Proceedings of Machine Learning Research 116:155–170, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

DermGAN: Synthetic Generation of Clinical Skin Images with Pathology

Amirata Ghorbani* AMIRATAG@STANFORD.EDU
Department of Electrical Engineering
Stanford, CA

Vivek Natarajan NATVIV@GOOGLE.COM
David Coz DCOZ@GOOGLE.COM
Yuan Liu YUANLIU@GOOGLE.COM
Google Health
Palo Alto, CA

Editors: Adrian V. Dalca, Matthew Mcdermott, Emily Alsentzer, Sam Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

Despite the recent success in applying supervised deep learning to medical imaging tasks, the problem of obtaining large and diverse expert-annotated datasets required for the development of high performant models remains particularly challenging. In this work, we explore the possibility of using Generative Adverserial Networks (GAN) to synthesize clinical images with skin condition. We propose DermGAN, an adaptation of the popular Pix2Pix architecture, to create synthetic images for a pre-specified skin condition while being able to vary its size, location and the underlying skin color. We demonstrate that the generated images are of high fidelity using objective GAN evaluation metrics. In a Human Turing test, we note that the synthetic images are not only visually similar to real images, but also embody the respective skin condition in dermatologists' eyes. Finally, when using the synthetic images as a data augmentation technique for training a skin condition classifier, we observe that the model performs comparably to the baseline model overall while improving on rare but malignant conditions.

1. Introduction

The combination of large scale data and compute has catalyzed the success of supervised deep learning in many domains including computer vision (Mahajan et al., 2018), natural language processing (Devlin et al., 2018) and speech recognition (Hannun et al., 2014). Over the last few years, several efforts have been made to apply supervised deep learning to various medical imaging tasks such as disease classification, detection of suspicious malignancy and organ segmentation on different imaging modalities including ophthalmology, pathology, radiology, cardiology, and dermatology (Esteva et al., 2017; Ghorbani et al., 2019; Gulshan et al., 2016; Ardila et al., 2019; Rajpurkar et al., 2017). Despite this progress, developing effective deep learning models for these tasks remain non trivial mainly due to the data hungry nature of such algorithms. Most previous efforts that report expert-level performance

* Work done while interning at Google Health.

required large amounts of expert annotated data (multiple thousands and sometimes even millions of training examples). However, the cost of obtaining expert-level annotations in medical imaging is often prohibitive. Moreover, it is near impossible to collect diverse datasets that are unbiased and balanced. Most of the data used in medical imaging and other healthcare applications come from medical sites which may disproportionately serve certain specific demographics. Such datasets also tend to have very few examples of rare conditions because they naturally occur sparingly in the real world. Models trained on such biased and unbalanced datasets tend to perform poorly on test cases drawn from under-represented populations or on rare conditions (Adamson and Smith, 2018). To ameliorate these issues, generative models present an intriguing alternative to fill this data void.

There has been remarkable progress in generative models in recent years. Generative Adverserial Networks (GAN) (Goodfellow et al., 2014), in particular, have emerged as the de facto standard for generating diverse and high quality samples, with many popular extensions such as CycleGAN (Zhu et al., 2017), StarGAN (Choi et al., 2018), StyleGAN (Karras et al., 2019) and BigGAN (Brock et al., 2018). GAN have been effectively used in many applications, including super resolution (Ledig et al., 2017), text-to-image generation (Zhang et al., 2017) and image in-painting (Pathak et al., 2016). In the medical domain, applications include generating medical records (Choi et al., 2017), liver lesion images (Frid-Adar et al., 2018), bone lesions (Gupta et al., 2019) and anomaly detection (Zenati et al., 2018).

In this work, we explore the possibility of synthesizing images of skin conditions that were taken by consumer grade cameras. We formulate the problem as an image to image translation task and use an adapted version of the existing GAN-based image translation architectures (Isola et al., 2017; Wang et al., 2018). Specifically, our model learns to translate a semantic map with a pre-specified skin condition, its size and location, and the underlying skin color, to a realistic image that preserves the pre-specified traits. In this way, images of rare skin conditions in minority demographics can be generated to diversify existing datasets for the downstream skin condition classification task. We demonstrate via both GAN evaluation metrics and qualitative tests that the generated images are of high fidelity and represent the respective skin condition. When we use the synthetic images as additional data to train a skin condition classifier, we observe that the model improves on rare malignant classes while being comparable to the baseline model overall.

Motivation and distinction from related work In dermatology, prior efforts (Baur et al., 2018a,b; Chi et al., 2018) on applying generative models to synthesize images have focused on datasets of dermoscopic images (Codella et al., 2018; Tschandl et al., 2018). Dermoscopic images are acquired using specialized equipment (dermatoscopes) in order to have a clean, centered, and zoomed-in image of the skin condition under normalized lighting. However, access to dermatoscopes is limited: they are often only available in dermatology clinics and are used to examine certain lesion conditions. On the other hand, clinical images are taken by consumer grade cameras (point-and-shoot cameras or smartphones), and are thus much more accessible to general users. Such images can be used either in a teledermatology setting, where patients or general practitioners can send such photographs to remote dermatologists for diagnosis, or to directly leverage AI based tools for self diagnosis. However, acquisition of such images is not part of the standard clinical workflow, leading to a data void to develop performant skin disease classification models (Yang et al., 2018; Mishra

et al., 2019). Last but not least, unlike dermoscopy images, clinical images of skin conditions have diverse appearances in terms of scale, perspective, zoom effects, lighting, blur and other imaging artifacts. In addition, presence of hair, various skin colors, and body parts, age induced artifacts (e.g., wrinkles), and background also contribute to the diversity of clinical data. Such diversity makes it challenging for generative models to learn the underlying image representation. Fig. 1(a) contrasts examples of a dermascopy dataset (on the left) (Tschandl et al., 2018) to that of a clinical dataset (on the right). To the best of our knowledge, no prior work has attempted to synthesize clinical images with skin pathology.

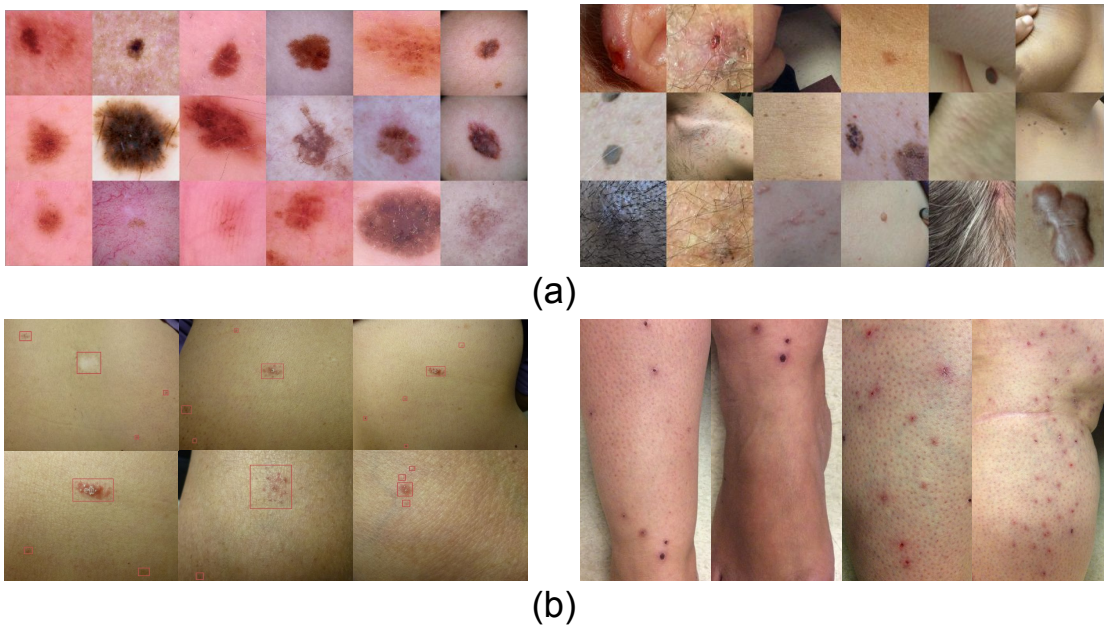

Figure 1: **Sample skin images in different datasets** Comparison between samples of dermatoscopic images (left) (Tschandl et al., 2018) and samples of cropped clinical images in our dataset (right) are shown in (a). Our original, uncropped images are shown in (b), collected from a real-world teledermatology service with varying size, scale, and quality.

2. Methods

Dataset For this work, we used a dataset provided by a teledermatology service, collected in 17 clinical sites in two U.S. states from 2010 to 2018. This dataset consisted of 9897 cases and 49920 images; each case contains one or more high resolution images (resolution range: 600×800 to 960×1280). Ground truth of the skin condition was established for each case by an aggregated opinion of several board-certified dermatologists to differentiate among 26 common skin conditions and an additional 'other' category. It is important to note that even though the 26 skin conditions are known to be highly prevalent, the dataset itself was

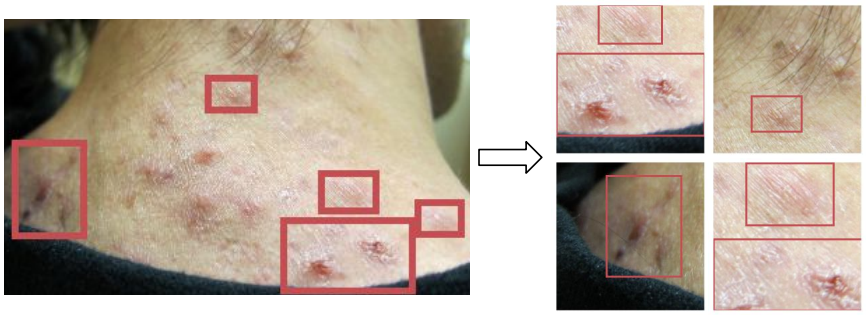

Figure 2: **Image pre-processing** Using the human-provided ROI annotation of skin conditions, we are able to create cropped images with clear skin condition in focus.

unbalanced, especially for certain malignant conditions like *Melanoma*, which has less than 200 examples. More details on the original dataset can be found in (Liu et al., 2019). In addition to the skin condition, we make use of two additional pieces of information: 1) For each condition, its presence in the image is marked by a Region of Interest (ROI) bounding box (Fig. 1(b)) and 2) the skin color given for each case based on the Fitzpatrick skin color scale that ranges from Type I ("pale white, always burns, never tans") to Type VI ("darkest brown, never burns") (Fitzpatrick, 1988). Both the ROI and the skin color annotations are determined by the aggregated opinions of several dermatologist-trained annotators.

Data Preprocessing Fig. 1(b) shows the heterogeneous nature of this dataset. As stated previously, the region occupied by the skin condition varies significantly and the backgrounds are non-uniform and unique to each individual image (walls, hospitals, clothing, etc). As a result, the signal to noise ratio is very low in most of the images. To alleviate this problem, using the annotated ROI bounding boxes, we create a more uniform version of the dataset where the skin conditions is prominent in each image (Fig. 2) We devise a simple heuristic that crops a random window around an ROI or a group of adjacent ROIs while removing the presence of background information. This results in 40000 images of size 256×256 for training the generative models and 24000 images for evaluation.

Problem Formulation Given a set of input-output pairs $\{(x_i, m_i)\}_{i=1}^{N}$, for each image $x_i \in R^{W \times H \times C}$, $m_i \in R^{W \times H \times C'}$ is its corresponding semantic map that encodes the skin color, the skin condition present in the image and the location of the condition in the image (Fig. 3). For a fully defined semantic map m, due to the possible variations (amount of hair on the skin, shooting angles, lighting conditions, morphology of the condition, etc), the corresponding image x is not unique. The variations can be modeled by a conditional probability distribution $P(x|m)$. Our goal is to be able to sample from $P(x|m)$ for arbitrary and valid m. This image to image translation problem can be addressed using the conditional GAN framework (Mirza and Osindero, 2014) which has been successfully used in similar settings (Isola et al., 2017; Wang et al., 2018; Chen and Koltun, 2017; Choi et al., 2018).

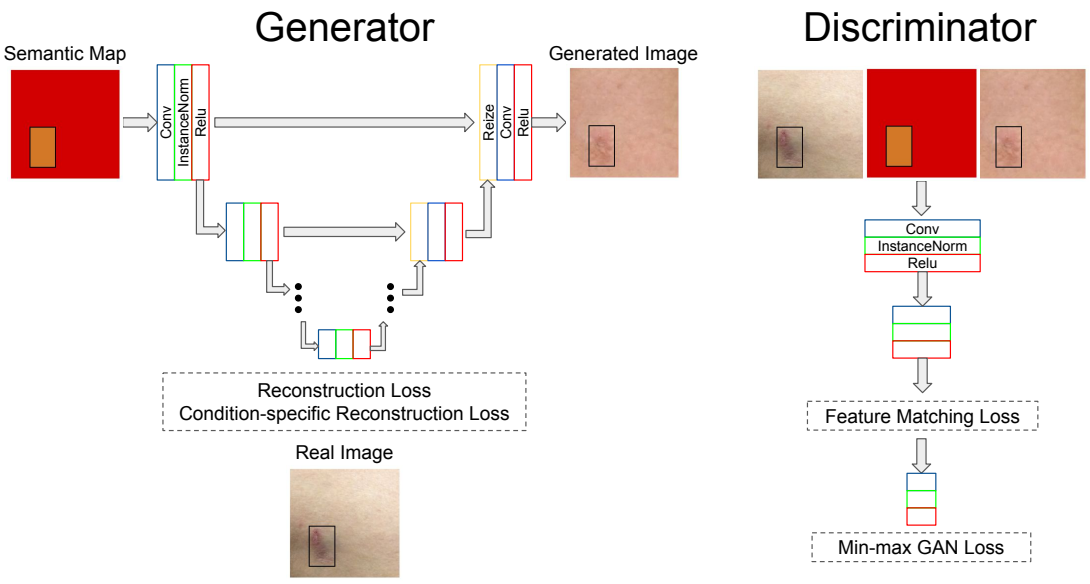

Figure 3: **DermGAN architecture** A semantic map encoding the skin condition and its region of presence (orange rectangle) and the skin color (red background) is passed through the generator to produce a synthetic image. The generator is a modified U-Net (Ronneberger et al., 2015) where the deconvolution layers are replaced with a resizing layer followed by a convolution to mitigate the checkerboard effect. The discriminator has a fully-convolutional architecture. The two architectures are trained to minimize four loss components: ℓ_1 reconstruction loss for the whole image, ℓ_1 reconstruction loss for the pathological region, feature matching loss for the second to last activation layer of the discriminator, and the min-max GAN loss.

For each image in our dataset, the semantic map is an RGB image. The R-channel encodes the skin color and the condition is encoded in the G & B channels by a non-zero value corresponding to its ROI bounding box(es). An example is shown in Fig. 3.

Given the pairs of preprocessed skin images and their semantic maps, the problem of synthetic image generation reduces to mapping any arbitrary semantic map to a corresponding skin condition image. Pix2Pix (Isola et al., 2017) model gives a two-fold solution to this problem: An encoder-decoder architecture such as U-Net (Ronneberger et al., 2015) is trained with an ℓ_1 reconstruction loss to reproduce a given real image from its semantic map. The main drawback, however, is that such a model produces blurry images that lack the details of a realistic image. Therefore, a second model is added to discriminate real images from synthetic ones. The addition of this min-max GAN loss results in generation of realistic images with fine-grained details. Later on, subsequent works improved the Pix2Pix method by applying various adaptations to the original algorithm: using several discriminator networks with various patch-sizes, progressively growing the size of generated images, using conditional normalization layers instead of instance normalization layers, and so forth (Park

et al., 2019; Choi et al., 2018; Lin et al., 2018). Similarly, in this work, based on the specifics of our data modality we apply three main adaptations to the original pix2pix algorithm:

- **Checkerboard effect reduction** The original pix2pix generator implementation makes use of transposed convolution layers. As discussed by Odena et al. (2016), using deconvolution layers for image generation can results in "checkerboard" effect. The problem was resolved by replacing each deconvolution layer with a nearest-neighbor resizing layer followed by a convolution layer.
- **Condition-specific loss** The original pix2pix loss function uses the ℓ_1 distance between the original and synthetic image as a loss function component. For skin condition images, generator model's reconstruction performance is more important in the condition ROI compared to its surrounding skin. Therefore, we add a condition-specific reconstruction term which is simply the ℓ_1 distance between the condition ROIs in the synthetic and real images. We should mention that unlike the reconstruction loss, adding an additional condition-specific discriminator model in order to include a condition-specific min-max GAN loss did not result in improvement.
- **Feature matching loss** Feature matching loss enforces the generated images to follow the statistics of the real data through matching the features of generated and real images in a chosen layer(s) of the discriminator. It is computed as the ℓ_2 distance between the activations of synthetic images in a chosen discriminator layer (or layers) and that of the real images. Apart from improving the quality of generated images, feature matching loss results in a more stable training trajectory (Salimans et al., 2016). We used the output of the discriminator's second to last convolutional layer to compute the feature matching loss.

All in all, the resulting model has four loss terms: reconstruction loss, condition-specific reconstruction loss, min-max GAN loss, and feature-matching loss. Grid-search hyperparameter selection was performed to choose the weighting coefficients for each loss component.

3. Experiments

Using the pre-processed dataset, we trained a DermGAN model to generate synthetic skin images with a chosen skin color, skin condition, as well as the size and region of the condition. In order to focus more on the critical and rare conditions, of the 26 classes in the original data, we choose 8 conditions that have fewer samples compared to other classes (17% of the dataset combined). In order to generate synthetic images, we first need to generate the corresponding semantic map. Different conditions have semantic maps with different statistics of bounding box size, shape, etc. As a result, in order to prevent domain shift between semantic maps the DermGAN is trained on and the ones it will see during test time, we use the maps in the validation set. For each synthetic image, we randomly sample a semantic map from the validation set and then apply random transformations on it: We apply random translations on the bounding boxes and if there are more than one bounding boxes in the map, we randomly select a subset of them. As a result, the same semantic map in the validation set could be used to generate diverse synthetic images. Examples of our generated images are shown in Fig. 4.

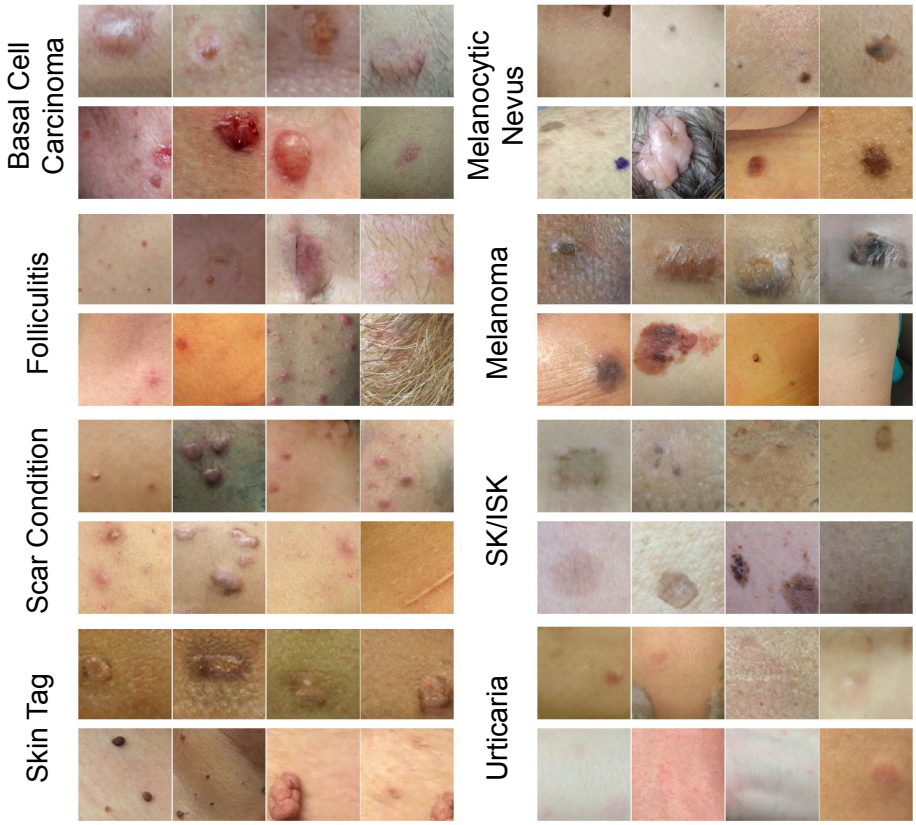

Figure 4: **Generated vs real images** For each condition, the top row shows samples of generated images and the bottom row shows samples of real images.

Synthetic images with different skin colors In this and the subsequent experiment, we train a DermGAN model on all of the 26 conditions of the dataset to use images of wider demographics. For a given semantic map in the test set, we vary the encoded background color and observe the respective changes in the generated image. Fig. 5 depicts examples of this experiment, in which the encoded skin color of a semantic map is replaced with each of the six types. As illustrated in the figure, the DermGAN model is able to change the background skin color while adjusting the condition itself to reflect this change. For instance, for *Melanocytic Nevus*, the generated image for the darker tones has also a darker mole, which mimics real data.

Synthetic images with different sizes of the skin condition For a given semantic map, we can vary the sizes of the pathological region for each skin condition and observe the respective changes in the generated image. Fig. 6 shows examples of this experiment, in which the size of the bounding box of a semantic map is gradually increased. We observe that as the size of the skin condition changes, the visual appearance also changes, which is consistent with real world occurrences. Note that in this experiment, the semantic maps we

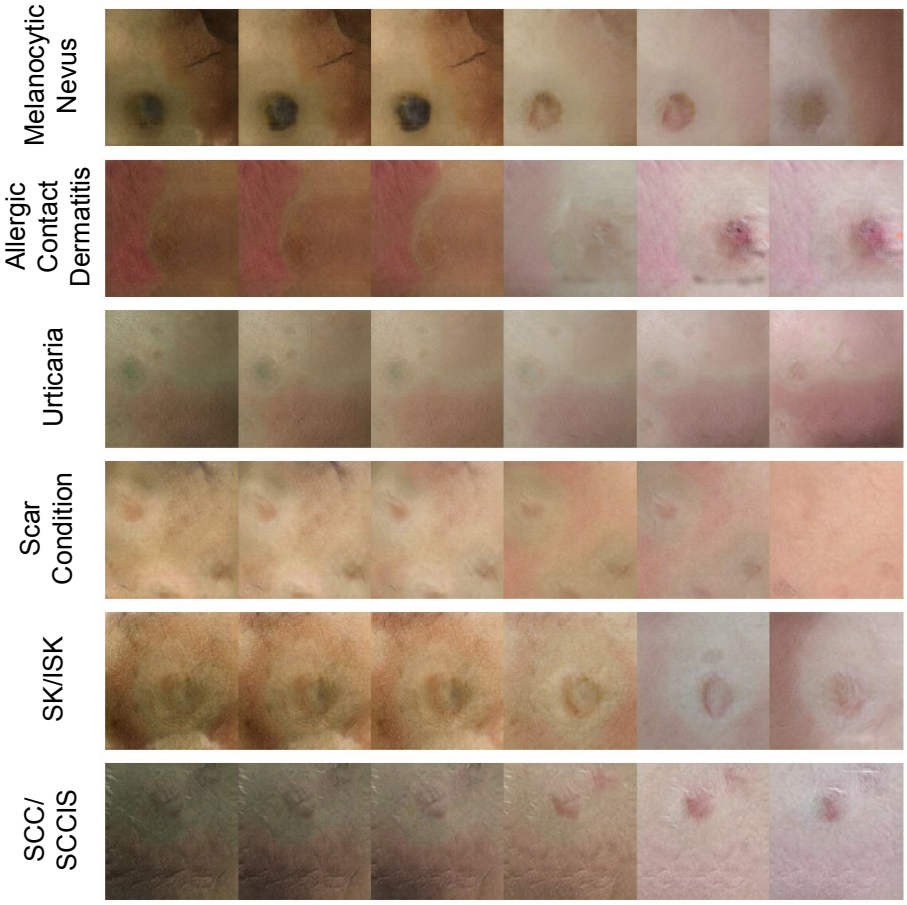

Figure 5: **Skin color variation** By changing the encoded skin color in the semantic map, we are able to create synthetic images with diverse skin colors. It should be noted that as the skin color varies, the change in surrounding visuals is consistent with real world occurrences.

fed into the model are generated synthetically (not by applying random transformations on semantic maps in the validation dataset.)

GAN evaluation metrics A perfect objective evaluation of GAN generated images remains a challenge (Theis et al., 2015). One widely-used measure is the inception score (Salimans et al., 2016) that works as a surrogate measure of the diversity and the amount of distinct information in the synthetic images. It is computed as the average KL-divergence between the class probabilities assigned to a synthetic sample by an Inception-V3 model (Russakovsky et al., 2015) trained on the ImageNet dataset (Szegedy et al., 2016) and the average class probabilities of all synthetic samples. The main drawback that makes the use of inception score inadmissible in our case is that it assumes the classes in the dataset at hand to be a subset of the 1000 ImageNet classes (Barratt and Sharma, 2018). Another widely-used measure is the Frechet Inception Distance (FID) (Heusel et al., 2017). FID

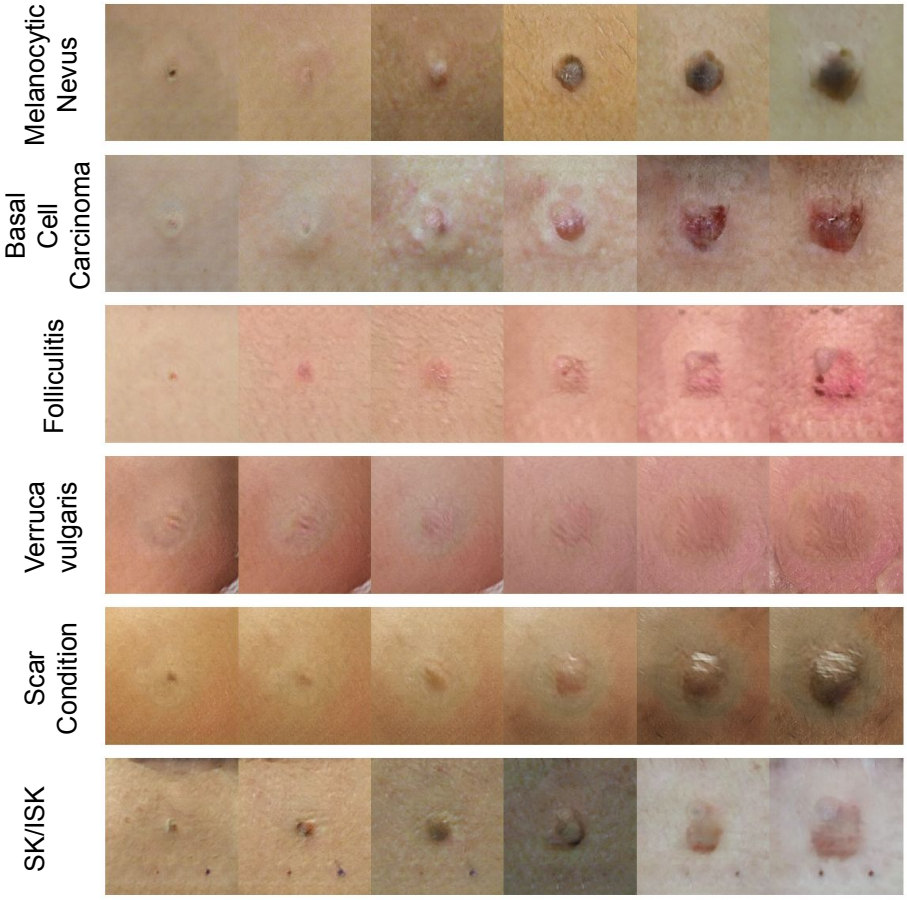

Figure 6: **Skin condition size variation** As DermGAN generates synthetic images given the condition's ROI, we are able to generate the same condition with various sizes. For specific conditions (e.g. *Basal Cell Carcinoma*), the change in size results in a change in the surrounding visuals, which is consistent with the real world occurrences.

directly measures the difference between the distribution of generated and real images in the activation space of the "Pool 3" layer of the Inception-V3 model. We perform an ablation study of the DermGAN model. Results of the FID scores on our test set (24000 images) are reflected in Table 1 (confidence intervals are for 50 trials).

Human Turing test For a subjective measure of how realistic the generated images are, we conducted two qualitative experiments. The first test was a Turing test with 10 participants. Each participant was asked to choose the skin images they found realistic in a collection of 80 real and 80 randomly selected synthetic images. On average the true positive rate (TPR) (the ratio of real images correctly selected) is 0.52 and the false positive rate (FPR) (the ratio of synthetic images detected as real) is 0.30. Results for each condition are demonstrated in Fig. 7(a), with average TPR ranging from 0.51 to 0.69 and average FPR

from 0.37 to 0.50. As expected, the TPR is higher than FPR for all conditions. However, the high FPR rate among all conditions indicate the high fidelity of synthetic images. The standard deviation of participants' performances is large and we hypothesize that it is due to the fact that they come from various backgrounds with different levels of experience with skin images.

The second experiment was designed to measure the medical relevance of the synthetic images. In this experiment, two board certified dermatologists answered a set of 16 questions. In each question, the participants were asked to choose the images relevant to a given skin condition among a combined set of real and randomly selected synthetic images. The average recall (ratio of related images correctly chosen) is 0.61 and 0.45 for the real and synthetic images respectively. Results for each condition are shown in Fig. 7(b), with recall ranging from 0.3 to 1.00 for real images and from 0.00 to 0.67 for synthetic images. For *Melanocytic nevus*, *Melanoma*, and *Seborrheic Keratosis / Irritated Seborrheic Keratosis (SK/ISK)*, synthetic images were identified to better represent the respective skin condition, indicating that our approach is able to preserve the clinical characteristics of those skin conditions.

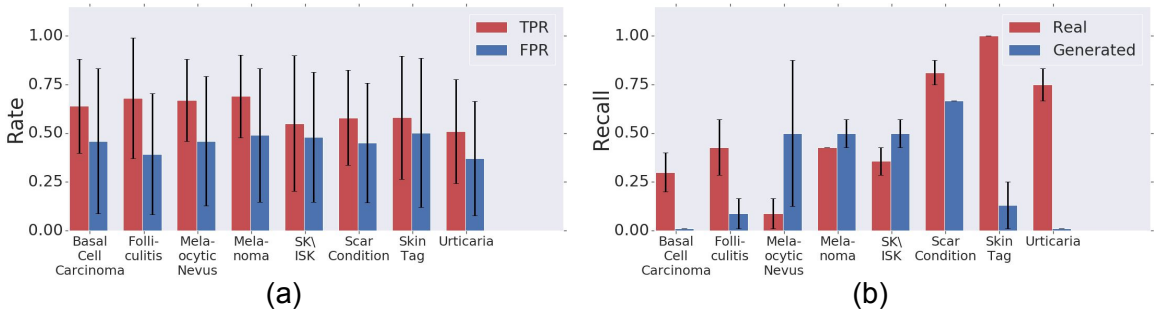

Figure 7: **Human Turing test** Results for discriminating between real and synthetic images are shown in (a), whereas results for whether images correctly describe the respective skin condition are shown in (b). Error bars represent standard deviation.

	Real Data	Derm GAN	No checkerboard effect mitigation	No condition-specific reconstruction loss	No feature matching loss
FID (± 1.96 STD)	83.6 ±2.5	122.4 ±3.4	151.6 ±3.4	174.0 ±4.7	140.7 ±2.5

Table 1: **Ablation study of GAN evaluation using FID scores.** Note that a lower FID score means a better performance.

Synthetic images as data augmentation for skin condition classification We first trained a MobileNet model (Howard et al., 2017) on our original (uncropped) data to differentiate between 27 skin condition classes (26 plus "other") from a single image. This baseline model achieves a top-1 accuracy of 0.496 on a test set of 5206 images, with poor performance on some of the rare conditions. To help alleviate this issue, we generate 20000 synthetic images using the 8-class DermGAN model and add them to the existing training data. We trained another MobileNet skin condition classifier using this enriched dataset and evaluated its performance on the same test set. While the top-1 accuracy remains relatively unchanged ($p = 0.56$ using paired T-test), performance improves for some of the malignant minority classes: *Melanoma* $F1$ score increases from 0.148 ([0.067, 0.193], 95% confidence interval using bootstrapping) to 0.282 ([0.110, 0.356]), whereas *Basal cell carcinoma* $F1$ score increases from 0.428 ([0.343, 0.439]) to 0.458 ([0.301, 0.534]), though at the cost of misclassifying *Melanocytic nevus* (0.113 decrease in $F1$). For the other 5 classes, the performances between the two models are comparable. (Fig. 8). Conventional data augmentation techniques (flipping, saturation, jitters) are used in both of the training setups.

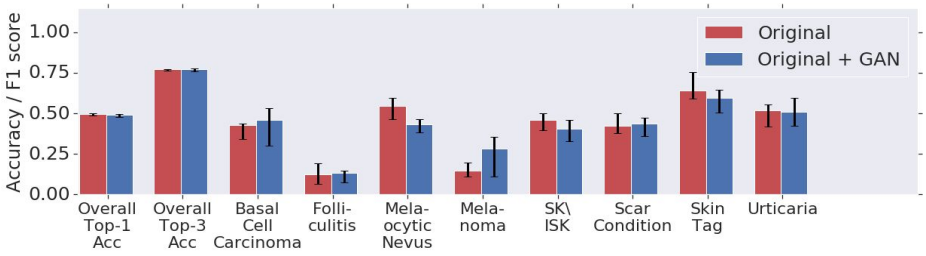

Figure 8: **Augmenting training data with synthetic images.** We added 20000 synthetic images to the original training data of size 49920. The overall performance is comparable to the baseline, but the performance on rare conditions like *Melanoma* and *Basal cell carcinoma* has noticeable improvement.

4. Discussion and Conclusion

In this work, we tackle the problem of generating clinical images with skin conditions as seen in a teledermatology setting. We frame the problem as an image to image translation task and propose DermGAN, an adaptation of the popular Pix2Pix GAN architecture. Using the proposed framework we are able to generate realistic images for pre-specified skin condition. We demonstrate that when varying the skin color or the size and location of the condition, the synthetic images can reflect such changes, while maintaining the characteristics of the respective skin condition. We further demonstrate that our generated images are of high fidelity using objective GAN evaluation metrics and qualitative tests. When using the synthetic images as data augmentation for training a skin condition classifier, the model is comparable to baseline, with improved performance on rare skin conditions. Further work is needed to improve the resolution and the diversity of the generated images, to effectively utilize such images for classification tasks by using techniques such as (Grover et al., 2019), and to explore the benefit of image synthesis in other clinical applications.

Acknowledgments

The authors would like to acknowledge Susan Huang and Kimberly Kanada for their clinical support. Thanks also go to Yun Liu, Greg Corrado, and Erica Brand for their feedback on the project. We also appreciate the help from the rest of the dermatology research team at Google Health for their engineering and operational support.

References

Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954, 2019.

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Christoph Baur, Shadi Albarqouni, and Nassir Navab. Melanogans: high resolution skin lesion synthesis with gans. *arXiv preprint arXiv:1804.04338*, 2018a.

Christoph Baur, Shadi Albarqouni, and Nassir 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*, pages 260–267. Springer, 2018b.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017.

Yucong Chi, Lei Bi, Jinman Kim, Dagan Feng, and Ashnil Kumar. Controlled synthesis of dermoscopic images via a new color labeled generative style transfer network to enhance melanoma segmentation. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2591–2594. IEEE, 2018.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint arXiv:1703.06490*, 2017.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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)*, pages 168–172. IEEE, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. Deep learning interpretation of echocardiograms. *bioRxiv*, page 681676, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *arXiv preprint arXiv:1906.09531*, 2019.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

Anant Gupta, Srivas Venkatesh, Sumit Chopra, and Christian Ledig. Generative image translation for data augmentation of bone lesion pathology. *arXiv preprint arXiv:1902.02248*, 2019.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5524–5532, 2018.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *arXiv preprint arXiv:1909.05382*, 2019.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Sourav Mishra, Hideaki Imaizumi, and Toshihiko Yamasaki. Interpreting fine-grained dermatological classification by deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

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

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1258–1266, 2018.

Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.