Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik
FH Zentralschweiz

Applications of Generative Adversarial Networks to Dermatologic Imaging

Fabian Furger and Ludovic Amruthalingam and Alexander A. Navarini and Marc Pouly

Abstract

Even though standard dermatological images are relatively easy to take, the availability and public release of such dataset for machine learning is notoriously limited due to medical data legal constraints, availability of field experts for annotation, numerous and sometimes rare diseases, large variance of skin pigmentation or the presence of identifying factors such as fingerprints or tattoos. With these generic issues in mind, we explore the application of Generative Adversarial Networks (GANs) to three different types of images showing full hands, skin lesions, and varying degrees of eczema. A first model generates realistic images of all three types with a focus on the technical application of data augmentation. A perceptual study conducted with laypeople con- firms that generated skin images cannot be distinguished from real data. Next, we propose models to add eczema lesions to healthy skin, respectively to remove eczema from patient skin using seg- mentation masks in a supervised learning setting. Such models allow to leverage existing unrelated skin pictures and enable non-technical applications, e.g. in aesthetic dermatology. Finally, we combine both models for eczema addition and removal in an entirely unsupervised process based on CycleGAN. Although eczema can no longer be placed in particular areas, we achieve convincing results for eczema removal without relying on ground truth annotations anymore.

Technical Report 5/2020

DOI: 10.5281/zenodo.3873159

The Technical Report Series

Technical Reports in this series publish research results and working papers from the School of Information Technology at Lucerne University of Applied Sciences and Arts covering a wide range of topics.

Contact

Hochschule Luzern – Lucerne University of Applied Sciences and Arts
Informatik – School of Information Technology

Suurstoffi 41b
Ch-6330 Rotkreuz
Switzerland
www.hslu.ch/informatik

Impressum

Edited by the School of Information Technology at Lucerne University of Applied Sciences and Arts.

Applications of Generative Adversarial Networks to Dermatologic Imaging

Fabian Furger [1] **Ludovic Amruthalingam** [1 2] **Alexander A. Navarini** [3] **Marc Pouly** [1]

Abstract

Even though standard dermatological images are relatively easy to take, the availability and public release of such dataset for machine learning is notoriously limited due to medical data legal constraints, availability of field experts for annotation, numerous and sometimes rare diseases, large variance of skin pigmentation or the presence of identifying factors such as fingerprints or tattoos. With these generic issues in mind, we explore the application of Generative Adversarial Networks (GANs) to three different types of images showing full hands, skin lesions, and varying degrees of eczema. A first model generates realistic images of all three types with a focus on the technical application of data augmentation. A perceptual study conducted with laypeople confirms that generated skin images cannot be distinguished from real data. Next, we propose models to add eczema lesions to healthy skin, respectively to remove eczema from patient skin using segmentation masks in a supervised learning setting. Such models allow to leverage existing unrelated skin pictures and enable non-technical applications, e.g. in aesthetic dermatology. Finally, we combine both models for eczema addition and removal in an entirely unsupervised process based on CycleGAN. Although eczema can no longer be placed in particular areas, we achieve convincing results for eczema removal without relying on ground truth annotations anymore.

1. INTRODUCTION

Generative Adversarial Networks (GANs) initially proposed (Goodfellow et al., 2014) have since then produced impressive results in a variety of synthetic data generation tasks. In contrast to other deep learning methods, which are notoriously data-intensive, GANs achieve good results even with relatively small data sets (Frid-Adar et al., 2018; Baur et al., 2018). This makes GANs attractive for domains where training data is difficult or expensive to obtain. A standard example is the medical field, where specialized machinery may be needed or occurrences of pathologies may be hard to find. Using data sets augmented with GAN-generated synthetic data to train machine learning models has improved performance in a variety of medical domains (Bissoto et al., 2019; Guibas et al., 2017; Hiasa et al., 2018).

Dermatology is one domain particularly suited for the application of deep learning models (Haenssle et al., 2018), but with far too few publicly-available data sets compared to the diversity of the cases encountered in clinical practice. Therefore, the idea to leverage the GAN framework to generate new samples is very promising. In this paper we present our results for two different types of skin lesions: eczema and moles. For eczema we use a private data set (due to identifying patient information) but for moles we use an established public data set for reproducibility and as an example of the generality of our approach.

Besides technical applications such as data augmentation or the creation of paired data, image transformation also enables domain-specific use cases such as prediction of a skin lesion evolution or the evaluation of aesthetic effects of treatment. With this in mind, we train our GAN models to add or remove eczema from skin pictures pursuing two different strategies: a supervised approach where we use ground truth lesion segmentation masks to target modifications to precisely defined areas as well as an unsupervised process entirely freed from the availability of training data.

2. RELATED WORK

2.1. Generative Adversarial Network

GANs distinguish themselves from other generative frameworks by combining a *generator* with a discriminative model, a *discriminator* (Goodfellow et al., 2014). Both models learn by playing an *adversarial* game against each other: the generator produces fake samples while the discriminator

[1]Department of Information Technology, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland [2]Department of Biomedical Engineering, University of Basel, Basel, Switzerland [3]Department of Dermatology, University Hospital Basel, Basel, Switzerland.

attempts to distinguish between real and generated samples. In its original formulation, the generator processes some input vector $\mathbf{z}$ to generate a sample $G(\mathbf{z})$. This vector is typically sampled randomly from a prior distribution, $p_{\mathbf{z}}$, often a standard normal distribution. The generator attempts to produce output that matches the empirical distribution of the real data, p_d.

On the other hand, the discriminator is a regular binary classification model that classifies a given sample into two classes, *real* and *generated*, referring to the source of the sample. During training, the discriminator is shown both real samples drawn from the real data set $\mathbf{x} \sim p_d$ and generated samples $G(\mathbf{z}), \mathbf{z} \sim p_z$ from the generator. For any input sample, the discriminator estimates a likelihood that it is a real and not generated. The discriminator's objective is to confidently determine the origin of a sample, while the generator attempts to produce samples that are mistaken for real data by the discriminator. Formally, the two models attempt to maximize their contrasting objectives:

$$\begin{aligned} \mathcal{L}_G &= \mathbb{E}_{p_z(\mathbf{z})} \log\left(D\left(G\left(\mathbf{z}\right)\right)\right) \quad (1) \\ \mathcal{L}_D &= \mathbb{E}_{p_d(\mathbf{x})} \log\left(D\left(\mathbf{x}\right)\right) + \mathbb{E}_{p_z(\mathbf{z})} \log\left(1 - D\left(G\left(\mathbf{z}\right)\right)\right) \end{aligned}$$

The generator's objective is stated for maximization, which does not saturate during early training and yields better gradients when the generator's samples are confidently rejected (Goodfellow et al., 2014; Jolicoeur-Martineau, 2018).

2.2. Image Translation with GANs

Beside unconditional generation, the generator's output may also be *conditioned* on some specific input, such as images. In this case, its task can be regarded as *image translation*, where images are translated from some input domain to another output domain. Without further restrictions, the generator is not encouraged to produce output that matches its input. The desired aspects of this match or pairing vary from task to task. For instance, when the goal is to modify only a part of the image, a *relevancy loss* may be employed (Andermatt et al., 2018). It penalizes changes to an input image $\mathbf{x}$ outside certain areas, as defined by a segmentation map $\mathbf{y}$. The changes are quantified by the *mean squared error* between the input and output images, both masked by the negative of the segmentation map, denoted by $\odot$:

$$\mathcal{L}_{REL} = \mathbb{E}_{\mathbf{xy}} \operatorname{MSE}\left((1-\mathbf{y}) \odot \mathbf{x}, (1-\mathbf{y}) \odot G\left(\mathbf{xy}\right)\right) \quad (2)$$

However, the relevancy loss relies on the availability of paired segmentations for the processed images. A more general approach for achieving paired translations is described as *CycleGAN* (Zhu et al., 2017). This extended GAN framework combines the two translation directions between two domains, X and Y, by employing two translation generators, G_{XY} and G_{YX}, and two discriminators for the two domains, D_X and D_Y. Paired translations are then achieved by including a *cycle consistency loss* in the generator objectives: when applying both generators in sequence – translating a sample to the other and then back to the original domain – the *reconstruction* of the input should match the original input. The difference between the two samples is typically assessed with MSE. For G_{XY}, this objective is stated as follows:

$$\mathcal{L}_{CYC_{XY}} = \mathbb{E}_{\mathbf{x}} \operatorname{MSE}\left(\mathbf{x}, G_{YX}(G_{XY}(\mathbf{x}))\right) \quad (3)$$

2.3. GANs for Dermatology Images

Machine learning approaches have gained a lot of popularity in the medical field, including the application of GANs. Their most common task is regular image synthesis commonly applied to MR images (Yi et al., 2018). On the other hand, applications in dermatology are less common. One such example is *MelanoGAN* (Baur et al., 2018), which generates images of skin lesions from ISIC 2017 (Codella et al., 2018). The authors compare the results of different GAN models by training a lesion classifier on synthetic data data only. In another work, (Bissoto et al., 2019) generate skin lesions from ISIC 2018 by translating lesion segmentation masks to images. The resulting images are thus directly associated with ground truth segmentations, which can be leveraged for further applications. However, the authors were not able to improve their classification score.

3. METHOD

3.1. Data Sets

Sets of Hands. The first set of experiments is conducted on photos of hands. Each of the 246 individual pairs of hands was photographed from the front and the back side, for a total of 492 photos. They were taken under uniform condition with green background and downscaled to 640×480 pixels. A sample is shown in Figure 1 (upper left).

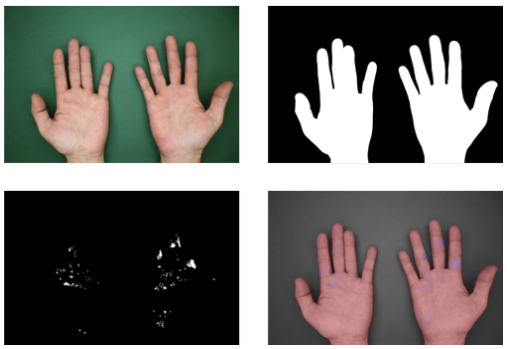

Figure 1. A sample of the EUSZ2 data set (Koller et al., 2018) showing 1) photo of a pair of hands; 2) hand segmentation mask; 3) eczema segmentation mask; overlay of photo and both masks

Patches of Skin. Most of the remaining experiments leverage high-resolution photos (3456×2304 pixels) of the back

side of hands from the EUSZ2 data set collected in the SkinApp project (Koller et al., 2018). There are 79 photos available for training and we use a test set of 52 photos to analyze the overfitting of the discriminator. The photos are annotated with segmentations marking the contour of the hands and eczema lesions as shown in Figure 1. From these photos, we extract patches of skin fulfilling the following criteria: a patch consists of skin only (no background) with a specified amount of skin being afflicted with eczema. We create a data set with *healthy skin* patches and a data set with *skin with eczema* patches, where 10-80% of the skin pixels are annotated as eczema. For these experiments, patches of 128×128 pixels are used. This procedure yields 51023 patches of healthy skin and 2872 patches of skin with eczema. Larger patch sizes yield smaller data sets, especially in the case of skin with eczema. Such smaller data sets significantly increase overfitting.

Skin Lesions. The final data sets consist of dermoscopic images of skin lesions from the ISIC archive 2018 (Tschandl et al., 2018; Codella et al., 2018). In particular, we generate new lesion images of *Dermatofibroma* (DF) and *Melanoma* (MEL) with 115 and 1113 samples available for training, respectively. These different data set sizes allow to analyze the effects on GAN performance. Samples of the two types of lesion are shown in Figure 2. The original images have varying sizes and are resized to a common resolution of 256×256 pixels.

3.2. Model Architecture

This section describes the architecture of the generator and discriminator models for the experiments. Some aspects of these models are based on the architecture of DCGAN (Radford et al., 2015). All models are optimized using Adam (Kingma & Ba, 2014) with learning rate $5e{-}5$ and default moment decays $\beta_1 = 0.9$, $\beta_2 = 0.999$.

3.2.1. Unconditional Generator

The generator for unconditional image synthesis receives a 100-dimensional input vector, which is first passed through a dense layer to produce 64 initial feature maps. The layer's output is reshaped based on the desired aspect ratio of the generated images with lower resolution. Then, a sequence of fractionally-strided convolutions (deconvolutions) increases the image size until the desired output resolution is achieved.

Following common practice, the number of feature maps per convolution are halved at each resolution stage. After each convolution, the output is passed through batch normalization (Ioffe & Szegedy, 2015) and activated with LeakyReLU (Maas et al., 2013). Finally, a regular convolution with 3 output feature maps is activated with $\tanh$ to produce the RGB-channels of the generated image.

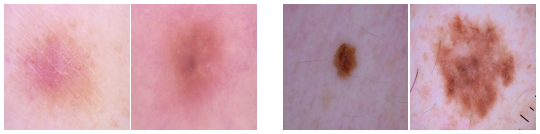

Figure 2. Sample ISIC skin lesions: DF (left) and MEL (right).

Experiment	Dense layers	Deconv	Res.
Full hands (4.1.1)	$20 \times 15 \times 64$	5	640×480
Skin patches (4.1.2)	$8 \times 8 \times 64$	4	128×128
Skin lesions (4.1.3)	$8 \times 8 \times 64$	5	256×256

Table 1. Unconditional generator: image resolution overview.

The hand images generator benefits from unstrided convolutions after each deconvolution to refine the intermediate representations. This is attributed to the comparatively large complexity of these images and does not help with the generation of patches of skin and skin lesions. The size of the initial dense layer and the number of deconvolutions determine the image resolution. Table 1 summarizes the model parametrizations.

3.2.2. Image Translation Generator

The image translation model is based on the U-Net architecture (Ronneberger et al., 2015): an encoder with increasing number of features, which reduces the image resolution, and a decoder to reverse the process. Additionally, the encoded representation is translated with a sequence of residual blocks (He et al., 2015). We find that 2 strided convolutions in the encoder and 2 deconvolutions in the decoder yield the best results. Consequently, the residual blocks translate features with a resolution of 32×32 pixels. We find that 4 residual blocks are ideal, which is surprisingly low but can be attributed to the fact that the skin images are small and relatively simple. Skip connections between the encoder and the corresponding decoder stages are used as suggested by (Isola et al., 2017). These connections forward intermediate features from the encoder that are combined with the decoder features by concatenation.

Finally, we task the image translation generator with *image modification*. To that end, the input image is added to the 3 output channels of the generator, so that it is essentially tasked with generating an image *residual*. The generated residual contains the information to modify the input photo in the desired way.

3.2.3. Discriminator

All experiments leverage the same *multi-scale discriminator* architecture (Wang et al., 2017): Two individual discriminators process an input image and a downscaled version of the image. Afterwards, their outputs are averaged. Thus, the discriminators are simultaneously sensitive to low-level details and high-level structures. We observed that more than

two discriminators do not improve results, which can be explained by our images' lower resolution when compared with (Wang et al., 2017).

Both discriminators have the same architecture: a sequence of strided convolutions with batch normalization and LeakyReLU activation, followed by a dense layer with one output neuron to produce the prediction. The features are doubled after each convolution and the number of convolution layers matches the deconvolution layers of the corresponding generators, as summarized in Table 1. All the image translation experiments operate on patches of skin image with 4-convolution discriminators. As the generators produce normalized images, the channels of the real images are also normalized before discrimination.

3.2.4. Model Balance and Selection

The balance between the generator and discriminator is difficult to maintain, as neither should overpower the other (Yi et al., 2018). Model balance is adjusted by selecting the number of *initial features* of the generator and discriminator. Table 2 summarizes the initial features of all models in this work's experiments. The ideal numbers of features are determined empirically with the restriction of the available GPU memory.

Beside visual inspection, we minimize the Fréchet Inception Distance (FID) (Heusel et al., 2017) to select the best model. The FID measures the dissimilarity between real and generated images, it is commonly used to quantitatively compare the results of GAN models. In our experiments, this metric works well with unconditional generation, but not with image translations. Furthermore, we observe that FID scores computed on different data sets should not be compared as the data set's inherent statistics and variability greatly influence the FID scores.

Model selection is additionally guided by the discriminator's predictions confidence and consistency, which indicate whether the discriminator requires additional capacity to adequately distinguish real and generated samples, and thus, to better guide generator learning.

4. EXPERIMENTS

4.1. Unconditional Dermatology Data Synthesis

The first experiments concern the unconditional generation of dermatology data. The objective is to explore the quality of generated images for different target data sets. The findings indicate the expected performance when the GAN task is not restricted and serves as a baseline for later comparisons with the results of restricted tasks.

For unconditional generation, the generator's input is drawn randomly from a prior distribution, which can be selected freely. We use the common choice of 100-dimensional vectors, where components are drawn independently from a standard Gaussian.

Experiment	Generator	Discriminator
Full hands (4.1.1)	512	32
Healthy patches (4.1.2)	1024	128
Eczema patches (4.1.2)	1024	256
Skin lesions (4.1.3)	512	64
Targeted eczema (4.2)	1024	256
Untargeted eczema (4.3)	1024	256

Table 2. Initial features for the generator and discriminator models.

4.1.1. Sets of Hands

There are two central aspects to the quality of the generated images: high-level structures like anatomy and low-level details like textures. Here, the multi-scale discriminator architecture proves useful, as the two discriminators each focus on one of these aspects. However, many of the generated images still contain visible defects such as hands with more than 5 fingers. These issues are linked to unlikely generator input vectors and can be mitigated using the *truncation trick* (Wang et al., 2017) to improve the quality of the generated images.

The truncation technique includes the truncation of the input below some a priori-defined threshold. Every component of the input vector that exceeds this threshold is re-sampled. Truncation trades sample variability for quality: aggressive truncation significantly reduces variability, while sample quality increases. We determine empirically that a threshold of 0.1 is suitable for the generation of hands, based on the generated samples and FID scores. These scores are summarized in Table 3. Figure 3 shows the results obtained with the selected truncation threshold of 0.1.

While the samples do not show great variability, their quality is generally high. The hands' textures look realistic, the side (front or back) of most pairs of hands can be determined in most samples and most hands consist of four fingers and a thumb. However, convincing samples would still need to be hand-picked, as individual images are not anatomically correct or contain defects near the wrists. These come from sleeves that are visible in a low fraction of the photos.

This application shows that high-resolution dermatology images can be generated with a relatively small data set. These images can be conceivably mistaken for real photos at short glance. The model obtains a FID score of 74.2 without truncation, a significantly lower value than in all other

Threshold	0.01	0.02	0.05	**0.1**	0.2	0.5	1	None
FID	111.4	94.5	75.0	**69.5**	69.5	70.3	74.1	74.2

Table 3. Truncation threshold selection with FID score.

Figure 3. Samples of the unconditional generation of hands.

experiments. This indicates that FID scores on different data sets should not be compared.

4.1.2. Patches of Skin

We further experiment with the unconditional generation of images of healthy skin and of skin that contains eczema. These experiments are a prerequisite for later eczema modification experiments.

Healthy skin. With the large data set of 51023 patches of skin that do not contain any eczema, our GAN is able to generate high-quality images. Samples are shown in Figure 4. The generated samples look very realistic and are also very diverse. Different types of skin, as well as creases and wrinkles are generated. The selected model achieves a FID score of 538.7.

Figure 4. Samples of the unconditional generation of healthy skin.

Skin with Eczema. We observe that the discriminator's task becomes more difficult when classifying patches of skin with eczema, so that the best results are achieved when the discriminator contains more feature maps. Sample results are shown in Figure 5. The quality of the generated images is comparable with the synthetic healthy skin shown in Figure 4. The skin is detailed and contains different kinds of wrinkles and eczema. Overall, there are more creases than in the patches of healthy skin, which is attributed to the increased prevalence of eczema in such areas of the hand. The model achieves a FID score of 599.6 for this task.

Figure 5. Samples of the unconditional gen. of skin with eczema.

Perceptual study. We further evaluate the generated images quantitatively in a perceptual study. The results are presented in Section 4.1.3 along with the analysis of synthetic skin lesion images.

Overfitting. Finally, we analyze the models' overfitting, quantitatively for the discriminator and qualitatively for the generator. For patches of skin with eczema, the discriminator increasingly overfits over the course of the training. Samples from the training set are predicted as real with high likelihood, while testing samples are increasingly being rejected as generated. This is not the case for the discriminator of healthy skin. Figure 6 compares the two models' overfitting.

As the discriminator for skin with eczema has greater capacity, it is more prone to overfitting. However, we find that overfitting is mainly linked to the data set size. Low-capacity discriminators also overfit to the set of 2872 images, while high-capacity discriminator do not overfit on larger data sets.

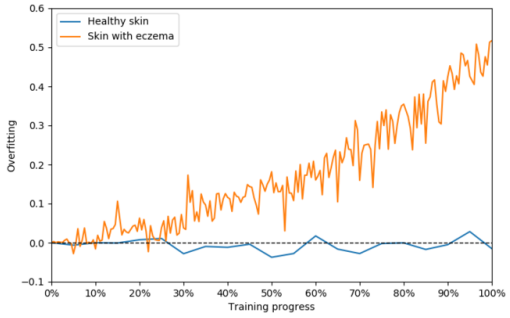

Figure 6. Skin patches discriminators overfitting during training. The lines indicate the difference of the mean predictions on training and test data. Training progress is marked as a percentage of total training epochs: 20 for healthy skin and 200 for skin with eczema.

We further investigate how the overfitting of the discrimina-

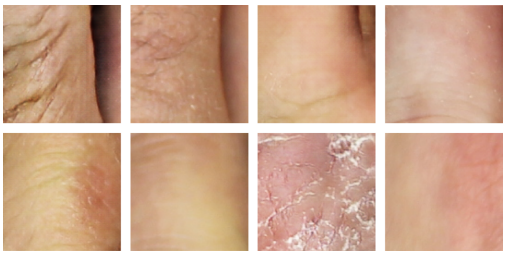

Figure 7. Generated samples (col. 1 and 3) and their nearest training sample (col. 2 and 4).

tor for patches of skin with eczema impacts the generator. We perform a qualitative assessment of the generator overfitting with the common method of comparing generated samples with their nearest training samples (Denton et al., 2015; Karras et al., 2017; Brock et al., 2018). In our experiments, the *structural similarity index* (Wang et al., 2004) yields more similar samples than the *mean squared error*. Sample results are shown in Figure 7. We find that the generated samples do not contain memorized parts of the training set, so we can conclude that the discriminator's overfitting is not leading the generator to overfit as well.

4.1.3. Skin Lesions

Finally, we use our GAN model to generate images of skin lesions. Samples of generated DF lesions are shown in Figure 8 and samples of MEL in Figure 9.

Figure 8. Samples of the unconditional generation of DF lesions.

Dermatofibroma. While they resemble the samples of the training set, they lack variability. Furthermore, they show clear tiling artifacts; patterns that are repeated within a generated image. In this case, the discriminator is trained with only 115 real samples and overfits severely. This visibly impacts the generator: we observe structures, such as lesion shapes or the hairs in the bottom left corners across different samples. With these negative aspects, the generator achieves a FID score of 822.9.

Melanoma. The generated images of MEL lesions contain far greater variability but also suffer from significant tiling. In this case, the generator's FID is 607.8. There is significantly less overfitting, as this data set contains 1113 samples. However, some of the hairs are still repeated. We hypothesize that such specific and distinctive hairs are prone to be copied, as they are rare among the real samples.

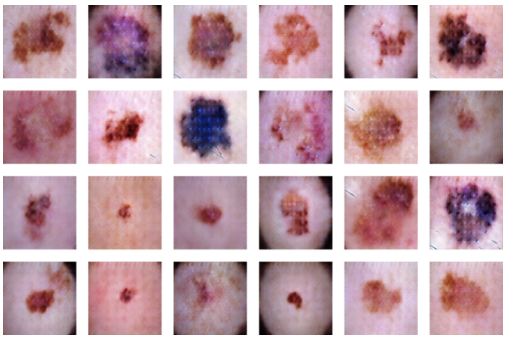

Figure 9. Samples of the unconditional generation of MEL lesions.

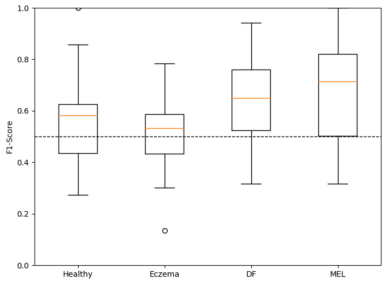

Figure 10. Perceptual study: the box plots show the three quartiles of the obtained F1-scores for each data set.

Perceptual study. We assess the realism of the generated patches of skin lesions with a perceptual study, where we ask 104 participants (laymen without prior training) to determine whether a given image is real or generated. The participants are asked to discriminate 20 images from one of four sets: *patches of healthy skin*, *patches of skin with eczema*, *DF lesions*, and *MEL lesions*. They have 2-3 seconds observation time per image and do not receive intermediate feedback. Such experiments are often conducted to assess if the generated images are easily identified (Salimans et al., 2016; Isola et al., 2017; Wang et al., 2017). The classifications are evaluated with the F1-score and the distribution of the results are visualized per data set in Figure 10. The majority of participants are unable to distinguish real and generated patches of skin, regardless of the presence of eczema: the mean F1-scores are just above random guessing, with 0.58 and 0.53. The third quartiles are also very low, with 0.63 and 0.59. This result confirms that the models are able to generate realistic skin patches. On the other hand, skin lesions are simpler to distinguish, with a mean F1-scores of 0.65 and 0.71. This reflects the observations of the qualitative analysis, where generated lesions look less realistic

than synthetic patches of skin. Interestingly, DF lesions are perceived as slightly more realistic than MEL lesions.

4.2. Targeted Eczema Modification

We formulate eczema addition and removal as an image translation task: the generator receives a skin photo and an eczema segmentation mask as input and should either remove or add eczema within the indicated areas. This is performed by generating a residual, which is added to the input image. To encourage pairing between the generator's input and output, its adversarial objective $\mathcal{L}_{ADV}$ is combined with the *relevancy loss* $\mathcal{L}_{REL}$ (Andermatt et al., 2018), as stated in Equation 2. The importance of the relevancy loss is weighted with λ_{REL}, so that the generator maximizes:

$$\mathcal{L}_G = \mathcal{L}_{ADV} - \lambda_{REL} \cdot \mathcal{L}_{REL} \tag{4}$$

In our experiments we use a weight of $\lambda_{REL} = 10$ for the relevancy loss. We find that this weight places sufficient emphasis on the relevancy objective, while it also maintains the adversarial aspect.

The translations are performed between the data sets of skin with and without eczema, two data sets with very different sample sizes. Thus, the set of patches of healthy skin is truncated to 2872 samples, to match the smaller data set. We use additional healthy skin images to train the discriminator for eczema removal, which effectively prevents overfitting. Furthermore, we use the same segmentation with multiple photos of healthy skin. This also helps with generalization, though the effects of this technique are less pronounced.

4.2.1. Eczema Removal

In figure 11 we show the translation results of removing eczema from afflicted skin. Columns 3 and 6 still show the same parts of hands as the input photos in columns 1 and 4, but they no longer contain the structures and skin disruptions associated with eczema. However, the generated patches generally lose some fine details such as creases, which are often less visible, compared to the inputs. We observe that the FID score applies poorly to the results of image translation. For these experiments, the FID is often oscillating, in this case between 600 and 1100. Thus, we rely on the visual qualitative evaluation of the generated samples.

4.2.2. Eczema Addition

We modify photos of healthy skin by adding eczema to specified areas. Figure 12 shows sample results of this translation. The generator again produces realistic images, as we show in columns 3 and 6. Generally, the structures of the skin are retained and fewer details are lost, compared to eczema removal. Further, realistic-looking eczema is placed in the desired parts of the images. These results show that convincing eczema can be in-painted accurately in the indicated locations, which enables applications such as simulating the progression of untreated eczema.

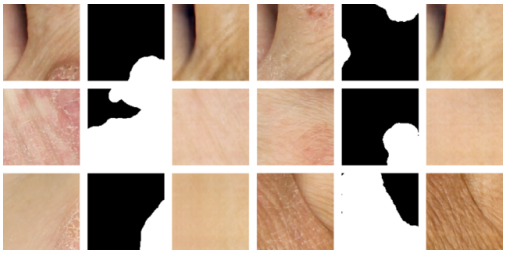

Figure 11. Eczema removal from afflicted skin: col. 1 and 4 show the input photos, col. 2 and 5 the input segmentations and col. 3 and 6 the generation results.

4.3. Untargeted Eczema Modification

We experiment the cyclic translation between patches of skin with and without eczema. No segmentation masks are used and the translations are learned with the completely unsupervised CycleGAN framework (Zhu et al., 2017). As before, the pairing between generator input and output is achieved with an additional generator loss term. In this case, the *cycle consistency loss* $\mathcal{L}_{CYC}$ (Zhu et al., 2017) penalizes differences between a generator's input and its reconstruction, as stated in Equation 3. The generators' combined objective is defined as follows:

$$\mathcal{L}_G = \mathcal{L}_{ADV} - \lambda_{CYC} \cdot \mathcal{L}_{CYC} \tag{5}$$

While placing a greater emphasis on cycle consistency does increase the pairing, this benefit comes at the cost of reduced sample quality. We find that the commonly-used value of $\lambda_{CYC} = 10$ strikes a reasonable balance, like λ_{REL} in the previous experiments.

The same considerations regarding data set size from the previous image translation experiments also apply in this case: The set of patches of healthy skin is reduced to match the set of patches of skin with eczema. Again, extra im-

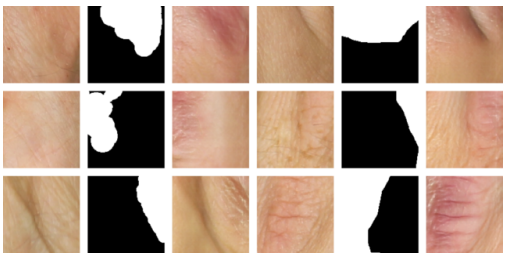

Figure 12. Eczema addition on healthy skin: col. 1 and 4 show the input photos, col. 2 and 5 the input segmentations, and col. 3 and 6 the generation results.

ages of healthy skin are used to train the corresponding discriminator in order to avoid overfitting. Sample results of unsupervised eczema modification are shown in Figure 13.

Figure 13. Cyclic eczema modification without segmentation. Each row shows the translation of corresponding samples: col. 1 and 4 show the sick and healthy input photos, col. 2 and 5 the generated translations without and with eczema and col. 3 and 6 the input reconstructions.

Overall, both generators obtain good results: the generated samples in columns 2 and 5 look realistic and the original inputs of columns 1 and 4 are reconstructed reasonably accurately in columns 3 and 6. We observe that the details of the reconstructed eczema in column 3 do not match the original eczema in column 1. This is to be expected, as the generated patches of healthy skin in column 2 should not contain any hints on where or how to in-paint specific eczema. On the other hand, the second generator properly learns to apply realistic-looking eczema lesions, as demonstrated in column 5. However, the addition of eczema is no longer targeted and can not always be clearly determined.

The loss of details observed in previous translation experiments is barely noticeable here. Indeed, the structures of the original images are mostly retained during both translations. This is likely a positive effect of the cycle consistency objective. The metrics of these cyclic translation experiments are more stable than those of the individual translations. For completeness, we mention that the synthetic patches of healthy skin have a FID of 654.7 to the real data, while the synthetic patches of skin with eczema have a FID of 690.2. These scores are reasonably similar to the scores of unconditional generation, with 538.7 and 599.6, respectively.

5. CONCLUSION

We present different applications of GANs on dermatologic images. First, the common task of unconditional image generation is performed successfully with photos of hands and patches of skin in particular. This is also shown for skin patches in the perceptual study. The validity of our approach is therefore confirmed and our initial objective to create realistic synthetic data achieved.

In the case of generated skin lesions, the results do not look as realistic. This could be corrected by further filtering of the images with rare features (such as hair in our particular case), when compared to the other images in the data set. Our analysis shows that the discriminator already overfits with data sets of several thousand images. On the other hand, we only notice overfitting in the generator when using smaller data sets of merely hundreds of samples. Thus, we conclude that the discriminator complexity should be especially controlled when working with small datasets

In the second part of this work, we explore the task of image modification, with eczema addition or removal within a specified area. The obtained results are again visually appealing but we observe that the FID score may be unsuitable to assess the quality of image translation experiments. In particular, we demonstrate the precise addition of eczema to the areas indicated by the segmentation mask. These results open the door for new applications in dermatology with great value for both doctors and patients, such as anomaly detection in a disease appearance or the visualization of the long term aesthetic effects of a disease.

Finally, we also perform domain translation between healthy skin and skin with eczema lesions in an entirely unsupervised experiment. In particular, the eczema removal results may be interesting for future applications, such as weakly-supervised eczema segmentation similar to (Andermatt et al., 2018). This is certainly the most probable case that researchers will encounter as labeling is a costly step. In practice, before labeling is even considered, it is often necessary to first get prototyping results which could be achieved following this approach.

References

Andermatt, S., Horváth, A., Pezold, S., and Cattin, P. C. Pathology segmentation using distributional differences to images of healthy origin. *CoRR*, abs/1805.10344, 2018. URL http://arxiv.org/abs/1805.10344.

Baur, C., Albarqouni, S., and Navab, N. Melanogans: High resolution skin lesion synthesis with gans. *CoRR*, abs/1804.04338, 2018. URL http://arxiv.org/abs/1804.04338.

Bissoto, A., Perez, F., Valle, E., and Avila, S. Skin lesion synthesis with generative adversarial networks. *CoRR*, abs/1902.03253, 2019. URL http://arxiv.org/abs/1902.03253.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018. URL http://arxiv.org/abs/1809.11096.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., 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.

Denton, E. L., Chintala, S., szlam, a., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1486–1494. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5773-deep-generative-image-models-using-a-laplacian-pyramid-of-adversarial-networks.pdf.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *CoRR*, abs/1803.01229, 2018. URL http://arxiv.org/abs/1803.01229.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Guibas, J. T., Virdi, T. S., and Li, P. S. Synthetic medical images from dual generative adversarial networks. *CoRR*, abs/1709.01872, 2017. URL http://arxiv.org/abs/1709.01872.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., study level I, R., and level II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 05 2018. ISSN 0923-7534. doi: 10.1093/annonc/mdy166. URL https://doi.org/10.1093/annonc/mdy166.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL http://arxiv.org/abs/1706.08500.

Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Prince, J. L., Sugano, N., and Sato, Y. Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size. *CoRR*, abs/1803.06629, 2018. URL http://arxiv.org/abs/1803.06629.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL http://arxiv.org/abs/1502.03167.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Jolicoeur-Martineau, A. Gans beyond divergence minimization. *CoRR*, abs/1809.02145, 2018. URL http://arxiv.org/abs/1809.02145.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL http://arxiv.org/abs/1710.10196.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Koller, T., Schnürle, S., vor der Brück, T., Christen, R., and Pouly, M. Skinapp deeplearning. Technical report, Lucerne University of Applied Sciences, 9 2018.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. URL http://arxiv.org/abs/1511.06434.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. URL http://arxiv.org/abs/1606.03498.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. *CoRR*, abs/1711.11585, 2017. URL http://arxiv.org/abs/1711.11585.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861.

Yi, X., Walia, E., and Babyn, P. Generative Adversarial Network in Medical Imaging: A Review. *ArXiv e-prints*, September 2018.

Zhu, J., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL http://arxiv.org/abs/1703.10593.