Improving Skin Lesion Analysis with Generative Adversarial Networks

Alceu Bissoto[1] Sandra Avila
RECOD Lab., Institute of Computing, University of Campinas (Unicamp)
Email: {alceubissoto, sandra}@ic.unicamp.br

***Abstract*—Melanoma is the most lethal type of skin cancer. Early diagnosis is crucial to increase the survival rate of those patients due to the possibility of metastasis. Automated skin lesion analysis can play an essential role by reaching people that do not have access to a specialist. However, since deep learning became the state-of-the-art for skin lesion analysis, data became a decisive factor in pushing the solutions further. The core objective of this M.Sc. dissertation is to tackle the problems that arise by having limited datasets. In the first part, we use generative adversarial networks to generate synthetic data to augment our classification model's training datasets to boost performance. Our method generates high-resolution clinically-meaningful skin lesion images, that when compound our classification model's training dataset, consistently improved the performance in different scenarios, for distinct datasets. We also investigate how our classification models perceived the synthetic samples and how they can aid the model's generalization. Finally, we investigate a problem that usually arises by having few, relatively small datasets that are thoroughly re-used in the literature: bias. For this, we designed experiments to study how our models' use data, verifying how it exploits correct (based on medical algorithms), and spurious (based on artifacts introduced during image acquisition) correlations. Disturbingly, even in the absence of any clinical information regarding the lesion being diagnosed, our classification models presented much better performance than chance (even competing with specialists benchmarks), highly suggesting inflated performances.**

I. Introduction

Melanoma is the most dangerous form of skin cancer. It causes the most deaths, representing about 1% of all skin cancers in the United States[2], and 3% in Brazil[3]. The crucial point for treating melanoma is early detection. The estimated 5-year survival rate of diagnosed patients rises from 15%, if detected in its latest stage, to over 97%, if detected in its earliest stages.

The medical diagnosing procedure for melanoma relies on pattern analysis. This enables machine learning to be a practical approach to this problem. It is essential to highlight that we do not see automated methods for diagnosis replacing specialists. It is quite the opposite: the technology would reach more risky patients, and they would need to consult with dermatologists. The outcome is the increase of the overall quality of the diagnosis, and life of specialists, enabling them to focus on positive cases.

Despite the possibilities of using this technology, we first need to achieve high confidence in our solutions output. Of course, false negatives are a huge problem, since it could potentially kill a patient by discouraging him from seeking proper treatment for such a time-dependent disease. Also, high amounts of false positives could be disastrous (especially in a scenario where the number of people reached is the highest), crowding hospitals with alarmed healthy patients seeking treatment (excision), wasting money, and specialists' time.

Deep Neural Networks are state-of-the-art for automated skin lesion analysis, and data is critical for improving those solutions [1]. For medical contexts, such as ours, the lack of annotated data is severe. Due to the high cost (both in money and time) of acquiring and labeling new samples, the datasets available are very limited.

This severe limitation in our data creates two main problems investigated in this M.Sc. dissertation: the inability to generalize, and dataset bias. For the first problem, we introduce a Generative Adversarial Networks (GANs)-based method for generating realistic synthetic data to improve generalization of skin lesion classification models [2]. GANs [3] aim to model the real image distribution by forcing the synthesized samples to be indistinguishable from real images. Our synthetic skin lesion generation process takes advantage of dermoscopic attributes. These attributes are local patterns in the lesion that are core to different medical algorithms [4], [5]. Their addition to the solution not only sharply increased the quality of the synthetic samples but also delivered meaningful information to the generation improving their clinical relevance.

For the second problem, we build upon dermoscopic attributes and medical algorithms [4], [6] to improve our understanding of our classification models. We verify if they are learning with clinically-meaningful information, or are exploiting artifacts in the skin lesion images. For this, we contrast two experiments: in the first, we build upon dermoscopic attributes, progressively adding information; in the second, we progressively destroy information according to the ABCD rule of dermoscopy [6], until a point where there is no clinically-meaningful information left. The results shocked ourselves and the community, showing that our classification models achieve high levels of performance without any lesion information.

We organized this paper as follows. In Sec. II, we review the GAN literature, dividing the advancements into topics to create a comprehensive view of the scenario. In Sec. III, we describe our data, including dermoscopy attributes. In Sec. IV,

[1]M.Sc. Dissertation
[2]http://www.cancer.net/cancer-types/melanoma/statistics
[3]https://www.inca.gov.br/tipos-de-cancer/cancer-de-pele-melanoma

we detail how we take advantage of these attributes to build a method for high-quality clinically-meaningful skin lesion image generation. We also show our methods and results to evaluate the synthetic images when used for data augmentation. In Sec. V, we investigate bias on skin lesion datasets, designing experiments to verify the network's performance when we expose it to correct or spurious correlations. In Sec. VI, we summarize our main achievements. Finally, in Sec. VII, we analyze our findings and propose future directions for the approached problems.

II. Literature Review

In this section, we review the literature of GANs. This story started in 2014 when Goodfellow et al. [3] introduced the GAN framework. This idea drew the attention of influential academics in machine learning such as Yan LeCunn (Turing Award 2018), which stated that "GANs is the most interesting idea in the last ten years in machine learning." Since 2014, the volume of works grew exponentially through the years, improving the GAN framework significantly through theoretical understanding, architecture enhancements, and applications.

In the M.Sc. dissertation, we divide the advancements in six fronts — Architectural, Conditional Techniques, Normalization and Constraint, Loss Functions, Image-to-image Translation, and Validation — providing a comprehensive notion of how the scenario evolved through the years, showing trends of thought that resulted in where we are today, where GANs are capable of generating face images that are almost indistinguishable from real photos.

Since we choose to give an evolutionary view of the GAN literature, sometimes the chronological information is inevitably lost in the process. To communicate the time dimension of the extensive GAN literature, we chronologically organize the GANs we comment during the review in Fig. 1 but also categorize them concerning their foremost contribution.

In this paper, we summarize the literature review made in the complete M.Sc. dissertation, focusing on the methods directly related to our experiments. For this, we overview the two main types of generation process: Plain Generation, and Image-to-Image Translation.

A. Plain Generation

The works that followed the original Goodfellow et al.'s paper compound the framework with architectural changes, enabling GANs to be explored in different contexts. At this time, GANs were capable only of generating low-resolution samples (32×32) from simpler datasets like MNIST [7] and Faces [8]. However, in 2016, crucial architectural changes were proposed by Deep Convolutional GAN (DCGAN) [9], boosting GANs research and increasing the complexity and quality of the synthetic samples. The proposals to remove pooling and fully-connected layers guided the future models' design, while the proposal of using batch normalization inspired other normalization techniques [10], [11] and still is used in modern GAN frameworks [12].

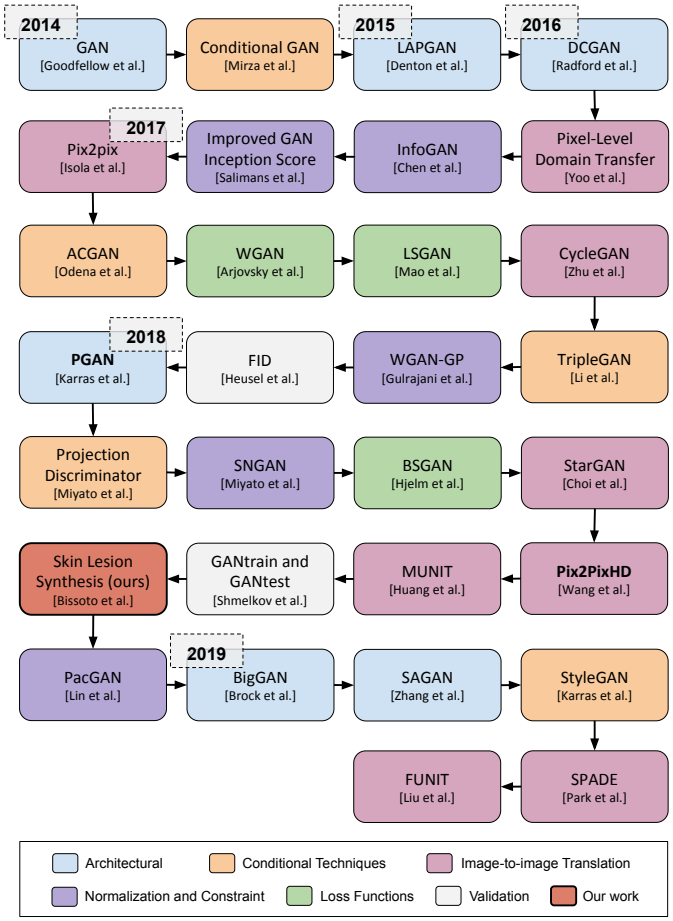


Fig. 1: Timeline of the GANs covered in the M.Sc. Dissertation' GAN literature review. We split it into six fronts, each represented with a color. Our work in skin lesion synthesis is in red.

In 2018, incremental architectures gained popularity, and it is still employed for improved stability and high-resolution generation. Progressive GAN (PGAN) [11] improved the incremental architecture to generate human faces of 1024×1024 resolution. While the spatial resolution of the generated samples increases, layers are progressively added for both generator and discriminator. Since older layers remain trainable, generation happens for different resolutions for the same image. It enables coarse/structural image details to be adjusted in lower resolution layers, and fine details in higher resolution ones.

B. Image-to-Image Translation

The addition of encoders in the GAN architecture enabled GANs for the task of image-to-image translation. In 2016, Yoo et al. [13] started using GANs for this task. The addition of the encoder to the generator's network transformed it into an encoder-decoder network (autoencoder). Now, the source image is first encoded into a latent code, which is then mapped to the target domain by the generator. The changes in the discriminator are not structural, but the task changed. In

addition to the traditional adversarial discriminator, the authors introduce a *domain discriminator* that analyzes pairs of source and target (real and fake) samples and judges if they are associated.

Until this time, the synthetic samples follow the same quality of plain generation: low quality and low resolution. This scenario changes with pix2pix [14]. Pix2pix employed a new architecture for both the generator and the discriminator, as well as a new loss function. It was a complete revolution! The generator is a U-Net-like network [15], where the skip connections allow to bypass information that is shared by the source-target pair. The authors also introduce a patch-based discriminator (which they called PatchGAN) to penalize structure at the scale of patches of a smaller size (usually 70×70), while accelerating evaluation. To compose the new loss function, the authors proposed a term that evaluates the L1 distance between synthetic and ground truth targets, constraining the synthetic samples without killing variability.

The next step towards high-resolution image-to-image translation is pix2pixHD (High-Definition) [16], which was widely used during this M.Sc. dissertation to generate realistic clinically-meaningful skin lesions. Pix2pixHD obviously is based upon pix2pix's work but includes several modifications while adopting changes brought by CycleGAN with respect to the generator's architecture.

The authors propose using two nested generators to enable the generation of 2048×1024 resolution images, where the outer "local" generator enhances the generation of the inner "global" generator. Just like CycleGAN, it uses [17]'s style transfer network as a global generator, and as a base for the local generator. The output of the global generator feeds the local generator in the encoding process (element-wise sum of global's features and local's encoding) to carry information on the lower resolution generation. They are also trained separately: first, they train the global generator, then the local, and finally, they fine-tune the whole framework together.

In pix2pixHD, the discriminator also receives upgrades. Instead of working with lower-resolution patches, pix2pixHD uses three discriminators that work simultaneously in different resolutions of the same images. This way, the lower-resolution discriminator will be more concerned about the general structure and coarse details, while the high-resolution discriminators will pay attention to fine details. So far, every image-to-image translation GAN generator has assumed the form of an autoencoder, where the source image is encoded into a reduced latent representation, that is finally expanded to its full resolution. The encoder plays an essential role in extracting information of the source image that will be kept in the output. Often, even multiple encoders are employed to extract different information, such as content and style.

III. Data

Due to the scarcity of good-quality, annotated skin lesion images, two datasets dominate research on automated skin lesion analysis: the Interactive Atlas of Dermoscopy [18] and the ISIC Archive [19]. The Atlas is an educational medical resource, with many standardized metadata over the cases it contains, while the ISIC Archive is a much larger, but also less controlled dataset, with images of different sources. Nowadays, nearly every reproducible work in the field of skin lesion analysis refers to these datasets for training, evaluating, or comparing its models [1], [2], [20], [21].

Two types of skin lesion images compose these datasets: dermoscopic and clinical. Clinical images — which are only present in the Atlas dataset — can be captured with standard cameras, while dermoscopic images can only be captured with a device called dermatoscope. This device normalizes the light influence on the lesion, allowing it to capture deeper details. Dermoscopic images enable the application of medical algorithms that support the specialists' decision. There are algorithms based on the lesion's characteristics, such as the ABCD rule (lesions' Asymmetry, Border, Color, and Diameter); and others such as the 7-points, that are built upon *dermoscopic attributes*.

Throughout this work, we rely on dermoscopic attributes to generate better skin lesion images and to measure bias in our skin lesion datasets. These attributes are present in the lesions in the form of visual patterns such as networks, globules, and streaks. There is a wide variety of dermoscopic attributes, and each of them can stratify with respect to their regularity, color, and other specific details. These malignancy markers are only visible in dermoscopic images and are crucial for specialists when diagnosing melanoma.

The annotation regarding dermoscopic attributes, despite crucial for human specialists, is present only for small subsets of data available for machine learning. Only recently, at ISIC 2017 and 2018 Challenges, the organizers made available a subset of dermoscopic images with this special annotation to support the task for dermoscopic features segmentation. The provided annotation is a map for each of the five interest attributes: pigment network, negative network, streaks, globules, and milia-like cysts. This way, we know if an attribute is present and the portion of the lesion that displays it. On Atlas, which is the only other source of skin lesion images that contains this annotation, the annotation is binary (present or not), not showing the lesion regions that display the patterns.

IV. Skin Lesion Image Synthesis

We proposed a GAN-based method for generating high-definition, visually-appealing, and clinically-meaningful synthetic skin lesion images. This work was the first that successfully generates realistic skin lesion images (Fig. 2). To evaluate synthetic images' relevance, we trained a skin cancer classification network with synthetic and real images, reaching an improvement of 1.3 percentage points. Our full implementation is available at https://github.com/alceubissoto/gan-skin-lesion.

We aim to generate high-resolution synthetic images of skin lesions with fine-grained detail to provide correct correlations to the classification networks, aiding at its generalization. For that, we directly feed the GAN's generator with maps that encode lesions' dermoscopic attributes and border's specificities. This way, instead of generating the image from pure noise

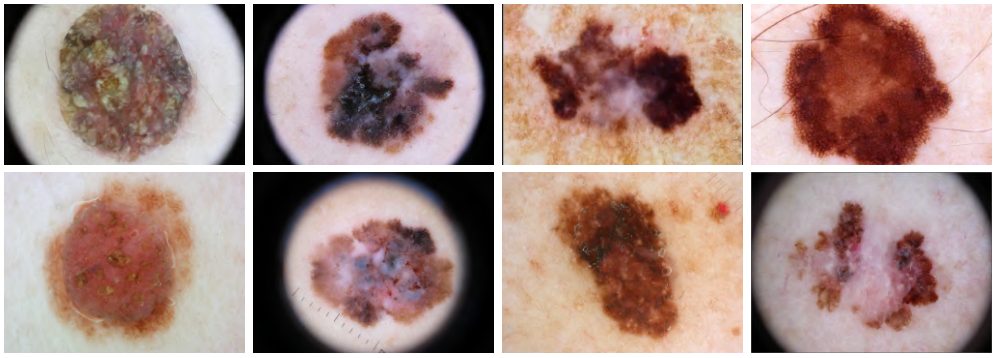
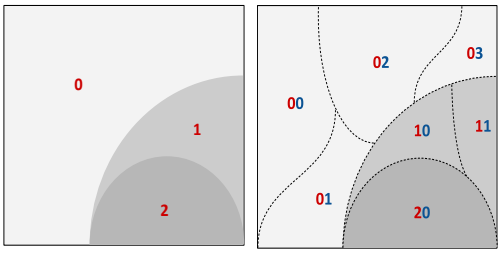


Fig. 2: Comparison between our synthetic samples (top row) and real samples from the ISIC Archive (bottom row).



(a) Semantic Map (b) Instance Map

Fig. 3: Simplification of our semantic and instance maps. While the semantic map's pixels' values are only ruled by the class, instance maps' take in consideration class and the individual instance defined by superpixels.

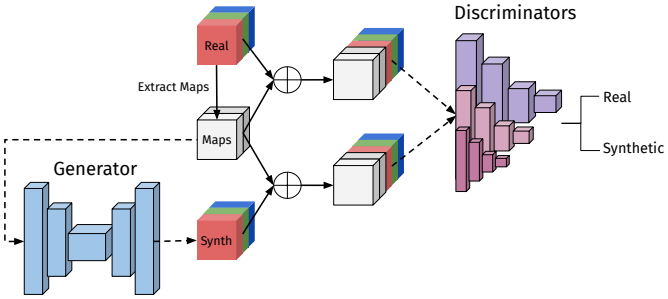


Fig. 4: Our Pipeline. We feed the generator with maps extracted from real images, resulting in the synthetic images. The discriminator is fed with batches combining real images and its maps, or synthetic images and the maps used to generate them. The output of the discriminators (there are three, each operating in a different resolution) is finally backpropagated to train the whole pipeline.

(usual procedure with GANs), we synthesize from a semantic label map and an instance map (Fig. 3). Because of this different objective, our problem of image synthesis specified to image-to-image translation. Fig. 4 summarizes our pipeline.

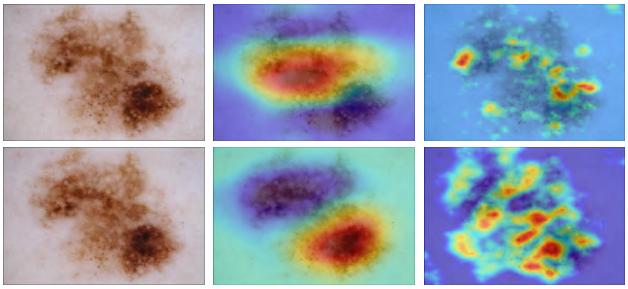
The semantic and instance maps are crucial for this process of generation. They encapsulate information about the skin lesions that simplify the generation process, enabling the synthesis of higher quality images. Both maps are based on segmentation masks and dermoscopic attributes annotations publicly provided by specialists as part of Task II from the ISIC 2018 Challenge. The annotation comprises five types of dermoscopic patterns (pigment network, negative network, streaks, milia-like cysts, globules), providing a binary mask for each.

Training Data	AUC (%)	Set Size	p-value
Real	83.4 ± 0.9	2,346	2.5×10^{-3}
Instance	82.0 ± 0.7	2,346	2.8×10^{-5}
Semantic	78.1 ± 1.2	2,346	6.9×10^{-8}
PGAN	73.3 ± 1.5	2,346	2.3×10^{-9}
Real+Instance	82.8 ± 0.8	4,692	1.1×10^{-4}
Real+Semantic	82.6 ± 0.8	4,692	1.2×10^{-4}
Real+PGAN	83.7 ± 0.8	4,692	2.6×10^{-2}
Real+2×PGAN	83.6 ± 1.0	7,038	2.0×10^{-2}
Real+Instance+PGAN	$\mathbf{84.7 \pm 0.5}$	7,038	–

TABLE I: Performance comparison of real and synthetic training sets for a skin cancer classification network. We train the network 10× with each set. The features present in the synthetic images are not only visually appealing but also contain meaningful information to classify skin lesions correctly.

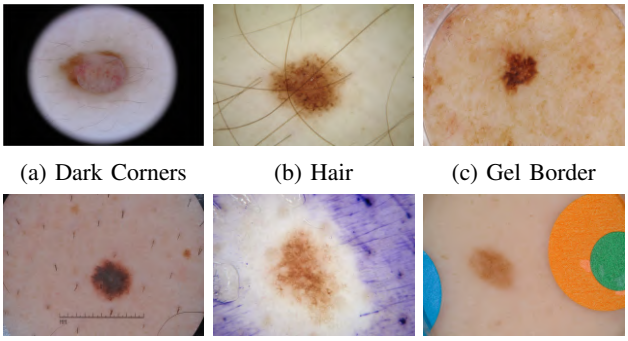
To evaluate the complete set of synthetic images, we train a skin classification network with different combinations of synthetic and real data (including only real, only synthetic, and combinations of both) to compose our training dataset. We compare the achieved area under the ROC curve (AUC), testing with augmented replicas of real only images. The results for training the classification network with synthetic images confirm they contain features that characterize a lesion as malignant or benign, and their addition to the training set of a classification network yields an improvement in the AUC performance by an average of 1.3 percentage points while keeping the network more stable (see Table I).

Finally, we need to investigate how our classification models receive synthetic information and make sure they are providing correct correlations to improve generalization. Although synthetic images are almost identical (at least to our eyes) to real ones, the network perceives differences between them,



(a) Real/Synthetic (b) GradCAM (c) Occlusion

Fig. 5: Saliency maps from (b) GradCAM [22] and (c) Occlusion [23] for real (first rows) and synthetic (second rows) images using the same model, trained only with real images. The saliency maps highlight (in hot colors) the portions of the image that contributed the most to the prediction. That is, when highlighted areas are perturbed or altered, the classification network's prediction was highly affected.



(a) Dark Corners (b) Hair (c) Gel Border

(d) Ruler (e) Ink markings and Gel bubbles (f) Patches

Fig. 6: Possible artifacts that may provide spurious correlations to our classification models.

causing the saliency maps to output differently between real and synthetic (Fig. 5). This raises awareness to researchers (including ourselves) when using synthetic lesions to augment the model's training datasets. We need to make sure the synthetic images included are contributing positively to the result, while not reinforcing any possible spurious correlation already present in the data. Nevertheless, our results, when augmenting our training datasets with synthetic images, show that this technique can significantly aid classification for small datasets.

We highlight that this M.Sc. dissertation's contribution was published at the ISIC Skin Image Analysis Workshop at MICCAI 2018 [2] and still is, to this day, state of the art for skin lesion images synthesis.

V. Bias in Skin Lesion Datasets

Dataset biases may inflate the performance of models (presenting them features that are not truthful to real-world data), or play down their performance (by destroying correlations that occur in real-world data, thus preventing models from exploiting them).

If we think of general datasets, there can be bias over the scenes (rural or urban), acquisition methods (professional or amateur), amount of objects in the scene, angles of views, among other factors [24]. If bias is present even in more significant and more diverse datasets [24] like ImageNet [25], it is naive to think it is not present in the smaller and harder to obtain skin cancer datasets, where we lack works identifying the possible sources of dataset bias. We know, however, that there are visible artifacts introduced during the image acquisition process (e.g., dark corners, marker ink, gel bubbles, color charts, ruler marks, skin hair) [26] that could inflate models performances due to spurious correlations (Fig. 6).

Despite being impossible to eliminate wholly, it is crucial to understand bias and its sources to improve our image acquisition processes and deep learning models further. A useful way to measure the first possible effect of a dataset bias (undue inflation of performances due to spurious correlations in the dataset), is a counterfactual experiment, that destroys the cogent information in the data, and measures how much the performance of models drops. Therefore, our first set of experiments follows that procedure, gradually removing information from skin lesion images, and measuring the network performance. We perform single- (training and testing on the same dataset) and cross-dataset (training on ISIC and testing on Atlas) experiments, and find that in both cases, the networks are able to maintain a surprising amount of accuracy, even after almost all cogent information has been destroyed (Fig. 7).

Measuring the second possible effect (inability to provide useful correlations for learning) is much harder, since we cannot, *a priori* prove those correlations exist in the real world, neither that the machine-learning model would learn from them if they were correctly represented in the dataset. The best we can do is provide additional evidence for the models that we expect would be useful for a human, and measure if that makes any difference.

Thus, in our second experiment set, we add progressively more features, based upon fine-grained dermoscopic attributes (pigment network, negative network, streaks, milia-like cysts, and globules) spatially located on the lesions. To provide those features, we employ the annotations available for Task II from ISIC 2018 Challenge. We expected that such clinically-meaningful skin lesion information would improve the network learning process. However, the performance fails to improve in all scenarios we tested, even when we feed the network with all the image's pixels with an additional channel containing extra clinically-meaningful information (Fig. 8).

When we hide meaningful lesion information from our models, should it still be able to learn patterns that differentiate benign from malignant lesions? We believe that when a model learns to classify malignant lesions by analyzing only the skin — without information on the borders, biological markers, or lesions' diameter — it strongly relies on patterns introduced during image acquisition and general dataset bias. That problem is critical for deploying automated skin lesion analysis. When performing in the real world, we want the network to

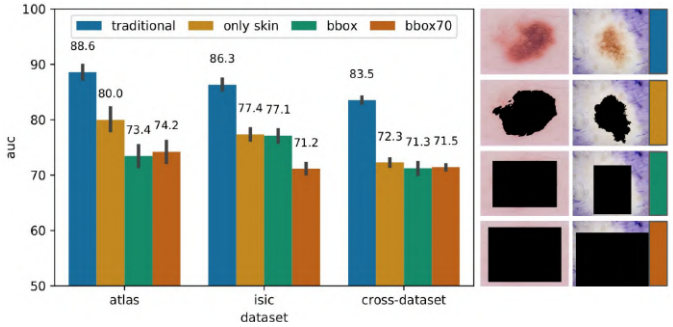


Fig. 7: Models' performance over the disturbed datasets. We first remove all the pixel colors inside the lesion (*only skin*), proceeding to remove border information (*bbox*), and finally, removing the size (diameter) of the lesion (*bbox70*). We show samples from each dataset in the right. Surprisingly, even when we destruct all clinical-meaningful information, the network finds a way to learn to classify skin lesion images much better than chance.

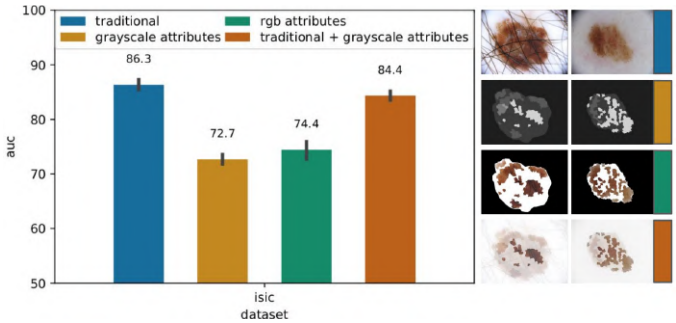


Fig. 8: Performance comparison of the different sets of images with the ISIC dataset. We show samples from each set in the right. Surprisingly, when we try to simplify the learning process, feeding the network with dermoscopic attributes, the result does not improve.

be as unbiased as possible to make decisions based on clinical features. Therefore, it is urgent to understand the current bias in the datasets used to train and evaluate our works.

We highlight that this M.Sc. dissertation' contribution was published at the ISIC Skin Image Analysis Workshop at CVPR 2019 [27], and our paper received the Best Paper Award. All our source code is readily available on https://github.com/alceubissoto/deconstructing-bias-skin-lesion.

VI. Research Accomplishments

We summarize our main achievements as follows:

- Two conference papers [2], [27] and a technical report [28].
- Second best poster award at the 2018 International Educational Symposium of The Melanoma World Society [29].
- Sixth place award in the ISIC 2018 Challenge [30].
- Best paper award at the ISIC Skin Image Analysis Workshop, at the Conference on Computer Vision and Pattern Recognition (CVPR) [27].
- Winner of the Google Latin America Research Awards for two years (Google LARA 2018 and 2019).
- Dissemination of our research findings on traditional media: Jornal Estadão https://tinyurl.com/v8fzqnl, Jornal Correio Popular https://tinyurl.com/sepv732, Jornal da Unicamp https://tinyurl.com/s76ggdo, TV Cultura https://tinyurl.com/vc7qbad, EPTV 2ª Edição https://tinyurl.com/tov4vj4, TVB Record TV Campinas https://tinyurl.com/rpabxod, TV BandMais https://tinyurl.com/u7cwr2w, Rádio CBN, Show da Notícia https://tinyurl.com/tve42ff.

VII. Conclusions and Future Work

GAN literature review: We provided a comprehensive GAN state-of-the-art, splitting it into six topics (Architecture, Conditional Techniques, Normalization, and Constraint Techniques, Loss Functions, Image-to-Image Translation Methods, and Validation) showing how works influenced each other until we arrive at the stage we are today [28].

Skin lesion image synthesis: We proposed a method for skin lesion synthesis that generates high-definition clinically-meaningful synthetic skin lesions. Our method is, to this day, the state-of-the-art for skin lesion synthesis, and we have many paths to increasing the quality and variability of our synthetic samples [2], [29].

Bias on skin lesion datasets: We investigated the data used for both synthesis and classification models. Even when no clinically-meaningful information is presented to it (according to medical algorithms), the performance of the model is shockingly high, surpassing benchmarks that quantify a specialist's performance [27]. This is not a good sign for AI research. Our work in this matter raised awareness in the community, and we hope it can be approached soon.

Cancer is already a public health challenge. It is predicted that 30 million people will die of cancer every year until 2030. The vast majority of deaths will occur in lower-middle-income countries, such as Brazil. Automated solutions, like the one explored in this research, can help change this horizon by reaching people and selecting the ones at risk and enabling specialists to focus on them.

We believe we can speed up this process by extracting the most of each sample while combining different domains (text from medical records, clinical and histopathologic images, and genomics). Gathering diversified quality data and learning how to combine and make sense of this data can be the next revolution for skin cancer analysis.

Acknowledgment

A. Bissoto was partially funded by CAPES (88887.388163/2019-00). S. Avila is partially funded by FAPESP (2017/16246-0, 2013/08293-7). A. Bissoto and S. Avila are also partially funded by Google LARA 2018 and 2019. RECOD Lab. is supported by projects from FAPESP, CNPq, and CAPES. The funding sources had no involvement in the data acquisition, study design, result analysis, or in the manuscript writing. Finally, we acknowledge the donation of GPUs by NVIDIA.

REFERENCES

[1] E. Valle, M. Fornaciali, A. Menegola, J. Tavares, F. V. Bittencourt, L. T. Li, and S. Avila, "Data, depth, and design: Learning reliable models for skin lesion analysis," *Neurocomputing*, vol. 383, pp. 303–313, 2020.

[2] A. Bissoto, F. Perez, E. Valle, and S. Avila, "Skin lesion synthesis with generative adversarial networks," in *ISIC Skin Image Analysis Workshop, International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[4] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives of Dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.

[5] S. W. Menzies, I. C., C. K. A., and M. W. H., "Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features," *Archives of Dermatology*, pp. 1178–1182, 1996.

[6] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt *et al.*, "The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions," *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551–559, 1994.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] J. Susskind, A. Anderson, and G. Hinton, "The toronto face dataset," University of Toronto, Tech. Rep., 2010.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations*, 2016.

[10] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[12] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *International Conference on Learning Representations*, 2018.

[13] D. Yoo, N. Kim, S. Park, A. Paek, and I. Kweon, "Pixel-level domain transfer," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 517–532.

[14] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[16] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8798–8807.

[17] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 694–711.

[18] G. Argenziano, H. P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino *et al.*, "Dermoscopy: a tutorial," *EDRA, Medical Publishing & New Media*, p. 16, 2002.

[19] "International Skin Imaging Collaboration: Melanoma Project," https://isic-archive.com.

[20] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer, "A state-of-the-art survey on lesion border detection in dermoscopy images," *Dermoscopy Image Analysis*, pp. 97–129, 2015.

[21] T. J. Brinker, A. Hekler, A. Hauschild, C. Berking, B. Schilling, A. H. Enk, S. Haferkamp, A. Karoglan, C. von Kalle, M. Weichenthal *et al.*, "Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark," *European Journal of Cancer*, vol. 111, pp. 30–37, 2019.

[22] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[23] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[24] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521–1528.

[25] O. Russakovsky, J. Deng, H. S., J. Krause, S. Satheesh, S. Ma *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[26] A. Bissoto, E. Valle, and S. Avila, "Debiasing skin lesion datasets and models? not so fast," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[27] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "(De)Constructing bias on skin lesion datasets," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[28] A. Bissoto, E. Valle, and S. Avila, "The six fronts of the generative adversarial networks," *arXiv preprint arXiv:1910.13076*, 2019.

[29] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "Generating high quality synthetic skin lesions for boosting automated screening," *International Educational Symposium of the Melanoma World Society*, pp. 43–43, 2018.

[30] A. Bissoto, F. Perez, V. Ribeiro, M. Fornaciali, S. Avila, and E. Valle, "Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD Titans at ISIC Challenge 2018," *arXiv preprint arXiv:1808.08480*, 2018.