IST-Africa 2023 Conference Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
IST-Africa Institute and IIMC, 2023
ISBN: 978-1-905824-71-7

Abdominal MRI Synthesis using StyleGAN2-ADA

Bernardo GONÇALVES[1,2], Pedro VIEIRA[1], Ana VIEIRA[3]
[1]*Physics Department, NOVA School of Science and Technology, Caparica Campus, Caparica, 2829-516, Portugal, Email: bernardo.goncalves@itgest.pt*
[2]*Bee2Fire SA, Edi. Inov. Point, Sala 2.16, TagusValley-Tecnopolo do Vale do Tejo, R. José Dias Simão, Alferrarede, Abrantes, 2200-062, Portugal*
[3]*Polytechnic Institute of Lisbon - School of Health Technology, Av.D.João II, Lote 4.69.01, Lisboa, 1990-096, Portugal*

Abstract: The lack of labelled medical data still poses as one of the biggest issues when creating Deep Learning models in the medical field. Modern data augmentation techniques like the generation of synthetic images have gained a special interest. In recent years there has been a significant improvement in GANs. StyleGAN2 achieves impressive results in the generation of natural images. StyleGAN2-ADA was created to respond to the lack of training data when training an image synthesis model, which is very frequent in the medical field. Some works used styleGAN to generate melanomas, breast cancer histological images, MR and CT images. In this work we apply, for the first time, a styleGAN2-ADA to a small dataset of abdominal MRI with 1.3k images. From the augmentation pipeline created by the authors of styleGAN2-ADA, we removed all augmentations except the geometric transformations and pixel blitting operations. We trained our network for 70 hours. Our generated dataset has a precision score of 59,33 % and a FID score of 18,14. We conclude that the styleGAN2-ADA is a viable solution to generate MRI using a small dataset.

Keywords: Magnetic Resonance Imaging, Generative Adversarial Networks, StyleGAN2, Image Synthesis, Medical Imaging

1. Introduction

The lack of labelled medical data is a major issue that hinders the creation of supervised Deep Learning (DL) models capable of generalising well in the medical field. Usually, finding a balanced medical imaging dataset with enough images is an impossible task [1]. Additionally, as the model complexity increases its need for data also increases, to avoid overfitting [2]. For that reason, data augmentation techniques have gained a special interest. Traditional methods of data augmentation are already implicitly implemented by many DL libraries. However, some works showed that the introduction of synthetic images in the training datasets, in addition to traditional augmentation, can further improve the results of the DL models, in tasks like segmentation or classification of medical images, as can be seen in the survey papers: [3]–[5].

Since their presentation [6], GANs have been the most studied network type for the generation of medical images. Many variants of the initial architecture were created. Each variant alters the objective function, the structure or the condition of the original GAN. Some even mix variants and therefore change more than one component of the original GAN. The styleGAN belongs to this last group [7].

The work done to create the styleGAN [8] was motivated by the lack of understanding of the GANs generator. Karras et al re-designed the generator architecture. The latent code

of the generator was always provided at the input layer. The styleGAN authors innovated by mapping the latent code into an intermediate latent space before the input layer. This intermediate space controls the generator at each convolutional layer, using the adaptative instance normalization (AdaIN) that creates \textit{styles}, applying affine transforms to the intermediate space. This procedure leads to an improved control over the strength of the image features at different resolutions. This network achieved state-of-the-art results in the unconditional generation of high-resolution natural images.

StyleGAN2 [8] was created by the same group with the objective of improving the quality of the images generated by styleGAN and fixing characteristic artefacts. In this work, Karras et al revised the architecture of the generator, changing the AdaIN operation, improving the progressive growth and adding regularization to the generator. These changes along with tackling the above-mentioned issues, also improve the training performance of the network.

Following the improvements made in styleGAN2 and motivated by the fact that training GANs with small datasets often leads to the overfitting of the discriminator, the same research group released a newly augmentation pipeline and mechanism of adaptive discriminator augmentation (ADA) [9]. This mechanism, jointly with the styleGAN2, can achieve impressive results with limited training data. The common solution to the lack of training data is to perform data augmentation. Training a GAN with traditional data augmentations, such as image rotations, noise addition, flips, etc, is an issue because those augmentations "leak" into the generated images. To avoid the leakage of augmentations, the authors applied those augmentations both to the training set and to the generated images (before the evaluation by the discriminator). Moreover, the performed augmentations need to verify one of the following two conditions: they are invertible or if not, they need to be skipped with a non-zero probability. Their augmentation pipeline consists of 18 transformations. ADA is the mechanism that updates the strength of those augmentations (same value for each) accordingly with the level of overfitting during the training process. Karras et al verified that its augmentation pipeline and ADA stabilize training and improve the quality of the generated images when training with limited data, for example with a training dataset of 2k images. ADA can be applied when training from scratch or while performing transfer learning.

The review papers [3]–[5] show that the usage of GANs for medical image analysis is popular and promising research field. GANs can be useful not only for data augmentation but also for image reconstruction, cross-modality transfer or segmentation [4]. Sorin et al stated that the major benefit of GANs is the ability to increase data quantity and quality at a low cost. GANs can reduce the time spent acquiring medical images by generating the necessary images with high quality [5]. However, there are still some major issues that need to be surpassed: the creation of synthesized data that can be trusted by the clinician; stabilisation of the training process and an improvement in the metrics used to evaluate the generated data [3].

Frequently, mode collapse and other training problems with GANs are related to the lack of training data. Most of the modern GANs are trained with 10^5 - 10^6 images which is an impossible number in the medical field [9]. In this paper we used a small dataset with verified quality of abdominal magnetic resonance images (MRI) - CHAOS Challenge dataset [10] to train a styleGAN2 ADA. Then we used the model to generate a synthetic MRI. Finally, we evaluate and discuss the quality of those images.

2. Related Work

In [4] we can see that GANs were applied with success to the synthesis of radiologic images. Most of the works focused on the synthesis of computed tomography (CT) images and the brain is the most studied organ. In the application of GANs solely for data

augmentation, the review paper, showed two works using abdominal images. [11] marked a breakthrough. They augment a CT dataset of liver lesion patches with synthetic patches created by a Deep Convolution GAN (DCGAN) and verified a classification improvement. Most of the existent works, that augment datasets for classification, synthesize only portions of the original radiologic images, for example, liver lesions, lung nodules, etc. This is beneficial because it decreases the amount of GPU time spent training the network [5].

Fetty et al trained a styleGAN with MR and CT images. The GAN was trained with 17k images for one month. The authors performed a manipulation of latent space, by doing this they were able to select the modality, the gender of the patient, or the slice position of the synthetic images. The manipulation was based on a predictive model of the features of interest (modality, gender, position) trained using the ground data of the original dataset [12].

Gonçalves performed a comparative analysis between traditional augmentation and generative adversarial augmentations to an image classification task. In that work multiple progressive GANs were used to synthesize melanomas images. The used dataset has about 25k images. The GAN with the best results in terms of image quality was the styleGAN2 ADA. However, the classification performance of the model trained with a dataset augmented with synthetic images from styleGAN2 ADA did not surpass the one augmented using classic augmentations. Some issues were identified in those synthetic images such as a checkerboard pattern covering the lesion and an image with two modes mixed [13].

Skandarani et al analysed multiple GANs applied to 3 different datasets: cardiac MRI with 2k training images (ACDC); liver CT with 4k training images (SLiver07) and a retinopathy dataset with 516 images. GANs were trained with a joint distribution of the mask and the image to enable the evaluation with a downstream segmentation task. The best GAN in terms of image quality was the styleGAN2 for the ACDC and SLiver dataset, for the IDRID dataset SPADE GAN was better. Despite the apparent good quality of the generated images none of the networks was able to produce a set of generated images that, when augmenting the original dataset, improved the final segmentation score. The authors also concluded that simpler GANs such as the DCGAN and Wasserstein GAN performed poorly in every dataset [14].

We propose the application of the styleGAN2-ADA, known for its impressive results with small datasets, to the generation of complete MR images. The generated images will be evaluated using well-reviewed evaluation metrics. This work will be the first, to the best of our knowledge, to apply a styleGAN2-ADA to radiologic images.

3. Methods

We used a public database called CHAOS - Combined Healthy Abdominal Organ Segmentation [10]. This database contains abdominal CT and MR images, not related. For this work, only the MR images were used, in particular, the T1 in phase studies. These images have data from 40 patients, each with 26 to 50 MRI slices. In total, our dataset had 1300 DICOM images. We rescaled the intensity of those images to range from 0 to 255 and converted them to png without compression. To ensure that all images had the same resolution, we opt to resize them to 256x256. These steps of data preparation were made with custom code written in Python, using the OpenCV[1] library for image processing and pydicom[2] to read dicom files.

[1] Source: https://pypi.org/project/opencv-python [accessed on 24 March 2023]
[2] Source: https://github.com/pydicom/pydicom [accessed on 24 March 2023]

The application of the styleGAN2-ADA was made with the following GitHub repository: *StyleGAN2-ADA — Official PyTorch implementation*[3]. Custom code was written to run the necessary commands:

1. to prepare the data to match the supported format (*dataset_tool*.py script);
2. to train the network (*train.py*);
3. to compute the evaluation metrics (*calc_metrics.py*);
4. to generate synthetic images (*generate.py*).

The network was trained in a computer with two *Nvidia GeForce RTX 2080*, 64Gb of RAM and an *Intel 9700K CPU*. The training script provides a variety of high and low-level options. Our best model had the base configuration used in [15] for the FFHQ and LSUN datasets. However, we made two modifications in the R1 gamma value and the augmentations. Following the recommendations in the styleGAN2-ADA paper, we experiment with different values of R1 gamma in a range defined by the resolution of our images and our batch size. The final value of R1 gamma was 1.5. For the augmentations, we chose to disable all augmentations except pixel blitting and geometric transformations (90º rotations and flips). With these augmentations, we double the number of images to train our network. Another type of augmentation would change the pixel intensity of the images, which is not beneficial in medical imaging tasks. Although the network paper reports an increase in performance when performing transfer learning, the same did not happen with our dataset, since the available pre-trained models were trained with natural images, a much distinct dataset of our own.

To avoid losing the progress of the training process and to perform post-training evaluation we took snapshots of the network every 40 kimgs. Kimgs is the number of images in thousands shown to the discriminator and it is used to define the duration of the training. The authors of the network state that, in typical cases, 25000 kimgs are necessary to achieve convergence but with 5000 kimgs the results were already reasonable.

To evaluate the generated images during training we chose two metrics: precision and recall. Those metrics were reformulated to the application in generative models in [15]. Precision is the probability of a randomly generated image belonging to the probability distribution of the real dataset. The recall is the reverse. In other words, precision can be seen as a measure of the realism of the generated images and recall as a measure of the variability of the generated images. In addition, we also performed the evaluation of the generated images using the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). KID is similar to FID; both measure the distance between the generated images distribution and the training dataset distribution using an inception network. However, KID is unbiased to the dataset size [16]. For that reason, Karras et al were able to show, experimentally, that KID was best suited for small datasets than FID [9]. All metrics were computed using the complete training dataset and 50k generated images.

4. Results

The network was trained for 70 hours, reaching a total of 7800 kimgs. At this stage, the metrics stabilized and the network stopped learning at a significant rate.

Figure 1 shows the evolution of the precision and recall of the model during the training process. It can be seen as an increasing tendency, especially in precision. Figure 2 and Figure 3 show the evolution of FID and KID across consequent checkpoints. It can be seen as a general decreasing tendency as expected since the optimal value of these metrics is 0. To improve the visualization of the FID and KID evolution we removed the first point of the plot. That point corresponds to the initial state of the network. At that stage, the

[3] Source: https://github.com/NVlabs/stylegan2-ada-pytorch [accessed on 9 March 2023]

values of FID and KID were 346.35, 443.23x10^3, respectively. The metrics were computed for each 200k generated images.

When choosing the best model, we can see that the best metrics do not correspond to the same checkpoint, in fact, for each metric we had a different best checkpoint. For FID and precision, the best checkpoints were in the last stages of our training, 7800 and 7400 kimgs, respectively. For recall, the best checkpoint was at 2600 kimgs. Finally, for KID the best checkpoint was at 1200 kimgs. We chose as our best model the network with the best precision value (Table 1). We used that network to generate synthetic samples of T1 in phase MRI. An example of those samples can be seen in Figure 4. The left column corresponds to images from the training dataset. The right column has generated images.

Table 1 - Evaluation metrics of the images generated by our best model.

Kimgs	Precision	Recall	FID	KID
7400	59.33	15.70	18.14	7.17x10^3

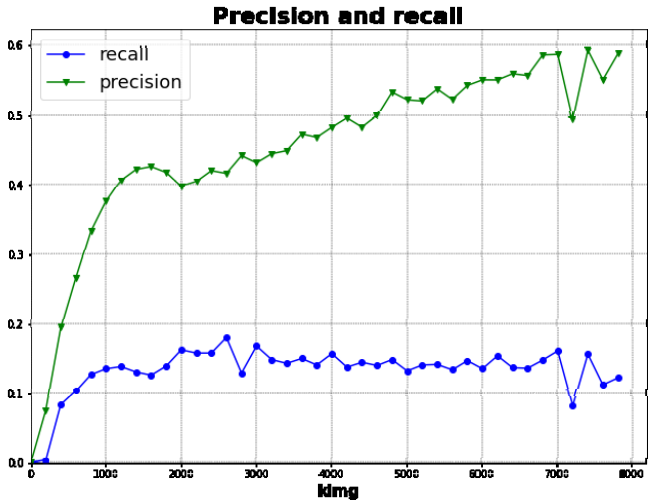

Figure 1 - Precision and recall values computed during the training process. The horizontal axis corresponds to the number of images in thousands shown to the discriminator, the measure of training progress used by the stylegan-ADA.

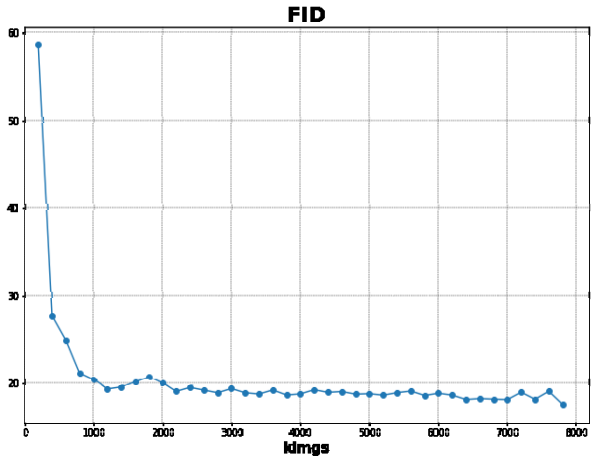

Figure 2 - FID value evolution computed across multiple checkpoints of the training process. The horizontal axis corresponds to the number of images in thousands shown to the discriminator, the measure of training progress used by the stylegan2-ADA.

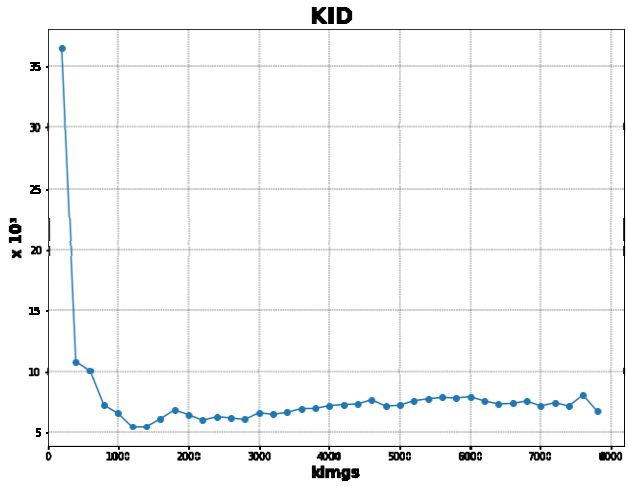

Figure 3 - KID value evolution computed across multiple checkpoints of the training process. The horizontal axis corresponds to the number of images in thousands shown to the discriminator, the measure of training progress used by the stylegan2-ADA. The vertical axis corresponds to the KID values multiplied by 10^3.

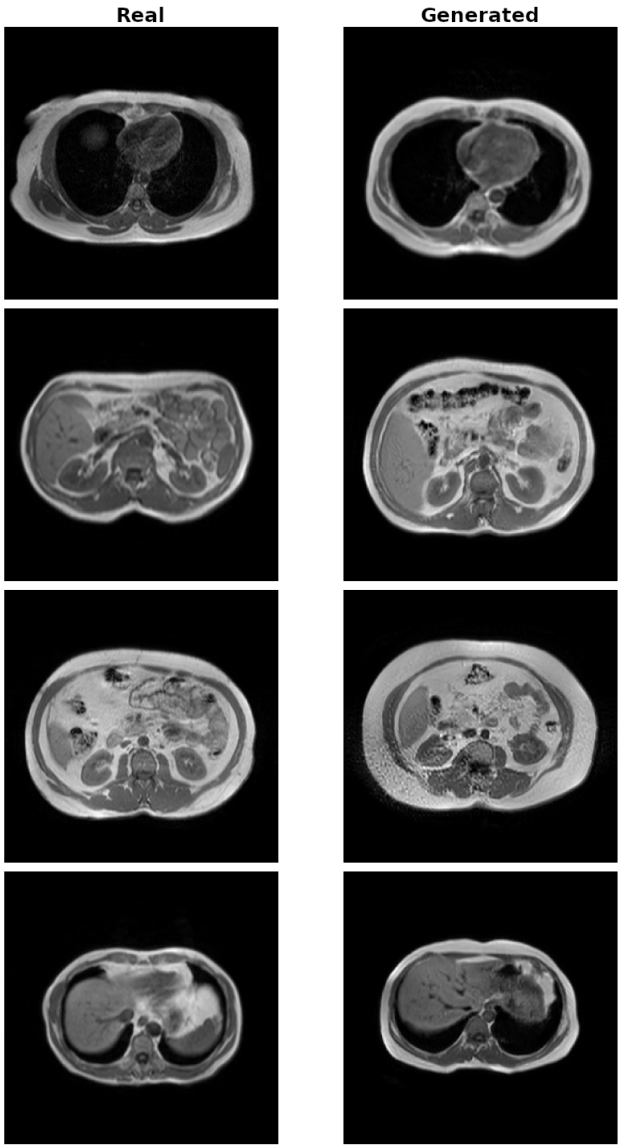

Figure 4 - A sample of real and generated images.

5. Discussion

All the metrics used have in common the fact that they are computed using the complete training dataset and 50k samples of generated images. These sampled images are randomly chosen. For that reason, the same metric computed multiple times in the same checkpoint will produce different results and, in some cases, this fact can even create some outliers, for example, Figure 1 at 7200 kimgs. Besides that, the randomness can also create some abnormal variance in the plots presented, like in Figure 2, from 7000 to 7800 kimgs. However, the general evolution of the metric is as expected in most of the plots. Both precision and recall have an increasing tendency (Figure 1) that suggests an improvement in terms of realism and variability of the generated images, which can be visually analysed in Figure 4. The KID plot, however, does not have a strictly decreasing tendency, in fact, after 1400 kimgs it increases, from 5,50 to 7,93 at 6000 kimgs and finally, it decreases, with an outlier, until the final checkpoint.

A low value of KID or FID could mean a generated dataset with high realism or variability, in other words, these metrics do not make a difference between precision and recall. Also, these metrics do not detect if the GAN is duplicating the training dataset [15]. For that reason, to produce high-quality synthetic MRI in terms of realism we chose our best model as the model with the best value of precision. That model even with a very small (1,3k images) dataset achieves evaluation results in the state-of-the-art range. Gonçalves reported a FID value of 30.79 when using a styleGAN2-ADA to create 256x256 melanomas images [13]. Skandarani et al reported their best FID value of 1.09 when applying a SPADE GAN to the IDRID dataset [14]. Finally, Fetty et al trained a styleGAN with MR and CT images and reported a FID score of 12.3 [12]. Karras et al tested the styleGAN2-ADA with a breast cancer histological dataset achieving a KID score of 2.41 x 10^3 and FID score of 18.22, training from scratch with 162 images reorganized in 1944 partially overlapping crops of 512x512 [9]. Our FID and KID scores of 18,14 and 7.17 x 10^3 fall in the range of the prior obtained values, even though we are using the smallest dataset and only MR sequences.

To show the realism and variability of the synthetic dataset we chose 4 generated images that represent 4 different slices of an abdominal MRI (right column). At the left column we have real slices of MRI that are like the generated ones. The generated images are blurry when compared to the real ones. The third (from top to bottom) synthetic image has some grain-like noise, representing an individual sample with less realism or a noise artefact. Overall, the synthetic images have a good intensity contrast and should be evaluated by the physician to assess their anatomical quality. However, this evaluation was not possible.

We could not find medical-related projects that reported the values of precision or recall of their synthetic datasets. However, in the field of natural image synthesis, in [15] the FID values were related to the values of precision and recall of a styleGAN2. The model optimized for FID (4,5) had a precision of 70 % and a recall of 40 %. On the other hand the model with the highest precision, 82 % (with non-null recall - 25 %) had a FID score of 16,9. The precision and recall values are well above the ones obtained in this project. Which is related to the difference in the dataset size and type mainly. Still, in the field of natural images, we find better results using the same network that we used in this project. For small datasets (about 2k images), styleGAN2-ADA achieve a FID score of 3.05 and a KID score of 0.45 x 10^3 [9]. In fact, the used metrics, are all more appropriated for the evaluation of natural images, as stated in [14] that understand by performing a downstream task that the images that achieved the best FID scores did not always improve the segmentation results when augmenting a medical dataset. These metrics compute the feature vectors of the images using a pre-trained classifier, VGG network for precision and

recall, an inception network for FID and KID. These classifiers are trained with big datasets of natural images and therefore their values when evaluating medical images do not pose as extremely correct.

To surpass the identified limitations, we would like to improve this work by presenting the synthetic dataset to a physician for evaluation and by performing a downstream task with the generated dataset. An extensive analysis of the best evaluation metrics for medical datasets should be made. For example, one could assess the employment of a pre-trained classifier in medical images to compute the feature vectors used to determine the evaluation scores. Furthermore, to improve our network, a thorough analysis of the augmentation types that most benefit from this type of image should be made.

6. Conclusion

With this work, we show that the styleGAN2-ADA is a viable solution to generate medical images with a small training dataset as we obtained state of art evaluation scores. This was the first time that this network was applied to radiologic images, and we expect the applications in this domain to increase in the following years.

Funding

This work was funded by FCT---Portuguese Foundation for Science and Technology and Bee2Fire SA under the PhD grant with reference PD/BDE/150624/2020.

References

[1] L. Lan *et al.*, "Generative Adversarial Networks and Its Applications in Biomedical Informatics," *Front Public Health*, vol. 8, no. May, pp. 1–14, 2020, doi: 10.3389/fpubh.2020.00164.

[2] A. Adadi, "A survey on data□efficient algorithms in big data era," *J Big Data*, vol. 8, no. 1, p. 24, Dec. 2021, doi: 10.1186/s40537-021-00419-9.

[3] S. Kazeminia *et al.*, "GANs for medical image analysis," *Artificial Intelligence in Medicine*, vol. 109. Elsevier B.V., Sep. 01, 2020. doi: 10.1016/j.artmed.2020.101938.

[4] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review," *Acad Radiol*, no. 9, 2020, doi: 10.1016/j.acra.2019.12.024.

[5] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med Image Anal*, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/j.media.2019.101552.

[6] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.2661

[7] S. W. Park, J. S. Ko, J. H. Huh, and J. C. Kim, "Review on generative adversarial networks: Focusing on computer vision and its applications," *Electronics (Switzerland)*, vol. 10, no. 10. MDPI AG, May 02, 2021. doi: 10.3390/electronics10101216.

[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.04958

[9] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training Generative Adversarial Networks with Limited Data," 2020, Accessed: Aug. 04, 2021. [Online]. Available: https://github.com/NVlabs/stylegan2-ada

[10] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data." Zenodo, Apr. 2019. doi: 10.5281/zenodo.3362844.

[11] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018, doi: 10.1016/j.neucom.2018.09.013.

[12] L. Fetty *et al.*, "Latent space manipulation for high-resolution medical image synthesis via the StyleGAN," *Z Med Phys*, vol. 30, no. 4, pp. 305–314, Nov. 2020, doi: 10.1016/j.zemedi.2020.05.001.

[13] G. Gonçalves, "A Comparative Study of Data AugmentationTechniques for Image Classification: GenerativeModels vs. Classical Transformations," 2020.

[14] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for Medical Image Synthesis: An Empirical Study," May 2021, [Online]. Available: http://arxiv.org/abs/2105.05318

[15] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved Precision and Recall Metric for Assessing Generative Models," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.06991

[16] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.01401