TarGAN: Target-Aware Generative Adversarial Networks for Multi-modality Medical Image Translation

Junxiao Chen[1], Jia Wei[1](✉), and Rui Li[2]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
cs_xiao@mail.scut.edu.cn, csjwei@scut.edu.cn
[2]Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY 14623
rxlics@rit.edu

Abstract. Paired multi-modality medical images, can provide complementary information to help physicians make more reasonable decisions than single modality medical images. But they are difficult to generate due to multiple factors in practice (e.g., time, cost, radiation dose). To address these problems, multi-modality medical image translation has aroused increasing research interest recently. However, the existing works mainly focus on translation effect of a whole image instead of a critical target area or Region of Interest (ROI), e.g., organ and so on. This leads to poor-quality translation of the localized target area which becomes blurry, deformed or even with extra unreasonable textures. In this paper, we propose a novel target-aware generative adversarial network called **TarGAN**, which is a generic multi-modality medical image translation model capable of (1) learning multi-modality medical image translation without relying on paired data, (2) enhancing quality of target area generation with the help of target area labels. The generator of TarGAN jointly learns mapping at two levels simultaneously — whole image translation mapping and target area translation mapping. These two mappings are interrelated through a proposed crossing loss. The experiments on both quantitative measures and qualitative evaluations demonstrate that TarGAN outperforms the state-of-the-art methods in all cases. Subsequent segmentation task is conducted to demonstrate effectiveness of synthetic images generated by TarGAN in a real-world application. Our code is available at https://github.com/2165998/TarGAN.

Keywords: Multi-modality translation · GAN · Abdominal organs.

1 Introduction

Medical imaging, a powerful diagnostic and research tool creating visual representations of anatomy, has been widely available for disease diagnosis and surgery planning [2]. In current clinical practice, Computed Tomography (CT)

and Magnetic Resonance Imaging (MRI) are most commonly used. Since CT and multiple MR imaging modalities provide complementary information, an effective integration of these different modalities can help physicians make more informative decisions.

Since it is difficult and costly to obtain paired multi-modality images in clinical practice, there is a growing demand for developing multi-modality image translations to assist clinical diagnosis and treatment [17].

Existing works can be categorized into two types. One is crossing-modality medical image translation between two modalities, which has scalability issues to the increasing number of modalities [18,19], since these methods have to train $n(n-1)$ generator models in order to learn all mappings between n modalities. The other is multi-modality image translation [1,7,16,17]. In this category, some methods [7,17] rely on paired data, which is hard to acquire in clinical reality. Other methods [1,16] can learn from unpaired data, however, they tend to lead to deformation in target area without prior knowledge, as concluded by Zhang *et al.* [19]. As demonstrated in Figure 1, the state-of-the-art multi-modality image translation methods give rise to poor quality local translations. The translated target area (For example, Liver, in red curves) is blurry, deformed or perturbed with redundant unreasonable textures. Comparing to them, our method can not only perform whole image translation in competitive quality but also achieve significantly better local translation for the target area.

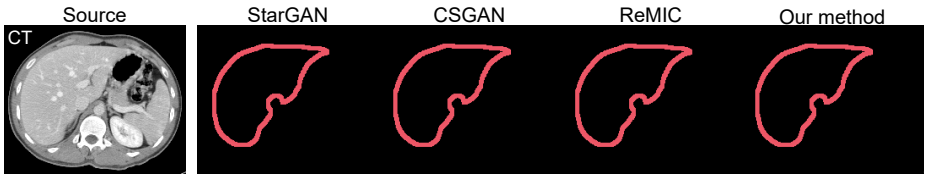

Fig. 1. Translation results (CT to T1w) of different methods are shown here. The target area (i.e., liver) is contoured in red.

To address the above issues, we present a novel unified general-purpose multi-modality medical image translation method named "Target-Aware Generative Adversarial Networks" (TarGAN). We incorporate target labels to enable the generator to focus on local translation of target area. The generator has two input-output streams. One stream translates a whole image from source modality to target modality, the other focuses on translating a target area. In particular, we combine the cycle-consistency loss [21] and the backbone of StarGAN [1] to learn the generator, which enables our model to scale up to modality increase without relying on paired data. Then, the untraceable constraint [20] is employed to further improve translation quality of synthetic images. To avoid the deformation of output images caused by untraceable constraint, we construct a shape-consistency loss [3] with an auxiliary network, namely shape controller. We further propose a novel crossing loss to allow the generator to focus on the

target area when translating the whole image to target modality. Trained in an end-to-end fashion, TarGAN can not only accomplish multi-modality translation but also properly retain the target area information in the synthetic images.

Overall, the contributions of this work are: (1) We propose TarGAN to generate multi-modality medical images with high-quality local translation on target areas by integrating global and local mappings with a crossing loss. (2) We show qualitative and quantitative performance evaluations on multi-modality medical image translation tasks with CHAOS2019 dataset [12], demonstrating our method's superiority over the state-of-the-art methods. (3) We further use the synthetic images generated from TarGAN to improve the performance of a segmentation task, which indicates that the synthetic images generated by TarGAN achieve the improvement by enriching the information of source images.

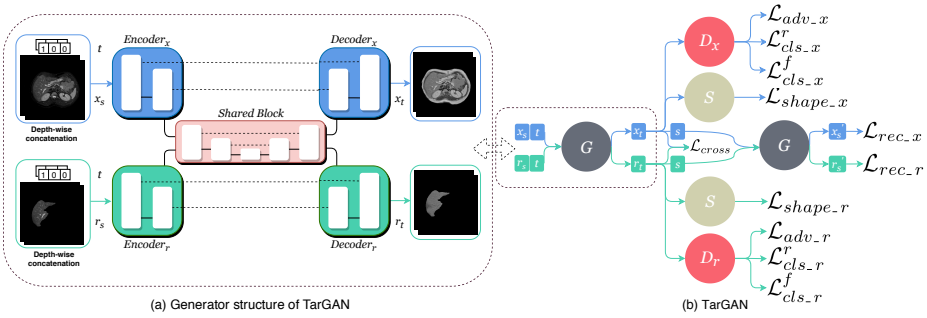

Fig. 2. The illustration of TarGAN. As in **(b)**, TarGAN consists of four modules (G, S, D_x, D_r). The generator G translates a source whole image x_s and a source target area image r_s to a target whole image x_t and a target area image r_t. The detailed structure of G is shown in **(a)**. The shape controller S preserves the invariance of anatomy structures. The discriminators D_x and D_r distinguish whether a whole image and its target area are real or fake and determine which modalities the source images come from.

2 Methods

2.1 Proposed framework

Given an image x_s from source modality s and its corresponding target area label y, we specify a target area image r_s which only contains the target area by binarization operation $y \cdot x_s$. Given any target modality t, our goal is to train a single generator G that can translate any input image x_s of source modality s to the corresponding output image x_t of target modality t, and translate the input

target area image r_s of source modality s to the corresponding output target area image r_t of target modality t simultaneously, denoted as $G(x_s, r_s, t) \rightarrow (x_t, r_t)$. Figure 2 illustrates the architecture of TarGAN, which is composed of four modules described below.

To achieve the aforementioned goal, we design a double input-output streams **generator** G consisting of a shared middle block and two pairs of encoder-decoder. Combining with the shared middle block, both encoder-decoder pairs translate an input image into an output image of the target modality t. One stream's input is the whole image x_s, and the other's input only includes the target area r_s. The shared middle block is designed to implicitly enable G to focus on target area in whole image translation. Note that target area label y of x_s is not available in test phase, so the input block $Encoder_r$ and output block $Decoder_r$ are not used at that time.

Given a synthetic image x_t or r_t from G, **the shape controller** S generates a binary mask which can represent the foreground area of the synthetic image.

Lastly, we use **two discriminators denoted as D_x and D_r** corresponding to two output streams of G. The probability distributions inferred by D_x distinguish whether the whole image is real or fake, and determine which modality the whole image comes from. Similarly, the D_r distinguish whether the target area image is real or fake, and to determine which modality the target area image comes from.

2.2 Training objectives

Adversarial loss. To minimize the difference between the distributions of generated images and real images, we define the adversarial loss as

$$\begin{aligned} \mathcal{L}_{adv\_x} &= \mathbb{E}_{x_s}[\log D_{src\_x}(x_s)] + \mathbb{E}_{x_t}[\log(1 - D_{src\_x}(x_t))], \\ \mathcal{L}_{adv\_r} &= \mathbb{E}_{r_s}[\log D_{src\_r}(r_s)] + \mathbb{E}_{r_t}[\log(1 - D_{src\_r}(r_t))]. \end{aligned} \tag{1}$$

Here, D_{src_x} and D_{src_r} represent the probability distributions of real or fake over input whole images and target area images.

Modality classification loss. To assign the generated image to their target modality t, we impose the modality classification loss on G, D_x and D_r. The loss consists of two terms: modality classification loss of real images which is used to optimize D_x and D_r, denoted as $\mathcal{L}^r_{cls\_(x/r)}$, and modality classification loss of fake images which is used to optimize G, denoted as $\mathcal{L}^f_{cls\_(x/r)}$. In addition, to eliminate synthetic images' style features from source modalities, the untraceable constraint [20] is combined into $\mathcal{L}^r_{cls\_(x/r)}$ as:

$$\begin{aligned} \mathcal{L}^r_{cls\_x} &= \mathbb{E}_{x_s,s}[-\log D_{cls\_x}(s|x_s)] + \lambda_u \mathbb{E}_{x_t,s'}[-\log D_{cls\_x}(s'|x_t)], \\ \mathcal{L}^r_{cls\_r} &= \mathbb{E}_{r_s,s}[-\log D_{cls\_r}(s|r_s)] + \lambda_u \mathbb{E}_{r_t,s'}[-\log D_{cls\_r}(s'|r_t)]. \end{aligned} \tag{2}$$

Here, D_{cls_x} and D_{cls_r} represent the probability distributions over modality labels and input images. s' indicates whether an input image is fake, and is

translated from a source modality s [20]. Besides, we define $\mathcal{L}^f_{cls\_(x/r)}$ as

$$\mathcal{L}^f_{cls\_x} = \mathbb{E}_{x_t,t}[-\log D_{cls\_x}(t|x_t)],\ \mathcal{L}^f_{cls\_r} = \mathbb{E}_{r_t,t}[-\log D_{cls\_r}(t|r_t)]. \tag{3}$$

Shape consistency loss. Since the untraceable constraint can affect the shape of anatomy structures in synthetic images by causing structure deformation, we correct it by adding a shape consistency loss [3] to G with shape controller S as

$$\mathcal{L}_{shape\_x} = \mathbb{E}_{x_t,b^x}[||b^x - S(x_t)||_2^2],\ \mathcal{L}_{shape\_r} = \mathbb{E}_{r_t,b^r}[||b^r - S(r_t)||_2^2], \tag{4}$$

where b^x and b^r are the binarizations (with 1 indicating foreground pixels and 0 otherwise) of x_s and r_s. S constrains G to focus on the multi-modality mapping in a content area.

Reconstruction loss. To allow G to preserve the modality-invariant characteristics of the whole image x_s and its target area image r_s, we employ a cycle consistency loss [21] as

$$\mathcal{L}_{rec\_x} = \mathbb{E}_{x_s,x'_s}[||x_s - x'_s||_1],\ \mathcal{L}_{rec\_r} = \mathbb{E}_{r_s,r'_s}[||r_s - r'_s||_1]. \tag{5}$$

Note that x'_s and r'_s are from $G(x_t, r_t, s)$. Given the paired synthetic image (x_t, r_t) and the source modality s, G tries to reconstruct the input images (x_s, r_s).

Crossing loss. To enforce G to focus on a target area when generating a whole image x_t, we directly regularize G with a crossing loss defined as

$$\mathcal{L}_{cross} = \mathbb{E}_{x_t,r_t,y}[||x_t \cdot y - r_t||_1], \tag{6}$$

where y is the target area label corresponding to x_s. By minimizing the crossing loss, G can jointly learn from double input-output streams and share information between them.

Complete objective. By combining the proposed losses together, our complete objective functions are as follows:

$$\mathcal{L}_{D_{(x/r)}} = -\mathcal{L}_{adv\_(x/r)} + \lambda^r_{cls}\,\mathcal{L}^r_{cls\_(x/r)}, \tag{7}$$

$$\mathcal{L}_G = \mathcal{L}_{adv\_(x/r)} + \lambda^f_{cls}\,\mathcal{L}^f_{cls\_(x/r)} + \lambda_{rec}\,\mathcal{L}_{rec\_(x/r)} + \lambda_{cross}\,\mathcal{L}_{crossing}, \tag{8}$$

$$\mathcal{L}_{G,S} = \mathcal{L}_{shape\_(x/r)}, \tag{9}$$

where λ^r_{cls}, λ^f_{cls}, λ_{rec}, λ_{cross} and λ_u (Eqs. (2)) are hyperparameters to control the relative importance of each loss.

Table 1. Quantitative evaluations on synthetic images of different methods. (↑ denotes higher is better, while ↓ denotes lower is better)

Method	FID↓			S-score(%)↑		
	CT	T1w	T2w	CT	T1w	T2w
StarGAN [1]	0.0488	0.1179	0.2615	42.89	29.23	42.17
CSGAN [19]	0.0484	0.1396	0.4819	56.72	45.67	69.09
ReMIC [16]	0.0912	0.1151	0.5925	51.03	32.00	69.58
Our method	**0.0418**	**0.0985**	**0.2431**	**57.13**	**65.79**	**69.63**

3 Experiments and results

3.1 Settings

Dataset. We use 20 patients' data in each modality (CT, T1-weighted and T2-weighted). They are from the Combined Healthy Abdominal Organ Segmentation (CHAOS) Challenge [11]. Detailed imaging parameters are shown in supplementary material. We resize all slices as 256×256 uniformly. 50% data from each modality are randomly selected as training data, while the rest as test data. Because CT scans only have liver labels, we set liver as the target area.

Baseline methods. Translation results comparisons are conducted against the state-of-the-art translation methods, StarGAN [1], CSGAN [19] and ReMIC [16]. Note that we implement an unsupervised ReMIC because of the lack of ground-truth images.

Target segmentation performances are also evaluated against the above methods. We train and test models using only real images of each modality, denoted as **Single**. We use the mean results of two segmentation models of each modality from CSGAN and use the segmentation model G_s from ReMIC. As for StarGAN and TarGAN, inspired by *'image enrichment'* [5], we extend every single modality to multiple modalities and concatenate multiple modalities within each sample, as [CT] → [CT, synthetic T1w, synthetic T2w].

Evaluation metrics. In the translation tasks, due to the lack of ground-truth images, we can not use the common metrics like PSNR, SSIM, etc. So we evaluate both the visual quality and the integrity of target area structures of generated images using Frechét inception distance (**FID**) [6] and segmentation score (**S-score**) [19]. We compute FID and S-score for each modality and report their average values. The details on above metrics are further described in supplementary material.

In the segmentation tasks, dice coefficient (**DICE**) and relative absolute volume difference (**RAVD**) are used as metrics. We compute each metric on every modality, and report their average values and standard deviations.

Implementation details. We use U-net [15] as the backbone of G and S. In G, only half of the channels are used for every skip connection. As for D_x and D_r, we implement the backbone with PatchGAN [9]. Details of above networks are included in the supplementary material. All the liver segmentation experiments are conducted with nnU-Net [8] except CSGAN and ReMIC.

To stabilize the training process, we adopt Wasserstein GAN loss with a gradient penalty [4,14] using $\lambda_{gp} = 10$ and two-timescale update rule (TTUR) [6] for G and D. The learning rates for G, S are set to 10^{-4}, while that of D is set to 3×10^{-4}. We set $\lambda^r_{cls} = 1$, $\lambda^f_{cls} = 1$, $\lambda_{rec} = 1$, $\lambda_{cross} = 50$ and $\lambda_u = 0.01$. The batch size and training epoch are set to 4 and 50, respectively. We use the Adam optimizer [13] with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$. All images are normalized to $[-1, 1]$ prior to the training and test. We use exponential moving averages over parameters [10] of G during test, with a decay of 0.999. Our implementation is trained on an NVIDIA GTX 2080Ti with PyTorch.

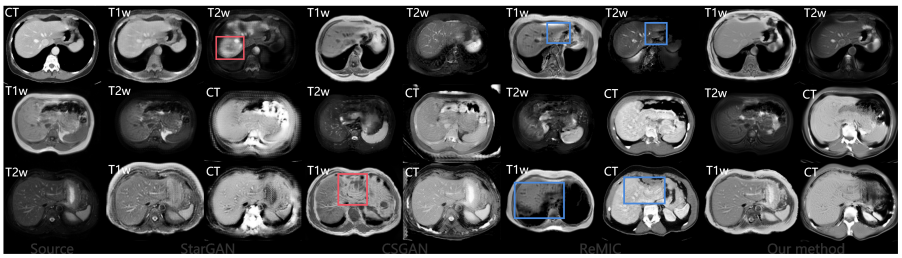

Fig. 3. Multi-modality medical image translation results. Red boxes highlight the redundant textures, and blue boxes indicate the deformed structures.

3.2 Results and analyses

Image Translation. Figure 3 shows qualitative results on each pair of modal image translation. As shown, StarGAN fails to translate image from CT to T1w and produces many artifacts in MRI to CT translation. CSGAN sometimes adds redundant textures (marked by the red boxes) in the target area while retaining the shape of target. ReMIC tends to generate relatively realistic synthetic images while deforming the structure of target area in most cases (marked by the blue boxes). Comparing to above methods, TarGAN generates translation results in higher visual quality and properly preserves the target structures. Facilitated by the proposed crossing loss, TarGAN can jointly learn the mappings of the target area and the whole image among different modalities, and further make G focus on the target areas to improve their quality. Furthermore, as shown in Table 1, TarGAN outperforms all the baselines in terms of FID and S-score, which suggests TarGAN produces the most realistic medical images, and the target area integrity of synthetic images derived from TarGAN is significantly better.

Table 2. Liver segmentation results (mean ± standard deviation) on different medical modalities.

Method	DICE(%)↑			RAVD(%)↓		
	CT	T1w	T2w	CT	T1w	T2w
Single	96.29±0.74	93.53±2.43	89.24±8.18	3.31±1.80	3.81±3.49	11.68±14.37
StarGAN [1]	96.65±0.34	92.71±1.66	86.38±4.95	3.07±1.53	5.40±2.87	15.71±9.85
CSGAN [19]	96.08±2.05	87.47±5.97	86.35±6.29	4.47±3.94	15.74±14.18	**8.23±8.55**
ReMIC [16]	93.81±1.43	86.33±8.50	82.70±4.36	5.33±3.55	8.06±8.80	10.62±6.80
Our method	**97.06±0.62**	**94.02±2.00**	**90.94±6.28**	**2.33±1.60**	**3.50±1.82**	9.92±11.17

Liver segmentation. The quantitative segmentation results are shown in Table 2. Our method achieves better performance than all other methods on most of the metrics. This suggests TarGAN can not only generate realistic images for every modality, but also properly retain liver structure in synthetic images. The high-quality local translation for the target areas plays a key role in the improvement of liver segmentation performance. By jointly learning from real and synthetic images, the segmentation models can incorporate more information on the liver areas within each sample.

Ablation test. We conduct an ablation test to validate effectiveness of different parts of TarGAN in terms of preserving target area information. For ease of presentation, we denote **shape controller**, **target area translation mapping** and **crossing loss** as **S**, **T** and **C**, respectively. As shown in Table 3, **TarGAN without (w/o) S, T, C** is closely similar to StarGAN except using our implementation. The proposed crossing loss plays a key role in TarGAN, which increases the mean of S-score from **TarGAN w/o C** 51.03% to 64.18%.

Table 3. Ablation study on different components of TarGAN. Note that **TarGAN w/o S, T** and **TarGAN w/o T** don't exist, since **T** is the premise of **C**.

Method	S-score(%)			
	CT	T1w	T2w	Mean
TarGAN w/o S, T, C	30.64	35.05	67.45	44.38
TarGAN w/o S, C	39.78	29.96	67.47	45.74
TarGAN w/o T, C	37.42	38.33	68.85	48.20
TarGAN w/o C	43.00	38.83	71.27	51.03
TarGAN w/o S	56.69	59.37	**71.89**	62.65
TarGAN	**57.13**	**65.79**	69.63	**64.18**

4 Conclusion

In this paper, we propose a novel general-purpose method TarGAN to mainly address two challenges in multi-modality medical image translation: learning multi-modality medical image translation without relying on paired data, and improving the quality of local translation on target area. A novel translation mapping mechanism is introduced to enhance the target area quality during generating the whole image. Additionally, by using the shape controller to alleviate the deformation problem caused by the untraceable constraint and combining a novel crossing loss in generator G, TarGAN addresses both challenges within a unified framework. Both the quantitative and qualitative evaluations show the superiority of TarGAN in comparison with the state-of-the-art methods. We further conduct a segmentation task to demonstrate effectiveness of synthetic images generated by TarGAN in a real application.

Acknowledgments. This work is supported in part by the Natural Science Foundation of Guangdong Province (2017A030313358, 2017A030313355, 2020A1515010717), the Guangzhou Science and Technology Planning Project (201704030051), the Fundamental Research Funds for the Central Universities (2019MS073), NSF-1850492 (to R.L.) and NSF-2045804 (to R.L.).

References

1. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
2. Ernst, P., Hille, G., Hansen, C., Tönnies, K., Rak, M.: A cnn-based framework for statistical assessment of spinal shape and curvature in whole-body mri images of large populations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–11. Springer (2019)
3. Fu, C., Lee, S., Joon Ho, D., Han, S., Salama, P., Dunn, K.W., Delp, E.J.: Three dimensional fluorescence microscopy image synthesis and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2221–2229 (2018)
4. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
5. Gupta, L., Klinkhammer, B.M., Boor, P., Merhof, D., Gadermayr, M.: Gan-based image enrichment in digital pathology boosts segmentation accuracy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 631–639. Springer (2019)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017)

7. Huang, P., Li, D., Jiao, Z., Wei, D., Li, G., Wang, Q., Zhang, H., Shen, D.: Cocagan: Common-feature-learning-based context-aware generative adversarial network for glioma grading. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 155–163. Springer (2019)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
10. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
11. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
12. Kavur, A.E., Selver, M.A., Dicle, O., Barıs, M., Gezer, N.S.: Chaos-combined (ct-mr) healthy abdominal organ segmentation challenge data. In: Proc. IEEE Int. Symp. Biomed. Imag.(ISBI) (2019)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015)
14. Martin Arjovsky, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia (2017)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
16. Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P., et al.: Multi-domain image completion for random missing input data. IEEE Transactions on Medical Imaging (2020)
17. Xin, B., Hu, Y., Zheng, Y., Liao, H.: Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1803–1807. IEEE (2020)
18. Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P.: Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis. IEEE transactions on medical imaging **38**(7), 1750–1762 (2019)
19. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9242–9251 (2018)
20. Zhu, D., Liu, S., Jiang, W., Gao, C., Wu, T., Wang, Q., Guo, G.: Ugan: Untraceable gan for multi-domain face translation. arXiv preprint arXiv:1907.11418 (2019)
21. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)