

# GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks

Christopher Bowles<sup>1</sup>, Liang Chen<sup>1</sup>, Ricardo Guerrero<sup>1,6\*</sup>, Paul Bentley<sup>2</sup>, Roger Gunn<sup>2,3</sup>, Alexander Hammers<sup>4</sup>, David Alexander Dickie<sup>5,7</sup>, Maria Valdés Hernández<sup>5</sup>, Joanna Wardlaw<sup>5</sup>, and Daniel Rueckert<sup>1</sup>

<sup>1</sup> Department of Computing, Imperial College London, UK

<sup>2</sup> Department of Medicine, Imperial College London, UK

<sup>3</sup> Imanova Ltd., London, UK

<sup>4</sup> PET Centre, Kings College London, UK

<sup>5</sup> Department of Neuroimaging Sciences, University of Edinburgh, UK

<sup>6</sup> Samsung AI Research Centre (SAIC), Cambridge, UK

<sup>7</sup> Institute of Cardiovascular and Medical Sciences, University of Glasgow, UK

## Abstract

One of the biggest issues facing the use of machine learning in medical imaging is the lack of availability of large, labelled datasets. The annotation of medical images is not only expensive and time consuming but also highly dependent on the availability of expert observers. The limited amount of training data can inhibit the performance of supervised machine learning algorithms which often need very large quantities of data on which to train to avoid overfitting. So far, much effort has been directed at extracting as much information as possible from what data is available. Generative Adversarial Networks (GANs) offer a novel way to unlock additional information from a dataset by generating synthetic samples with the appearance of real images. This paper demonstrates the feasibility of introducing GAN derived synthetic data to the training datasets in two brain segmentation tasks, leading to improvements in Dice Similarity Coefficient (DSC) of between 1 and 5 percentage points under different conditions, with the strongest effects seen fewer than ten training image stacks are available.

## 1 Introduction

Data augmentation is commonly used by many deep learning approaches in the presence of limited training data. Increasing the number of training examples through the rotation, reflection, cropping, translation and scaling of existing images is common practice during the training of learning algorithms, allowing for the number of samples in a dataset to be increased by factors of thousands [12]. Populating the training data with realistic, if synthetic, data in this way can significantly reduce overfitting and thus not only improve the accuracy but also

---

\* All work done while at Imperial College London

the generalisation ability of deep learning approaches. This is of particular importance in Convolutional Neural Networks (CNNs) which cannot easily learn rotationally invariant features unless there are sufficient examples at different rotations in the training data. This paper investigates using a GAN to model the underlying distribution of training data to allow for additional synthetic data to be sampled and used to augment the real training data.

First proposed in [6], GANs are a class of neural networks which learn to generate synthetic samples with the same characteristics as a given training distribution. In the case of images, this involves learning to produce images (via a generator) which are visually so similar to a set of real images that an adversary (the discriminator) cannot detect them. The original formulation has been built on to address problems such as training stability [16], low resolution [10], and the absence of a true image quality based loss function [2], and applied to tasks such as super resolution [13], reconstructing images from a minimal data [22] and anomaly detection [19].

Various methods for using GANs to expand training datasets have been recently proposed. In [20], the authors use an adversarial network to improve the quality of simulated images, and use these for further training. In [1], the authors train a conditional GAN on unlabelled data to generate alternative versions of a given real image, and in [23], the authors use a similar GAN to impose emotions on neutral faces to expand underrepresented classes. However, the use of non-conditional GANs to augment training data directly as a preprocessing step with no additional data has only very recently been explored [5,15], with promising results in medical image classification tasks.

## 1.1 Motivation

Imaging features can be divided into two categories, measuring either *pertinent* or *non-pertinent* variance. Pertinent features are those which are important to whatever information the user wishes to extract. In medical imaging, these are features such as the size, shape, intensity and location of key components such as organs or lesions. Non-pertinent features are those which vary between images but are unrelated to the information the user wishes to extract. Examples of these are global intensity differences, position within the image field of view and appearance of unrelated anatomy. Exactly which features are pertinent or non-pertinent will depend on the application, and may not be known a-priori.

A lot of non-pertinent variance can easily be removed from a dataset. Common methods include intensity normalisation, cropping, and registration to a standard space. These processes substantially simplify the data distribution, and importantly, can be applied to test instances. Keeping too much non-pertinent variance can not only occlude the diagnostically important information, but also lead to overfitting, especially in the small datasets often used in medical imaging.

Data augmentation is an alternative to removing non-pertinent variance. One of the goals of data augmentation is to populate the data with a large amount of synthetic data in the directions of these non-pertinent sources of variance.

The aim of this is to reduce this variance to noise, removing any coincidental correlation with labels and preventing its use as a discriminative feature.

As noted in [11], there is a tendency within medical imaging to remove non-pertinent variance rather than use augmentation. This due to both the ease with which much of the non-pertinent variance can be removed, and the lack of suitable augmentation procedures for many sources of non-pertinent variance. This is reflected in [9], where the authors choose to only employ reflection and intensity augmentation for brain lesion segmentation, with even the latter omitted when using larger datasets. On the other hand, in [18] the authors benefit from extensive augmentation in their application of microscopy images, particularly through random elastic deformations. This demonstrates how careful consideration of the application will inform which types of augmentation are appropriate. While random elastic deformations may be an appropriate model for microscopy images, in which the objects of interest (cells) are generally fluid and unconstrained, applying the same procedure to brain images could lead to certain anatomical constraints such as symmetry, rigidity, and structure being disregarded. In addition, some sources of non-pertinent variance can be neither removed nor augmented by traditional means. For example, patient specific variation in non-relevant anatomy, where it may not be possible to remove this anatomy through cropping, or to define an accurate enough model to augment this variance with realistic cases.

GANs offer a potentially valuable addition to the arsenal of augmentation techniques which are currently available. One of the main potential advantages of GANs is that they take many decisions away from the user, in much the same way that deep learning removes the need for “hand crafted” features. An ideal GAN will transform the discrete distribution of training samples into a continuous distribution, thereby simultaneously applying augmentation to each source of variance within the dataset. For example, given a sufficient number of training examples at different orientations, a GAN will learn to produce examples at any orientation, replicating the effects of applying rotation augmentation. While orientation is a source of variance which can easily be augmented or removed using traditional methods, consider instead a more challenging source of variance such as ventricle size in brain imaging. Again, given a sufficient number of training examples of patients with different discrete ventricle sizes, a trained GAN will be able to produce examples along the continuum of all sizes. To perform the same kind of augmentation using deformations would involve a complex model of realistic ventricle size, shape and impact on the surrounding anatomy. By simultaneously learning the distribution of all sources of variance, the GAN infers this model directly from the available data.

One potential limitation of using GANs for augmentation is their ability to generate images with a high enough quality. While improvements have been made, GANs cannot be relied upon to produce images with perfect fidelity. This is not a problem for traditional augmentation procedures which do not significantly degrade the images. However, both [4] and [17] demonstrate that complete realism is not necessary to improve results with synthetic data. Whether the ad-

vantage of additional data is outweighed by the disadvantage of lower quality images is one of the questions we address in this paper.

## 1.2 Contribution

The results reported in [5,15] suggest that GANs can have a significant benefit when used for data augmentation in some classification tasks. In this paper we thoroughly investigate this use of GANs in different domains for the purpose of medical image segmentation. An in depth investigation into the effects of GAN augmentation is first carried out on a complex multi-class Computed Tomography (CT) Cerebrospinal Fluid (CSF) segmentation task using two segmentation architectures. By choosing not to co-register the images in this dataset, we are able to examine how GAN augmentation compares and interacts with rotation augmentation. The transferability of the method is then evaluated by applying it to a second dataset of Fluid-Attenuated Inversion Recovery (FLAIR) Magnetic Resonance (MR) images for the purpose of single-class White Matter Hyperintensity (WMH) segmentation. This is a well studied problem, and poses challenges typical to medical image segmentation tasks.

Aside from establishing whether GAN augmentation can lead to an improvement in network performance, we answer the following five important questions:

- *Does the choice of segmentation network affect this improvement?*
- *How does GAN augmentation compare to rotation augmentation?*
- *Does the amount of synthetic data added affect this improvement?*
- *Does the amount of available real data affect this improvement?*
- *Does the approach generalise to multiple datasets?*

We also explore the distribution of generated images to better understand what modes of augmentation are provided. This allows us to confirm that the GANs are producing images which are different to those in the dataset. We show how images are generated with the same pathology, but different unrelated anatomy, and vice versa, demonstrating the ability to perform these particularly challenging forms of augmentation.

## 2 Methods

We use a Progressive Growing of GANs (PGGAN) network [10] to generate synthetic data. PGGAN was chosen on the basis of its training stability at large image sizes and apparent robustness to hyperparameter selection. Whether the choice of GAN architecture will affect the quality of the augmentation is unclear, however there is evidence [14] to suggest that different GAN architectures produce results which are, on average, not significantly different from each other.

We train a PGGAN on 80k patches sampled from the available training data set as a preprocessing step prior to training a segmentation CNN. The PGGAN is trained on multi-channel image patches containing both the acquired image and manual segmentation label, thereby learning the manifold containing this joint data distribution. Synthetic examples are then sampled randomly from

this manifold using the trained generator and used to augment the same 80k patches, forming the training data used when training the subsequent segmentation network. The only alteration to the default PGGAN architecture was to concatenate a 32x32 layer of Gaussian noise at the start of the fourth (32x32) resolution level when training on CT data. This change was found empirically to produce CT images with a more realistic noise pattern. The networks were configured to produce images with a size of 128-by-128px with 6 resolution levels.

Segmentation networks were evaluated using training, validation, and test sets. Performance (measured by DSC) on the validation set was monitored during training with the best model at the conclusion of training applied to the test set.

A set of experiments were designed to assess effect of introducing GAN derived synthetic data to a segmentation task. In these experiments, a number of key variables were modified:

*Amount of available real data:* To simulate a situation with limited training data, the amount of training images was artificially reduced by randomly selecting a percentage of the available images, prior to sampling the 80k training patches. We performed experiments with percentage reductions in available data ranging from 10% to 90%. Note that this reduction is enforced for both the GAN and segmentation network training stages, ensuring the GAN is never exposed to more labelled data than the corresponding segmentation network.

*Amount of additional synthetic data:* To investigate whether the amount of synthetic data added to the real data affects the performance of a segmentation network, experiments were run with different amounts of additional synthetic data. To ensure equal access to the information available in the real data between experiments, synthetic data is added to the real data, increasing the size of the dataset, rather than replacing real data. The amount of additional patches is expressed as a percentage of the real patches. For example an experiment with +50% synthetic data would use 120k patches (80k real and 40k synthetic).

*Dataset:* Two different datasets are explored to assess the ability for GAN augmentation to generalise across segmentation tasks. The first dataset contains CT images with manually delineated CSF labels split into into 3 classes: cortical CSF, brain stem CSF and ventricular CSF. Data is split in the same way as in [3], using the same preprocessing and sampling procedures. This provides 500 manually labelled training image slices, with an additional 282 validation slices, from 101 subjects. For these experiments, the average DSC is used as the primary measure of performance, though results across each class are also analysed. The second dataset contains FLAIR images with manual WMH segmentations. 147 FLAIR image stacks were acquired as described in [21]. These were manually segmented, before being bias corrected, brain extracted, rigidly co-registered and intensity normalised as in [8], and randomly split into equal sized training, validation and test sets. By selecting two dissimilar tasks (multi- and single-class segmentation) across two modalities (CT and MR) we cover a wide range of likely applications for GAN augmentation.

*Segmentation network:* We investigate three different segmentation networks across the experiments. In [3], the authors show that both UNet and Residual

UNet (UResNet) [7] architectures perform well on this CT dataset, we therefore choose to explore both of these. The same hyperparameters were used as in [3]. DeepMedic [9] is a popular general purpose segmentation algorithm which has been shown to perform well in many applications, and was therefore chosen as a third network to explore. DeepMedic was modified only so as to accept 128x128 2-dimensional (2D) patches. Between these three, we represent the most popular CNN architectures currently in use.

*Augmentation:* As discussed in Section [11] extensive augmentation, beyond simple reflection, is rarely used in brain imaging due to the variety of preprocessing options available and anatomical constraints of the brain. However, in order to examine the interaction of GAN and rotation augmentation we elect not to perform coregistration on the CT dataset. Of the other common forms of augmentation, reflection augmentation is routinely performed in all experiments, translation augmentation is encapsulated in the patch based approach, intensity augmentation is obviated by intensity normalisation, and deformations are not considered due to anatomical constraints (preserving shape, symmetry etc.).

Table 1: Summary of experiments

% of available real % data sampled from	% added synthetic data	syn- Segmentation network	Dataset	Augmentation type
100, 50, 10	0, 50, 100	UNet, UResNet	CT	Rotation+GAN
100, 50, 10	0, 100	UNet	CT	None, GAN, rotation, rotation+GAN
100, 50, 10	0, 12.5, 25, 37.5, 50, 100	UNet	CT	Rotation+GAN
100, 90...20, 10	0, 50	UNet	CT	Rotation+GAN
100, 50, 10	0, 50, 100	DeepMedic	MR	GAN

Table 1 summarises the five sets of experiments which were carried out to answer the questions posed earlier. In each experiment, the segmentation network is treated as a black box and unchanged. This provides a fair platform upon which to observe the effects of GAN augmentation by ensuring that any changes in performance are as a result of the additional synthetic data, and not of changes in the network itself. GAN training took 36 hours, each UNet took 4 hours, each Res-UNet took 24 hours and each DeepMedic network took 24 hours on an Nvidia GTX 1080 Ti or similar GPU. All segmentation experiments on CT were repeated 8 times, while those on MR were repeated 14 times to compensate for a higher observed variance. Examples of real and synthetic patches generated for each dataset can be seen in Figure 1.

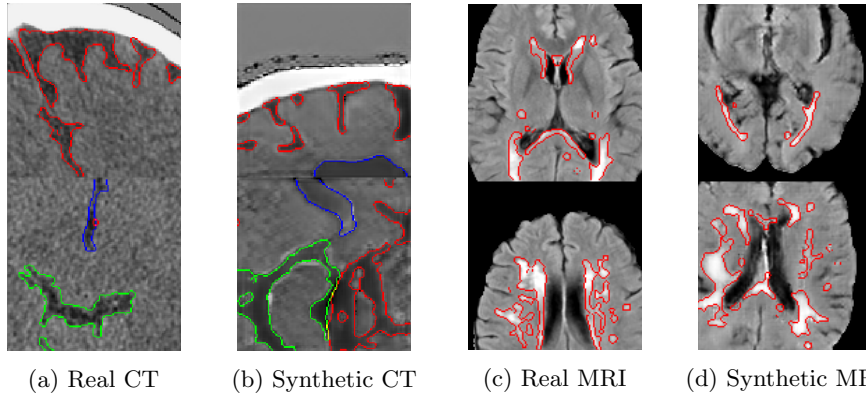


Fig. 1: Examples of real and GAN generated synthetic patches. *Left* CSF. Red: Cortical CSF. Green: Brain stem CSF. Blue: Ventricular CSF. *Right*: WMH.

### 3 Results

#### 3.1 Segmentation results

The following tables and graphs show the results over the two sets of experiments. All tables show the average DSC, with the standard deviation in brackets. Results which are statistically different from the baseline (2-tailed t-test, 5% significance level) are shown in bold.

Table 2: **CSF segmentation on CT**: Results with different proportions of the available training data and varying amounts of additional synthetic data using UNet and UResNet architectures.

		Available data					
		UNet			UResNet		
		100%	50%	10%	100%	50%	10%
Additional Data	0%	88.9 (0.51)	86.0 (0.50)	76.9 (0.58)	86.8 (0.82)	82.7 (1.55)	72.5 (1.98)
	50%	89.2 (0.30)	<b>87.3</b> (0.46)	<b>78.6</b> (1.04)	86.3 (1.44)	<b>84.3</b> (1.31)	74.3 (1.63)
	100%	89.3 (0.39)	<b>86.9</b> (0.36)	<b>78.4</b> (0.99)	86.3 (1.24)	84.1 (1.32)	<b>74.7</b> (1.18)

Table 3: **CSF segmentation on CT**: UNet results with different proportions of the available training data and different augmentation techniques.

		Available data		
		100%	50%	10%
	No augmentation	88.1 (0.32)	85.0 (0.58)	75.1 (0.60)
	GAN augmentation	88.4 (0.41)	85.6 (1.33)	76.3 (1.77)
	Rotation augmentation	<b>88.9</b> (0.51)	<b>86.0</b> (0.50)	<b>76.9</b> (0.58)
	GAN + Rotation augmentation	<b>89.3</b> (0.39)	<b>86.9</b> (0.36)	<b>78.4</b> (0.99)

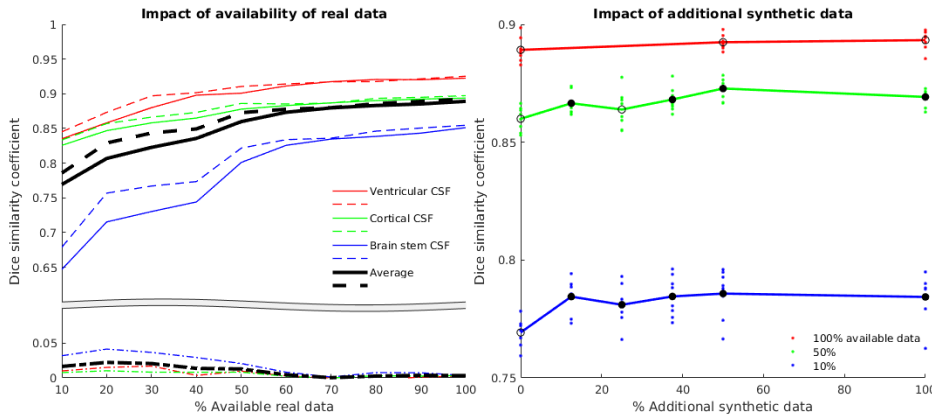


Fig. 2: **CSF segmentation on CT**: *Left*: Average DSC for each class (coloured) and mean across classes (black) as availability of real data varies. Solid lines show performance without GAN augmentation, dashed lines show performance with +50% synthetic data, and dot/dashed lines show the difference, indicating the improvement seen with GAN augmentation. *Right*: Average DSC observed using a UNet as synthetic data is added, when 100%, 50% and 10% of the total amount of real data is used. Each coloured dot represents an experiment. Black circles show the mean with filled circles indicating results significantly different from the baseline.

Table 4: **WMH segmentation on MRI**: Results with different proportions of the available training data and varying amounts of additional synthetic data.

		Available data		
		100%	50%	10%
Additional Data	0%	66.0 (1.26)	61.4 (2.67)	52.2 (6.65)
	50%	65.5 (1.21)	<b>63.7 (0.69)</b>	<b>57.2 (4.09)</b>
	100%	<b>64.8 (1.34)</b>	62.8 (1.17)	55.7 (4.26)

### 3.2 Qualitative evaluation

As well as the quantitative segmentation results, the generated MR images were also compared to their nearest neighbour in the training set to elucidate what extra information GAN augmentation provides. These images, a subset of which are shown in Figure 3, were examined looking for cases where: lesions were duplicated on different anatomy; lesions were changed whilst anatomy stays the same; the nearest neighbour is substantially different. The latter indicates the GAN has learned a smooth manifold leading to potentially novel anatomy.



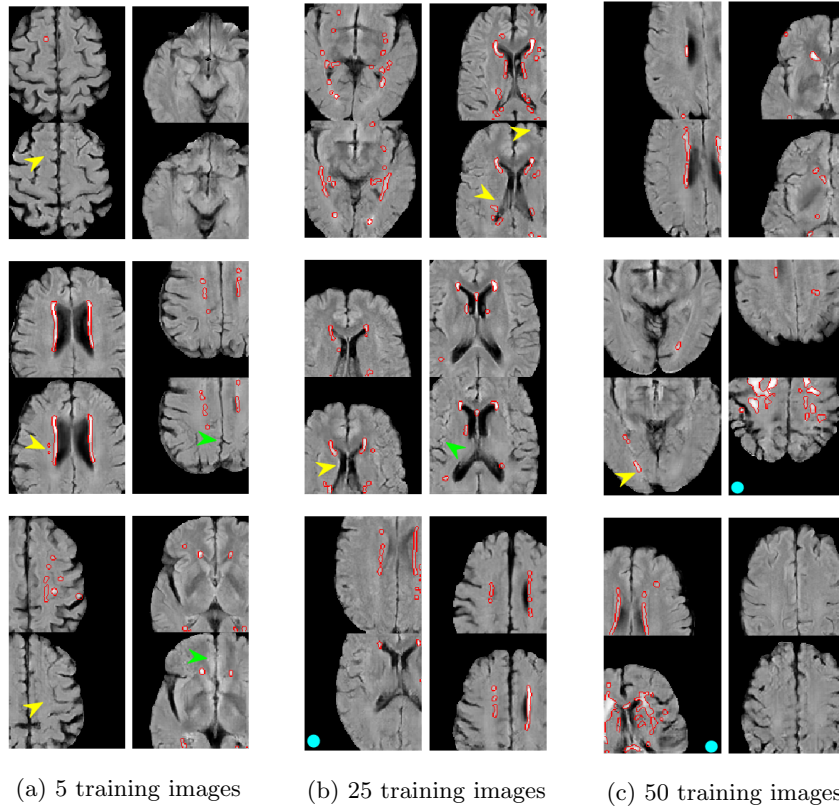


Fig. 3: Synthetic images (top of pair) with their nearest neighbours in the training set (bottom of pair) from GANs trained on patches from 5, 25 and 50 real MR images. Some local signs of successful augmentation are indicated using green (same lesions, different anatomy) and yellow (same anatomy, different lesions) arrows, and novel images (new anatomy and lesions) are shown with blue dots.

## 4 Discussion

It can be seen from across all of the results that GAN augmentation can provide a modest but significant improvement in segmentation performance in many cases. By far the strongest factor controlling the improvement seen is the amount of real data available for training. There is a clear trend across all results that the greatest improvements can be seen in the cases where real data is the most limited. However, Figure 2 suggests that there is perhaps a drop in improvement seen at the very lowest levels of available data, likely due to there being too little data to properly train the GAN. Results on the CT data suggest that there are no circumstances in which using synthetic data leads to worse results even when large amounts of real data is available. However, this is not reflected in the MR

results in Table 4, where a loss in DSC is observed when all the data is used. This suggests a tipping point associated with the amount of available data, beyond which GAN augmentation harms rather than helps.

Another benefit of GAN augmentation can be seen in the DSC observed on the individual CSF classes in Figure 2. A ratio of 1.35:4.35:1 between ventricular, cortical and brain stem CSF classes in the training set indicates a moderate class imbalance, with examples of the former and latter being relatively limited. Figure 2 shows that it is these two classes which benefit most from GAN augmentation. Brain stem CSF segmentation appears to benefit the most, though this can be attributed to ventricular CSF segmentation being an inherently easier proposition, and therefore being consistently well segmented anyway.

Table 2 shows that there is little difference in the effect of GAN augmentation when using different segmentation networks. This, coupled with the WMH results, suggests that GAN augmentation may benefit any segmentation network, regardless of architecture. Similarly, Figure 2 shows that a similar level improvement is seen across a broad range of additional synthetic data quantities. The amount of synthetic data is therefore not an additional parameter which needs to be finely tuned. This, coupled with the earlier observation that synthetic data rarely impairs performance, makes GAN augmentation a practical proposition.

It is interesting to note that the improvements given by using both traditional and GAN augmentation, as seen in Table 3, are consistently more than the sum of the improvements given by using the two methods separately. This provides strong evidence that the additional information provided by the two augmentation methods are independent. It also suggests that when used together they are potentially synergistic, an observation which agrees with the results in [5]. This could be due to the two methods acting in different ways, with GANs providing an effective alternative to traditional augmentation when attempting to interpolate within the training distribution, but cannot extrapolate beyond its extremes without the aid of traditional augmentation like rotation.

Figure 3 provides an interesting insight into what additional information is being provided by GAN augmentation. In the case of 5 training images, it is clear that each generated image is based heavily on an image from the training set. This is unsurprising as there are very few images to train on, and little variation which can be learned. However, there are subtle differences present in the majority of synthetic images. There are cases where lesions present in the real image are not reproduced in the synthetic image, as well as cases where the shape and number of lesions present in the synthetic image differ from those in the real image. Both of these effects can be extremely valuable to prevent overfitting when training a model - the former decoupling the presence of lesions from the surrounding anatomy, and the latter providing more variety of pathology. When the number of training images increases to 25, we begin to see cases where there are no close matches in the training set, in addition to the cases of differing anatomy and pathology seen previously. This trend gets even stronger in Figure 3 where all 50 training images are used. There are often substantial differences between the synthetic images and their closest real image, suggesting that the

GAN has learned to produce data substantially beyond what was provided to it. We also observe that these modifications appear reasonable in all cases, with no obvious unrealistic lesions or anatomy being synthesised.

#### 4.1 Conclusion

This paper has investigated augmenting training data using GAN derived synthetic images, and demonstrated that this can improve results across two segmentation tasks. The approach has been shown to work best in cases of limited data, either through a lack of real data or as a result of class imbalance. GAN augmentation requires little overhead, involving only the training of a single out-of-the-box GAN, does not involve optimising additional parameters, and has been shown to be low-risk by not hurting performance when training data is limited. A conservative interpretation of the results from the typical tasks explored here suggests that in cases where 5 – 50 labelled image volumes are available, augmenting these with an additional 10 – 100% GAN derived synthetic patches has the potential to lead to significant improvements in DSC.

One major advantage that traditional augmentation has over GAN augmentation is the ability to extrapolate. GANs can provide an effective way to fill in gaps in the discrete training data distribution and augment sources of variance which are difficult to augment in other ways, but will not extend the distribution beyond the extremes of the training data. In general, appropriate traditional augmentation procedures should be used to extrapolate and extend the manifold of semantically viable images. GANs can then be used to interpolate between the discrete points on this manifold, providing an additional data driven source of augmentation. Future work will involve investigating GAN augmentation in other areas, and to evaluate the impact of different GAN architectures.

#### References

1. Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2017)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint arXiv:1701.07875 (2017)
3. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: DRINet for medical image segmentation. *IEEE Transactions on Medical Imaging* (2018)
4. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: *International Conference on Computer Vision*. pp. 2758–66 (2015)
5. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using GAN for improved liver lesion classification. arXiv preprint arXiv:1801.02385 (2018)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in neural information processing systems*. pp. 2672–80 (2014)

7. Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M., Dickie, D., Wardlaw, J., et al.: White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *Neuroimage: Clinical* **17**, 918–34 (2018)
8. Huppertz, H.J., Wagner, J., Weber, B., House, P., Urbach, H.: Automated quantitative FLAIR analysis in hippocampal sclerosis. *Epilepsy research* **97**(1-2), 146–56 (2011)
9. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis* **36**, 61–78 (2017)
10. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017)
11. Krivov, E., Pisov, M., Belyaev, M.: MRI augmentation via elastic registration for brain lesions segmentation. In: *International MICCAI Brainlesion Workshop*. pp. 369–80. Springer (2017)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. pp. 1097–105 (2012)
13. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *CVPR*. p. 4. No. 3 (2017)
14. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are GANs created equal? a large-scale study. *arXiv preprint arXiv:1711.10337* (2017)
15. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: *Medical Imaging 2018: Image Processing*. vol. 10574, p. 105741M. International Society for Optics and Photonics (2018)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015)
17. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: *European Conference on Computer Vision*. pp. 102–18. Springer (2016)
18. 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. 234–41. Springer (2015)
19. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *arXiv preprint arXiv:1703.05921* (2017)
20. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: *CVPR*. p. 5. No. 4 (2017)
21. Valdés Hernández, M.d.C., Armitage, P.A., Thrippleton, M.J., Chappell, F., Sandeman, E., Muñoz Maniega, S., Shuler, K., Wardlaw, J.M.: Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain and behavior* **5**(12), e00415 (2015)
22. Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: *CVPR*. p. 4. No. 3 (2017)
23. Zhu, X., Liu, Y., Qin, Z., Li, J.: Data augmentation in emotion classification using generative adversarial networks. *arXiv preprint arXiv:1711.00648* (2017)