcal{X}}$, $E_c^{\mathcal{Y}}$, $E_s^{\mathcal{Y}}$, $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ are obtained and concatenated to the registration network. Due to network T is capable of translating an image in domain $\mathcal{X}$ or $\mathcal{Y}$ into an image in domain $\mathcal{Y}$ or $\mathcal{X}$, intra-modality metrics can be adopted to train the registration network. In this work, we use the MSD to evaluate the dissimilarities between moving images and translated transformed images:

$$MSD = \frac{1}{2|\Omega|}\sum_{\mathbf{q}\in\Omega}[\tilde{x}(\mathbf{q}) - x(\mathbf{q})]^2 + [\tilde{y}(\mathbf{q}) - y(\mathbf{q})]^2. \tag{10}$$

In addition, to improve the learning ability and accelerate the convergence of the registration network, we introduce a content consistency metric in the loss function. It bases on the intuition that the content information of the transformed moving images should be equal to the content information of the target images. Therefore, the content consistency metric L_c is defined as

$$L_c = ||E_c^{\mathcal{X}}(\hat{x}) - E_c^{\mathcal{Y}}(y)||_1 + ||E_c^{\mathcal{Y}}(\hat{y}) - E_c^{\mathcal{X}}(x)||_1. \tag{11}$$

After introducing deformation fields smoothing loss $L_{smooth}(\phi_i) = \sum_{\mathbf{q}\in\Omega}||\nabla\phi_i(\mathbf{q})||^2$, $(i = 1, 2)$ and $\nabla\phi_i(\mathbf{q}) = \left(\frac{\partial\phi_i(\mathbf{q})}{\partial X}, \frac{\partial\phi_i(\mathbf{q})}{\partial Y}\right)$, the overall loss function L of the registration network is defined as

$$L = \gamma_1 L_{smooth}(\phi_1) + \gamma_2 L_{smooth}(\phi_2) + \gamma_3 L_c + \gamma_4 MSD. \tag{12}$$

D. Implementation Detail

The RDM is built based on the work presented in [12] and pre-trained using paired T1w and T2w images. We set $\lambda_{rec} = 10$, $\lambda_{lat} = 1$, $\lambda_{cc} = 10$, $\lambda_{adv} = 1$, $\gamma_{rec} = 20$, $\gamma_{lat} = 10$, $\gamma_{cc} = 20$ and $\gamma_{adv} = 1$ in our experiments. Then, all parameters of the $E_s^{\mathcal{X}}$, $E_s^{\mathcal{Y}}$, $E_c^{\mathcal{X}}$, $E_c^{\mathcal{Y}}$, $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ are fixed and used in the image-to-image translation network T. To train the registration network R, we empirically set $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 1$ and $\gamma_4 = 5$. Our networks are implemented in PyTorch using Adam as optimizer with the learning rates of 1×10^{-3} for training RDM and 1×10^{-4} for training R.

III. Experimental Results and Evaluation

A. Dataset and Evaluation Metric

We use two datasets to evaluate the proposed method. The first dataset is public available, know as IXI dataset[1], consists of almost 600 subjects and for each subject T1w and T2w paired images were collected (image size: $256 \times 256 \times 150$ mm^3). The second dataset consists of 16 T1w and T2w image pairs[2] (image size: $190\times230\times180$ mm^3) and for each image 12 subcortical structures were manually delineated. We performed standard sequentially steps for the first dataset including spatially adaptive non-local denoising, N4 bias correction, and skull-stripping to extract brains [13].

Then, these two datasets were affinely aligned to the MNI512 space. We extracted a total of 6000 paired 2D slices on the transverse plane from 300 randomly selected image pairs of the first dataset to train the networks T and R (5400 for training and 600 for validation). 176 paired 2D slices on the transverse plane of the second dataset were extracted and used as a testing dataset. The dice similarity coefficient (DSC) was adopted to quantify the registration accuracy on six structures including the bilateral caudate, bilateral putamen, and bilateral thalamus. Wilcoxon signed-rank tests were performed to quantify the significance of all group comparison differences.

B. Results and Discussion

To demonstrate the performance of our method, we conducted the comparison experiments with other two conventional state-of-the-art registration algorithms including symmetric diffeomorphic image registration (SyN) [14] and Elastix [15]. Mutual information was used as the cost function in these two algorithms. Testing experiments were performed by registering 176 T1w slices to 176 T2w slices. Group comparison results on six structures are listed in Table I. Evidently, tpahe registration accuracies for six structures are superior to the SyN and the Elastix.

Specifically, for all of the six structures, the DSCs of our proposed method are significantly higher than those of Elastix with p-value$<$ 1.65×10^{-3}. Also, compared the results of our method to those of SyN, the DSCs of our method are significantly higher (p-value$<$ 7.11×10^{-23}). In addition, the mean DSC calculated across all six structures, our method is significantly higher than those of SyN and Elastix with p-values of 5.14×10^{-14} and 4.53×10^{-30}. To further demonstrate the ability of our framework in registering T2w slices to T1w slices, we used the same trained network R to predict the deformation fields. The statistic of DSCs from the registration results of the three methods is reported in Fig. 2, which also shows that our approach

[1] http://brain-development.org/ixi-dataset/

[2] https://www.predict-hd.net/

TABLE I

THE MEAN DSC FOR EACH OF THE SIX STRUCTURES AS WELL AS THE MEAN DSC ACROSS ALL SIX STRUCTURES OBTAINED FROM SYN, ELASTIX AND OUR METHOD. BOLD FONT INDICATES STATISTICALLY SIGNIFICANT GROUP DIFFERENCE.

	SyN	Elastix	Our method
R Caudate	0.640±0.228	0.815±0.124	**0.847±0.051**
L Caudate	0.653±0.207	0.829±0.107	**0.863±0.058**
R Putamen	0.508±0.303	0.764±0.187	**0.868±0.052**
L Putamen	0.553±0.278	0.742±0.169	**0.845±0.054**
R Thalamus	0.742±0.105	0.843±0.076	**0.889±0.051**
L Thalamus	0.771±0.136	0.840±0.095	**0.890±0.061**
mean	0.645±0.179	0.806±0.107	**0.867±0.035**

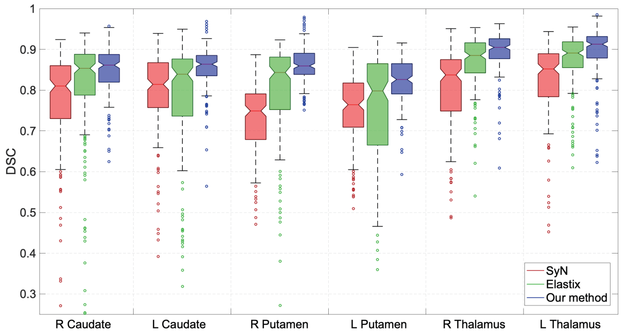

Fig. 2. A comparison of the mean DSCs of the three different methods on the six brain structures.

achieved the highest registration accuracies. The superior registration accuracies and the same network simultaneously realizing the registration from T1w to T2w and T2w to T1w demonstrated the effectiveness of the proposed image-to-image translation network and the proposed training loss function. In terms of the mean run time of these 352 registrations, our method is much less than SyN and Elastix (0.07s for our method, 6.93s for SyN, and 2.81s for Elastix).

Registration results of the three methods on two representative images, as well as the moving images and target images, are illustrated in Fig. 3. We can note that the registration results of our method are the closest to the target images with the most local anatomical details preserved. A potential reason is that the image-to-image translation network converted the cross-modality registration problem into intra-modality one, makes the intra-modality metric MSD can be introduced to capture more local anatomical details. Furthermore, the content consistency metric enhanced the consistencies between the transformed images and the target images resulting in better registration results. For future work, we will further extend our method to 3D MRIs registration.

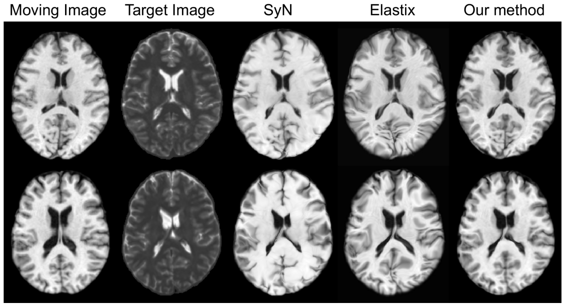

Fig. 3. Visualization results of the proposed method against baseline methods.

IV. CONCLUSION

In this paper, we proposed a novel disentangled representations based unsupervised deformable framework for cross-modality MRI registration. Experimental results showed the superiority of our proposed framework against other conventional approaches in terms of both accuracy and speed. This work provides new insight for other cross-modality image registration tasks.

REFERENCES

[1] Z. Tang, P.-T. Yap, and D. Shen, "A new multi-atlas registration framework for multimodal pathological images using conventional monomodal normal atlases," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2293–2304, 2018.

[2] F. Maes, A. Collignon, *et al.*, "Multimodality image registration by maximization of mutual information," *IEEE transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[3] N. D. Cahill, J. A. Schnabel, *et al.*, "Revisiting overlap invariance in medical image alignment," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2008, pp. 1–8.

[4] S. Roy, A. Carass, *et al.*, "Mr to ct registration of brains using image synthesis," in *Medical Imaging 2014: Image Processing*, vol. 9034. International Society for Optics and Photonics, 2014, p. 903419.

[5] C. Wachinger and N. Navab, "Entropy and laplacian images: Structural representations for multi-modal registration," *Medical image analysis*, vol. 16, no. 1, pp. 1–17, 2012.

[6] J. E. Iglesias, E. Konukoglu, *et al.*, "Is synthesizing mri contrast useful for inter-modality analysis?" in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 631–638.

[7] M. Chen, A. Carass, *et al.*, "Cross contrast multi-channel image registration using image synthesis for mr brain images," *Medical image analysis*, vol. 36, pp. 2–14, 2017.

[8] A. Creswell, T. White, *et al.*, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[9] Q. Yang, N. Li, *et al.*, "Mri cross-modality image-to-image translation," *Scientific reports*, vol. 10, no. 1, pp. 1–18, 2020.

[10] R. Han, C. K. Jones, *et al.*, "Deformable mr-ct image registration using an unsupervised end-to-end synthesis and registration network for endoscopic neurosurgery," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 11598. International Society for Optics and Photonics, 2021, p. 1159819.

[11] C. Qin, B. Shi, *et al.*, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 249–261.

[12] X. Huang, M.-Y. Liu, *et al.*, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.

[13] J. Wu, Y. Zhang, and X. Tang, "Simultaneous tissue classification and lateral ventricle segmentation via a 2d u-net driven by a 3d fully convolutional neural network," in *EMBC 2019*. IEEE, 2019, pp. 5928–5931.

[14] B. B. Avants, C. L. Epstein, *et al.*, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[15] S. Klein, M. Staring, *et al.*, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.