Synthesizing skin lesion images using CycleGANs – a case study

Sondre Fossen-Romsaas[1,*], Adrian Storm-Johannessen[1,*], and Alexander Selvikvåg Lundervold[1,2]

[1]Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[2]Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland University Hospital, Bergen, Norway
[*]These authors contributed equally to the work

Abstract

Generative adversarial networks (GANs) have seen some success as a way to synthesize training data for supervised machine learning models. In this work, we design two novel approaches for synthetic image generation based on CycleGANs, aimed at generating realistic-looking, class-specific dermoscopic skin lesion images. We evaluate the images' usefulness as additional training data for a convolutional neural network trained to perform a difficult lesion classification task. We are able to generate visually striking images, but their value for augmenting the classifier's training data set is low. This is in-line with other researcher's investigations into similar GAN models, indicating the need for further research into forcing GAN models to produce samples further from the training data distribution, and to find ways of guiding the image generation using feedback from the ultimate classification objective.

1 Introduction

Figure 1: Examples of synthetic images of skin lesions generated by our models. From left to right: *Nevus*, *Melanoma*, *Nevus*, *Melanoma*, *Nevus*. A color version of the image can be found here: `https://tinyurl.com/GAN-NIK2020-Fig1`

Deep learning has shown great potential across a variety of medical domains, especially within medical imaging, where convolutional neural networks (CNNs) now form the state-of-the-art approach to many core problems in the field [1, 2]. However, there are many difficult challenges that must be overcome to unlock the full value of these methods [1]. One of which is the models insatiable appetite for training data.

This paper was presented at the NIK-2020 conference; see `http://www.nik.no`.

In practice it is often expensive and difficult to produce large amounts of high-quality labelled data, which is exactly what's needed to construct deep neural network models of practical utility. The problem is particularly severe in medical settings because of strict privacy regulations and the relative rarity of pathological findings. Furthermore, it is not just about quantity: it is crucial that the training data is representative of what the machine learning models will be faced with after deployment. If the training samples are taken from a distribution that differ significantly from the one met in the real world, the models will fail to generalize. Considering the large difference between the high-quality medical images one typically work with when doing research and the messiness of the real, clinical world, this can be a major obstacle to putting deep learning systems into production. See Zech et al. [3] for a recent exploration of this issue. Here, models trained on data pooled from a fixed set of sites were shown to perform significantly worse on new, unseen sites, illustrating the need for domain adaption of machine learning models. Note that this differs from the later results of [4], showing good generalizability for similar, but higher-performing, multi-task models used for analogous tasks.

Recently, *generative adversarial networks* (GAN) [5] have been proposed as a way to adapt models for medical imaging tasks to new domains (*domain adaptation*) and to generate synthetic training data (*data augmentation*). This is motivated by the rapid and impressive progress for GAN-based natural image synthesis. See e.g. BigGAN [6] for natural looking synthetic images, and e.g. [7, 8] for GANs as a data augmentation tool.

There's a growing interest and literature on the subject [9], with multiple success stories across medical imaging domains, e.g. [10, 11, 12, 13, 14, 15]. How well GANs will perform as a general data augmentation tool in their present forms is still far from clear [16], but research progress is rapid. Most attempts at using GANs for data augmentation assume that aiming for visual realism results in synthetic images that are valuable as additional training data for the task at hand, e.g. for disease classification.

In this paper, we demonstrate that modern, state-of-the-art generative models can be used to create realistic-looking synthetic medical images, and investigate their value for data augmentation.

Our target application is skin lesion analysis using dermoscopic images, focusing on skin cancer. As cancer is the second leading cause of death globally [17] and skin cancer is the most common form [18], this is an important area that has seen a lot of attention from the computer vision community lately [19, 20, 21]. We show that using a combination of ACGANs [22] and CycleGANs [23] it is possible to generate images that are close to indistinguishable from real images to an untrained eye. We perform two experiments aimed at generating synthetic class-specific skin lesion images: (i) generating images from random noise using a novel combination of ACGAN and CycleGAN, and (ii) generating images of a specific, rare and important class (melanoma) from another more common class (nevus) using an image transfer approach based on CycleGAN and the so-called *Path-Rank-Filter* of [12]. As melanoma often develops from nevi [24, 25], our hypothesis is that the nevus class provides useful inputs to a pipeline generating melanoma images. We then assess the usefulness of adding the synthetic data to the training data set in a difficult classification task.

Related work

Generative adversarial networks have many potential practical uses in medical imaging. The paper [9] provides a review of results and an overview of the main current applications, i.e. reconstruction, segmentation, classification and abnormality detection.

For the purposes of our work, the most relevant applications are those dealing with image synthesis in the context of data augmentation. Data augmentation is typically based on simple transformations of the images, e.g. scaling, rotations, etc [26], but approaches based on GAN models have recently been proposed.

The work reported in [12] used image-to-image GANs for data augmentation. Using cycle-consistent generative adversarial networks (CycleGANs) in an unpaired image-to-image translation setting, they transformed normal colonic mucosa images (the innermost layer of the colon) to synthetic colonic mucosa images containing an uncommon class of colorectal polyp (an abnormal tissue growth that can lead to colon cancer). The images that where generated was of such a good quality that two out of four gastrointestinal pathologists could not tell the synthesized images apart from the real ones. Additionally, they found that the generated images were useful for data augmentation, as they led to an improved classification model. When the generated images where used in combination with real images the classifier's ROC-AUC score improved from 0.78 on only real images to 0.89 when combined with the augmented images. When only synthetic images where used the AUC dropped down to 0.68, leading to some interesting ideas regarding the usefulness of using a few generated images as opposed to using a fully generated data set.

GANs can also be used to mitigate data imbalances by generating additional images for the classes with low representation. In [27] different GAN models where used to generate realistic high quality images of melanoma lesions. The models where used in a skin lesion classification experiment. They trained a classifier on three different classes, showing that generated images helped by improving accuracy for cases with high class imbalance. To test the utility of the generated images they constructed two baseline models, B_{Full} using the entire training data set and B_{Imb} where they reduced the number of melanoma images to artificially create class imbalance. B_{Full} had a accuracy of 0.9809 on the training set and 0.7160 on validation, while B_{Imb} got 0.8503 and 0.6394. When the generated melanoma images where used alongside the images used for B_{Imb}, they got an accuracy of 0.9929 on the training set and 0.7400 on validation set. Meaning that not only did it improve B_{Imb}, it even surpassed the results obtained in B_{Full}.

Main contributions

1. Our novel combinations of ACGAN, CycleGAN and the Path-Rank-Filter results in realistic-looking synthetic images of skin lesions. In particular, we are able to generate class-specific images of a rare class from images of a more common class, providing a possible way to tackle imbalanced data sets.

2. We investigate to what extent the visually appealing generated images are useful for data augmentation by expanding the training data set of a CNN-based classifier with synthetic images, measuring the effect on the classification performance. We find that the synthetic images have a minor impact, indicating the need for other objectives rather than visual realism when using GANs for data augmentation.

2 Methods and materials

Data and data preparation

We used the training data set from the International Skin Imaging Collaboration (ISIC) Challenge 2019 [28, 29, 30], consisting of 25.331 images, each classified as either Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic

keratosis (AK), Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis) (BKL), Dermatofibroma (DF), Vascular lesion (VASC) or Squamous cell carcinoma (SCC). See Fig. 7 a) for some example images.

For preprocessing the data we used the transforms module in the PyTorch library [31]. First the images were resized to 128x128 for ACGAN training and 256x256 for the CycleGAN model, using bicubic interpolation for image resampling. This was chosen to replicate the setup of the model papers. The images where then converted to tensors before being normalized. Because of the tanh activation in the models, the normalization was done by subtracting 0.5 from the mean and standard deviation for all three color channels, resulting in (close to) zero mean and pixel values in the range of $[-1,1]$.

A ResNet-based skin lesion classifier

To both test the effect of GAN-based data augmentation and to provide a necessary component in the Path-Rank-Filter discussed below, a lesion-type image classification model was needed. The `fastai` library [32] built on top of PyTorch provides an efficient way to create state-of-the-art image classifiers, as the library incorporates a number of modern tricks and techniques for effective model construction and training. In this project we trained a 50 layer ResNet model. Residual networks were introduced in 2015 by Kaiming He et. al. [33], and are based on adding so-called *skip connections* to CNNs. Their model achieved an easy win in the 2016 ImageNet Large Scale Visual Recognition Challenge, and to this day, models based on their ideas form the state-of-the-art architectures for image classification. The ResNet model we used in our work was already pretrained on the ImageNet data set and made available in `fastai`. We used the Adam optimizer [34] during training, and the learning rate finder of `fastai` to find a good base learning rate. The training also employed the 1cycle policy of [35], which first progressively increase the learning rate while at the same time progressively decreasing the momentum, and then does the exact opposite.

Generative adversarial networks

Generative models aim to model data distributions and provide a way to sample from them. The *generative adversarial networks* (GAN) introduced in [5] is a particularly powerful class of such models that has received a lot of attention in recent years. In the original GAN model of [5] there are two deep neural networks, the *generator G* and the *discriminator D*. The generator is fed a vector z, often a random noise vector, producing "fake" data $G(z)$ by sampling from the distribution p_{data} of the training data, attempting to fool the discriminator D, tasked with distinguishing real samples y from those produced by G. During training, the discriminator provides guidance to the generator, each making the other better. See Fig. 2 a) for an illustration. If the training process is successful, the generator can produce synthetic samples with similar properties as the training data.

More precisely, the objective of basic GAN models can be expressed as follows, using the notation in [23]:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G,D) = {} & E_{y\sim p_{\text{data}}(y)}\big[\log(D(y))\big] \\ & + E_{z\sim p_{\text{data}}(z)}\big[\log(1-D(G(z))\big]\big),\end{aligned}$$

where $D(\cdot)$ denotes the probability that the discriminator assigns to its input. The first term on the right represents the real images, while the second accounts for the generated

distribution. G and D together tries to solve the minimax problem

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G,D),$$

aiming for a Nash equilibrium for this two-player non-cooperative game. In other words, the minimax solution is reached when the discriminator cannot differentiate between the generated images and the real images, i.e. $\bar{D} = 1/2$.

Multiple modifications and extensions to this basic setup have been proposed, aimed at providing more stable training and making the generators produce higher quality and more diverse samples [36].

Conditional GANs and the auxiliary classifier GAN

The basic GAN setup can be adapted to the generation of images from specific image classes by providing both the generator and discriminator class labels. I.e. conditioning on class information c when providing random noise inputs:

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(G,D) = & E_{y\sim p_{\text{data}}(y|c)}\big[\log(D(y))\big] \\ & + E_{z\sim p_{\text{data}}(z)}\big[\log(1-D(G(z|c))\big])),\end{aligned}$$

This leads to what is called *conditional GANs* [37] (CGAN), of which there are many variants. In our experiments we used the so-called *auxiliary classifier GAN* (ACGAN) of [22].

ACGANs [22] modifies the CGAN approach by not providing the class information to the discriminator, leaving it to reconstruct that information by itself. This is done using an “auxiliary classifier network" as part of the discriminator, trained on the real images in the training set. This auxiliary network is tasked with reconstructing the class labels in the images it is presented. The generator in ACGANs is as for conditional GANs, aiming to synthesize images of a specific class given a class label and a noise vector.

See Fig. 2 for an illustration of the relation between GANs, CGANs and ACGANs.

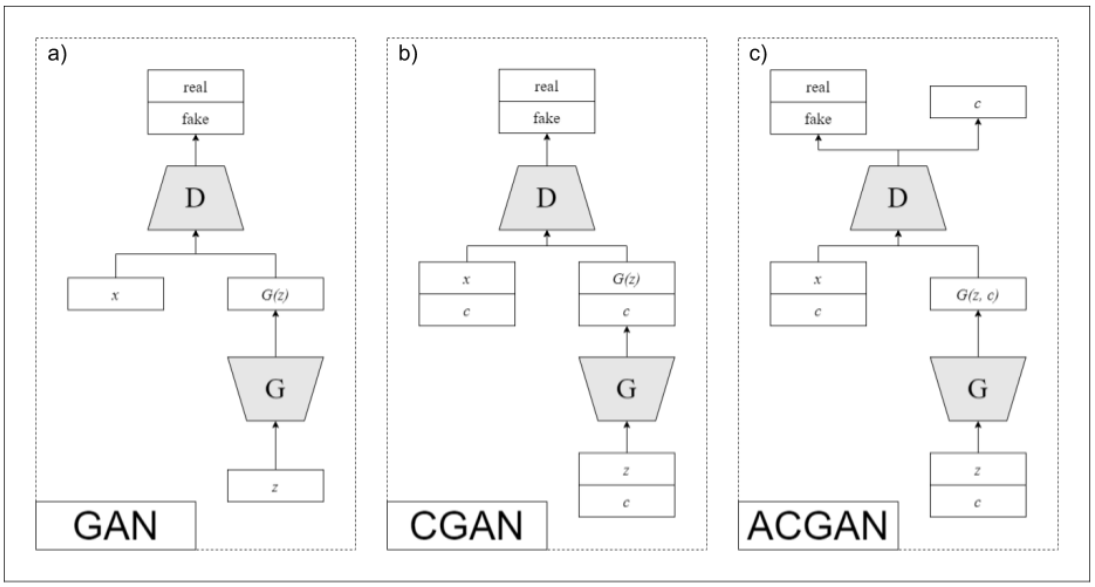

Figure 2: An illustration of the architectural differences between a) the original GAN, (b) the CGAN, and c) the ACGAN.

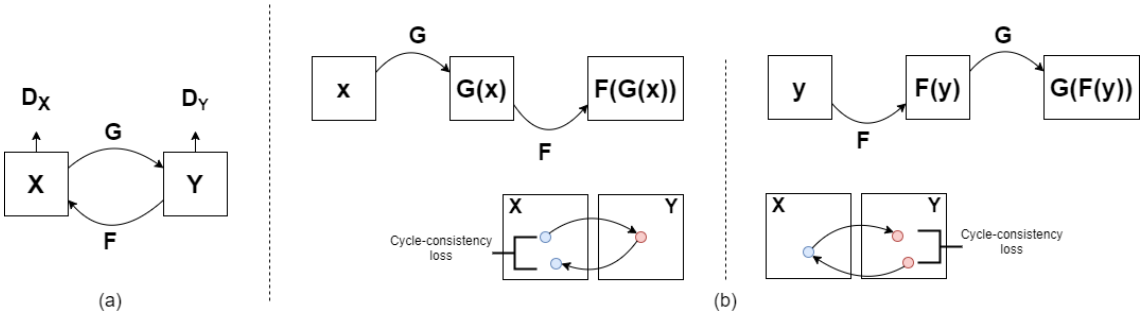

Figure 3: Fig. (a) illustrates the basic adversarial setup, while (b) illustrates Cycle-Consistency Loss. This figure is inspired by Fig. 3 in [23]. Color version available here: `https://tinyurl.com/GAN-NIK2020-Fig3`

Image-to-image translation with Cycle-Consistent GANs

CycleGAN [23] is an *image-to-image* GAN model that generates images not from random noise but from a given image. There are two pairs of generators and discriminators, (G, D_X) and (F, D_Y), enabling image translation back and forth between two image domains X and Y. The generators $G : X \to Y$ and $F : Y \to X$ are trained simultaneously using two paired discriminators and an adversarial loss functions $\mathcal{L}_{\text{GAN}}$, with the goal of having them produce images from the distributions of Y and X, respectively. To preserve information from the source images in the generated images, the transformations are trained to become approximate inverses, $F(G(x)) \approx x$, $G(F(y)) \approx y$, for all $x \in X$, $y \in Y$, i.e. $x \mapsto G(x) \mapsto F(G(x)) \approx x$. This property is called *cycle consistency*, illustrated in Fig. 3. To obtain approximate cycle consistency the so-called *cyclic consistency loss* $\mathcal{L}_{\text{cyc}}$ is used during training. A CycleGAN has the combined objective function

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = &\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{\text{GAN}}(F, D_X, X, Y) \\ &+ \lambda \mathcal{L}_{\text{cyc}}(G, F),\end{aligned}$$

where λ controls the importance given to each type of objective [23].

Experiment 1: Generating images from random noise

From random noise X we use ACGAN to construct a generator G_{ACGAN} that can sample from the class-specific image distribution Y_{c}, resulting in synthetic images for each class.

To further improve image quality, we use CycleGANs trained on each class. This gives us a generator $G : Y_{\text{c}} \to Z$ which we compose with G_{ACGAN} to generate class-specific images (Fig. 4).

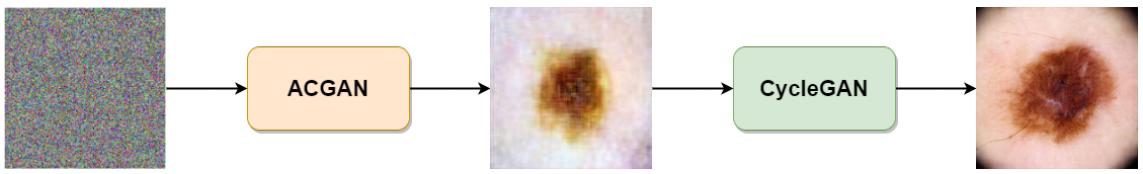

Figure 4: We train an ACGAN model to produce class-specific images from random noise. To further enhance image quality we train CycleGAN models on each class separately and use the ACGAN images of each class as inputs. Color version: `https://tinyurl.com/GAN-NIK2020-Fig4`

For this experiment we used data from the ISIC 2019 Classification Task. We separated the data randomly into a training, validation and a test set (Table 1), and the training and validation sets were used during the construction of the GANs. The test set was used to evaluate the performance of the classifier after it had been trained on the generated images.

Label	Train	Validation	Test
NV	10300	1287	1288
MEL	3618	452	452
BCC	2658	332	333
BKL	2099	262	263
AK	694	87	86
SCC	502	63	63
VASC	202	26	25
DF	191	24	24
Sum	20264	2533	2534

Table 1: Table showing the number of images per class for each data set in Experiment 1.

To evaluate this approach as a data augmentation technique, we trained a baseline model on the real images, then ACGAN and CycleGAN models on five sets of synthetic image-enhanced training sets, with 2000, 4000, 8000, 16000 and 24000 added images, respectively. Each of these 11 models were evaluated on the same test set, consisting of real images not used during construction of the model.

Experiment 2: Generating melanoma from melanocytic nevus

Correctly identifying images corresponding to melanoma is particularly important, as melanoma is behind the vast majority of deaths from skin cancer [38]. As melanoma is relatively rare compared to other kinds of lesions, it is challenging to create classifiers of high recall for this specific class. The sparsity of such samples are also reflected in the ISIC data set, as seen in Fig. 5.

As melanoma images share many characteristics with the much more prevalent melanocytic nevus, the present experiment aims to synthesize samples from the Melanoma class using CycleGAN with Melanocytic nevus as inputs.

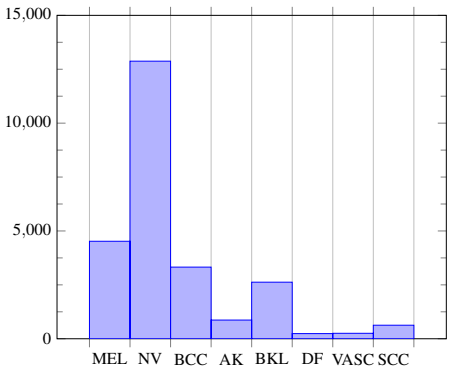

Figure 5: The class distribution in the ISIC 2019 training data set. Note the high number of images from the Melanocytic nevus class (NV) compared to Melanoma (MEL).

To increase the chances that the generator is trained on melanoma images that clearly show melanomas, we use the idea of a *Path-Rank-Filter* from [12]. The confidence for the predictions of the Melanoma class as assigned by the lesion classifier are used to select the training set for a CycleGAN model. The fraction of confident images to include is controlled by a parameter $\alpha \in (0, 1]$. I.e. $\alpha = 0.5$ means that 50% of the Melanoma images are used while the 50% having lower softmax outputs are discarded. See Fig. 6 and [12] for further details. For this experiment, we only use the melanoma and nevus images for

training the CycleGAN (Table 1), while all classes are used to train the lesion classifier used to evaluate the synthetic images data augmentation applicability.

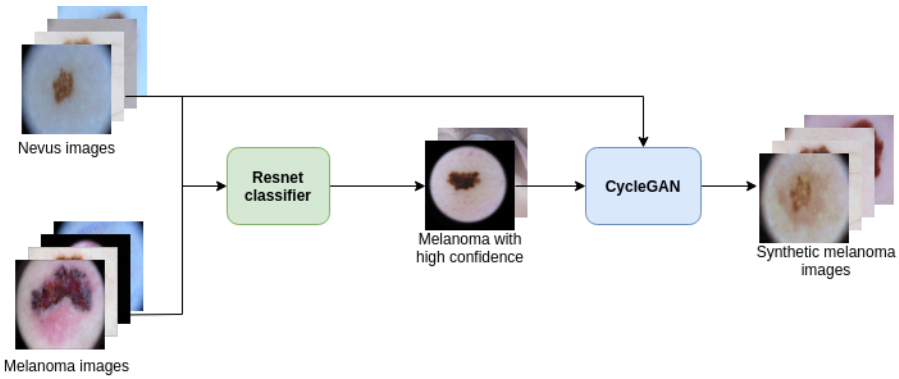

Figure 6: We train a ResNet classifier to construct the set Y of melanoma images predicted with high confidence (Path-Rank-Filter). A CycleGAN model is then trained to obtain a generator $G : X \rightarrow Y$, where X is a set of Melanocytic nevus images. A color version available here: `https://tinyurl.com/GAN-NIK2020-Fig6`

We train four CycleGAN models with four different parameters $\alpha = 1/8, 1/4, 1/2, 1$. To evaluate the effect of including synthetic data when training the lesion classifier, we generated eight different sets of synthetic data from the four different α, resulting in 32 lesion classifiers.

Performance evaluation

Our first objective is to create visually realistic dermoscopic images from each class in the data set. The second objective is to evaluate the synthetic images' usefulness for data augmentation for the lesion classification model. We assess this using confusion matrices, accuracy, sensitivity and specificity, putting particular emphasis on the model's ability to distinguish the *Melanocytic nevus* and *Melanoma* classes.

3 Results

Experiment 1 generating images from random noise produced the results in Fig. 7. A selection of synthetic melanoma images generated in Experiment 2 is shown in Fig. 8.

Adding synthetic images to the training set for the lesion classifier produced the results shown in Table 2 and Table 3.

4 Discussion

In this work we aimed to (i) use GANs to generate synthetic, class-specific realistic-looking dermoscopic skin lesion images, and (ii) investigate their value for data augmentation in a classification setting. Our experiments showed that it is possible to generate images of quite high quality, that for the untrained eye could be taken to be real, directly from random noise and using an image translation approach from one image class to another. In our experiments we were however not able to measure any significant benefit in using these images to increase classifier performance. This differs from some of the findings of other researchers pursuing related approaches, but is in-line with what a number of researchers have discovered during their investigations of GANs for data augmentation, e.g. [39, 40].

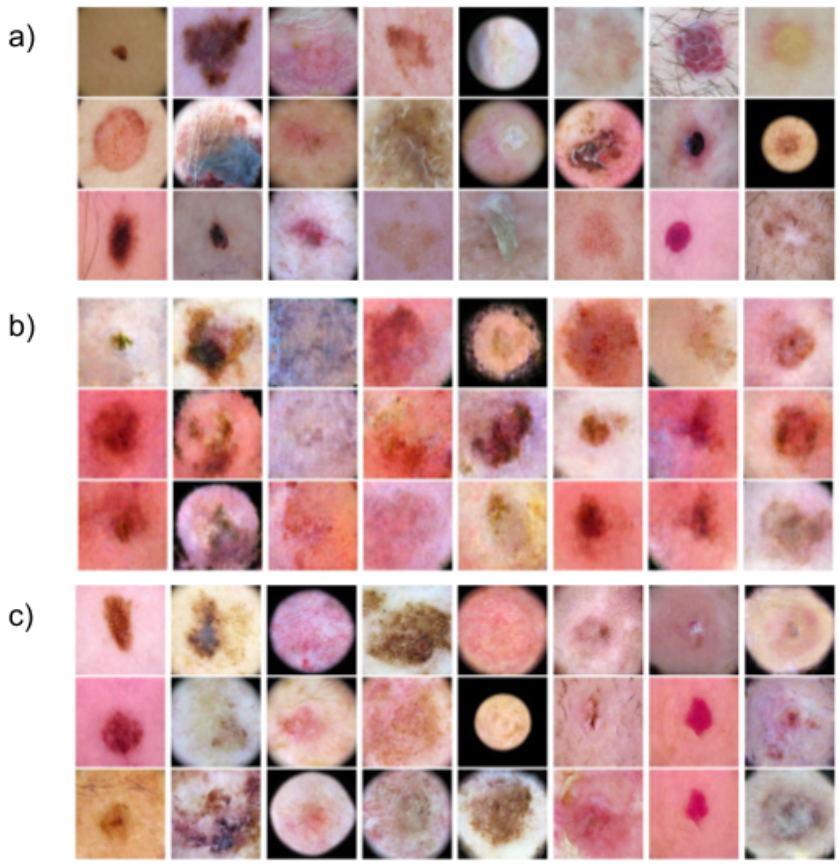

Figure 7: A set of original images from the ISIC 2019 data set are shown in a). Figure b) and c) shows images generated by the ACGAN model and after improvement by the CycleGAN model, respectively. The image classes, from left to right, are *NV, MEL, BCC, BKL, AK, SCC, VASC, DF*. A color version is available here: `https://tinyurl.com/GAN-NIK2020-Fig7`.

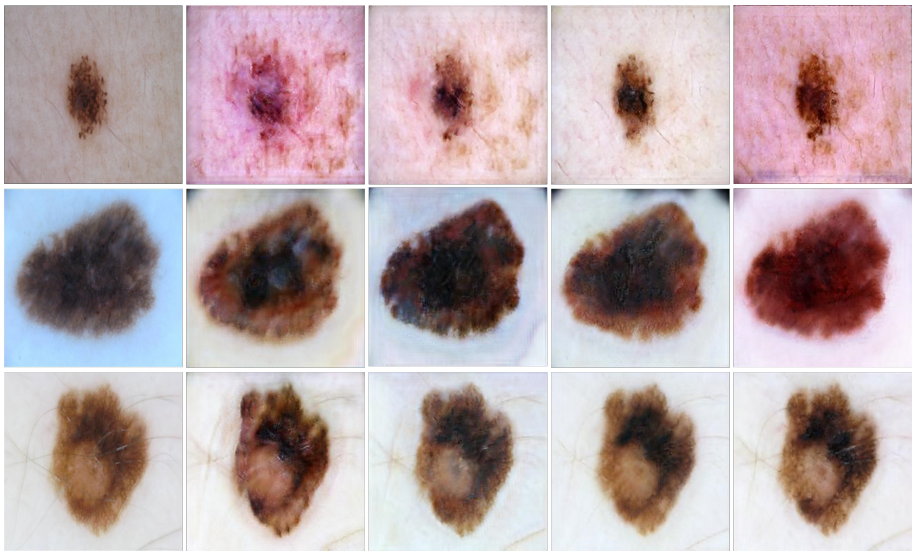

Figure 8: The first column shows three nevus images from the original data set. The four columns to its right shows generated melanoma images with $\alpha = 1/8, 1/4, 1/2, 1$, respectively. Higher α value tend to increase the resemblance with the original image. A color version is available here: `https://tinyurl.com/GAN-NIK2020-Fig8`.

Model	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	Accuracy
ISIC-0	0.823/0.721	0.894/0.939	0.854/0.877	0.761/0.628	0.786/**0.783**	0.708/0.708	0.913/0.84	0.683/0.683	0.856
acgan-0.09	0.763/0.757	0.909/0.911	0.824/**0.925**	0.736/0.616	0.752/0.692	0.727/0.667	0.957/0.88	0.738/**0.714**	0.845
cyclegan-0.09	0.782/0.739	0.903/0.925	0.853/0.889	0.679/0.64	0.78/0.753	0.692/**0.75**	1.0/0.8	0.712/0.667	0.85
cyclegan-0.165	0.803/0.748	0.894/0.932	0.864/0.898	0.651/0.651	0.826/0.741	0.857/**0.75**	0.846/0.88	0.714/0.635	0.856
acgan-0.165	0.811/0.701	0.883/0.94	0.845/0.883	0.675/0.605	0.795/0.738	0.783/**0.75**	0.95/0.76	0.695/0.651	0.847
cyclegan-0.283	0.797/0.748	0.892/0.93	0.861/0.895	0.671/0.64	0.785/0.722	0.762/0.667	0.84/0.84	0.804/0.651	0.851
acgan-0.283	0.812/0.743	0.9/0.938	0.846/0.889	0.671/0.616	0.802/0.753	0.842/0.667	0.88/0.88	0.707/0.651	0.856
acgan-0.441	0.839/0.701	0.885/**0.946**	0.839/0.889	0.685/**0.709**	0.793/0.73	0.783/**0.75**	0.95/0.76	0.698/0.587	0.852
cyclegan-0.441	0.813/**0.759**	0.907/0.942	0.842/0.913	0.738/0.686	0.807/0.73	0.857/**0.75**	0.952/0.8	0.741/0.635	**0.864**
acgan-0.542	0.806/0.706	0.896/0.932	0.844/0.91	0.638/0.698	0.785/0.734	0.8/0.667	1.0/0.8	0.644/0.603	0.848
cyclegan-0.542	0.785/0.752	0.895/0.936	0.845/0.886	0.641/0.581	0.824/0.711	0.762/0.667	0.958/**0.92**	0.696/0.619	0.85

Table 2: Precision/recall for the classifiers. The accuracy is the $F1$ score across all classes. The number of generated images added to the training set is indicated under *Model*. ISIC-0 was built from the original training data.

	$\alpha = \frac{1}{8}$			$\alpha = \frac{1}{4}$			$\alpha = \frac{1}{2}$			$\alpha = \frac{1}{1}$		
		MEL			**MEL**			**MEL**			**MEL**	
Num. gen.	**acc.**	**prec**	**recall**	**acc.**	**prec**	**recall**	**acc.**	**prec**	**recall**	**acc.**	**prec**	**recall**
500 (0.035)	0.852	0.797	0.712	0.853	0.807	0.712	0.857	0.820	0.724	0.856	0.819	0.710
1000 (0.067)	0.851	0.824	0.717	0.848	0.797	0.719	0.861	0.843	0.701	0.851	0.825	0.708
2000 (0.1256)	0.857	0.815	0.730	0.847	0.817	0.719	0.856	0.836	0.732	0.856	0.828	0.712
3000 (0.177)	0.858	0.837	0.715	0.857	0.837	0.728	0.857	0.833	0.726	0.859	0.825	0.746
4000 (0.223)	0.854	0.801	0.746	0.856	0.802	0.728	0.864	0.855	0.715	0.854	0.807	0.730
5000 (0.264)	0.858	0.848	0.717	0.856	0.804	0.754	0.855	0.818	0.737	0.858	0.811	0.741
6000 (0.301)	0.865	0.851	0.759	0.852	0.816	0.724	0.856	0.825	0.730	0.858	0.819	0.741
7000 (0.335)	0.854	0.781	**0.748**	0.864	0.810	**0.763**	0.849	0.783	**0.743**	0.856	0.819	**0.752**
ISIC-0 (0.0)	0.856	0.823	0.721	0.856	0.823	0.721	0.856	0.823	0.721	0.856	0.823	0.721

Table 3: Classification results from experiment 2 across the various number of generated images added to the training set.

In future work it would be natural to investigate whether the visual assessment of image quality is a good metric to use in a data augmentation setting, or whether a direct optimization of classification performance when guiding the image generation would be more advantageous, as in the approach taken by [41].

In general, more research into GANs is needed, both to create models able to produce diverse samples by moving further from the training data distribution, and to make the objective of fooling the discriminator in-line with the ultimate objective of the task the GAN is brought to bear upon.

Acknowledgments

The work was done while S.F-R. and A.S-J. were MSc students at the Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences. S.F-R. and A.S-J. would like to thank the Mohn Medical Imaging and Visualization Centre at the Dept. of Radiology, Haukeland University Hospital, for hosting us while we worked on our MSc thesis project.

ISIC images used in the article are courtesy of the following sources: BCN_20000 Dataset: © Department of Dermatology, Hospital Clinic de Barcelona. HAM10000 Dataset: © by ViDIR Group, Department of Dermatology, Medical University of Vienna; `https://doi.org/10.1038/sdata.2018.161`. MSK Dataset: © Anonymous; `https://arxiv.org/abs/1710.05006`; `https://arxiv.org/abs/1902.03368`

References

[1] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[2] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[3] J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS medicine*, vol. 15, no. 11, 2018.

[4] P. Rajpurkar *et al.*, "Chexpedition: Investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting," *arXiv preprint arXiv:2002.11379*, 2020.

[5] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[6] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," in *International Conference on Learning Representations*, 2018.

[7] A. Antoniou, A. Storkey, and H. Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv preprint arXiv:1711.04340*, 2017.

[8] G. Mariani *et al.*, "BAGAN: Data Augmentation with Balancing GAN," *arXiv preprint arXiv:1803.09655*, 2018.

[9] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, p. 101552, 2019.

[10] M. Frid-Adar *et al.*, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[11] H.-C. Shin *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*, pp. 1–11, Springer, 2018.

[12] J. Wei *et al.*, "Generative Image Translation for Data Augmentation in Colorectal Histopathology Images," *arXiv preprint arXiv:1910.05827*, 2019.

[13] A. Gupta *et al.*, "Generative image translation for data augmentation of bone lesion pathology," *arXiv preprint arXiv:1902.02248*, 2019.

[14] C. Bass *et al.*, "Image synthesis with a convolutional capsule generative adversarial network," *Medical Imaging with Deep Learning*, 2019.

[15] A. C. Quiros, R. Murray-Smith, and K. Yuan, "Pathology GAN: Learning deep representations of cancer tissue," *arXiv preprint arXiv:1907.02644*, 2019.

[16] S. Ravuri and O. Vinyals, "Seeing is Not Necessarily Believing: Limitations of bigGANs for Data Augmentation," *OpenReview*, 2019.

[17] World Health Organization, "Cancer." `https://www.who.int/news-room/fact-sheets/detail/cancer`, 2012. Online; accessed 17 September 2019.

[18] The Skin Cancer Foundation, "Skin Cancer Facts & Statistics." `https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/`, 2019. Online; accessed 17 September 2019.

[19] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[20] H. A. Haenssle *et al.*, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[21] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019.

[22] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2642–2651, JMLR. org, 2017.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[24] W. Damsky and M. Bosenberg, "Melanocytic nevi and melanoma: unraveling a complex relationship," *Oncogene*, vol. 36, no. 42, pp. 5771–5792, 2017.

[25] R. Pampena, A. Kyrgidis, A. Lallas, E. Moscarella, G. Argenziano, and C. Longo, "A meta-analysis of nevus-associated melanoma: Prevalence and practical implications," *Journal of the American Academy of Dermatology*, vol. 77, no. 5, pp. 938–945, 2017.

[26] P. Y. Simard *et al.*, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," in *Icdar*, vol. 3, 2003.

[27] C. Baur, S. Albarqouni, and N. Navab, "MelanoGANs: High Resolution Skin Lesion Synthesis with GANs," *arXiv preprint arXiv:1804.04338*, 2018.

[28] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[29] N. C. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.

[30] M. Combalia *et al.*, "BCN20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[31] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[32] J. Howard and S. Gugger, "fastai: A Layered API for Deep Learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[33] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[36] K. Kurach *et al.*, "A Large-Scale Study on Regularization and Normalization in GANs," in *International Conference on Machine Learning*, pp. 3581–3590, 2019.

[37] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.

[38] American Cancer Society, "American Cancer Society Facts & Figures 2019," 2019.

[39] C. Baur, S. Albarqouni, and N. Navab, "Generating Highly Realistic Images of Skin Lesions with GANs," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 260–267, Springer, 2018.

[40] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv preprint arXiv:1712.04621*, 2017.

[41] F. P. Such *et al.*, "Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data," *arXiv preprint arXiv:1912.07768*, 2019.