Lennart Maack*, Lennart Holstein, and Alexander Schlaefer

GANs for generation of synthetic ultrasound images from small datasets

https://doi.org/10.1515/cdbme-2022-0005

Abstract: The task of medical image classification is increasingly supported by algorithms. Deep learning methods like convolutional neural networks (CNNs) show superior performance in medical image analysis but need a high-quality training dataset with a large number of annotated samples. Particularly in the medical domain, the availability of such datasets is rare due to data privacy or the lack of data sharing practices among institutes. Generative adversarial networks (GANs) are able to generate high quality synthetic images. This work investigates the capabilities of different state-of-the-art GAN architectures in generating realistic breast ultrasound images if only a small amount of training data is available. In a second step, these synthetic images are used to augment the real ultrasound image dataset utilized for training CNNs. The training of both GANs and CNNs is conducted with systematically reduced dataset sizes. The GAN architectures are capable of generating realistic ultrasound images. GANs using data augmentation techniques outperform the baseline StyleGAN2 with respect to the Fréchet Inception distance by up to 64.2%. CNN models trained with additional synthetic data outperform the baseline CNN model using only real data for training by up to 15.3% with respect to the F1 score, especially for datasets containing less than 100 images. As a conclusion, GANs can successfully be used to generate synthetic ultrasound images of high quality and diversity, improve classification performance of CNNs and thus provide a benefit to computer-aided diagnostics.

Keywords: deep learning, medical image analysis, image classification, ultrasound imaging, generative adversarial networks (GANs), synthetic image generation, small datasets

1 Introduction

Ultrasound imaging is among the most cost-effective and portable modalities to acquire medical images today, making it one of the most important tools for diagnosing various diseases such as breast cancer [1]. The acquired images contain information that must be comprehensively analysed by medical experts in a short time. A typical application of medical image analysis is the classification of diseases in ultrasound images. Through the support of different algorithms, additional information is provided to the physician. This increases the chances of accurately identifying incidental findings in an automated manner and can lead to an improved clinical workflow.

Especially deep learning methods like convolutional neural networks (CNNs) gained significant importance due to their superior performance in medical image analysis compared to many explicit algorithms [2]. To be successful, CNNs need a high-quality training dataset with a large number of annotated samples, which are particularly scarce in the medical field. To artificially enlarge the training dataset, typical data augmentation techniques are used [3]. These techniques are limited in creating completely new patterns in the dataset since they use a finite set of known invariances that are easy to invoke [4]. Generative adversarial networks (GANs) showed significant results in the generation of realistic images and can be used to extend the training dataset with synthetic images [5]. In the ultrasound image domain, GANs have been used to generate images to extend a training dataset, which lead to an improved classification performance of fetal brain anomalies [6]. However, GANs require sufficient amounts of training data to be able to synthesise realistic images. The influence of the amount of training data on GAN performance has not been considered in the previous work.

In this work, we analyse the performance of GANs in the case of smaller available medical ultrasound datasets. For this purpose, we systematically reduce the amount of images used for training state-of-the art GANs. Furthermore the GANs' performance and the influence of the corresponding generated synthetic images on the performance of CNNs are evaluated.

2 Methods and Materials

2.1 Dataset

The Breast Ultrasound dataset (BUS) consists of 780 grayscale images with an average image size of 500×500 pixels, collected among 600 female patients aged between 25 and 75

***Corresponding author: Lennart Maack,** Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Am Schwarzenberg-Campus 3, Hamburg, Germany, e-mail: lennart.maack@tuhh.de
Lennart Holstein, Alexander Schlaefer, Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany

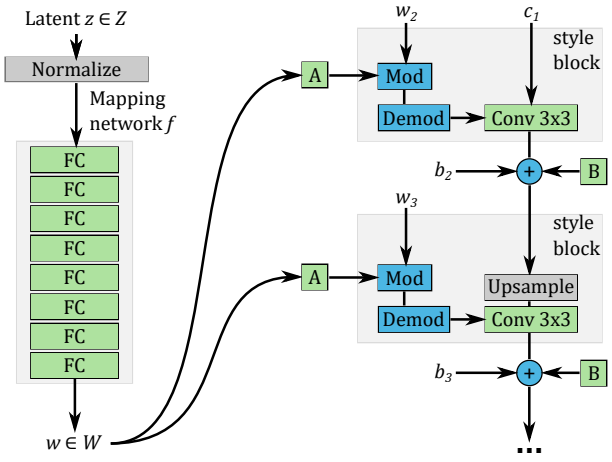

Fig. 1: StyleGAN2 architecture with its two main components mapping network and style block explained in the main text.

years [7]. The images are categorized into the three different classes: normal, benign and malignant. Before a systematic reduction of the BUS dataset can be applied, we split the dataset into a test set with 280 images and a training set of 500 images, with the same distribution of classes across the different splits. All images are resized to 256 $\times$ 256 and normalized.

2.2 GAN architectures and evaluation

The baseline architecture to generate synthetic ultrasound images is StyleGAN2 [8]. The generator architecture is visualized in Figure 1 and utilizes two main components. The first component is the noise mapping network f which consists of eight fully connected layers and takes in a noise vector z as input and maps it into an intermediate noise vector w to get a more disentangled latent space. The second component is the so called style block. It takes the vector w through a learned affine transformation A and converts it to a parameter that scales the initial convolutional weights w_i of the input feature maps and controls the style details of the generated image. After this modulation, a demodulation step to remove the effect of scaling from the statistics of the convolution output feature map is applied. Additionally, bias b_i and a random noise tensor B are inserted after each style block.

To minimize the discriminator's chance of overfitting and prevent the leaking of augmentations to the generated images, we use adaptive discriminator augmentation (ADA) as the first GAN method in this work [9]. ADA implements an adaptive part that dynamically tunes the augmentation strength during the training using the overfitting heuristic r_t. The augmentations used for training consist of pixel blitting and general geometric transformations. All augmentations are invertible and differentiable. The second GAN method used in this work is Differentiable Augmentation (DiffAug), another strategy to circumvent the "leaking" problem by updating the generator with transformed samples using the generator loss L_G [10]:

$$L_G = \mathbb{E}_{z\ p(\mathbf{z})}[f_G(-D(T(G(\mathbf{z}))))]. \quad (1)$$

The following augmentations are used: translation within [-1/8, 1/8] of the image size and padded with zeros, as well as cutout. For all GAN experiments in this work, we adapt the implementation details of StyleGAN2 that achieved state-of-the-art results for the LSUN datasets.

For evaluating the different GAN methods, we use the Fréchet Inception Distance (FID) as the main metric [11]. The FID is determined by measuring the distance in terms of mean and covariance matrix between two data distributions that are calculated from the image features of the real and synthetic images, respectively. In order to further assess the quality of the synthetic images, the mean structural similarity index (SSIM), as well as the Jensen-Shannon distance (JS distance) between the gray value distributions of the real images from the testset and the generated images are calculated.

2.3 Classification evaluation

Synthetic images generated by the different GAN methods are used to improve the classification performance of CNNs with EfficientNetb2 architecture [12]. Our baseline model is pretrained on ImageNet and finetuned with the real dataset only. Other setups to compare with the baseline consist of pretrained models that are finetuned on combined real and synthetic data. The amounts of added synthetic data are 50%, 100% and 200% relative to the real image dataset size. During the experiments, we show that pretrained models with 100% additional synthetic data achieve the best results. Therefore, only these results are presented in section 3.2. All CNN models are trained using stratified k-fold cross validation and evaluated with the F1-score, a harmonic mean of the precision and recall. For our multiclass problem, the F1-score is micro-averaged, i.e. globally counting the total true positives, false negatives and false positives over the three classes.

3 Results

3.1 Image Generation

Figure 2 shows the FID scores of the different GAN models for each dataset size used for training. The FID scores of the baseline range from 140.6 ± 6.9 for the models trained with 500 images to 219.8 ± 27.62 for the models trained with 50

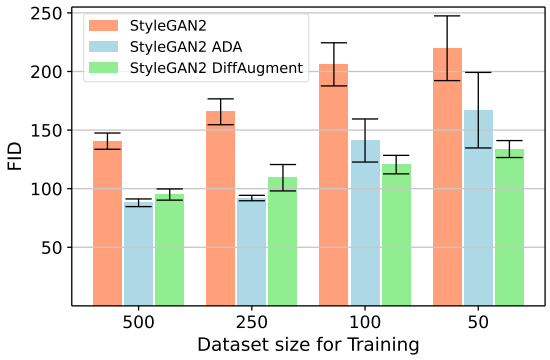

Fig. 2: FID's mean and std. of the different GAN models for each dataset size. Low FID values indicate better GAN performance.

Tab. 1: JS distance and SSIM of synthetic images generated by the GAN models trained with different number of training images.

# images	GAN model	JS (↓)	SSIM (↑)
500	StyleGAN2	0.123	$\mathbf{0.175 \pm 0.05}$
	StyleGAN2 ADA	**0.096**	0.156 ± 0.05
	StyleGAN2 DiffAug	0.1	0.16 ± 0.05
250	StyleGAN2	0.133	0.151 ± 0.04
	StyleGAN2 ADA	**0.091**	0.155 ± 0.04
	StyleGAN2 DiffAug	0.1	$\mathbf{0.163 \pm 0.05}$
100	StyleGAN2	0.163	0.147 ± 0.04
	StyleGAN2 ADA	0.1	$\mathbf{0.158 \pm 0.05}$
	StyleGAN2 DiffAug	**0.097**	0.15 ± 0.04
50	StyleGAN2	0.184	0.135 ± 0.03
	StyleGAN2 ADA	**0.097**	0.138 ± 0.04
	StyleGAN2 DiffAug	0.1	$\mathbf{0.15 \pm 0.04}$

images. For the ADA and DiffAug models, the FID scores range from 87.97 ± 3.28 and 95 ± 4.78 to 166.98 ± 32.23 and 133.80 ± 7.23, respectively. ADA and DiffAug outperform the baseline in terms of FID score. The SSIM score, as well as the JS distance values are displayed in Table 1.

Sample images of each GAN method trained with different dataset sizes are displayed in Figure 3. Whereas ADA and DiffAug generate images with high fidelity, even with only 50 training images available, the baseline GAN synthesises images with a more blurry and wavy pattern the less training images are available. Furthermore, the baseline generates images with lower diversity in comparison to ADA and DiffAug when trained with smaller dataset sizes.

3.2 Image Classification

Table 2 shows the classification results in terms of the F1 score of the CNN models trained with different dataset sizes. The CNN models that use extra synthetic images for training outperform the baseline model for all dataset sizes in terms of the

Tab. 2: Classification results on BUS for CNN models with the respective number of images and type of additional synthetic data generated by different GANs used for training.

# images	Extra synthetic data	F1 score [%]	p-value
500	No synthetic data	81.14 ± 5.69	/
	StyleGAN2	$\mathbf{87.14 \pm 1.95}$	0.084
	StyleGAN2 ADA	83.71 ± 3.49	0.478
	StyleGAN2 DiffAug	85.57 ± 2.12	0.253
250	No synthetic data	75.86 ± 3.19	/
	StyleGAN2	79.29 ± 2.92	0.06
	StyleGAN2 ADA	80.36 ± 2.52	< 0.01
	StyleGAN2 DiffAug	$\mathbf{80.86 \pm 4.31}$	0.078
100	No synthetic data	62.79 ± 5.41	/
	StyleGAN2	71.51 ± 6.51	< 0.01
	StyleGAN2 ADA	$\mathbf{74.14 \pm 4.76}$	0.02
	StyleGAN2 DiffAug	72.79 ± 3.85	< 0.01
50	No synthetic data	59.36 ± 5.22	/
	StyleGAN2	$\mathbf{69.64 \pm 2.42}$	< 0.01
	StyleGAN2 ADA	66.71 ± 4.81	0.03
	StyleGAN2 DiffAug	67.07 ± 3.44	< 0.01

F1 score. There is no indication, that synthetic images generated by a specific GAN result in an increased improvement of classification performance. To check if the F1 scores between the baseline model and the models trained with an extended dataset differ significantly, the pairwise t-test is conducted. A p-value below 0.05 indicates a significant difference.

4 Discussion

The examined GAN models are able to generate realistic ultrasound images that show high quality details and reproduce the typical speckle pattern in ultrasound images. Our qualitative assessment of artefacts and mode collapse detectable in the synthetic images generated by the different GAN models correlates with the corresponding FID scores. The JS distance metric indicates the same trend as the FID score for all models, whereas the SSIM metric shows slightly different results. The FID scores of all models decrease when trained with less than or equal to 100 images, which might be due to the discriminator's overfitting and the lack of useful information fed back to the generator. StyleGAN2 is not able to generate synthetic images with the same quality as ADA or DiffAug. Especially for lower training dataset sizes, the baseline generates synthetic ultrasound images with less fidelity and diversity compared to the two used augmentation methods. The use of data augmentation in GANs leads to the generation of diverse and high quality synthetic samples even with only 50 training images available. All CNN models trained with additional synthetic data outperfom the baseline in terms of mean

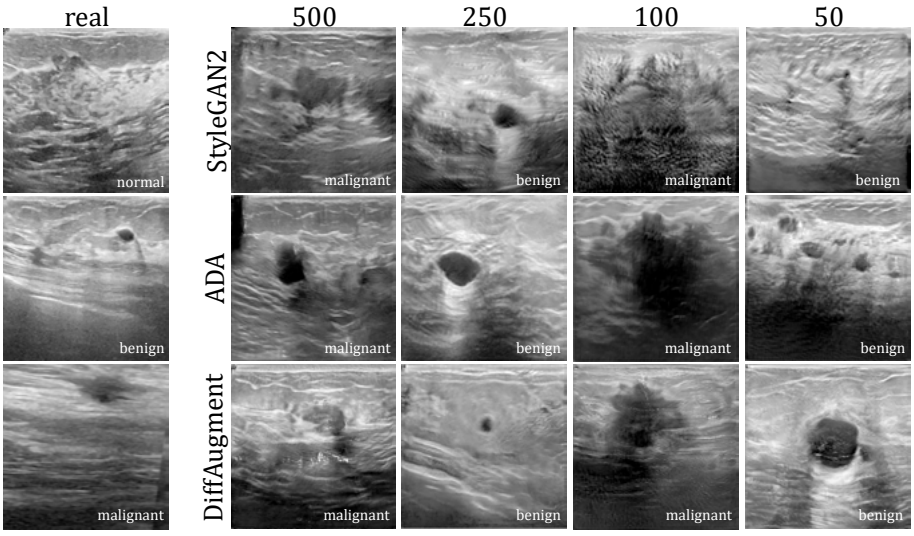

Fig. 3: Sample images from the BUS dataset and synthetic images generated by the different GANs StyleGAN2, ADA, DiffAug (numbers above the column indicate the number of images used for training). The class of the image is displayed in the lower right corner.

F1-score. Especially for dataset sizes 100 and 50, the improvements with synthetic images are statistically significant. However, the higher quality and diversity in the synthetic datasets generated by ADA and DiffAug do not increase the classification performance of the CNNs in comparison to synthetic datasets generated by the baseline GAN. A possible reason for this may be that the synthetic images with lower quality still contain useful new features to successfully enhance the training dataset. Another reason might be, that the particular augmentations applied in ADA and DiffAug do not add much value to the feature space needed to improve medical ultrasound image classification.

5 Conclusion

In this work, we investigate the capabilities of GANs for generating high quality synthetic breast ultrasound images from small datasets. We show that especially data augmentation techniques such as ADA and DiffAug improve the image quality and diversity when only small datasets are available. Synthetic ultrasound images can improve the performance of CNNs used for classification. However, our results also indicate that higher visual quality of synthetic data does not directly correlate with added value for training CNNs.

Author Statement

Research funding: The author state no funding involved.
Conflict of interest: Authors state no conflict of interest.
Informed consent / Ethical approval: Not applicable, since a publicly available dataset was used.

References

[1] Cheng, H.D., Shan, J., Ju, W., Guo, Y., Zhang, L.: Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition 43(1), 299–317 (2010)
[2] Domingues, I., Pereira, G., Martins, P. et al. Using deep learning techniques in medical imaging: a systematic review of applications on CT and PET. AI Rev 53, 4093–4160 (2020).
[3] Shorten C, Khoshgoftaar TM A survey on image data augmentation for deep learning. J Big Data 6:50 (2019)
[4] Antoniou, A., Storkey, A., Edwards, H.: Data augmentation generative adversarial networks. American Economic Journal: Applied Economics (2018)
[5] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Bengio, Y. et al.: Generative adversarial networks. Advances in Neural Information Processing Systems (NeurIPS) 27 (2014)
[6] Montero, A.; Bonet-Carne, E.; Burgos-Artizzu, X.P. Generative Adversarial Networks to Improve Fetal Brain Fine-Grained Plane Classification. Sensors 21, 7975 (2021)
[7] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief 28, 104863 (2020)
[8] Karras, T., Laine, S., Aittala, M., Aila, T. et al.: Analyzing and improving the image quality of stylegan. CVPR pp. 8110–8119 (2020)
[9] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. NeurIPS 33 (2020)
[10] Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. NeurIPS 33 (2020)
[11] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
[12] Tan, M., Le V, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. ICML (2019)