Adversarial Pseudo Healthy Synthesis Needs Pathology Factorization

Tian Xia[1] TIAN.XIA@ED.AC.UK
Agisilaos Chartsias[1] AGIS.CHARTSIAS@ED.AC.UK
Sotirios A. Tsaftaris[1,2] S.TSAFTARIS@ED.AC.UK
[1] *School of Engineering, University of Edinburgh, West Mains Rd, Edinburgh EH9 3FB, UK*
[2] *The Alan Turing Institute, London, UK*

Abstract

Pseudo healthy synthesis, i.e. the creation of a subject-specific 'healthy' image from a pathological one, could be helpful in tasks such as anomaly detection, understanding changes induced by pathology and disease or even as data augmentation. We treat this task as a factor decomposition problem: we aim to separate what appears to be healthy and where disease is (as a map). The two factors are then recombined (by a network) to reconstruct the input disease image. We train our models in an adversarial way using either paired or unpaired settings, where we pair disease images and maps (as segmentation masks) when available. We quantitatively evaluate the quality of pseudo healthy images. We show in a series of experiments, performed in ISLES and BraTS datasets, that our method is better than conditional GAN and CycleGAN, highlighting challenges in using adversarial methods in the image translation task of pseudo healthy image generation.

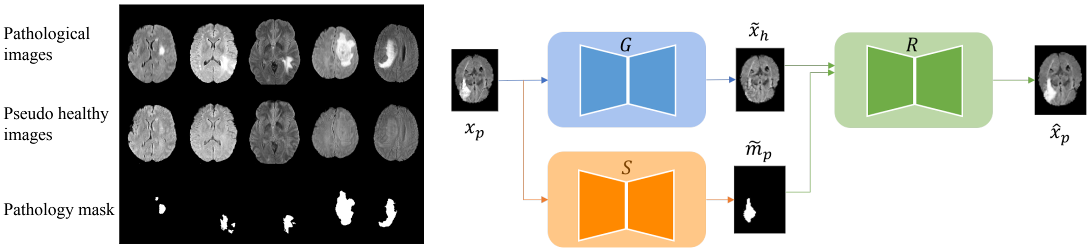
Keywords: pseudo healthy synthesis, GAN, cycle-consistency, factorization

1. Introduction

The aim of pseudo healthy synthesis is to synthesize a subject-specific 'healthy' image from a pathological one. Generating such images can be valuable both in research and in clinical applications. For example, these images can be used as a means to perform pathology segmentation (Bowles et al., 2017; Ye et al., 2013), detection (Tsunoda et al., 2014), to help with the visual understanding of disease classification networks (Baumgartner et al., 2018) and to aid experts with additional diagnostic information (Sun et al., 2018).

A challenge with pseudo healthy synthesis is the lack of paired pathological and healthy images for training, i.e. we do not have images of the same patient moments before and after pathology has appeared. Thus, methods based on pure supervised learning are not fit for our purpose. While longitudinal observations could perhaps partially alleviate this problem, the time difference between observations is an additional factor that may complicate learning. Thus, it is imperative to overcome this lack of paired data. One approach is to learn distributions that characterize the domains of healthy and pathological images, for example by learning a compact manifold of patch-based dictionaries (Ye et al., 2013; Tsunoda et al., 2014), or alternatively by learning mappings between the two domains with the use of adversarial training (Sun et al., 2018).

We follow a similar approach here but focus on factorizing the pathology. Simple schematic and examples are shown in Figure 1. We aim to separate what appears to be healthy out of a disease image. We let neural networks decompose an input image into a healthy image (one factor) via a generator, and a binary map that aims to localize disease (the other factor) via a segmentor.



(a) Example results

(b) Simple illustration of our method

Figure 1: Example results and simple illustration of our method. The three rows of (a) show input pathological images, corresponding pseudo healthy images, and pathology segmentation masks, respectively. Images are taken from the BraTS dataset. In (b) a pseudo healthy image $\tilde{x}_h$ and a pathology mask $\tilde{m}_p$ are generated from a pathological image x_p, and then finally a reconstructed image $\hat{x}_p$ is generated from $\tilde{x}_h$ and $\tilde{m}_p$.

These two factors are then composed together to reconstruct the input via another network. The pathological map is necessary as a factor to solve the one-to-many problem[1] (Chu et al., 2017): the healthy image must by definition contain 'less information' than the disease image.

We can train the segmentor in a supervised way using '*paired*' pathological images and their corresponding masks. However, since annotations of pathology are not easy to acquire, we also propose an '*unpaired*' training strategy. We take advantage of several losses including a cycle-consistency loss (Zhu et al., 2017), but use a modified second cycle where we enforce healthy-to-healthy image translation to approach the identity. Finally, since most pseudo healthy methods focus on applications of the synthetic data, results are either evaluated qualitatively or by demonstrating improvements on downstream tasks. A direct quantitative evaluation of the quality of pseudo healthy images has been largely ignored. In this paper, we propose two numerical evaluation metrics for characterizing the '*healthiness*' (i.e. how close to being healthy) and '*identity*' (i.e. how close to corresponding to the input identity) of synthetic results.

Our **contributions** in this work are three-fold:

1. We propose a 2D method that factorizes anatomical and pathological information.

2. We consider two training settings: (a) *paired*: when we have paired images and ground-truth pathology masks; (b) *unpaired*: when such pairs are not available.

3. We propose numerical evaluation metrics to explicitly evaluate the quality of pseudo healthy synthesized images, and compare our method with conditional GAN (Mirza and Osindero, 2014) and CycleGAN (Zhu et al., 2017) on ISLES and BraTS datasets.

2. Related work on pseudo healthy synthesis

Medical image synthesis is an active research topic in medical image analysis (Frangi et al., 2018) with an active community and dedicated workshops (e.g. the SASHIMI MICCAI series). For brevity

[1] There could be many disease images that could originate from the same healthy image, e.g. consider the simple setting of a lesion in many different locations on the same brain.

here we focus on methods related to pseudo healthy image synthesis with adversarial mechanisms. Image synthesis (translation) can be solved by a conditional GAN that learns a mapping between image domains (e.g. A to B). However, preservation of ‘identity’ is not guaranteed: there are no explicit costs to enforce that an image from domain A to be translated to the same image in domain B. CycleGAN uses a cycle-consistency loss to promote identity and has been profoundly adopted in medical image analysis (Huo et al., 2018; Zhang et al., 2018; Wolterink et al., 2017; Chartsias et al., 2017; Wang et al., 2018).

Baumgartner et al. (2018) used Wasserstein GAN (Arjovsky et al., 2017) to generate disease effect maps, and used these maps to synthesize pathological images. Andermatt et al. (2018) combined the idea of Baumgartner et al. (2018) with CycleGAN to perform pseudo healthy synthesis for pathology detection. Yang et al. (2016) used a Variational Auto-encoder to learn a mapping from pathological images to quasi-normal (pseudo healthy) images to improve atlas-to-image registration accuracy with large pathologies. Schlegl et al. (2017) and Chen and Konukoglu (2018) trained adversarial auto-encoder networks only on normal data, then used the trained model to synthesize normal data from abnormal data as a way of detecting the anomaly. Sun et al. (2018) proposed a CycleGAN-based method to perform pseudo healthy synthesis treating ‘pathological’ and ‘healthy’ as two domains.

The majority of these works use pseudo healthy images to achieve improvements in downstream tasks. While the performance on such downstream tasks relies on pseudo healthy image quality, it is not explicitly evaluated. Herein, we pay particular attention to consistently evaluate how ‘healthy’ the synthetic images look, and whether they correspond to the same ‘identity’ of the input. All methods rely on some form of adversarial training to approximate a distribution. However, as we will detail below, when one of the domains has less information the one-to-many problem can appear and CycleGAN may collapse. Our method treats pathology as a ‘residual’ factor: it factorizes anatomical and pathological information using adversarial and cycle-consistent losses to bypass the one-to-many problem.

3. Methodology

3.1. Problem overview

We denote a pathological image as x_p and a healthy image as x_h, drawn from $\mathscr{P}$ and $\mathscr{H}$ distributions, respectively, i.e $x_p \sim \mathscr{P}$ and $x_h \sim \mathscr{H}$. Our task is to generate a pseudo healthy image $\tilde{x}_h$ for a sample x_p, such that $\tilde{x}_h$ lies in the distribution of healthy images, i.e. $\tilde{x}_h \sim \mathscr{H}$. In the meantime, we also want the generated image $\tilde{x}_h$ to maintain the identity of the original image x_p, i.e. to come from the same subject as x_p. Therefore, pseudo healthy synthesis can be formulated as two major objectives: *remove* the disease of pathological images, and *maintain* the identity and realism as good as possible.

3.2. The one-to-many problem: motivation for factorization

CycleGAN has to somehow invent (or hide) information when one domain contains less information than the other. In our case domain $\mathscr{P}$ does contain disease information that should not be present in $\mathscr{H}$, which leads to failure cases as shown in Figure 2. When CycleGAN cannot invent information, Chu et al. (2017) in fact showed that it hides information within an image to be able to solve the one-to-many mapping. Recently, several papers (Chartsias et al., 2018; Almahairi et al., 2018;

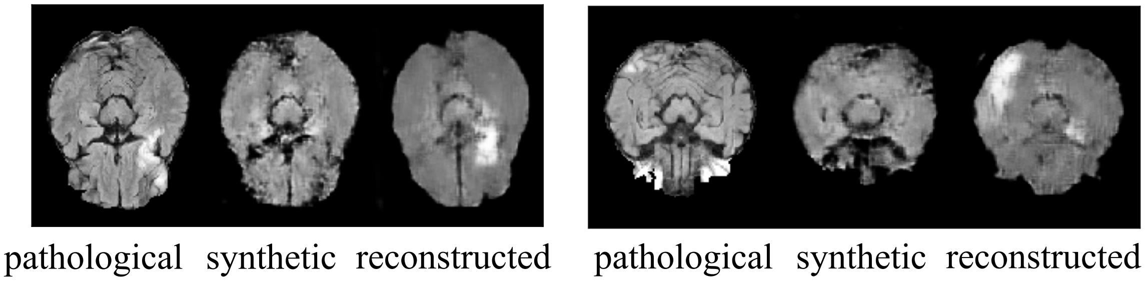


Figure 2: CycleGAN failure cases caused by the one-to-many problem. Each subfigure from left to right shows the input, the pseudo healthy and the input reconstruction. The lesion location in the reconstruction differs from the original one, since an accurate pseudo healthy image has no information to guide the reconstruction process. Images taken from ISLES.

Huang et al., 2018; Lee et al., 2018; Esser et al., 2018) have shown that one needs to provide auxiliary information in the form of a style or modality specific code (actually a vector) to guide the translation and allow many-to-many mappings. Our paper does follow this practice, but instead of providing a vector we consider the auxiliary information of where the disease could be, such that the decoder does *not* have to invent where things should go, and conversely the encoder does *not* have to hide information. We thus achieve that pseudo healthy images are of high quality, correspond to the identity of the same input, and also produce realistic disease maps.

3.3. Proposed approach

A schematic of our proposed method is illustrated in Figure 3. Recall that our task is to transform an input pathological image x_p to a disease-free image $\tilde{x}_h$ whilst maintaining the identity of x_p. Towards this goal, our method uses the cycle-consistency losses and treats 'pathological' and 'healthy' as two image domains. To solve the one-to-many mapping problem, we estimate a disease map from a pathological image using a segmentation network, and then use the map to provide information about disease location. Specifically, there are three main components: '*G*' the 'generator'; '*S*' the 'segmentor'; and '*R*' the 'reconstructor' trained using two cycles: *Cycle P-H* and *Cycle H-H*.

Cycle P-H, we perform pseudo healthy synthesis, where '*G*' takes a pathological image x_p as input and outputs a 'healthy' looking image $\tilde{x}_h$: $\tilde{x}_\mathrm{h} = G(x_\mathrm{p})$. '*S*' takes x_p as input and outputs a pathology map $\tilde{m}_p$: $\tilde{m}_\mathrm{p} = S(x_\mathrm{p})$. '*R*' takes the synthesized 'healthy' image $\tilde{x}_h$ and the segmented mask $\tilde{m}_p$ as input and outputs a 'pathological' image $\hat{x}_p$: $\hat{x}_\mathrm{p} = R(\tilde{x}_\mathrm{h}, \tilde{m}_\mathrm{p}) = R(G(x_\mathrm{p}), S(x_\mathrm{p}))$.

Cycle H-H utilizes healthy images and stabilizes the training. It starts with a healthy image x_h and a null 'healthy' mask m_h. First, *'R'* generates a fake 'healthy' image $\tilde{x}_h$: $\tilde{x}_\mathrm{h} = R(x_\mathrm{h}, m_\mathrm{h})$, which is then segmented into a healthy mask $\hat{m}_\mathrm{h}$: $\hat{m}_\mathrm{h} = S(\tilde{x}_\mathrm{h})$ and transformed to a reconstructed healthy image $\hat{x}_h$: $\hat{x}_\mathrm{h} = G(\tilde{x}_\mathrm{h})$.

There are several reasons why we design *Cycle H-H* in such a way. First, a pathology mask for a real healthy image is, by definition, a black mask. Second, we want to prevent the reconstructor '*R*' from inventing pathology when the input disease map is black. Third, we want to guide the generator *'G'* and segmentor *'S'* to preserve identity when the input (to both) is a 'healthy' image, such that the synthesized 'healthy' image is as similar to the input 'healthy' image as possible.

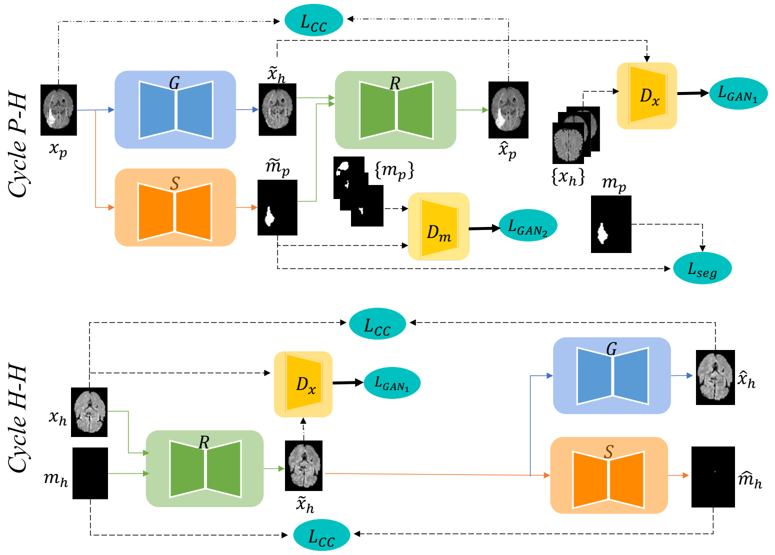


Figure 3: Training flowchart. *Cycle P-H* is the translation path from ‘pathological’ to ‘healthy’ and then back to ‘pathological’; *Cycle H-H* is the path from a healthy image and a black mask to a fake healthy image, then back to the reconstructed image and mask.

Similarly, when the input to the segmentor *‘S’* is a ‘healthy’ image, it should output a ‘healthy’ (no disease) map, i.e. a black mask.

3.4. Losses

The training losses are $\mathscr{L}_{\mathrm{CC}}$, $\mathscr{L}_{\mathrm{GAN}_1}$ and $\mathscr{L}_{\mathrm{Seg}}$ and $\mathscr{L}_{\mathrm{GAN}_2}$.

$\mathscr{L}_{\mathrm{CC}}$ is the cycle-consistency loss:

$$\mathscr{L}_{\mathrm{CC}} = \mathbb{E}_{x_p \sim \mathscr{P}}[||R(G(x_\mathrm{p}), S(x_\mathrm{p})) - x_\mathrm{p}||_1] \\ + \mathbb{E}_{x_h \sim \mathscr{H}, m_h \sim \mathscr{H}_m}[||G(R(x_\mathrm{h}, m_\mathrm{h})) - x_\mathrm{h}||_1] + \mathbb{E}_{x_h \sim \mathscr{H}, m_h \sim \mathscr{H}_m}[||S(R(x_\mathrm{h}, m_\mathrm{h})) - m_\mathrm{h}||_1],$$

where the first term is defined in *Cycle P-H* and the last two terms are defined in *Cycle H-H*. Note that the third term uses *Mean Average Error* instead of *Dice*, because if the target mask is black, then given any result mask, *Dice loss* will always produce 1.

$\mathscr{L}_{\mathrm{GAN}_1}$ is the least squares discriminator loss over synthetic images (Mao et al., 2017):

$$\mathscr{L}_{\mathrm{GAN}_1} = \max_{D_\mathrm{x}} \min_{G} \frac{1}{2} \mathbb{E}_{x_p \sim \mathscr{P}}[||D_x(G(x_\mathrm{p})) - 1||_2] + \max_{D_\mathrm{x}} \frac{1}{2} \mathbb{E}_{x_h \sim \mathscr{H}}[||D_x(x_h)||_2] \\ + \max_{D_\mathrm{x}} \min_{R} \frac{1}{2} \mathbb{E}_{x_h \sim \mathscr{H}, m_h \sim \mathscr{H}_m}[||D_x(R(x_\mathrm{p}, m_\mathrm{h})) - 1||_2] + \max_{D_\mathrm{x}} \frac{1}{2} \mathbb{E}_{x_h \sim \mathscr{H}}[||D_x(x_h)||_2],$$

where the first two terms correspond to *Cycle P-H* and the last two for *Cycle H-H*.

To train *‘S’*, we use two different training settings whether we have *paired* or *unpaired* data, and use a supervised or a GAN loss, respectively.

In the *paired* setting, we use manually annotated pathology masks corresponding to pathological images in $\mathscr{L}_{\text{Seg}} = \mathbb{E}_{x_p\sim\mathscr{P},m_p\sim\mathscr{P}_m}[Dice(S(x_p) - m_p)]$, with a differentiable *Dice* (Milletari et al., 2016) loss.

In the *unpaired* setting, since pathological images lack paired annotations, we replace $\mathscr{L}_{\text{Seg}}$ with a discriminator D_m which classifies real pathology masks from inferred masks:

$$\mathscr{L}_{\text{GAN}_2} = \max_{D_\text{m}} \min_{S} \frac{1}{2}\mathbb{E}_{x_p\sim\mathscr{P}}[\|D_m(S(x_\text{p})) - 1\|_2] + \max_{D_\text{m}} \frac{1}{2}\mathbb{E}_{m_p\sim\mathscr{P}_m}[\|D_m(m_p)\|_2],$$

where a pathological image x_p and a mask m_p come from different volumes.

4. Experiments

4.1. Experimental settings

Dataset and preprocessing: We demonstrate our method on two datasets. We use the FLAIR data of the *Ischemic Lesion Segmentation* (ISLES) 2015 dataset (Maier et al., 2017), which contains images of 28 volumes that are skull stripped and re-sampled to an isotropic spacing of $1mm^3$ (SISS) resp. We also use FLAIR data from MRI scans of glioblastoma (GBM/HGG), made available in the *Brain Tumour Segmentation* (BraTS) 2018 (Menze et al., 2015) challenge. The BraTS data contain images of 79 volumes that are skull-striped, and interpolated to $1mm^3$ resolution. Both datasets are released with segmentation masks of the pathological regions. For each dataset, we normalize each volume by clipping the intensities to [0, $V_{99.5}$], where $V_{99.5}$ is the 99.5% largest pixel value of the corresponding volume, then we normalize the resulting intensities to [0, 1]. We choose the middle 60 slices from each volume and label a slice as ‘healthy’ if its corresponding pathology mask is black, and as ‘pathological’ otherwise. We divide the datasets into a training and a testing set of 22 and 6 volumes for ISES, and 50 and 29 volumes for BraTS respectively.

Training and implementation details: The method is implemented in Python using Keras (Chollet et al., 2015). The loss function for the paired data option is defined as $L_{paired} = \lambda_1\mathscr{L}_{\text{CC}} + \lambda_2\mathscr{L}_{\text{GAN}_1} + \lambda_3\mathscr{L}_{\text{Seg}}$, where $\lambda_1 = 10$, $\lambda_2 = 1$, and $\lambda_3 = 10$ (same values as Chartsias et al. (2018)). The loss function for the unpaired data option is defined as $L_{unpaired} = \lambda_1\mathscr{L}_{\text{CC}} + \lambda_2\mathscr{L}_{\text{GAN}_1} + \lambda_3\mathscr{L}_{\text{GAN}_2}$, where $\lambda_1 = 10$, $\lambda_2 = 2$, and $\lambda_3 = 10$ (λ_2 has been increased to focus the attention on synthesis). Architecture details are in the Appendix.

Baselines: We consider two pseudo healthy synthesis baselines for comparison: a *conditional GAN* (Mirza and Osindero, 2014) (that is deterministic and is conditioned on an image) that consists of a pseudo healthy generator, trained with unpaired data and an adversarial loss against a discriminator that classifies real and fake healthy images; and a *CycleGAN* which considers two domains for healthy and unhealthy and is trained as in Zhu et al. (2017) to learn a domain translation using unpaired data.

4.2. Evaluation metrics

We propose, and use, numerical evaluation metrics to quantitatively evaluate the synthesized pseudo healthy images in terms of *‘healthiness’* and *‘identity’* i.e. how healthy do they look and how close to the input they are (as a proxy to identity).

‘Healthiness’ is not easy to directly measure since we do not have ground-truth pseudo healthy images. However, given a pathology segmentor applied on a pseudo healthy synthetic image, we can

measure the size of the segmented pathology as a proxy. To this end, we first train a segmentor to predict disease from pathological images, and then use the pre-trained segmentor to predict disease masks of synthetic pseudo healthy images and check how large the predicted disease areas are. Formally, 'healthiness' can be defined as:

$$h = 1 - \frac{\mathbb{E}_{\hat{x}_h \sim \mathscr{H}}[N(f_{\mathrm{pre}}(\hat{x}_{\mathrm{h}}))]}{\mathbb{E}_{m_p \sim \mathscr{P}_m}[N(m_{\mathrm{p}})]} = 1 - \frac{\mathbb{E}_{x_p \sim \mathscr{P}}[N(f_{\mathrm{pre}}(G(x_{\mathrm{p}})))]}{\mathbb{E}_{m_p \sim \mathscr{P}_m}[N(m_{\mathrm{p}})]},$$

where f_{pre} is the pre-trained segmentor whose output is a pathology mask, and $N(m)$ is the number of pixels which are labeled as pathology in the mask m. We normalize by the average size of all ground-truth pathological masks. Then we subtract the term from 1, such that h increases when the images have smaller pathology.

'Identity' is measured using a masked *Multi-Scale Structural Similarity Index* (MS-SSIM) with window width 11, defined as MS-SSIM$[(1-m_{\mathrm{p}}) \odot \hat{x}_{\mathrm{h}}, (1-m_{\mathrm{p}}) \odot x_{\mathrm{p}}]$. This metric is based on the assumption that a pathological image and its corresponding pseudo healthy image should look the same in regions not affected by pathology.

4.3. Experiments on ISLES and BraTS datasets

We train our proposed method in both *paired* and *unpaired* settings on ISLES and BraTS datasets, and compare with the baselines of Section 4.1. Some results can be seen in Figure 4, where we observe that all synthetic images visually appear to be healthy. However, the pseudo healthy images generated by *conditional GAN* are blurry and to some degree different from the original samples, i.e. the lateral ventricles (cavities in the middle) change: a manifestation of loss of *'identity'*. Similarly, we observe changes of lateral ventricles in the synthetic images generated by *CycleGAN*. These changes are probably due to the fact that *CycleGAN* needs to hide information to reconstruct the input images. We also observe that our methods preserve more details of the original samples. Together, these observations imply that our proposed methods maintain better *'identity'* than the baselines.

We also use the proposed evaluation metrics to measure the quality of synthetic images generated by our method and baselines, respectively. The numerical results are shown in Table 1. We can see that our proposed method (paired) when trained using pathological image and mask pairs achieves the best results, followed by our proposed method (unpaired). Both *paired* and *unpaired* versions outperform conditional GAN and CycleGAN in both the BraTS and ISLES datasets. The improvements of our method are due to the factorization of pathology, which ensures maintaining information of the pathology during the pseudo healthy synthesis such that the synthetic images do not need to hide information.

5. Conclusion

In this paper, we propose an adversarial network for pseudo healthy synthesis with factorization of pathology. Our proposed method is composed of a pseudo healthy synthesizer to generate pseudo healthy images, a segmentor to predict a pathology map, i.e. as a way of factorizing pathology, and a reconstructor to reconstruct the input pathological image conditioned on the map. Our method can be trained in (a) *paired* mode when we have paired pathological images and masks; or (b) *unpaired* mode for when we do not have image and mask pairs. We also propose two numerical evaluation

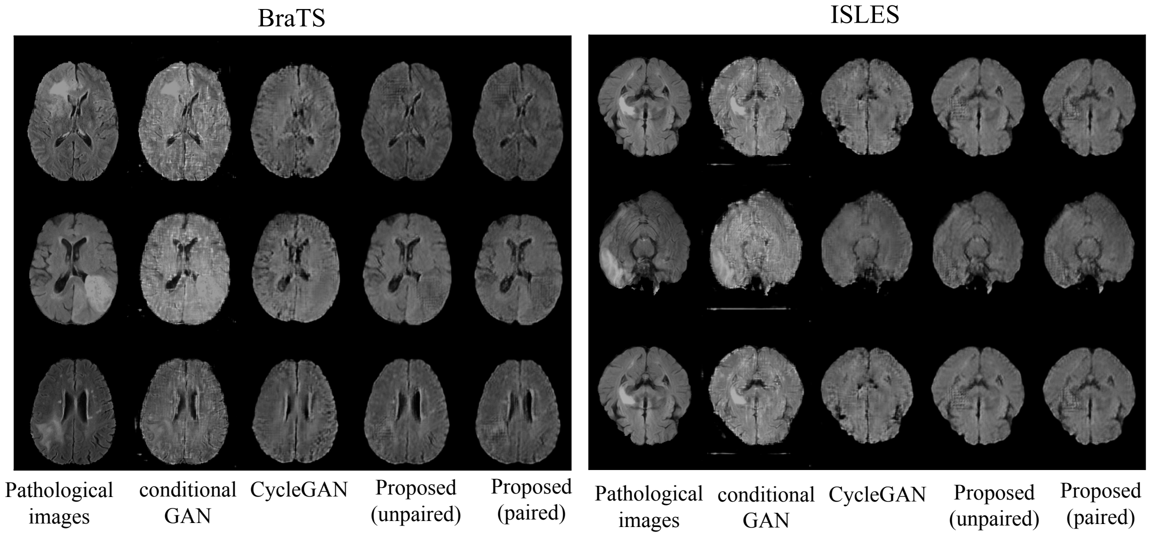


Figure 4: Experimental results for BraTS and ISLES data are shown in the *left* and *right* part respectively. Each part shows three samples (in three rows). The columns from left to right show the ground-truth pathological images, and pseudo healthy images generated by *conditional GAN*, *CycleGAN*, and the two proposed methods, respectively. A larger version of these results are shown in Appendix.

metrics to explicitly measure the quality of the synthesized images. We demonstrate on ISLES and BraTS datasets that our method outperforms the baselines both quantitatively and qualitatively.

Metrics that enforce or even measure identity is a topic of considerable interest in computer vision (Antipov et al., 2017). Our approach here is simple (essentially measures the fidelity of the reconstructed signal) but it does assume that changes due to disease are only local. This assumption is also adopted by several methods (Andermatt et al., 2018; Sun et al., 2018; Baumgartner et al.,

Table 1: Evaluation results on BraTS and ISLES of our proposed method trained with and without *pairs*, as well as of the baselines used for comparison. The best mean values for each defined metric (identity, healthiness) are shown in bold. Statistical significant results (5% level), of our methods compared to the best baseline are marked with a star (*).

	BraTS		**ISLES**	
Methods	**'Identity'**	**'Healthiness'**	**'Identity'**	**'Healthiness'**
conditional GAN	0.74 ± 0.05	0.82 ± 0.03	0.67 ± 0.02	0.86 ± 0.13
CycleGAN	0.80 ± 0.03	0.83 ± 0.04	0.78 ± 0.02	0.85 ± 0.11
proposed (unpaired)	0.83 ± 0.03	0.98 ± 0.07*	0.82 ± 0.03	0.94 ± 0.11*
proposed (paired)	**0.88 ± 0.03***	**0.99**±0.02*	**0.93 ± 0.02***	**0.98 ± 0.04***

2018). When disease globally affects an image, new approaches must be devised which is seen as future work.

Acknowledgments

This work was supported by the University of Edinburgh by a PhD studentship. This work was partially supported by EPSRC (EP/P022928/1) and by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was supported in part by the US National Institutes of Health (R01HL136578). We also thank Nvidia for donating a Titan-X GPU.

References

Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, 2018.

Simon Andermatt, Antal Horváth, Simon Pezold, and Philippe Cattin. Pathology Segmentation using Distributional Differences to Images of Healthy Origin. *Brain-Lesion workshop (BrainLes). MICCAI*, 2018.

Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018.

Christopher Bowles, Chen Qin, Ricardo Guerrero, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 16: 643–658, 2017.

Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *International Workshop on Simulation and Synthesis in Medical Imaging (MICCAI)*, pages 3–13. Springer, 2017.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 490–498, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2.

Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *Internatinal Conference on Medical Imaging with Deep Learning*, 2018.

François Chollet et al. Keras. https://keras.io, 2015.

Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN: a Master of Steganography. *NIPS 2017, Workshop on Machine Deception*, 2017.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A Variational U-Net for Conditional Appearance and Shape Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

A. F. Frangi, S. A. Tsaftaris, and J. L. Prince. Simulation and Synthesis in Medical Imaging. *IEEE Transactions on Medical Imaging*, 37(3):673–679, March 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2800298.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, volume 11207, pages 179–196. Springer International Publishing, 2018.

Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman. SynSeg-Net: Synthetic Segmentation Without Target Modality Ground Truth. *IEEE Transactions on Medical Imaging*, pages 1–1, 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2876633.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse Image-to-Image Translation via Disentangled Representations. In *European Conference on Computer Vision*, volume 11205, pages 36–52. Springer International Publishing, 2018.

Oskar Maier, Bjoern H. Menze, Janina von der Gablentz, Levin Häni, Mattias P. Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, Daan Christiaens, Francis Dutil, Karl Egger, Chaolu Feng, Ben Glocker, Michael Götz, Tom Haeck, Hanna-Leena Halme, Mohammad Havaei, Khan M. Iftekharuddin, Pierre-Marc Jodoin, Konstantinos Kamnitsas, Elias Kellner, Antti Korvenoja, Hugo Larochelle, Christian Ledig, Jia-Hong Lee, Frederik Maes, Qaiser Mahmood, Klaus H. Maier-Hein, Richard McKinley, John Muschelli, Chris Pal, Linmin Pei, Janaki Raman Rangarajan, Syed M.S. Reza, David Robben, Daniel Rueckert, Eero Salli, Paul Suetens, Ching-Wei Wang, Matthias Wilms, Jan S. Kirschke, Ulrike M. Krämer, Thomas F. Münte, Peter Schramm, Roland Wiest, Heinz Handels, and Mauricio Reyes. "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". *Medical Image Analysis*, 35:250 – 269, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.07.009.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, T. Arbel, B. B. Avants, N. Ayache,

P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024, Oct 2015. ISSN 0278-0062. doi: 10.1109/TMI.2014.2377694.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

Liyan Sun, Jiexiang Wang, Xinghao Ding, Yue Huang, and John Paisley. An Adversarial Learning Approach to Medical Image Synthesis for Lesion Removal. *arXiv preprint arXiv:1810.10850*, 2018.

Yuriko Tsunoda, Masayuki Moribe, Hideaki Orii, Hideaki Kawano, and Hiroshi Maeda. Pseudo-normal image synthesis from chest radiograph database for lung nodule detection. In *Advanced Intelligent Systems*, pages 147–155. Springer, 2014.

Chengjia Wang, Gillian Macnaught, Giorgos Papanastasiou, Tom MacGillivray, and David Newby. Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 52–60. Springer, 2018.

Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017.

Xiao Yang, Xu Han, Eunbyung Park, Stephen Aylward, Roland Kwitt, and Marc Niethammer. Registration of pathological images. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 97–107. Springer, 2016.

Dong Hye Ye, Darko Zikic, Ben Glocker, Antonio Criminisi, and Ender Konukoglu. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–613. Springer, 2013.

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision*, 2017.

Appendix A. Architecture details

The detailed architecture of our generator *'G'* is shown in Table 2. IN stands for Instance Normalization. The detailed architecture of our reconstructor *'R'* is shown in Table 3. The detailed architecture of our discriminator 'D_x' and 'D_m' is shown in Table 4.

Table 2: Detailed architecture of generator *'G'*.

Layer	Input	filter size	stride	IN	activation	Output
conv2d	(208,160,1)	7	1	Yes	ReLu	(208,160,32)
conv2d	(208,160,32)	3	2	Yes	ReLu	(104,80,64)
conv2d	(104,80,64)	3	2	Yes	ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
upsampling2d	(52,40,128)	-	-	-	-	(104, 80, 128)
conv2d	(104, 80, 128)	3	1	Yes	ReLu	(104,80,64)
upsampling2d	(104,80,64)	-	-	-	-	(208, 160, 64)
conv2d	(208, 160, 64)	3	1	Yes	ReLu	(208, 160, 32)
conv2d	(208, 160, 32)	3	1	No	sigmoid	(208, 160, 1)

Table 3: Detailed architecture of reconstructor *'R'*.

Layer	Input	filter size	stride	IN	activation	Output
conv2d	(208,160,2)	7	1	Yes	ReLu	(208,160,32)
conv2d	(208,160,32)	3	2	Yes	ReLu	(104,80,64)
conv2d	(104,80,64)	3	2	Yes	ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
upsampling2d	(52,40,128)	-	-	-	-	(104, 80, 128)
conv2d	(104, 80, 128)	3	1	Yes	ReLu	(104,80,64)
upsampling2d	(104,80,64)	-	-	-	-	(208, 160, 64)
conv2d	(208, 160, 64)	3	1	Yes	ReLu	(208, 160, 32)
conv2d	(208, 160, 32)	3	1	No	sigmoid	(208, 160, 2)

Table 4: Detailed architecture of discriminator ‘D_x’ and ‘D_m’.

Layer	Input	filter size	stride	IN	activation	Output
conv2d	(208,160,2)	4	2	Yes	Leaky ReLu	(104,80,32)
conv2d	(104,80,32)	4	2	Yes	Leaky ReLu	(52,40,128)
conv2d	(52,40,128)	4	2	Yes	Leaky ReLu	(26,20,256)
conv2d	(26,20,256)	4	2	Yes	Leaky ReLu	(13,10,512)
conv2d	(13,10,512)	4	1	No	sigmoid	(13,10,1)

The detailed architecture of our segmentor *‘S’* is a U-Net, and follows the structure of Ronneberger et al. (2015). We change the activation function from ‘ReLu’ to ‘Leaky ReLu’. We also found that using residual connection on each layer slightly improved the results.

The pre-trained segmentor f_{pre} which is used for evaluation uses the same structure as *‘S’*. We train the segmentor f_{pre} on the ISLES and BraTS training datasets (see Section 4.1) respectively, and then use it to evaluate synthetic images generated from samples in ISLES and BraTS testing datasets. The Dice loss of the segmentor on ISLES and BraTS testing datasets are 0.12 and 0.16, respectively.

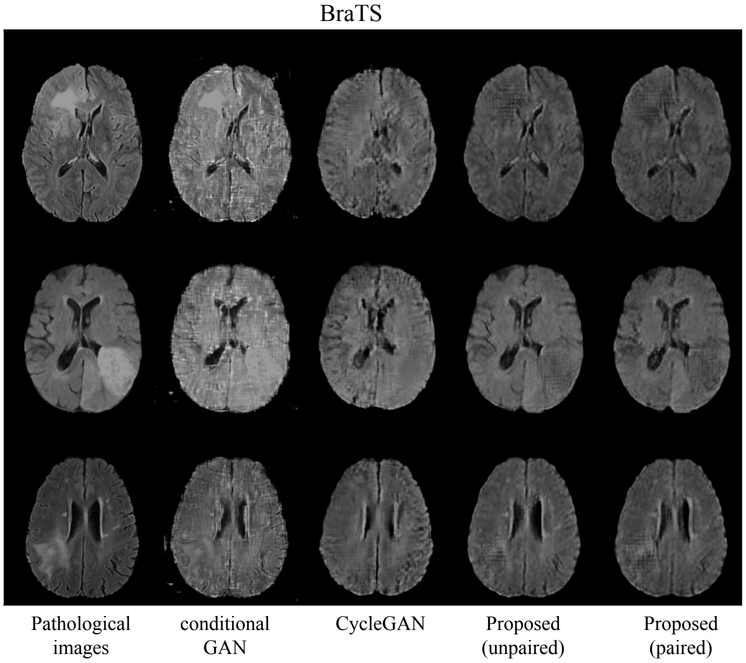


Figure 5: Experimental results for BraTS. The columns from left to right show the ground-truth pathological images, and pseudo healthy images generated by *conditional GAN*, *CycleGAN*, and the two proposed methods, respectively.

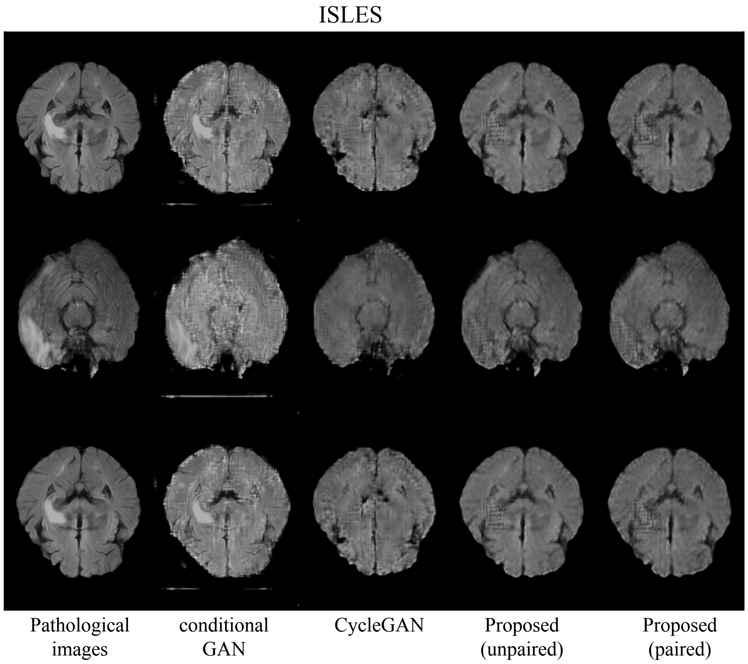


Figure 6: Experimental results for ISLES. The columns from left to right show the ground-truth pathological images, and pseudo healthy images generated by *conditional GAN*, *CycleGAN*, and the two proposed methods, respectively.