Universidade Estadual de Campinas
Instituto de Computação

Alceu Emanuel Bissoto

Improving Skin Lesion Analysis with Generative Adversarial Networks

Análise de Lesões de Pele usando Redes Generativas Adversariais

CAMPINAS
2019

Alceu Emanuel Bissoto

Improving Skin Lesion Analysis with Generative Adversarial Networks

Análise de Lesões de Pele usando Redes Generativas Adversariais

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientadora: Profa. Dra. Sandra Eliza Fontes de Avila

Este exemplar corresponde à versão final da Dissertação defendida por Alceu Emanuel Bissoto e orientada pela Profa. Dra. Sandra Eliza Fontes de Avila.

CAMPINAS
2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

B545i

Bissoto, Alceu Emanuel, 1994-
Improving skin lesion analysis with generative adversarial networks / Alceu Emanuel Bissoto. – Campinas, SP : [s.n.], 2019.

Orientador: Sandra Eliza Fontes de Avila.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Melanoma. 2. Aprendizado de máquina. 3. Aprendizado profundo. 4. Redes neurais convolucionais. 5. Diagnóstico por imagem. I. Avila, Sandra Eliza Fontes de, 1982-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Análise de lesões de pele usando redes generativas adversariais
Palavras-chave em inglês:
Melanoma
Machine learning
Deep learning
Convolutional neural networks
Diagnostic imaging
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Sandra Eliza Fontes de Avila [Orientador]
Roberto de Alencar Lotufo
Jefersson Alex dos Santos
Data de defesa: 10-09-2019
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-2293-6160
- Currículo Lattes do autor: http://lattes.cnpq.br/6982744466245113

Universidade Estadual de Campinas
Instituto de Computação

Alceu Emanuel Bissoto

Improving Skin Lesion Analysis with Generative Adversarial Networks

Análise de Lesões de Pele usando Redes Generativas Adversariais

Banca Examinadora:

- Profa. Dra. Sandra Eliza Fontes de Avila
  UNICAMP
- Prof. Dr. Roberto de Alencar Lotufo
  UNICAMP
- Prof. Dr. Jefersson Alex dos Santos
  UFMG

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 10 de setembro de 2019

Acknowledgements

First, I want to wholeheartedly thank Prof. Sandra Avila, who is directly responsible for the person that I've transformed to during this M.Sc. degree. During the whole process, she not only helped me to learn the technologies described in this thesis but also provided the opportunities for me to grow as a person. She is undoubtedly the best teacher I have ever had and one of the most hardworking, fair, courageous, and inspiring people I know. I feel very fortunate to have had her as an advisor during my M.Sc., and currently as my Ph.D. advisor.

I also want to thank the Learning Titans research group: Prof. Eduardo Valle, Michel Fornaciali, Eduardo Seiti, Fábio Perez, Vinícius Ribeiro, and all the other members. All highly contributed to this work, and made the process a lot easier with their valuable friendship. I want to especially thank Prof. Eduardo Valle, who, although was not a direct advisor, contributed significantly with brilliant ideas and taught me how to become a better researcher.

I thank UNICAMP for the unique opportunities and memories of the last years. In special, I am very thankful to the RECOD Lab (Reasoning for Complex Data, our lab at the Institute of Computing (IC) at Unicamp), particularly its members, whose rapidly became friends. The daily companionship at the lab, laughs, and fruitful discussions significantly contributed to making the countless hours spent in this research very joyful. The RECOD Lab also provided the much-needed infrastructure to run my experiments, enabling this research.

I want to thank my parents, José Alceu Bissoto and Vera Lucia Bissoto, for the unconditional support during my studies and for always encouraging me to continue doing what I love. I also thank all my friends, for all the help and the good times we spent together.

I gratefully thank CNPq for the Master's scholarship (Project 134271/2017-3), and Google for the Latin America Research Award (LARA 2018). This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Finally, I thank FAPESP, Microsoft and NVIDIA for the all infrastructures investments, which enabled this research.

Resumo

Melanoma é a forma mais letal de câncer de pele. Devido a possibilidade de metástase, o diagnóstico precoce é crucial para aumentar a taxa de sobrevivência dos pacientes. A análise automatizada de lesões de pele pode ter um papel importante ao alcançar pessoas sem acesso a especialistas. Porém, desde que técnicas de aprendizado profundo se tornaram o estado-da-arte para análise de lesões de pele, os dados se tornaram um fator decisivo para avançar as soluções. O objetivo principal dessa tese de mestrado é tratar dos problemas que surgem por lidarmos com poucos dados nesse contexto médico. Na primeira parte, usamos redes generativas adversariais para gerar dados sintéticos para aumentar o conjunto de treino dos nossos modelos de classificação para elevar a performance. Nosso método é capaz de gerar imagens de lesão de pele em alta-resolução com significado clínico, que quando usadas para compor o conjunto de treino de redes de classificação, consistentemente melhoram a performance em diferentes cenários, para diferentes dados. Também investigamos como nossos modelos de classificação interpretam as amostras sintéticas, e como elas são capazes de ajudar na generalização do modelo. Finalmente, analisamos um problema que surge por termos poucos, relativamente pequenos conjuntos de dados que são reusados repetidamente na literatura: bias. Para isso, planejamos experimentos para estudar como nossos modelos usam os dados, verificando como ele explora correlações corretas (com base em algoritmos médicos), e espúrias (com base em artefatos introduzidos durante a aquisição das imagens). Surpreendentemente, mesmo sem contar com nenhuma informação clínica sobre a lesão sendo diagnosticada, nossos modelos de classificação apresentaram performance muito melhor que o acaso (competindo até mesmo com *benchmarks* de especialistas), sugerindo performances altamente infladas.

Abstract

Melanoma is the most lethal type of skin cancer. Due to the possibility of metastasis, early diagnosis is crucial to increase the survival rate of those patients. Automated skin lesion analysis can play an important 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 to push the solutions further. The core objective of this Master thesis is to tackle the problems that arise by having limited datasets. In the first part, we use generative adversarial networks (GANs) to generate synthetic data to augment our classification model's training datasets to boost performance. Our method is able to generate 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 are able to 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 of data, verifying how it exploits correct (based on medical algorithms), and spurious (based on artifacts introduced during image acquisition) correlations. Disturbingly, even in 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.

List of Figures

List of Tables

Contents

Chapter 1

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[1], and 3% in Brazil[2]. 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 [4].

The standard method to evaluate a skin growth to rule out melanoma is by biopsy, followed by a histopathological examination. The challenge lies in identifying the lesions that have the highest probability of being melanoma. Such lesions should be biopsied, and their histopathology appropriately evaluated at the earliest possible time in their development [81].

Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions (see Figure 1.1).

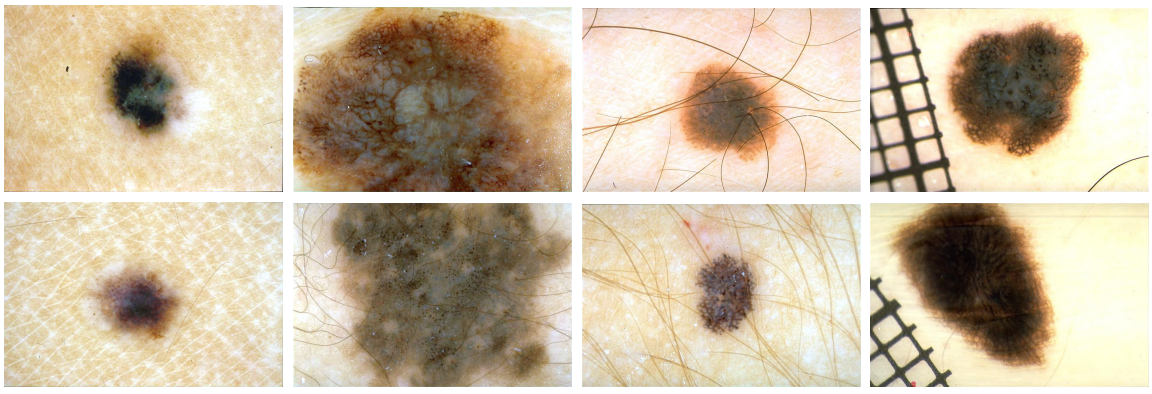


Figure 1.1: Extracts of skin lesions from the Interactive Atlas of Dermoscopy dataset [6]. Melanomas (top row) are difficult to differentiate from benign lesions (bottom row) for both human and machine. Reproduced from Fornaciali et al. [31].

An automated solution for classification of skin lesions could be beneficial in multiple scenarios. If we think of a device to identify patients in risk, it could reach people that do not have access to specialists (by geographic, or financial reasons). Since time is a critical factor for treating skin cancer, the highest the reach of the technology, the higher

[1]http://www.cancer.net/cancer-types/melanoma/statistics
[2]https://www.inca.gov.br/publicacoes/livros/estimativa-2018-incidencia-de-cancer-no-brasil

are the benefits for the people. Also, it could be operated by non-specialists with little training to properly collect images of the lesions, broadening its benefits even more. A device with the same capabilities used in the hospital can alleviate specialists' stress and fatigue, redirecting to them only risky patients.

Finally, an explainable, transparent, and accountable device can aid specialists by exposing a different point of view for difficult cases, highlighting portions of the lesion that it identifies as a sign of malignancy.

In any of those scenarios, we see the technology replacing specialists. It is quite the opposite: more risky patients would consult with dermatologists, and the overall quality of the diagnosis and life of specialists could improve, enabling them to focus on positive cases.

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

Since the adoption of Deep Neural Networks (DNNs), the state of the art improved rapidly for skin cancer classification [15,28,32,34,95]. The ISIC Challenge on Skin Lesion Analysis is responsible for this progress [24, 25, 62]. It is an annual event (organized by the International Skin Imaging Collaboration (ISIC)) that started in 2016, where different teams compete to achieve the best performance under the controlled supervision of the organizers. For every edition, the organization of the challenge makes more data publicly available, and by designing the tasks of the challenge, help to guide the directions of skin lesion research. Since its creation, challengers' works helped to boost skin lesion analysis, establishing state-of-the-art solutions.

Some techniques are already consensual among researchers in this area. Due to the smaller datasets when compared to general-purpose datasets, all winning solutions employ techniques to mitigate this limitation. Transfer learning [64] and data augmentation [78] are the most common and successful ones: Transfer learning attempts to take advantage of the generalization achieved by a DNN in a more general and vast dataset like ImageNet [83] to smaller and specific datasets like skin lesions'; and data augmentation increases the dataset size by inflating it with transformations (*e.g.*, rotations, flips, scales, color alterations) of the original training set.

When designing those solutions, the architecture choice used to play an important role. However, the gap between the computer vision state-of-the-art and skin lesion solutions decreased significantly since 2016 when the challenge started, and today, it is possible to affirm that it is not as important. As long as the chosen architecture is deep enough and is state-of-the-art of a main computer vision classification task, it is suitable for skin lesion classification [77].

The core objective of this Master Thesis is to boost both the performance and our understanding of skin lesion classification models. The first problem we decided to approach is the lack of annotated data [95], which is expensive and require much effort from specialists. To bypass this problem, we propose to use Generative Adversarial Networks

(GANs) [35] for generating realistic synthetic skin lesion images.

GANs aim to model the real image distribution by forcing the synthesized samples to be indistinguishable from real images. In Chapter 2, we review the GAN literature, describing the evolution of this growing family of techniques. Skin lesion images that compose datasets for classification are high-definition and ideally, offer high variability to guide the network's learning. Also, they present multiple fine-grained patterns called dermoscopic attributes, that are crucial for dermatologists to diagnose melanoma. Ideally, synthetic samples must display the same level of detail and variability. However, skin lesion images are very distinct in comparison to mainstream general-purpose datasets and are very scarce, making the generation task extra challenging.

Built upon these generative models, many methods were proposed to generate synthetic images based on GANs [46, 80, 85]. In 2018, we generated synthetic skin lesion images using different GANs architectures and inflated the training dataset of a skin lesion classification network [15]. 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 [5, 66]. 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 (see Figure 1.2).

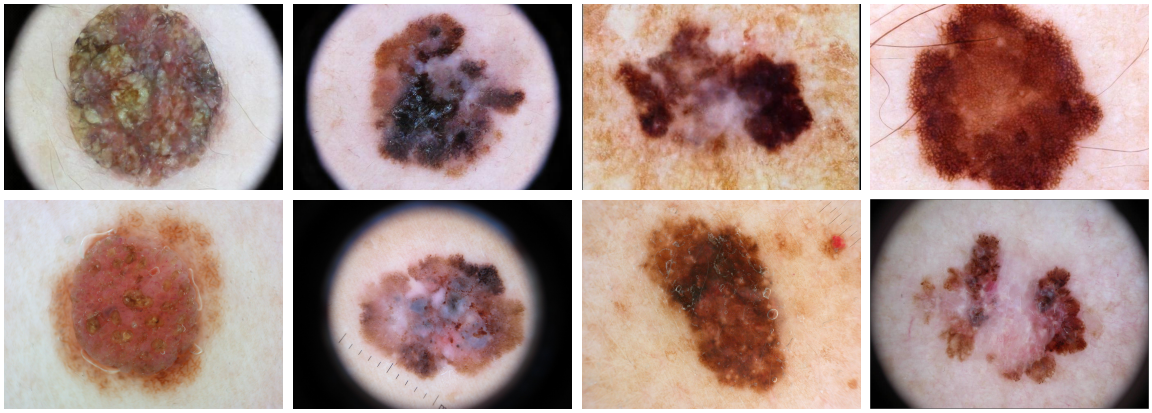


Figure 1.2: Comparison between our synthetic samples (top row) and real samples from the ISIC Archive (bottom row).

We achieved the improvement of 1 p.p. over the baseline Area under the ROC curve (AUC) (network trained only with real images) after augmenting our classification networks' training set. The included synthetic images were generated by different methods, enabling us to combine variety (from PGAN [50]) with fine-grained details (from Pix2pixHD [96]). We detail our method and evaluate our synthetic images in Chapter 3.

To validate the synthetic samples and their influence when inflating the training set of a classification network, we employed visualization methods, and dermoscopic attributes analysis. Only by using this diversified evaluation, we can identify and benchmark the synthetic images' qualities and flaws.

Finally, we build upon dermoscopic attributes and medical algorithms [5, 71] 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 [71], 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 detail this procedure and analyze the results in Chapter 4.

1.1 Contributions

We summarize our main contributions as follows:

- We provide a comprehensive up-to-date GAN state-of-the-art.
- Our work "Skin Lesion Synthesis with Generative Adversarial Networks" [15] is the state-of-the-art for skin lesion synthesis. In this work, we investigated the effects of augmenting our classification network's training data with synthetic images.
- We identified an urgent problem of bias in skin lesion datasets, raising awareness to this fact in the research community [13].
- All our work is publicly available on GitHub, enabling other people to reproduce our results: `https://github.com/alceubissoto/gan-skin-lesion` and `https://github.com/alceubissoto/deconstructing-bias-skin-lesion`.

1.2 Achievements

We summarize our main achievements as follows:

- First author of *Skin Lesion Synthesis with Generative Adversarial Networks* [15], published at the ISIC Skin Image Analysis Workshop, at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018).
- First author of *(De)Constructing Bias on Skin Lesion Datasets* [13], published at the ISIC Skin Image Analysis Workshop, at the Conference on Computer Vision and Pattern Recognition (CVPR 2019). This work received the Best Paper Award.
- Winner of the Google Latin America Research Awards (Google LARA 2018) with the project *Improving Skin Cancer Classification with Generative Adversarial Networks.*
- 2$^{\text{nd}}$ Best Poster Award at the International Educational Symposium of The Melanoma World Society, with the poster *Generating High Quality Synthetic Skin Lesions for Boosting Automated Screening.*
- Main contributor of RECOD Titans participation [14] in the Task 2 – Lesion Attribute Detection of the ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection challenge [25], ranking the 6$^{\text{th}}$ best submission.

1.3 Outline

We organized this Master thesis as follows:

- **Chapter 2 – GAN Literature Review:** We review the extensive GAN literature, dividing the advancements into topics to create a comprehensive view of the scenario, enabling the reader to understand how the techniques evolved and combined to get to its current state.
- **Chapter 3 – Skin Lesion Image Synthesis:** We describe dermoscopy attributes, and detail how we take advantage of this annotation 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 concerning its quality, and how they impact classification when used for data augmentation.
- **Chapter 4 – Bias in Skin Lesion Datasets:** We investigate bias on skin lesion datasets, designing experiments to verify the network's performance when we expose it to correct or spurious correlations.
- **Chapter 5 – Conclusion:** Finally, we summarize and analyze our findings and propose future directions for the approached problems.

Chapter 2

GAN Literature Review

In this chapter, we review the literature of Generative Adversarial Networks (GANs). This story started in 2014 when Goodfellow et al. [35] 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.

We divide the advancements in six fronts — Architectural (Section 2.3), Conditional Techniques (Section 2.4), Normalization and Constraint (Section 2.5), Loss Functions (Section 2.6), Image-to-image Translation (Section 2.7), and Validation (Section 2.8) — providing a comprehensive notion of how the scenario evolved through the years, showing trends of thought that resulted to where we are today, where GANs are capable of generating face images that are almost indistinguishable from real photos (see Figure 2.1).

Figure 2.1: Synthetic human face generated with StyleGAN. Reproduced from Karras et al. [51].

Since we choose to give an evolutionary view of the GAN literature, sometimes the chronological information is inevitably lost in the process. To also communicate the time dimension of the extensive GAN literature, we chronologically organize the GANs we comment during the review in Figure 2.2 but also categorize them with respect to their main contribution, linking them to a section of this chapter.

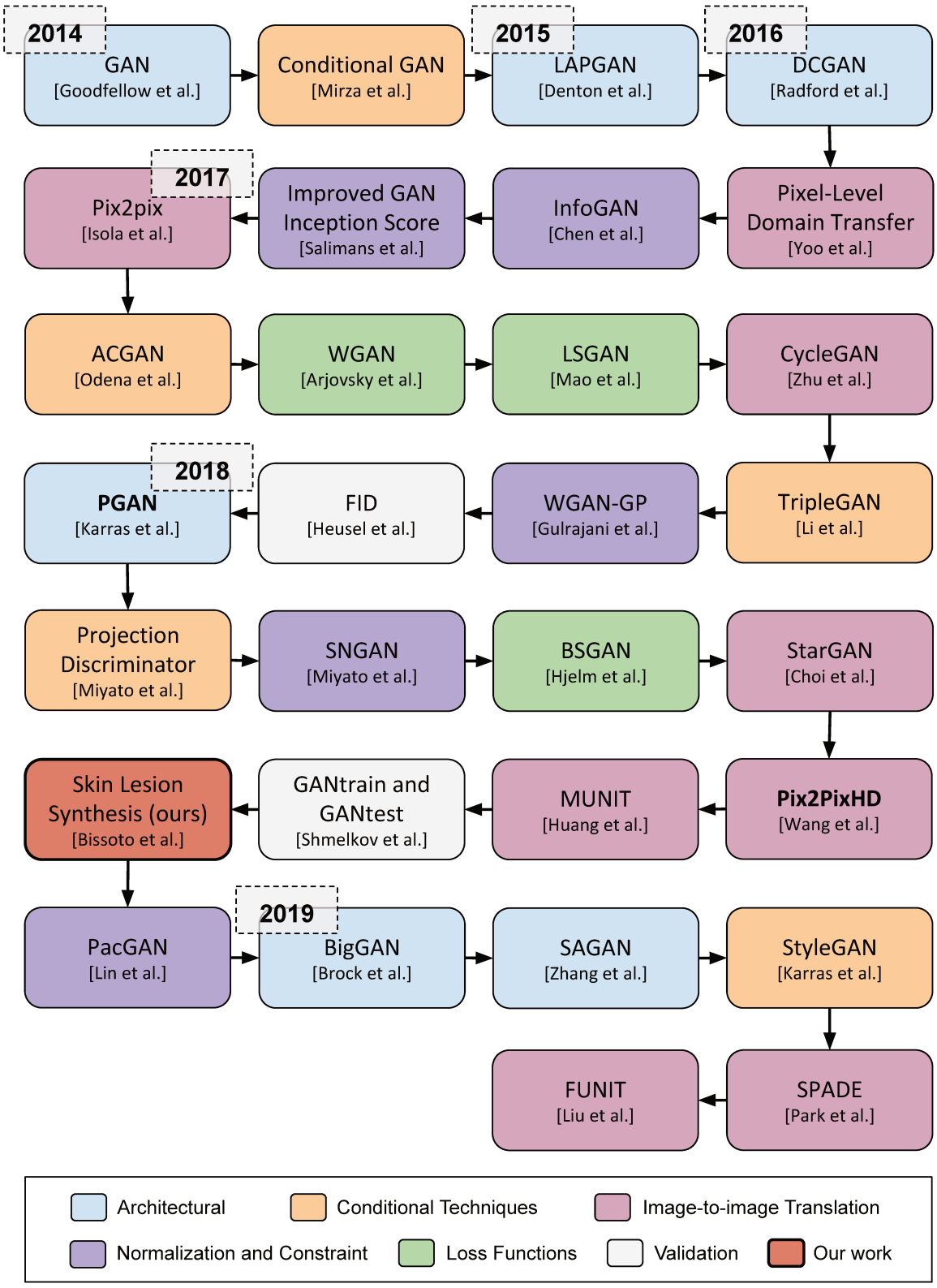


Figure 2.2: Timeline of the GANs covered in this GAN literature review. Just like our text, we also split it in six fronts (architectural, conditional techniques, normalization and constraint, loss functions, image-to-image translation and validation), each represented with a color. Our work in skin lesion synthesis is in orange.

2.1 Basic Concepts

Before introducing the formal concept of GANs, we start with an intuitive analogy proposed by Dietz [27]. The scenario is a boxing match, with a coach and two boxers. Let's call the fighters **Gabriel** and **Daniel**. Both boxers learn from each other during the match, but only **Daniel** has a coach. **Gabriel** only learns from **Daniel**. Therefore, at the beginning of the boxing, **Gabriel** keeps himself focused, observing his adversary and his moves, trying to adapt every round, guessing the teachings the coach gave to **Daniel**. After many rounds, **Gabriel** was able to learn the fundamentals of boxing during the match against **Daniel** and ideally, the boxing match would have 50/50 odds.

In the boxing analogy, **Gabriel** is the generator (G), **Daniel** is the discriminator (D), and the coach is the real data — larger the data, more experienced the coach. Goodfellow et al. [35] introduced GANs as two deep neural networks (the generator G and the discriminator D) that play a minimax two-player game with value function $V(D, G)$ as follows:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log (1 - D(G(z)))], \tag{2.1}$$

where z is the noise, p_z is the noise distribution, x is the real data, and p_{data} is the real data distribution.

The goal of the generator is to create samples as they were from the real data distribution p_{data}. To accomplish that, it learns the distribution of the real data and applies the learned mathematical function to a given noise z from the distribution p_z. The goal of the discriminator is to be able to discriminate between real (from the real data distribution) and generated samples (from the generator) with high precision.

During training, the generator receives feedback from the discriminator's decision (that classify the generated sample as real or fake), learning how to fool the discriminator better next time. In Figure 2.3, we show a simplified GAN pipeline.

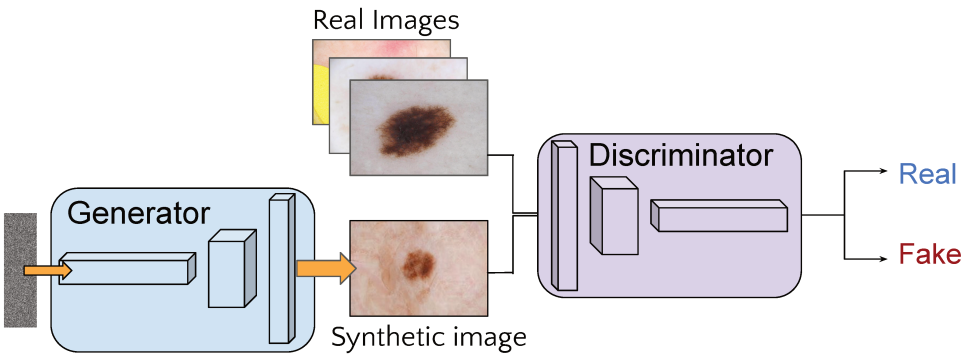


Figure 2.3: Simplified GAN Architecture. GANs are composed of two networks that are trained in a competition. While the generator learns to transform the input noise into samples that could belong to the target dataset, the discriminator learns to classify the images as real or fake. Ideally, after enough training, the discriminator can not differentiate between real and fake images, and the generator is synthesizing good quality images with high variability.

2.2 GAN's Challenges

In 2017, GANs suffered from high instability and were considered hard to train [7]. Since then, different architectures, loss functions, conditional techniques, and constrain methods were introduced, easing the convergence of GAN models. However, there are still hyperparameter choices that may highly influence training. The batch size and layers width were recently in the spotlight after BigGAN [18] showed state-of-the-art results for ImageNet image synthesis by increasing radically these, among other factors. Aside from its impressive results, it showed directions to improve the GAN framework further.

Is the gradient noise introduced when using small mini-batches more impactful than the problems caused by the competition between discriminator and generator? And of course, how much can GAN benefit (if it can at all) from scaling to very deep architectures and parallelism, and vast volumes of data? It seems that the computational budget can be critical for the future of GANs. Lucic et al. [60] showed that given enough time for hyperparameter tuning and random restarts, despite multiple proposed losses and techniques, different GANs can achieve the same performance.

However, even in normal conditions, GANs achieve incredible performance for specific tasks such as face generation, but the same quality is not perceived when dealing with more general datasets such as ImageNet. The characteristics of the dataset that favor GANs performance still uncertain. A possibility is the regularity (using the same object poses, placement on the image, or using different objects with the same characteristics) are easier than others where variation is high such ImageNet.

A factor that is related to data regularity and variability is the number of classes. GANs suffer to cover all class possibilities of a target dataset if it is unbalanced (*e.g.*, medical datasets) or if the class count is too high (*e.g.*, ImageNet). This phenomenon is called "mode collapse" [7], and despite efforts from the literature to mitigate the issue [58, 85], modern GAN solutions still display this undesired behavior. If we think in the use case where we want to augment a training dataset with synthetic images, which is the case for this Master thesis, it is even more impactful once the object classes we want to generate are usually the most unbalanced ones.

Validating synthetic images is also challenging. Qualitative evaluation, where grids of low-resolution images are compared side-by-side, was (and still is) one of the most used methods for performance comparison between different works. Authors also resort to services like the Amazon Mechanical Turk (AMT), which enable to perform statistical analysis on multiple human annotators' choices over synthetic samples. However, except for the cases where the difference is massive, the subjective nature of this approach may lead to wrong decisions.

Ideally, we want quantitative metrics that consider different aspects of the synthetic images, such as the overall structure of the object, the presence of fine details, and variability between samples. The most accepted metrics, IS and FID, both rely on the activations of an ImageNet pre-trained Inceptionv3 network to output its scores. This design causes scores to be unreliable, especially for contexts that are not similar to ImageNet's.

Other metrics such as GANtrain and GANtest can analyze those mentioned aspects, however we have to consider the possible flaws of the used classification networks —

especially bias — which could reinforce its presence in the synthetic images. Thus, only by employing diversified metrics, evaluation methods, and content-specific measures, we can truly assess the quality of the synthetic samples.

2.3 Architectural Methods

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 [56] and Faces [89]. However, in 2016, crucial architectural changes were proposed, boosting GANs research and increasing the complexity and quality of the synthetic samples.

Deep Convolutional GAN (DCGAN) [80] proposed detailed architectural guidelines that stabilized GAN's training, enabling the use of deeper models and achieving higher resolutions (see Figure 2.4). 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 [50, 69] and still is used in modern GAN frameworks [70]. DCGAN caused such an impact in the community that it is still used nowadays when working with simple low-resolution datasets and as an entry point when applying GANs in new contexts.

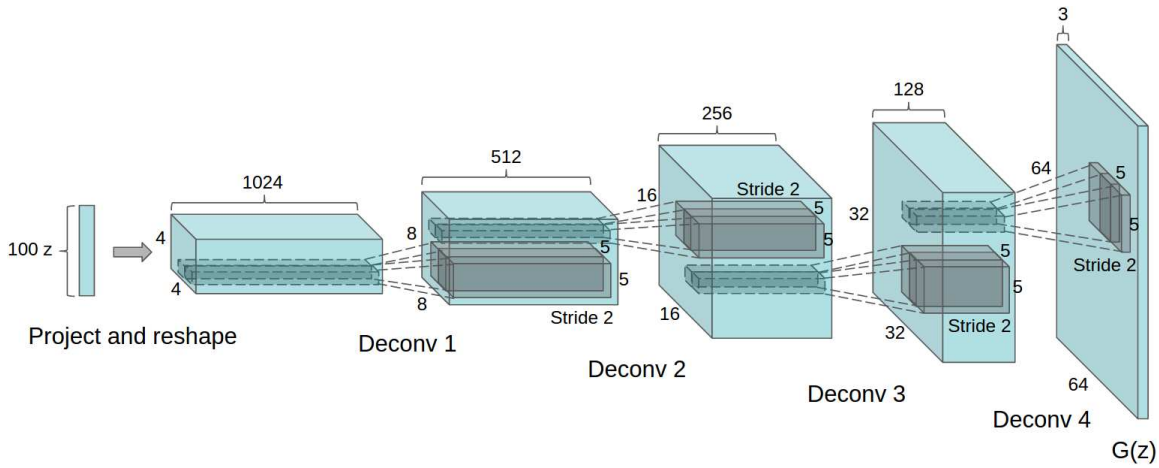


Figure 2.4: DCGAN's generator architecture reproduced from Radford et al. [80]. The replacement of fully-connected and pooling layers with (strided and fractional-strided) convolutional layers enabled GANs architecture to be deeper and more complex.

At the same period, Denton et al. [26] proposed Laplacian Pyramid GAN (LAPGAN): an incremental architecture where the resolution of the synthetic sample is progressively increased across the generation pipeline. This modification enabled the generation of synthetic images up to 96×96 resolution.

In 2018, this type of architecture gained popularity, and it is still employed for improved stability and high-resolution generation. Progressive GAN (PGAN) [50] improved the incremental architecture to generate human faces of 1024×1024 resolution. While the spatial resolution of the generated samples increases, layers are progressively

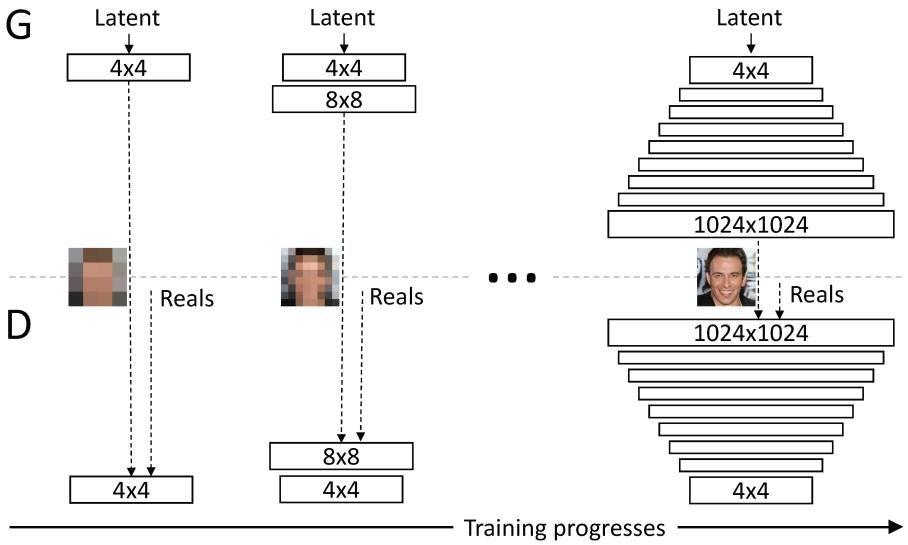


Figure 2.5: Simplified PGAN's Incremental Architecture. PGAN starts by generating and discriminating low-resolution samples, adjusting the generation's coarse details. Progressively, the resolution is increased, enhancing synthetic image's fine details. Reproduced from Karras et al. [50].

added for both Generator and Discriminator (see Figure 2.5). 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 layers.

Very recently, StyleGAN [51] showed remarkable performance when generating human faces, dethroning PGAN in this task. Despite keeping the progressive training procedure, several changes were proposed (see Figure 2.6). The authors changed how information is usually fed into the generators: In other works, the main input of noise passes through the whole network, being transformed layer after layer until becoming the final synthetic sample. In StyleGAN, the generator receives information directly in all its layers. Before feeding the generator, the input label data goes through a mapping network (composed of several sequential fully-connected layers) that extract class information. This information, called "style" by the authors, is then combined with the input latent vector (noise). Then, it is incorporated in all the generator's layers by providing the Adaptive Instance Normalization (AdaIN) [42] layer's parameters for scale and bias.

Also, authors added extra independent sources of noise that feeds different layers of the generator. By doing this, authors expect that each source of noise can control a different stochastic aspect of the generation (*e.g.*, placement of hair, skin pores, and background) enabling more variation and higher detail level.

An effect of these modifications, specially the mapping network, is the disentanglement of the relations between the noise and the style. This enables subtle modifications on the noise to reflect on subtle changes to the generated sample while keeping it plausible, improving the quality of the generated images, and stabilizing training.

The same idea of disentangling the class and latent vectors was first explored by InfoGAN [20], where the network was responsible for finding characteristics that could control

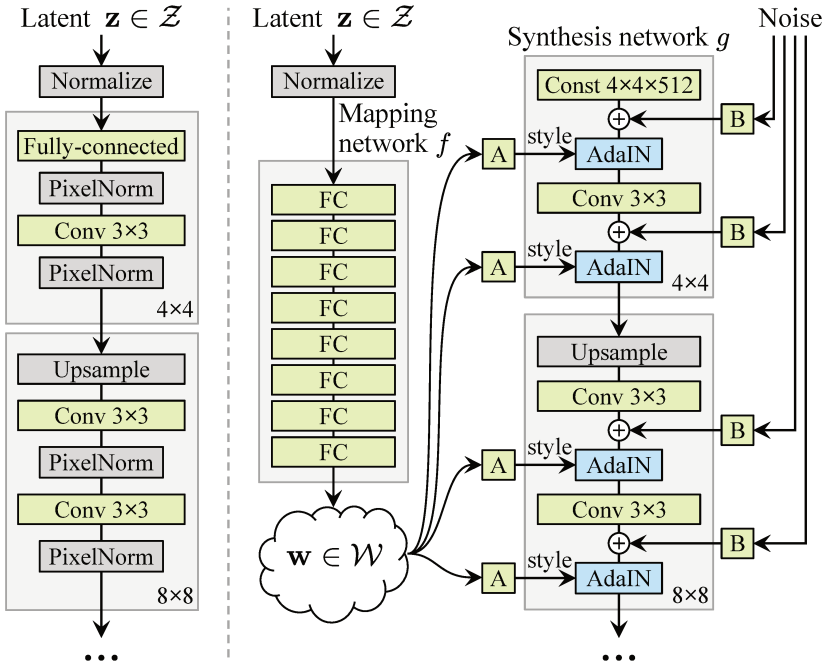


Figure 2.6: Comparison between generators on PGAN [50] (on the left) and StyleGAN (on the right). First, a sequence of fully-connected layers is used to extract style information and disentangle the noise and class information. Next, style information feeds AdaIN layers, providing its parameters for shift and scale. Also, different sources of noise feed each layer to simplify the task of controlling stochastic aspects of generation (such as placement of hair and skin pores). Reproduced from Karras et al. [51].

generation and disentangle this information from the latent vector. This disentanglement allows discovered (by the network itself) features, which are embedded in the latent vector, to control visual aspects of the image, making the generation process more coherent and tractable.

Another class of architectural modifications consists of GANs that were expanded to include different networks in the framework. Encoders are the most common component aside from the original generator and discriminator. Their addition to the framework created an entire new line of possibilities doing image-to-image translation [47] (detailed later in Section 2.7). Also, super-resolution and image segmentation solutions use encoders in their architecture. Since encoders are present in other generative methods such as Variational Autoencoders (VAE) [52], there were also attempts of extracting the best of both generative methods [55] by combining them. Finally, researchers also included classifiers to be trained jointly with the generator and discriminator, improving generation [72], and semi-supervised capabilities [22].

2.4 Conditional Techniques

In 2014, Mirza et al. [67] introduced a way to control the class of the generated samples: concatenate the label information to the generator's input noise. Despite simple, this method instigated researchers to be creative with GAN inputs, and enabled GANs to solve different and more complex tasks.

The next step was to include the label in the loss function. Salimans et al. [85] proposed to increase the output size of the discriminator, to include classes, for semi-supervised classification purposes. Although having a different objective, it inspired future works. Odena et al. [73] focused on generation with Auxiliary Classifier GAN (ACGAN), splitting both tasks (label and real/fake classification) for the discriminator, while still feeding the generator with class information. It greatly improved generation, creating 128×128 synthetic samples (surpassing LAPGAN's 96×96 resolution) for all 1000 classes of ImageNet. Also, authors showed that GANs benefit from higher resolution generation, which increases discriminability for a classification network.

The next changes happened again modifying how the class information is fed to the network. On TripleGAN [22] the class information is concatenated in all layers' feature vectors, except for the last one, on both discriminator and generator. Note that sometimes the presented approaches can be combined: concatenation became the standard method of feeding conditional information to the network, while using ACGAN's loss function component.

Next, Miyato et al. [70] introduced a new method to feed the information to the discriminator. The class information is first embedded, and then integrated to the model through an inner product operation at the end of the discriminator. In their solution, the generator continued concatenating the label information to the layers' feature vectors. Although the authors show the exact networks used in their experiments, this technique is not restricted and can be applied to any GAN architecture for different tasks.

Recently, an idea used for style transfer was incorporated to the GAN framework. In style transfer the objective is to transform the source image in a way it resembles the style (especially the texture) of a given target image, without losing its content components. The content comprehends information that can be shared among samples from different domains, while style is unique to each domain. When considering paintings for example, the "content" represent the objects in the scene, with its edges and dimensions; while "style" can be the artists' painting (brushing) technique and colors used. Adaptive Instance Normalization (AdaIN) [42] is used to infuse a style (class) into content information. The idea is to *first* normalize the feature maps according to its dimensions' mean and variance evaluated for each sample, and each channel. Intuitively, Huang et al. explained that at this point (which is called simply Instance Normalization), the style itself is normalized, being appropriate to receive a new style. Finally, AdaIN uses statistics obtained from a style encoder to scale and shift the normalized content, infusing the target style into the instance normalized features.

This idea was adapted and enhanced for current state-of-the-art methods for plain generation [43,51], image-to-image translation [75], and to few-shot image-translation [59]. Authors usually employ an encoder to extract style that feeds a multi-layer perceptron (MLP) that outputs the statistics used to control scale and shift on AdaIN layers.

In Spatially-Adaptive Normalization (SPADE) [75], a generalization of previous normalizations (*e.g.*, Batch Normalization, Instance Normalization, Adaptive Instance Normalization) is used to incorporate the class information into the generation process, but it focuses at working with input semantic maps for image-to-image translation. The parameters used to shift and scale the feature maps are tensors that preserve spatial information

from the input semantic map. Those parameters are obtained through convolutions, then multiplied and summed element-wise to the feature map. Since this information is included in most of the generator's layers, it prevents the map information to fade away during the generation process.

2.5 Normalization and Constraint Techniques

On DCGAN [80], authors advocated the use of batch normalization [45] layers on both generator and discriminator to reduce internal covariate shift. It was the beginning of the use of normalization techniques, which also developed with time, for increased stability. In batch normalization, the output of a previous activation layer is normalized using the current batch statistics (mean and standard variation). Then, it adds two trainable parameters to scale and shift the normalization result.

For PGAN [50] authors, the problem is not the internal covariate shift, but the signal magnitudes that explode due to competition between the generator and discriminator. To solve this problem, they moved away from batch normalization, and introduced two techniques to constrain the weights. On *Pixelwise Normalization* (see Equation 2.2), they normalize the feature vector in each pixel. This approach is used in the generator, and do not add any trainable parameter to the network.

$$b_{x,y} = a_{x,y} / \sqrt{\frac{1}{N}\sum_{j=0}^{N-1}(a_{x,y}^{j})^2 + \epsilon}, \tag{2.2}$$

where $\epsilon = 10^{-8}$, N is the number of feature maps, and $a_{x,y}$ and $b_{x,y}$ are the original and normalized feature vector in pixel (x, y), respectively.

The other technique introduced by Karras et al. [50] is called *Equalized Learning Rate* (see Equation 2.3).

$$\hat{w}_i = w_i / c, \tag{2.3}$$

where w_i are the weights and c is the per-layer normalization constant from He's initializer [37]. During initialization, weights are all sampled from the same distribution $\mathcal{N}(0, 1)$ at runtime, and then scaled by c. It is used to avoid weights to have different dynamic ranges across different layers. This way, the learning rate impacts all the layers by the same factor, avoiding it to be too large for some, and too little for others.

Spectral Normalization [69] also acts constraining weights, but only at the discriminator. The idea is to constrain the Lipschitz constant of the discriminator by restricting the spectral norm of each layer. By constraining its Lipschitz constant, it limits how fast the weights of the discriminator can change, stabilizing training. For this, every weight W is scaled by the largest singular value of W. This technique is currently present in state-of-the-art networks [75, 101], being applied on both generator and discriminator.

So far, the presented constraint methods concern about normalizing weights. Differently, *Gradient penalty* [36] enforces 1-Lipschitz constraint to make all gradients with norm at most 1 everywhere. It adds an extra term to the loss function to penalize the

model if the gradient norm goes beyond the target value 1.

Some methods did not change loss functions or added normalization layers to the model. Instead, those methods introduced subtle changes in the training process to deal with GAN's general problems, such as training instability and low variability on the generated samples. *Minibatch Discrimination* [85] gives the discriminator information around the mini-batch that is being analyzed. Roughly speaking, this is achieved by attaching a component[1] to an inner layer of the discriminator. With this information, the discriminator can compare the images on the mini-batch, forcing the generator to create images that are different from each other.

Similarly with respect of giving more information to the discriminator, PacGAN [58] packs (concatenates in the width axis) different images from the same source (real or synthetic) before feeding the discriminator. According to the authors, this procedure helps the generator to cover all target labels in the training data, instead of limiting itself to generate samples of a single class that are able to fool the discriminator (a problem called *mode collapse* [7]).

2.6 Loss Functions

Theoretical advances towards understanding GANs training and the sources of its training instabilities [7] pointed that the Jensen-Shannon Divergence (JSD) (which is used in the GAN's formulation to measure similarity between the real data distribution and the generator's) is responsible for vanishing the gradients when the discriminator is already well trained. This theoretical understanding contributed to motivate next wave of works, that explored alternatives to the JSD.

Instead of JSD, authors proposed using the Pearson χ^2 (Least Squares GAN) [61], the Earth-Mover Distance (Wasserstein GAN) [8], and Cramér Distance (Cramér GAN) [11]. One core principle explored was to penalize samples even when it is on the correct side of the decision boundary, avoiding the vanishing gradients problem during training.

Other introduced methods choose to keep the divergence function intact and introduce components to the loss function to increase image quality, training stability, or to deal with mode collapse and vanishing gradients. Those methods often can be employed together (and with different divergence functions) evidencing the many possibilities of tuning GANs when working on different contexts.

An example that shows the possibility of joining different techniques is the Boundary-Seeking GAN (BSGAN) [40], where a simple component (which must be adjusted for different f-divergence functions) tries to guide the generator to generate samples that make the discriminator output 0.5 for every sample.

Feature Matching [85] comprises a new loss function component for the generator that induces it to match the features that better describe the real data. Naturally, by training the discriminator we are asking it to find these features, which are present in its intermediate layers. Similarly to Feature Matching, *Perceptual Loss* [49] also uses statistics from a neural network to compare real and synthetic samples and encourage

[1] Vector of differences between the sample being analyzed and the others present in the mini-batch.

them to match. Differently though, it employs ImageNet pre-trained networks (VGG [88] is often used), and add an extra term to the loss function. This technique is commonly used for super-resolution methods, and also image-to-image translation [96].

Despite all the differences between the distinct loss function, and methods used to train the networks, given enough computational budget, all can reach comparable performance [60]. However, since solutions are more urgent than ever, and GANs have the potential to impact multiple areas, from data augmentation, image-to-image translation, super-resolution and many others, gathering the correct methods that enable a fast problem-solver solution is essential.

2.7 Image-to-image Translation Methods

The addition of encoders in the GAN architecture enabled GANs for the task of image-to-image translation. In 2016, Yoo et al. [99] 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 analyze pairs of source and target (real and fake) samples and judge if they are associated or not.

Up until this time, the synthetic samples follow the same quality of plain generation's: low quality and low resolution. This scenario changes with pix2pix [47]. Pix2pix employed a new architecture for both the generator and the discriminator, as well as a new loss function. It was a complete revolution! We reproduce a simplification of the referenced architecture in Figure 2.7. The generator is a U-Net-like network [82], where the skip connections allow to bypass information that is shared by the source-target pair. Also, the authors introduce a patch-based discriminator (which they called PatchGAN) to penalize structure at scale of patches of a smaller size (usually 70×70), while accelerating evaluation. To compose the new loss function, authors proposed the addition of a term that evaluates the L1 distance between synthetic and ground truth targets, constraining the synthetic samples without killing variability.

Despite advances on conditional plain generation techniques such as ACGAN [73], which allowed generation of samples up to 128×128 resolution, the quality of synthetic samples reached a new level with its contemporary pix2pix. This model is capable to generate 512×512 resolution synthetic images, comprising state-of-the-art levels of detail for that time. Overall, feeding the generator with the source sample's extra information simplify and guide generation, impacting the process positively.

The same research group responsible for pix2pix later released CycleGAN [102], further improving the overall quality of the synthetic samples. The new training procedure (see Figure 2.8) enforces the generator to make sense of two translation processes: from source to target domain, and also from target to source. The cyclic training also uses separate discriminators to deal with each one of the translation processes. The architectures on the discriminators are the same from pix2pix, using 70×70 patches, while the generator

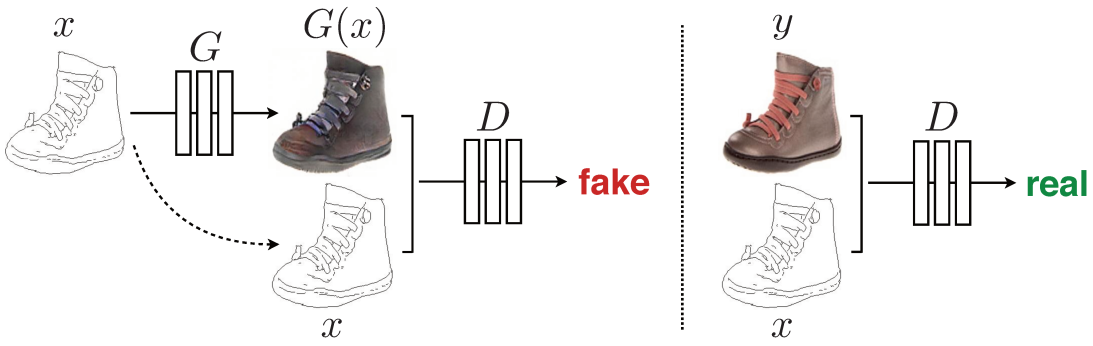


Figure 2.7: Pix2pix's training procedure. Source-target domain pairs are used for training: the generator is an autoencoder that translates the input source domain to the target domain; the discriminator critic pairs of images composed of the source and (real or fake) target domains. Reproduced from Isola et al. [47].

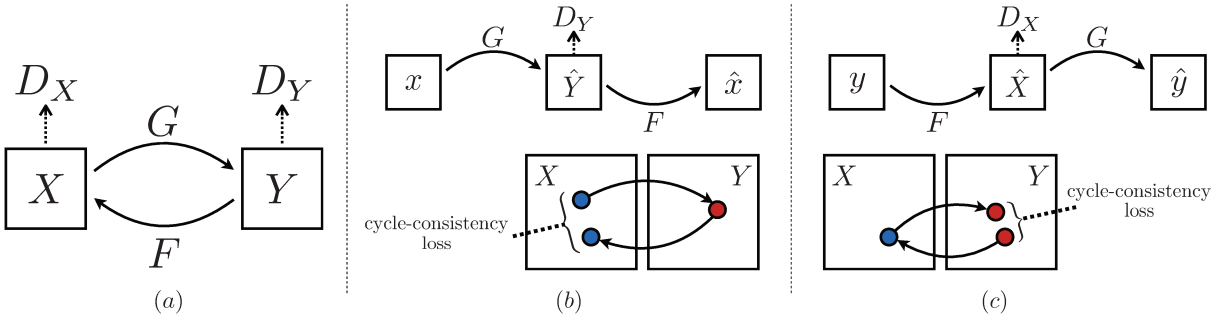


Figure 2.8: (a) CycleGAN uses two Generators (G and F) and two discriminators (D_X and D_Y) to learn to transform the domain X into the domain Y and vice-versa. To evaluate the cycle-consistency loss, authors force the generators to be able to reconstruct the image from the source domain after a transformation. That is, given domains X and Y, generators G and F should be able to: (b) $X \rightarrow G(X) \rightarrow F(G(X)) \approx X$ and (c) $Y \rightarrow F(Y) \rightarrow G(F(Y)) \approx Y$. Reproduced from Zhu et al. [102].

receives the recent architecture proposed by Johnson et al. [49] for style transfer.

CycleGAN increased the domain count managed (generated) by a single GAN to two, making use of two discriminators (one for each domain) to enable translation between both learned domains. Besides the architecture growth needed, another limitation is the requirement of having pairs of data connecting both domains. Ideally, we want to increase the domain count without scaling the number of generators or discriminators proportionally, and have partially-labeled datasets (that is, not having pairs for every source-target domains). Those flaws motivated StarGAN [21]. Apart from the source domain image, StarGAN's generator receives an extra array containing labels' codification that informs the target domain. This information is concatenated depth-wise to the source sample before feeding the generator, which proceeds to perform the same cyclic procedure from CycleGAN, making use of a reconstruction loss. To deal with multiple classes, without increasing the discriminator count, it accumulates a classification task to evaluate the domain of the analyzed samples.

The next step towards high-resolution image-to-image translation is pix2pixHD (High-Definition) [97], which was widely used during this Master thesis to generate realistic

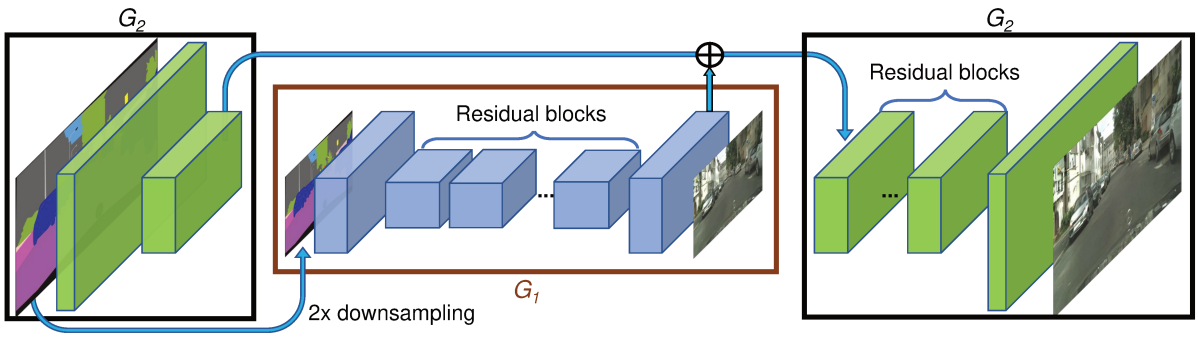


Figure 2.9: Summary of Pix2pixHD generators. G_1 is the global generator and G_2 is the local generator. Both generators are first trained separately, starting from the lower-resolution global generator, and next, proceeding to the local generator. Finally, both are fine-tuned together to generate images up to 2048×1024 resolution. Reproduced from Wang et al. [97].

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 (see Figure 2.9), where the outer "local" generator enhances the generation of the inner "global" generator. Just like CycleGAN, it uses Johnson et al. [49] style transfer network as global generator, and as 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 of 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 concern more about the general structure and coarse details, while the high-resolution discriminators will pay attention to fine details. The loss function also became more robust: besides the traditional adversarial component for each of the discriminators and generators, it comprises feature matching and perceptual loss components.

Other less structural aspects of generation, but maybe even more important, were explored by Wang et al. [97]. Usually, the input for image-to-image translation networks is semantic maps [57]. It is an image where every pixel has the value of its object class and it is commonly seen as a result of pixelwise segmentation tasks. During evaluation, the user can decide and pick the desired attributes of the result synthetic image by crafting the input semantic map. However, pix2pixHD authors noticed that sometimes this information is not enough to guide the network's generation. For example, let us think of the semantic map containing a queue of cars in the street. The blobs corresponding to each car would be connected, forming a strange format (for a car) blob, making it difficult for the network to make sense of it.

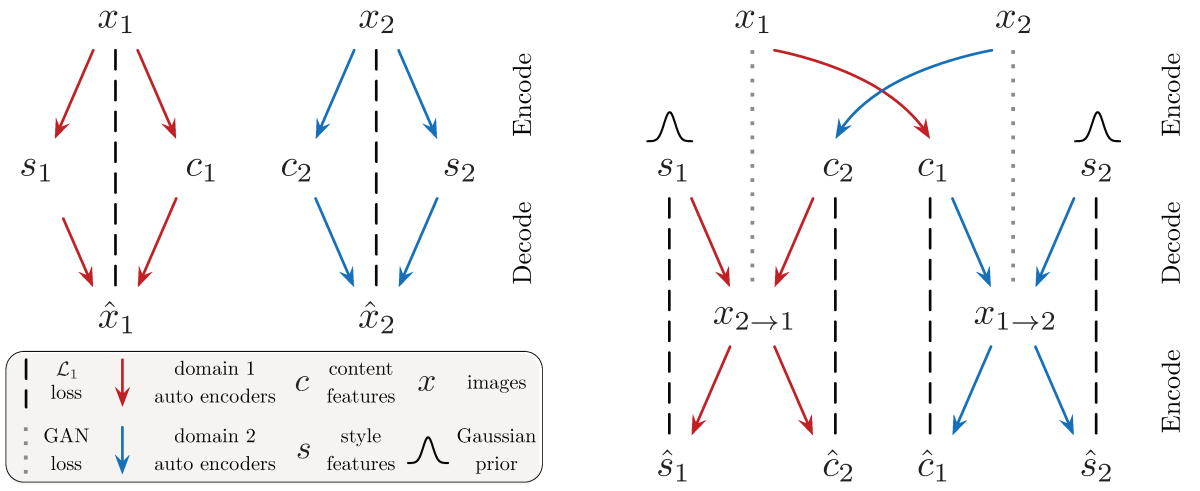
The authors' proposed solution is to add an instance map to the input of the networks. The instance map [57] is an image where the pixels combine information from its object

class and its instance. Every instance of the same class receives a different pixel value. The addition of instance maps is one of the factors that affected the most our skin lesion generation, as well as other contexts showed in their paper.

A drawback of pix2pixHD and the other methods so far is that generation is deterministic (*e.g.*, in test time, for a given source sample, the result is always the same). This is an undesired behavior if we plan to use the synthetic samples to augment a classification model's training data, for example.

Huang et al. introduce Multimodal Unsupervised Image-to-image Translation (MUNIT) [43] to generate diverse samples with the same source sample. For each domain, the authors employ an encoder and a decoder to compose the generator. The main assumption is that it is possible to extract two types of information from samples: content, that is shared among instances of different domains, controlling general characteristics of the image; and style, that controls fine details that are specific and unique to each domain. The encoder learns to extract content and style information, and the decoder to take advantage of this information.

During training, two reconstruction losses are evaluated: the image reconstruction loss, which measure the ability to reconstruct the source image using the extracted content and style latent vectors; and the latent vectors reconstruction loss, which measure the ability to reconstruct the latent vectors themselves, comparing a pair of source latent vectors sampled from a random distribution, with the encoding of a synthetic image created using them (see Figure 2.10).



(a) Within-domain reconstruction

(b) Cross-domain translation

Figure 2.10: MUNIT's training procedure. (a) Reconstruction loss with respect to the images of the same domain. Style "s" and content "c" information are extracted from the real image, and used to generate a new one. The comparison of both images composes the model's loss function. (b) For cross-domain translation, the reconstruction of the latent vectors containing style and content information also compose the loss function. The content is extracted from a source real image, and the style is sampled from a Gaussian prior. Reproduced from Huang et al. [43].

MUNIT's decoder (see Figure 2.11) incorporate style information using AdaIN [42]

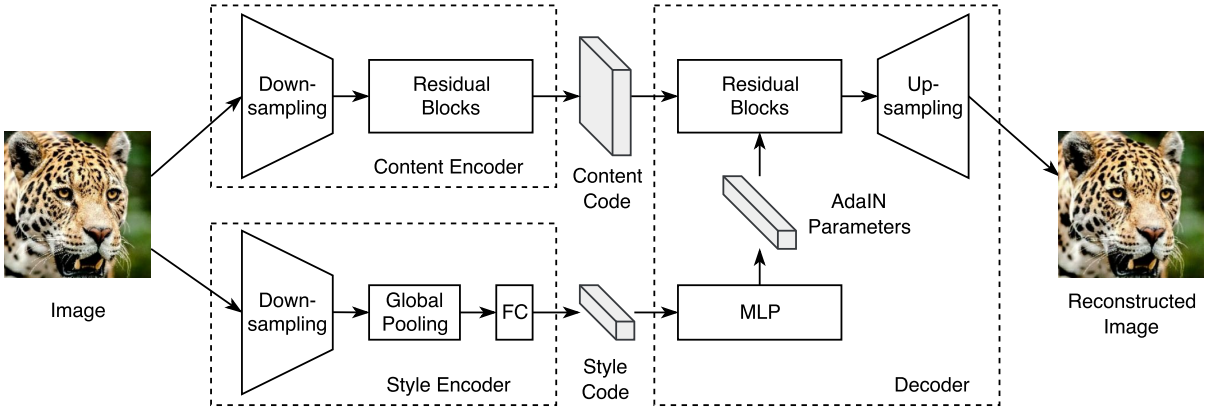


Figure 2.11: The generator's architecture from MUNIT. The authors employ different encoders for content and style. The style information feeds a Multi-Layer Perceptron (MLP) that provide the parameters for the AdaIN normalization. This process incorporates the style information to the content, creating a new image. Reproduced from Huang et al. [43].

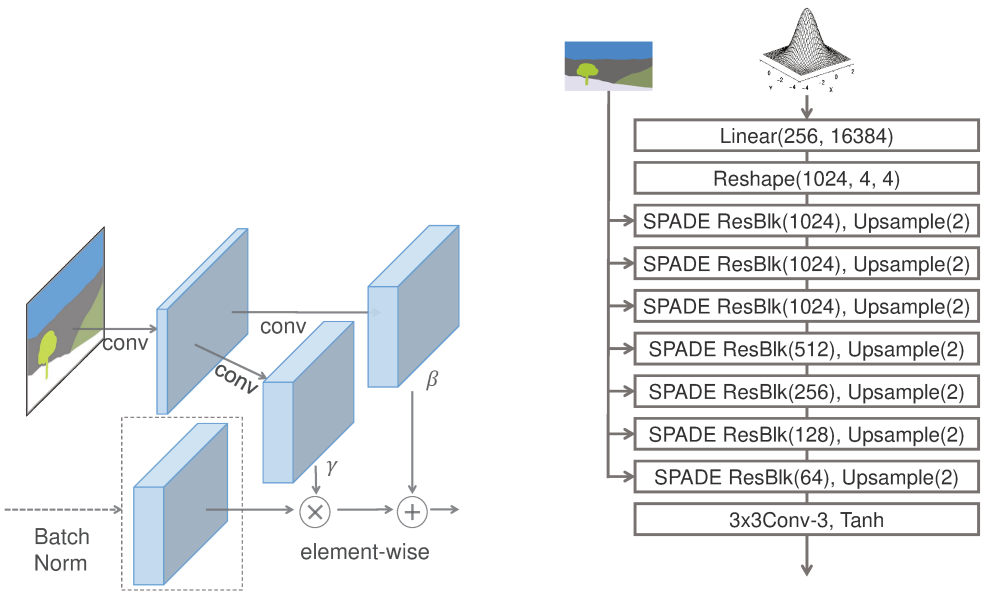
(we detailed AdaIN, and the intuitions behind using this normalization in Section 2.4), and it directly influenced current state-of-the-art GAN architectures [51, 59, 75].

One of the influenced works deals with a slightly different task: few-shot image-to-image translation. Few-shot translation attempts to translate a source image to a new unseen (but related) target domain after looking into just a few (2 or so) examples during test time (*e.g.*, train with multiple dog breeds, and test for lions, tigers, cats, wolves). To work on this problem, Liu et al. [59] introduce Few-shot Unsupervised Image-to-Image Translation (FUNIT). It combines and enhances methods from different GANs we already described. Authors propose an extension of CycleGAN's cyclic training procedure to multiple source classes (authors advocate that the higher the amount, the best the model's generalization); adoption of MUNIT's encoders for content and style, that are fused through AdaIN layers; enhancement of StarGAN's procedure to feed class information to the generator in addition to the content image, where instead of simple class information, the generator receives a set of few images of the target domain. The discriminator also follows StarGAN's, in a way that it performs an output for each of the source classes. This is an example of how works influence each other, and of how updating an older idea with enhanced recent techniques can result in a state-of-the-art solution.

So far, every image-to-image translation GAN generator's 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 important role to extract 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.

On Spatially-Adaptive Normalization (SPADE) [75] authors introduce a method for semantic image synthesis (*e.g.*, image-to-image translation using a semantic map as the generator's input). It can be considered pix2pixHD [97] successor, in a way that it deals with much of the previous work flaws. Although the authors call their GAN as SPADE in the paper (they call it GauGAN now), the name refers to the introduced normalization process, which generalizes other normalization techniques (*e.g.*, Batch Normalization,

Instance Normalization, AdaIN). Like AdaIN, SPADE is used to incorporate the input information into the generation process, but there is a key difference between both methods. On SPADE, the parameters used to shift and scale the feature maps are tensors that contain spatial information preserved from the input semantic map. Those parameters are obtained through convolutions, then multiplied and summed element-wise to the feature map (see Figure 2.12a). This process takes place in all the generator's layers, except for the last one, which outputs the synthetic image. Since the input of the generator's decoder is not the encoding of the semantic map, authors use noise to feed the first generator's layer (see Figure 2.12b). This change enables SPADE for multimodal generation, that is, given a target semantic map, SPADE can generate multiple different samples using the same map.



(a) SPADE normalization process.

(b) SPADE's generator.

Figure 2.12: (a) At the SPADE normalization process, the input semantic map is first projected into a feature space. Next, different convolutions are responsible for extracting the parameters that perform element-wise multiplication and sum to the normalized generator's activation. (b) A sampled noise is modified through the residual blocks. After each block, the output is shifted and scaled using AdaIN. This way, the style information of the input semantic map is present in every stage of the generation, without killing the output variability that comes with the sampled noise at the beginning of the process. Reproduced from Park et al. [75].

2.8 Validation of Synthetic Methods

Inception Score (IS) [85]: it uses an Inceptionv3 network pre-trained on ImageNet to compute the logits of the synthetic samples. The logits are used to evaluate Equation 2.4. The authors say it correlates well with human judgment over synthetic images. Since the network was pre-trained on ImageNet, we rely on its judgment of the synthetic images with respect to the ImageNet classes. This is a big problem because skin lesion images

do not relate to any class on ImageNet. Thus, this method is inappropriate to evaluate synthetic samples from any dataset that is not ImageNet.

$$\begin{aligned} \text{IS}(G) = \exp\big(\ _{\mathbf{x}\sim p_g}\ D_{KL}(\ p(y|\mathbf{x})\ \|\ p(y)\ )\ \big), \\ p(y) = \int_{\mathbf{x}} p(y|\mathbf{x}) p_g(\mathbf{x}), \end{aligned} \tag{2.4}$$

where $\mathbf{x} \sim p_g$ indicates that $\mathbf{x}$ is an image sampled from p_g, p_g is the distribution learned by the generator, $D_{KL}(p\|q)$ is the Kullback Leibler divergence between the distributions p and q, $p(y|\mathbf{x})$ is the conditional class distribution (logit of a given sample), and $p(y)$ is the marginal class distribution (mean logits over all synthetic samples).

Frechèt Inception Distance (FID) [39]: like the Inception Score, the FID relies on Inception's evaluation to measure quality of synthetic samples and suffer from the same problems. Differently though, it takes the features from the Inception's penultimate layer from both real and synthetic samples, comparing them. The FID uses Gaussian approximations for these distributions, which makes it less sensitive to small details (which are abundant in high-resolution samples).

Sliced Wasserstein Distance (SWD) [50]: Karras et al. introduced the SWD metric to deal specifically with high-resolution samples. The idea is to consider multiple resolutions for each image, going from 16×16 and doubling until maximum resolution (Laplacian Pyramid). For each resolution, slice 128 $7 \times 7 \times 3$ patches from each level, for both real and synthetic samples. Finally, use the Sliced Wasserstein Distance [79] to evaluate an approximation to the Earth Mover's distance between both.

GANtrain and GANtest [87]: the idea behind these metrics align with our objective of using synthetic images as part of a classification network, like a smarter data augmentation process. GANtrain is the accuracy of a classification network trained on the synthetic samples, and evaluated on real images. Similarly, GANtest is the accuracy of a classification network trained on real data, and evaluated on synthetic samples. The authors compare the performance of GANtrain and GANtest with a baseline network trained and tested on real data. We explored this idea of employing a classification network to evaluate the synthetic samples before Shmelkov et al. formally proposed it, and we reported our results in our work [15] presented at the ISIC Skin Image Analysis Workshop at MICCAI 2018.

Borji [16] analyzed the existing metrics in different criteria: discriminability (capacity of favoring high-fidelity images), detecting overfitting, disentangled latent spaces, well-defined bounds, human perceptual judgments, sensitivity to distortions, complexity and sample efficiency. After an extensive review of the metrics literature, the author compares the metrics concerning the presented criteria, and among differences and similarities, can not point the definitive metric to be used. The author suggests future studies to rely on different metrics to better assess the quality of the synthetic images.

Theis et al. [93], in a study of quality assessment for synthetic samples, highlighted that the same model may have very different performance on different applications, thus, a proper assessment of the synthetic samples must consider the context of the application. In this Master thesis, we follow in the same direction, employing tests that are specific for skin lesions to evaluate their clinical information.

Chapter 3

Skin Lesion Image Synthesis

GANs aim to model the real image distribution by forcing the synthesized samples to be indistinguishable from real images. Just recently works have shown promising results for high-resolution image generation [50, 51, 75, 96]. At the time of this work, the scenario was wholly dominated by low-resolution GANs (32×32 data generation, *e.g.*, [8, 61, 80]). However, for skin cancer classification, the images must have a higher level of detail (high-resolution) to be able to display malignancy markers (Section 3.1) that differ a benign from a malignant skin lesion. The scenario started to shift towards high-resolution generation, especially with Karras et al.'s Progressive GAN (PGAN) [50], and architectural advances on image-to-image translation [96].

PGAN's [50] progressive training procedure generates celebrity faces up to 1024×1024 pixels. At the beginning of the training phase, authors feed the network with low-resolution samples. Progressively, the network receives increasingly higher resolution training samples while amplifying the respective layers' influence to the output. In the same direction, Pix2pixHD [96] generates high-resolution images from semantic and instance maps. The authors propose to use multiple discriminators and generators that operate in different resolutions to evaluate fine-grained detail and global consistency of the synthetic samples. We investigate both networks for skin lesion synthesis, comparing the achieved results.

In this Master thesis, 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 (for illustration, see Figure 3.1). To evaluate the relevance of synthetic images, we trained a skin cancer classification network with synthetic and real images, reaching an improvement of 1 percentage point. Also, we analyze the synthetic images, comparing our modifications of PGAN and Pix2pixHD. Our full implementation is available at `https://github.com/alceubissoto/gan-skin-lesion`.

We highlight that the main contribution of this chapter was published at the ISIC Skin Image Analysis Workshop at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018), and still is up today the state-of-the-art for skin lesion images synthesis.

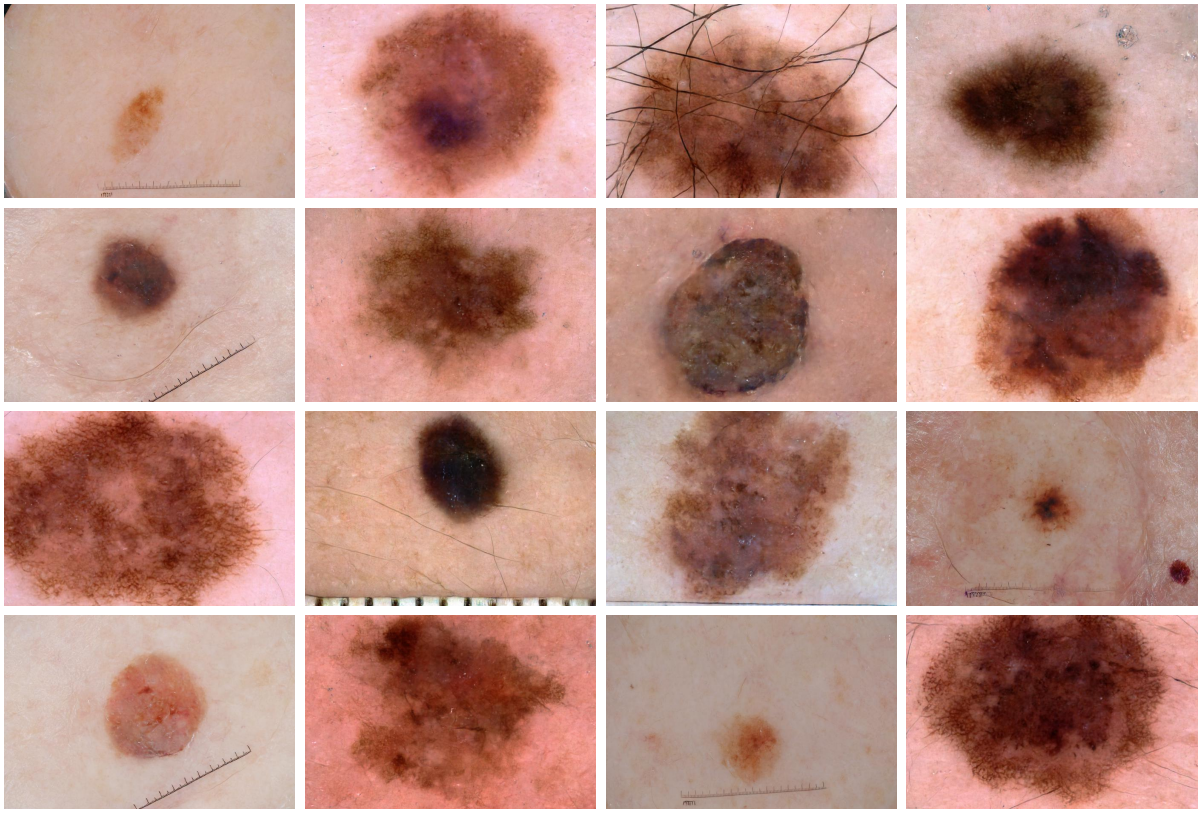


Figure 3.1: Our approach successfully generates high-definition, visually-appealing, clinically-meaningful synthetic skin lesion images. All samples are synthetic. Details can be found in Section 3.3.

3.1 Related Concepts – Dermoscopy

In this Master thesis, we take advantage of a very special annotation called dermoscopic attributes to our generation process. In this section, we give some details of its capabilities, and about the way it is currently presented on skin lesion datasets.

Dermoscopic attributes are only visible in dermoscopic images. Differently to clinical images, which can be captured with standard cameras, dermoscopic images are captured with a device called dermatoscope, that normalize the light influence on the lesion, allowing to capture deeper details (see Figure 3.2).

This special image enables the application of medical algorithms to skin lesions. Medical algorithms are used to evaluate a score that can support the diagnostic. There are algorithms based on more straightforward characteristics of the lesion. The ABCD rule [71], for example, takes into consideration the Asymmetry, Border, Color, and Diameter of the lesion.

Specialists diagnose melanoma with a technique called *Dermoscopy*, which analyzes the dermoscopic attributes present in the lesion. These attributes are only visible in dermoscopic images. Algorithms such as the 7-points are built upon the dermoscopic attributes. The algorithm accumulates the assigned score of each present feature, and compare the final number to a threshold, which determines if the lesion should be excised or not.

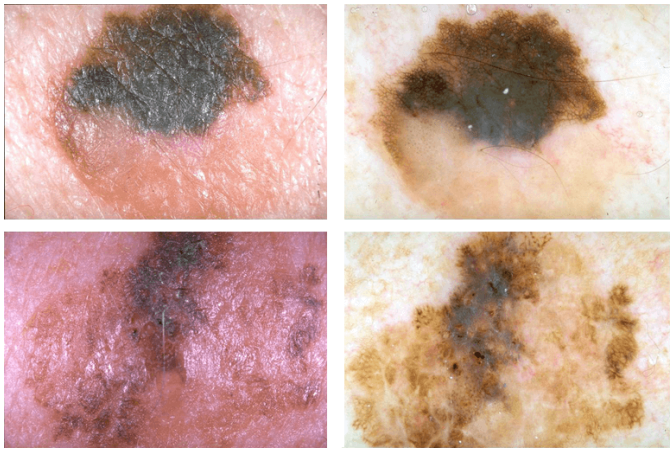


Figure 3.2: Clinical (left) versus dermoscopic (right) images. In dermoscopic images the details are enhanced, easing the process of analyzing colors and borders, and enabling dermoscopic attribute analysis. Reproduced from Argenziano et al. [6].

3.1.1 Dermoscopic Attributes

There is a big variety of dermoscopic attributes — also called *local features* — and each of them can stratify with respect to their regularity, color, and other specific details. We show the dermoscopic attributes annotation of the Atlas Dataset [6] in Table 3.1 and some examples in Figure 3.3.

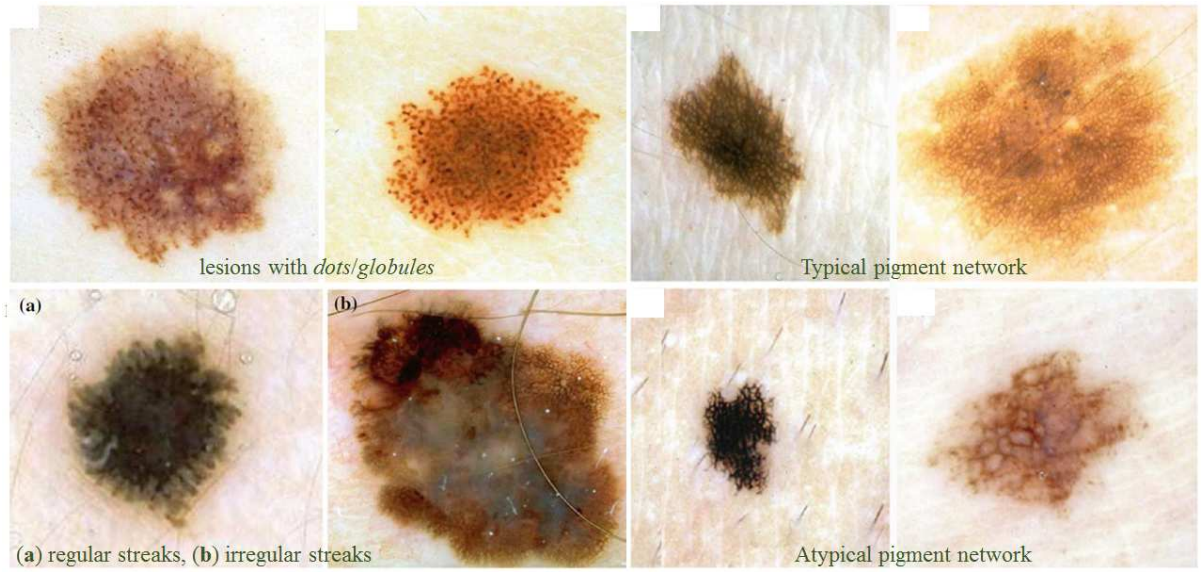


Figure 3.3: Example of dermoscopic attributes present in skin lesion images. Specialists rely on these patterns to diagnose melanoma. Image adapted from [84].

This annotation, despite crucial for human specialists, is present only for small subsets of data available for machine learning. Only recently, at ISIC 2017 and ISIC 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 not only know if an attribute

Atlas Dermoscopic Attributes Annotation	
Global feature	cobblestone, globular, homogeneous, lacunar, multicomponent, parallel, reticular, starburst, unspecific
Dots and globules	regular, irregular
Streaks	regular, irregular
Blue white veil	present
Pigmentation	diffuse irregular, diffuse regular, localized irregular, localized regular
Hypopigmentation	diffuse, focal, multifocal
Regression structures	blue areas, white areas, combination
Vascular structures	arborizing, comma, dotted, hairpin, linear, irregular, within regression, wreath
Other criteria	central white patch, comedo-like openings, exophytic papillary structures, leaf-like areas, milia-like cysts, red lacunas

Table 3.1: Atlas dataset dermoscopic attributes annotation. The annotation is very detailed, enabling the application of pattern-based medical algorithm such as 7-point checklist [5].

is present but also the portion of the lesion that displays it. Despite rich, this metadata is not sufficient for the application of the 7-points [5] medical algorithm because it does not contain other critical local features, and it lacks detail about each attribute. For example, to apply the 7-points algorithm, we need to know if the pigment network is regular or irregular. If it is not sufficient to apply medical algorithms, maybe deep learning models could also benefit from a more detailed annotation. However, we do not expect to see this improvement soon, since the ISIC 2019 Challenge did not include a similar task. Thus, the number of lesions annotated with their dermoscopic features did not increase.

Atlas, which is the only other source of skin lesion images that contains this annotation, was created as an educational source for dermatologists, enabling them to diagnose melanoma using dermoscopy. However, because of the original purpose of Atlas, it is biased with respect to the presented dermoscopic features. For educational purposes, the dermoscopic features are very well-defined, or the data include several lesions that are rare exceptions to the general rule. Differently from ISIC annotation, Atlas contains enough details of each dermoscopic attribute to apply the 7-points. Also, the annotation is binary (present or absent), not showing the lesion regions that display the patterns.

The dominant presence of a local feature and specific combinations or arrangement of them form global features. There are different types of global features: Globular Pattern contains numerous, diverse-sized globules; The Reticular Pattern presents pigment network covering most parts of the lesion; Starbust Pattern presents streaks arranged in a radial form; Homogeneous Pattern presents brown/gray/blue/reddish-black pigmentation with absence of local patterns; Unspecific patterns cannot be categorized into any global pattern.

Multiple global patterns can be present in a single lesion. Lesions that display more

than three global features are highly suggestive of melanoma. Although some patterns are characteristic of a specific diagnostic, there is almost always a particular case which compounds the difficulty of diagnosing skin cancer.

The characteristic of having multiple patterns (local or global) that can be combined within each other and be identified through visual inspection, is what makes skin lesion analysis a good candidate of a problem to be solved with deep learning. However, deep learning solutions benefit from voluminous data with rich metadata, and in this particular matter, we have a long way to thrill. In this Master thesis, we present a method for mitigating this requirement, pushing the results further.

3.2 PGAN Conditional Update

In this work, we want to augment a classification network's training dataset with labeled synthetic data. For this, we need to perform conditional generation of synthetic images. PGAN's [50] original implementation insert class information in the GAN pipeline by following ACGAN's procedure [73], where the discriminator accumulates the task to classify the images it receives. This way, the class information is incorporated through the loss function only.

However, when we performed preliminary tests with CIFAR-10 [53] to verify the network's conditional generation capabilities, the generator failed to learn the class representations even in this simpler dataset. In our literature review (Section 2.4) we detailed different methods to control the class of the synthetic images. In this case, we keep the ACGAN's loss and concatenate the class information in every layer of the generator and discriminator (except for the last one), similarly to TripleGAN [22].

Thus, we incorporate the class information in the generator after every pixel normalization that is applied after a convolution. In the discriminator, we apply it after every average pooling. For both networks, we concatenate it to the activation map channel-wise. Since the label is a single integer, and the network's activation maps are tensors, we reshape the label to match the activation map dimensions, repeating the label over all the matrix positions. This modification enabled conditional generation with PGAN on CIFAR-10, and also for skin lesion datasets.

3.3 Proposed Approach

We aim to generate high-resolution synthetic images of skin lesions with fine-grained detail. To explicitly teach the network the malignancy markers while incorporating the specificities of a lesion border, we feed this information directly to the network as input. Instead of generating the image from noise (usual procedure with GANs), we synthesize from a semantic label map (an image where each pixel value represents the object class) and an instance map (an image where the pixels combine information from its object class and its instance). Therefore, our problem of image synthesis specified to image-to-image translation. We detail our idea in the following sections.

3.3.1 GAN Architecture: The pix2pixHD Baseline

We employ Wang's et al. [96] pix2pixHD GAN, which improve the pix2pix network [46] (a conditional image-to-image translation GAN) by using a coarse-to-fine generator, a multi-scale discriminator architecture, and a robust adversarial learning objective function (see Section 2.7). The proposed enhancements allowed the network to work with high-resolution samples.

For generating 1024×512 resolution images, we only take advantage of the global generator from pix2pixHD. This generator's output resolution fits with the minimum common size of our dataset images. It is composed of a set of convolutional layers, followed by a set of residual blocks [38] and a set of deconvolutional layers.

To handle global and finer details, we employ three discriminators as Wang et al. [96]. Each of the three discriminators receives the same input in different resolutions. This way, for the second and third discriminator, the synthetic and real images are downsampled by 2 and 4 times, respectively. Figure 3.4 summarizes the architecture of the GAN network.

The loss function incorporates the feature matching loss [85] (Section 2.6) to stabilize the training. It compares features of real and synthetic images from different layers of all discriminators. The generator learns to create samples that match these statistics of the real images at multiple scales. This way, the loss function is a combination of the conditional GAN loss and feature matching loss.

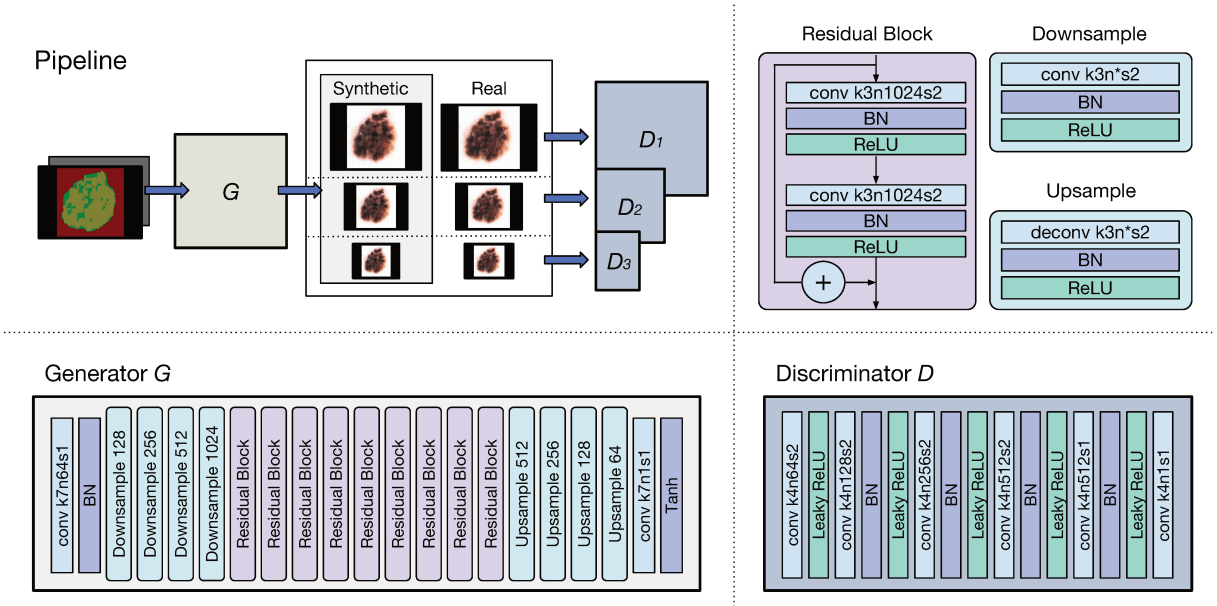


Figure 3.4: Summary of the GAN architecture. In the bottom-left, we show the pipeline. We detail both discriminator and generator, and the blocks that compose them. We show the parameters for each convolutional layer: k is the kernel size; n is the number of channels; and s is the stride. The number that follows both Downsample and Upsample blocks are the numbers of channels. Reproduced from Bissoto et al. [15].

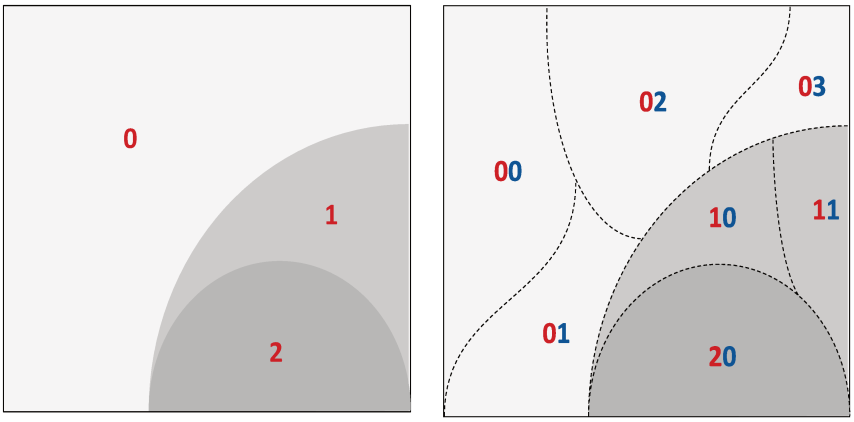
3.3.2 Modeling Skin Lesion Knowledge

Modeling meaningful skin lesion knowledge is the crucial condition for synthesizing high-quality and high-resolution skin lesions images. In the following, we show how we model the skin lesion scenario into semantic and instance maps for image-to-image translation.

Semantic map [57] is an image where every pixel has the value of its object class and is commonly seen as a result of pixel-wise segmentation tasks.

To compose our semantic map, we propose to use masks that show the presence of five malignancy markers and the same lesions' segmentation masks. The skin without lesion, the lesion without markers, and each malignancy marker are assigned a different label (Figure 3.5a). To keep the aspect ratio of the lesions, while keeping the size of the input constant as the same of the original implementation by Wang et al. [96], we assign another label to the borders, which do not constitute the skin image.

Instance map [57] is an image where the pixels combine information from its object class and its instance (Figure 3.5b). Every instance of the same class receives a different pixel value. When dealing with cars, people, and trees, this information is straightforward, but to structures within skin lesions, it is subjective.



(a) Semantic Map (b) Instance Map

Figure 3.5: 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.

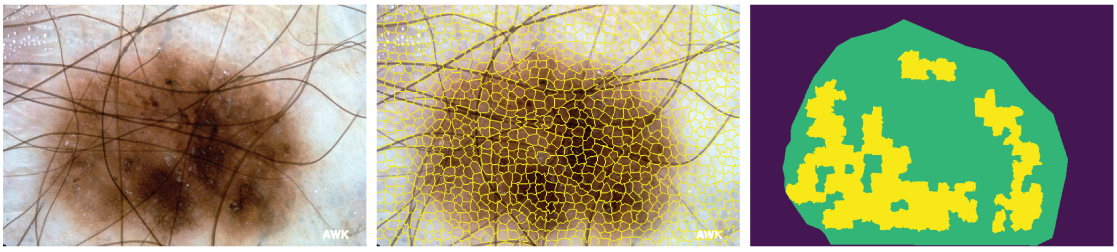
To compose our instance maps, we take advantage of superpixels [3]. *Superpixels* capture redundancy in the image, creating visually meaningful instances by clustering similar pixels. Naturally, the algorithm splits objects according to their boundaries into different superpixels.

In our case, we want to feed the generator with information on individual instances of each of our classes. However, this is hard due to the subjectivity and differences in the shape and size of different dermoscopic attributes. For example, it is easier to think about splitting different units of the globules pattern, since they are often contained in small circled-shaped structures; however, it is harder to define or contain individual instances of pigmented networks, since the whole structure is usually connected. An alternative found during the annotation process of Task 2 in the ISIC 2017 Challenge is to use superpixels

to group similar regions of the image.

If we are using a specialist's time to annotate medical images, we want to make sure this process has the perfect balance between efficiency (more images) and quality (more precise annotations). If we transform the pixel-annotation task into superpixel annotation, we are sacrificing a bit of detail to be able to have more annotated images.

Since SLIC (Simple Linear Iterative Clustering) superpixels [3] are used during the annotation process to create instances to be individually annotated by specialists, we thought they were the perfect candidate to help the generator to make sense of these highly specific attributes. In Figure 3.6 we show a lesion's semantic map, and its superpixels representing its instance map.



(a) Real image (b) Superpixels (c) Semantic label map

Figure 3.6: A lesion's semantic map, and its superpixels representing its instance map. Note how superpixels change its shape next to hairs and capture information of the lesion borders, and interiors.

We summarize our pipeline in Figure 3.7.

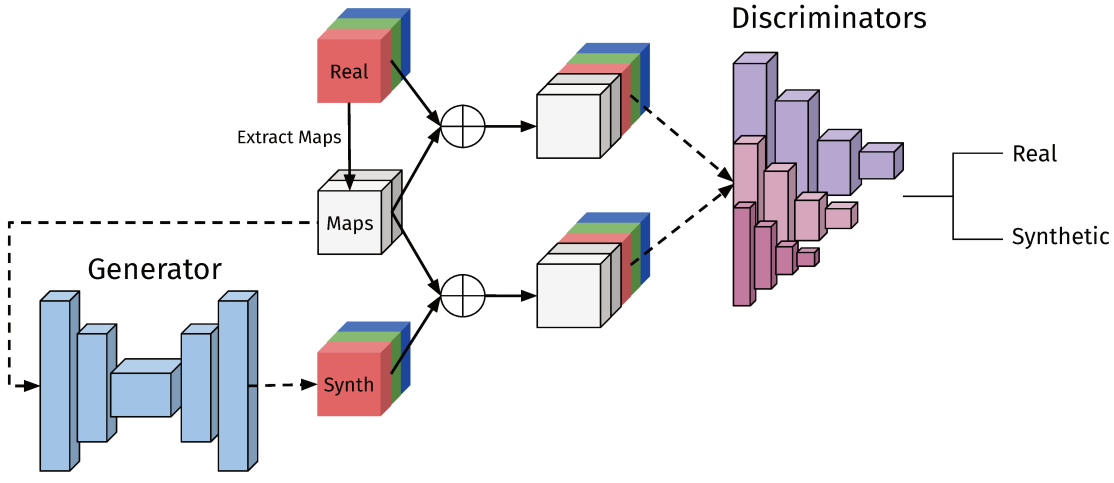


Figure 3.7: Our Pipeline. We feed the generator with maps extracted from real images, resulting the synthetic images. The discriminator is fed with batches combining real images and its maps, or synthetic images and the maps used to generate them. The output of the discriminators (there are three, each operating in a different resolution) is finally backpropagated to train the whole pipeline.

Next, we conduct experiments to analyze our synthetic images and compare the different approaches introduced to generate them.

3.4 Experiments

In this section, we evaluate GAN-based approaches for generating synthetic skin lesion images: 1) DCGAN [80], 2) our conditional version of PGAN [50], and 3) our versions of pix2pixHD [96] using only semantic map, and 4) using semantic and instance maps. We choose DCGAN to represent low-resolution GANs because of its traditional architecture. Results for other low-resolution GANs do not show much of an improvement.

3.4.1 Datasets

For the following experiments in this chapter, we use different gold-standard datasets. Although they contain different classes, we always consider a melanoma versus benign (others, except for basal cell carcinoma) scenario for classification.

- ISIC 2018 Challenge[1] – Task 2 (Lesion Attribute Detection). This dataset is composed of 2,594 images from the ISIC Archive (nevus, melanoma, and seborrheic keratosis), and all contain masks for each of five dermoscopic attributes (pigment network, negative network, streaks, milia-like cysts, and globules). This annotation with map representation is unique to this set.

- ISIC 2018 Challenge – Task 1 (Lesion Boundary Segmentation). This dataset is composed of the same images from the same year's ISIC challenge, containing each lesion segmentation mask.

- ISIC 2017 Challenge with 2,000 dermoscopic images [23] (nevus, seborrheic keratosis, and melanoma).

- ISIC Archive with 13,000 dermoscopic images (nevus, seborrheic keratosis, melanoma, actinic keratosis, basal cell carcinoma, squamous cell carcinoma, dermatofibroma, vascular lesion.) This dataset is also the most general, being collected by different institutions around the world, with different devices.

- Dermofit Image Library [9] with 1,300 images from different classes (nevus, seborrheic keratosis, melanoma, actinic keratosis, basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, pyogenic granuloma, haemangioma, dermatofibroma.)

- PH2 dataset [63] with 200 dermoscopic image (nevus, melanoma).

- Interactive Atlas of Dermoscopy [6]. The Atlas is an educational source of skin lesion images (nevus, seborrheic keratosis, melanoma, basal cell carcinoma, dermatofibroma, vascular lesion, lentigo). It contains clinical and dermoscopic versions of the same lesions (but we use only dermoscopic in this Master thesis.)

To train and evaluate pix2pixHD, we need specific masks that show the presence or absence of clinically-meaningful skin lesion patterns. For this reason, we use lesion images

[1] `https://challenge2018.isic-archive.com`

and masks from the training dataset of task 2 (2,594 images) of 2018 ISIC Challenge. The lesions' segmentation masks, that are used to compose both semantic and instance maps, were obtained from task 1 of ISIC 2018 Challenge. We split the data into train (2,346 images) and test (248 images). We use the "test" set to generate images using masks the network has never seen before.

For training DCGAN and our version of PGAN, we use the ISIC 2017 Challenge with 2,000 dermoscopic images [23], ISIC Archive, Dermofit Image Library [9], and PH2 dataset [63].

For training the classification network, we only use the 'train' set (2,346 images). For testing in a cross-dataset scenario, we use all 900 dermoscopic images from the Interactive Atlas of Dermoscopy [6].

3.4.2 Experimental Setup

For pix2pixHD, DCGAN (official PyTorch implementation) and PGAN (except for the modifications listed below), we keep the default parameters of each implementation.

We modified PGAN by concatenating the label (benign or melanoma) in every layer except the last on both discriminator and generator (Section 3.2). For training, we start with 4×4 resolution, always fading-in to the next resolution after 60 epochs, from which 30 epochs are used for stabilization. To generate images of resolution 256×256, we trained for 330 epochs. We ran all the experiments using the original Theano version.

For skin lesion classification, we employ the network (Inception-v4 [90]) ranked first place for melanoma classification [65] (our research group) at the ISIC 2017 Challenge. As Menegola et al. [65], we apply random vertical and horizontal flips, random rotations, and color variations as data augmentation. Also, we keep test augmentation with 50 replicas but skip the meta-learning Support Vector Machine (SVM).

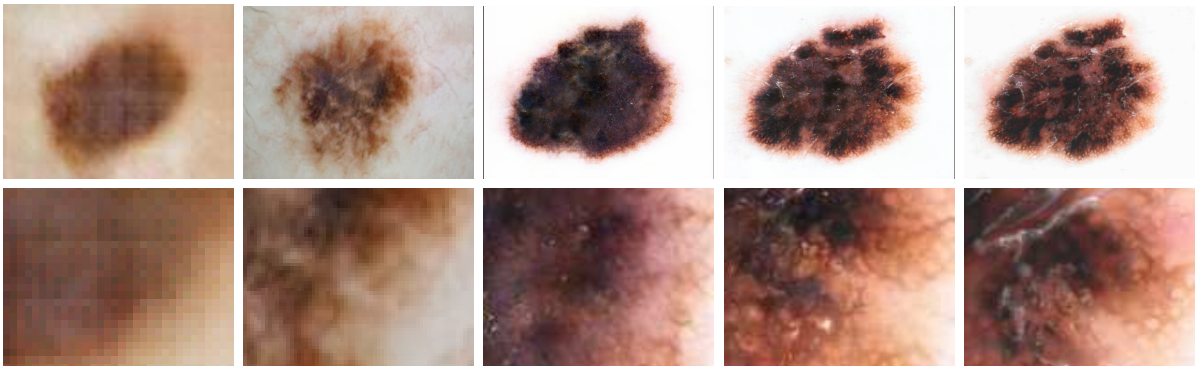
3.4.3 Qualitative Evaluation

In Figure 3.8 we visually compare the samples generated by GAN-based approaches.

DCGAN (Figure 3.8a) is one of the most employed GAN architectures. We show that samples generated by DCGAN are far from the quality observed on our models. It lacks fine-grained detail, being inappropriate for generating high-resolution samples.

Despite the visual result for PGAN (Figure 3.8b) is better than any other work we know of (at June/2018), it lacks cohesion, positioning malignancy markers without proper criteria. We cannot pixelwise compare the PGAN result with the real image. This synthetic image was generated from noise and had no connection with the sampled real image, except it was part of the GAN's training set. But, we can compare the sharpness, the presence of malignancy markers, and their fine-grained details.

When we feed the network with semantic label maps (Figure 3.8c) that inform how to arrange the malignancy markers, the result improves remarkably. When combining both semantic and instance maps (Figure 3.8d), we simplify the learning process, achieving the overall best visual result. The network learns patterns of the skin, and of the lesion itself.



(a) DCGAN (b) PGAN (Ours) (c) Pix2PixHD Semantic (Ours) (d) Pix2PixHD Instance (Ours) (e) Real

Figure 3.8: Results for different GAN-based approaches: (a) DCGAN [80], (b) Our version of PGAN, (c) Our version of pix2pixHD using only semantic map, (d) Our version of pix2pixHD using both semantic and instance map, (e) Real image. In the first row, we present the full image while in the second we zoom-in to focus on the details.

3.4.4 Quantitative Evaluation

To evaluate the complete set of synthetic images, we train a skin classification network with different combinations of synthetic and real data (including only reals, only synthetics, and combinations of both) to compose our training dataset. We compare the achieved area under the ROC curve (AUC), testing always with only real images. It is interesting to note that the procedure of training with synthetic only data and testing with reals, was later formalized by Shmelkov et al. [87] and called "GANtrain". We use three different synthetic images for this comparison: **Instance** are the samples generated using both semantic and instance maps with our version of pix2pixHD [96]; **Semantic** are the samples generated using only semantic label maps; **PGAN** are the samples generated using our conditional version of PGAN [50]. For statistical significance, we run each experiment 10 times.

For every individual set, we use 2,346 images, which is the size of our training set (containing semantic and instance maps) for pix2pixHD. For PGAN, we can generate unlimited amounts of data, but we keep it the same maintaining the ratio between benign and malignant lesions. Our results are in Table 3.2. To verify statistical significance (comparing 'Real + Instance + PGAN' with other results), we include the p-value of a paired samples t-test. With a confidence of 95%, all differences were significant (p-value < 0.05).

The synthetic samples generated using instance maps are the best among the synthetics. The AUC follows the visual quality perceived.

The results for synthetic images confirm they contain features that characterize a lesion as malignant or benign. Even more, the results suggest the synthetic images contain features that are beyond the boundaries of the real images, which improves the classification network by an average of 1.3 percentage point and keeps the network more stable.

To investigate the influence of the instance images over the achieved AUC for 'Real + Instance + PGAN', we replace the instance images with new PGAN samples ('Real

Training Data	AUC (%)	Training Data 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 3.2: Performance comparison of real and synthetic training sets for a skin cancer classification network. We train the network 10 times 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.

+ 2×PGAN'). Although both training sets have the same size, the result did not show improvements over its smaller version 'Real + PGAN'. Hence, the improvement over the AUC achieved suggests it is related to the variations the 'Instance' images carry, and not (only) by the size of the train dataset.

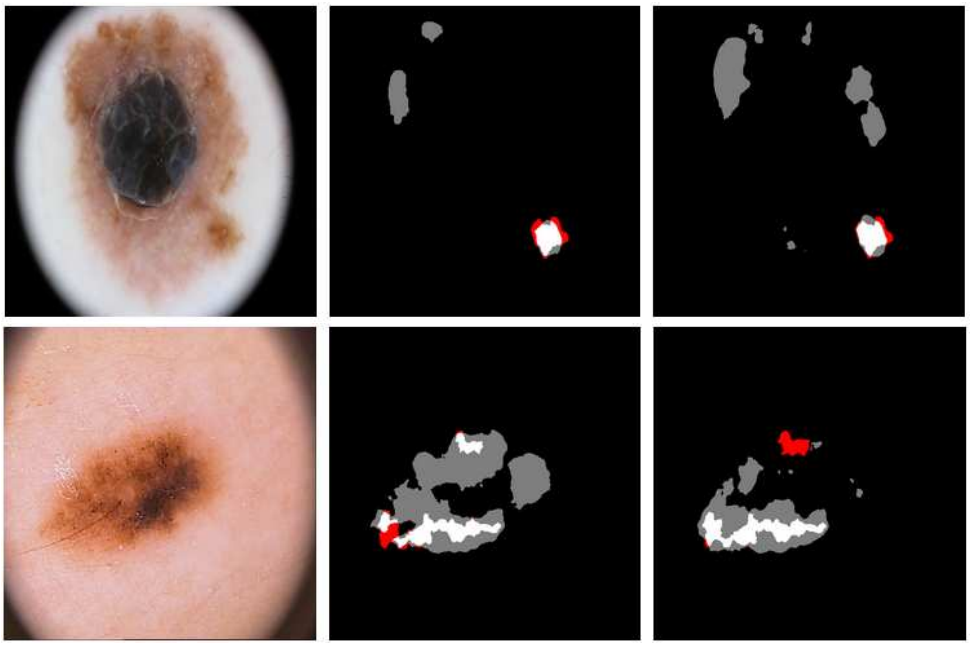
3.4.5 Synthetic Images Evaluation

One advantage of our method is that it explicitly learns about dermoscopic attributes, which are biological structures visible in dermoscopic images that are core to many diagnose methods [5, 66]. To assess the presence of these features in our synthetic samples, we employed winning solutions of Task 2 of ISIC 2018 Challenge for lesion attribute detection. Since our synthetic samples are generated based on maps extracted from real images, we use the same map used to generate them as ground-truth for the same image in the semantic segmentation task.

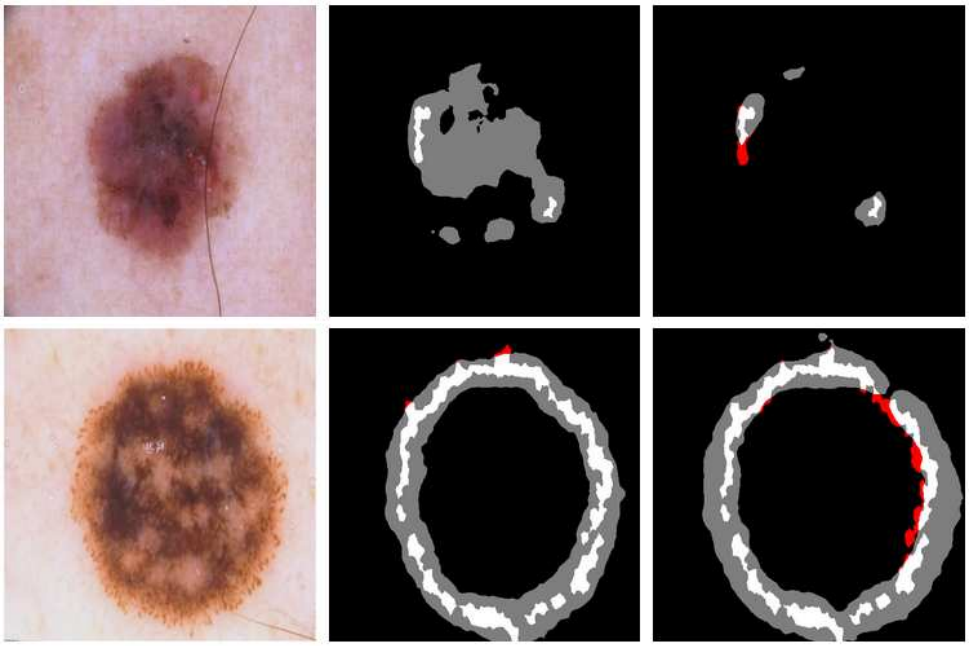
Thus, we employed a Deeplabv3+ network [54], which is the state-of-the-art for skin lesion segmentation to compare the output synthetic images, and their real counterparts annotation. We show the results in Table 3.3.

Attribute	Real	Synthetic
Pigment Network	0.469	0.424
Negative Network	0.205	0.001
Streaks	0	0
Milia-like Cysts	0.178	0.071
Globules	0.201	0.188

Table 3.3: Jaccard between Deeplab's output mask when analyzing real and synthetic samples, using the same ground-truth.



(a) Evaluation of Pigment Network in real (middle column) and synthetic images (last column).



(b) Evaluation of Globules in real (middle column) and synthetic images (last column).

Figure 3.9: Comparison between Deeplabv3+'s [54] semantic segmentation network of real and synthetic images with respect to dermoscopic attributes. The image read as follows. White: true positive regions; Gray: false positive regions; Red: false negative regions; Black: true negative regions. We show the real image in the first column, and Deeplabv3+'s evaluation on the next two — real image on the second column, and synthetic image on the last. We show that synthetic images can display correctly placed, distinctive dermoscopic attributes for the two classes (pigment network and globules).

The dermoscopic attribute segmentation proved to be a very hard task: In the ISIC 2018 Challenge, the best result for the "Lesion Attribute Detection" task achieved a 0.302 Jaccard mean over the 5 attributes. For comparison, the best result for lesion segmentation in the same challenge ("Lesion Boundary Segmentation" task) achieved 0.802 Jaccard.

Analyzing the results, it is clear that the generation process worked a lot better for pigment network and globules than the other attributes (mode collapse). This is no surprise, since those attributes are the most present in this very unbalanced dataset, while streaks and negative network are present in less than 5% of the images. This not only makes it difficult for the GAN to generate these attributes, but also for the semantic segmentation network to learn these patterns.

3.4.6 ISIC 2018 Challenge Participation

In our ISIC 2018 Challenge participation, we included the synthetic images described in this master thesis (from both PGAN and Pix2pixHD) to our training datasets. The experiments contained much more data than our initial experiments, and also was used in a context to solve a harder task since we are exploring multiple sources of skin lesion images. For a complete review of our participation, please refer to our report [14]. In this section, we focus on the classification task, which featured a highly unbalanced dataset with seven classes (melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesion).

For this challenge, we explored three different factors to design our models: three architectures (Inception-v4 [90], ResNet-152 [38], and DenseNet-161 [41]), three data sources (full, challenge only, and synthetic), and five splits, resulting in 45 models.

Our data sources have a significant influence on our results since they are very different between each other. We describe each set below:

- **Full**, composed of 30,324 images from different sources: ISIC 2018 Challenge training set [25], ISIC Archive [1], Interactive Atlas of Dermoscopy [6], Dermofit Image Library [9], PH2 Dataset [63]. To attempt to alleviate the influence of the high unbalancing between classes of the dataset, we also resorted to online public sources: the web sources were Dermatology Atlas (www.atlasdermatologico.com.br), Derm101 (www.derm101.com), DermIS (www.dermis.net/dermisroot). The online sources sum 631 images, being 414 basal cell carcinoma, 26 actinic keratosis, 132 dermatofibroma, and 59 vascular lesions.

- **Challenge Only**, composed only by the 10,015 images from the ISIC 2018 Challenge Task 3 training set.

- **Synthetic**, composed of the full dataset, adding the synthetic images in a 1:1 proportion, for each class. Thus, we doubled the training set and kept the class distribution the same.

We trained the networks using SGD with momentum 0.9, batch size of 32, starting learning rate of 10^{-3}, which is then multiplied by 0.1 when the validation loss failed to improve for 10 epochs, until it reaches a minimum value of 10^{-5}. To deal with the unbalanced

nature of our data, we employed weighted loss. We normalize our data according to ImageNet's mean and standard variation, and apply the augmentation procedure described in [78] (scenario J): random crops (preserving $0.4 - 1.0$ of the original area, and 3/4-4/3 of the original aspect ratio); random vertical/horizontal flips; rotation $(0 - 90°)$; shear $(0 - 20°)$; area scaling $(0.8 - 1.2)$; random color transformations on saturation,brightness, contrast, and hue. We applied the transformations to the validation (single replica), holdout (32 replicas), and final test (128 replicas), taking the decision as the average of the replicas.

To measure the performance of all our models, we used a holdout set. The holdout set contains 10% of the full dataset images, being selected before splitting the training/validation sets. In Table 3.4, we report the mean accuracy of our results for each training dataset, on both the challenge validation and on our holdout set.

Train Dataset	Train Dataset Size	Challenge Validation	Holdout
Full	$30,064$	0.69 ± 0.07	0.59 ± 0.16
Challenge Only	$10,015$	$\mathbf{0.77 \pm 0.09}$	0.43 ± 0.09
Synthetic	$60,128$	0.70 ± 0.09	$\mathbf{0.76 \pm 0.09}$

Table 3.4: Mean accuracy achieved by our models trained with each of our datasets, tested on the challenge's validation and our holdout. The differences between the result on challenge's validation and our holdout show how different both data distributions are. Nevertheless, augmenting our training data in a 1:1 proportion with synthetic images leaded to the best result in our holdout set by far.

A problem that became visible with our experiments is the difference between the challenge validation and our holdout. Our holdout contains more images (3,000 versus 150 from validation) of different sources, being more general, closer to a real-life scenario.

However, despite the success in our holdout set, our models trained with synthetics were not the best ranked among our test submissions. The challenge's validation set distribution seems to be closer to test's, causing our ensemble containing "Challenge Only" trained models to be our best submission, ranking 9^{th} in the overall leaderboard. Our ensemble of our eight best models in our holdout set (all used synthetic data) ranked only 39^{th}, while "Full" ranked 32^{th}.

This result shows the potential of using synthetic images to augment classification network training sets, and also the importance of using good quality (gold-standard annotation, from different sources, captured with different devices) datasets to evaluate the performance. Finally, with this complementary result, we show that augmenting the training dataset with synthetic images can significantly impact the performance in a scenario (more) similar to the real world.

3.5 Feature Visualization

Visualization is a process used to explain deep neural network results, often translating its internal features to images. Neural networks are often seen as black boxes, and it is

essential to make sense of their results. For the deployment of deep learning solutions in the real world, it is often necessary to gain the trust of the users of these solutions. This is especially true for contexts that may put a human's life at risk, such as the medical (which is the focus of this Master thesis). Visualizations can also lead to a deeper understanding of the models, being useful for identifying models' flaws and biases, ultimately leading to better, more robust solutions.

Next, we briefly describe three methods for visualization and show results when they are applied on a skin lesion classification network, analyzing real and synthetic samples.

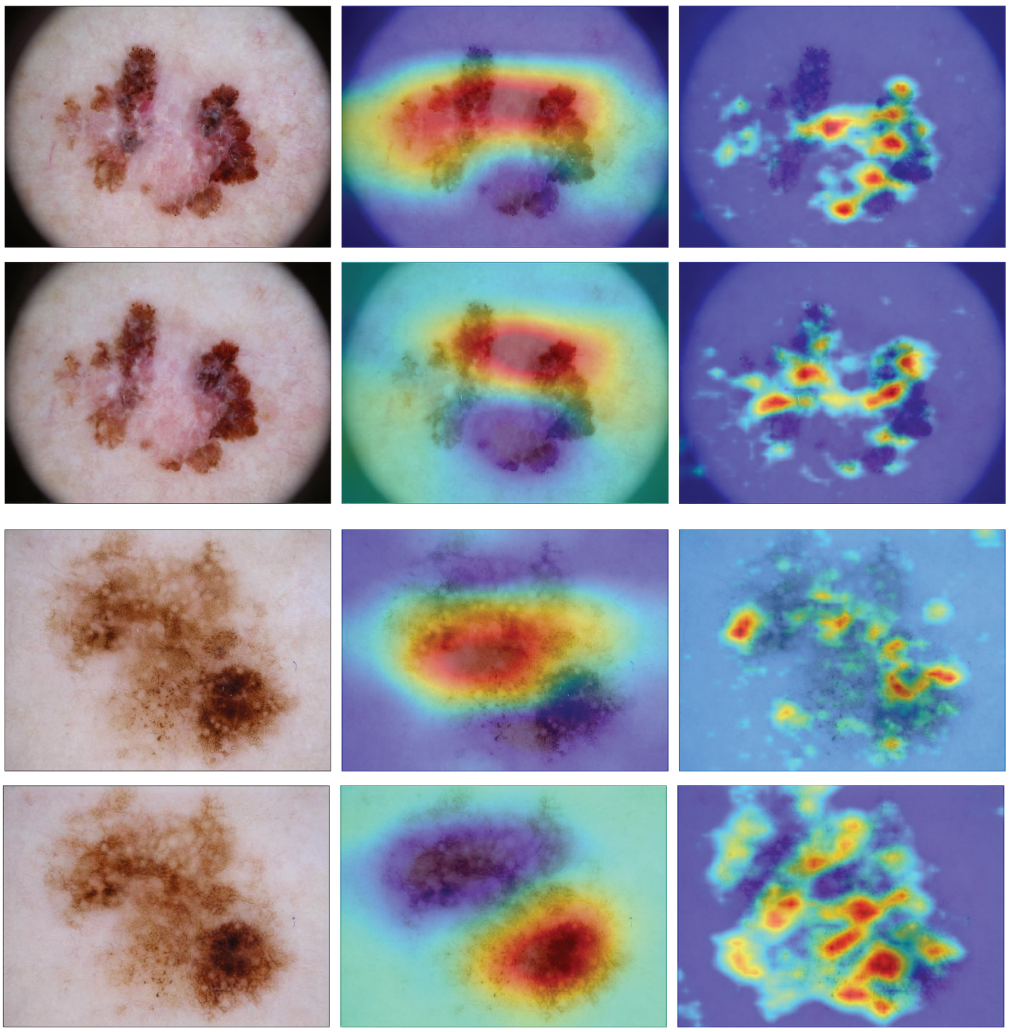
GradCAM [86]: Gradient-weighted Class Activation Mapping (GradCAM) evaluates the gradients on a target layer (with backpropagation) according to a target label. The gradients in the target layer multiply the same layer's activation map, and the result is accumulated through the channels, forming a two-dimensional saliency map that is the same size as the target layer's result activation map. Since layers of interest are usually next to the end of the network, where the concepts encapsulated are directly related to the target classes, the output is often in a lower resolution than the input of the network. Thus, the information presented is not precise, only providing general, coarser details of the network's decision. Also, since the gradients are accumulated through the channels, much information is lost in the process.

Occlusions [100]: Occlusion methods attempt to find the regions of an input image that when perturbed, affect the most the classification result. The particular method used for these preliminary experiments was introduced by Fong et al. [29]. The occluded region is learned through an iterative process that blurs the perturbed image, and the objective is to reach 99% of the score of a fully-blurred image.

Feature visualization [74, 100]: Feature visualization methods attempt to create artificial images that maximally activate a target neuron/filter in the network. By analyzing this artificial image, and the training set images that also maximally activate the target neuron, it is possible to learn concepts being encapsulated by different filters. Because of the enormous amount of filters in the network, it is almost impossible to analyze every filter qualitatively. Also, the same concept is often encapsulated in multiple filters [30], and some concepts even require multiple complimentary filters to be fully encapsulated, complicating this analysis. The comparison between methods is hard because it lacks ways to evaluate them better. Since the evaluation of the visualization images is often qualitative, it can be damaged by humans expectations of the results. Quantitative analysis is rare. It often depends on a benchmark dataset containing extra information about the scenes and concepts being learned by the network, such as BRODEN [10]. This requirement limits the reach of quantitative analysis, since most applications, including ours, do not have this kind of information available.

In the following, we show our results when using the before mentioned techniques to analyze how our classification networks perceive real and synthetic samples.

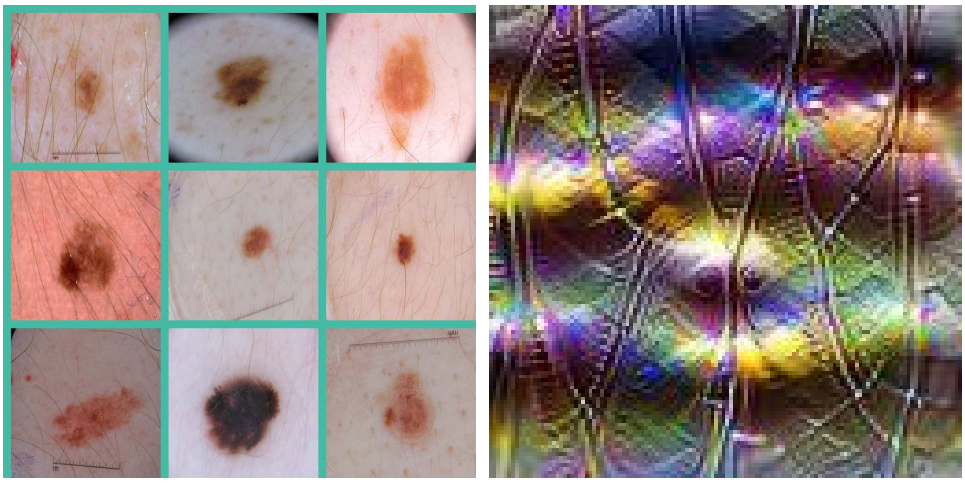
In Figure 3.10 we show how our classification model, trained only with real images, perceive real and synthetic samples. Even though they look very similar, real and synthetic images are perceived differently. This experiment sheds a light on how could the synthetic samples improved our classification results despite looking so similar to the training set.



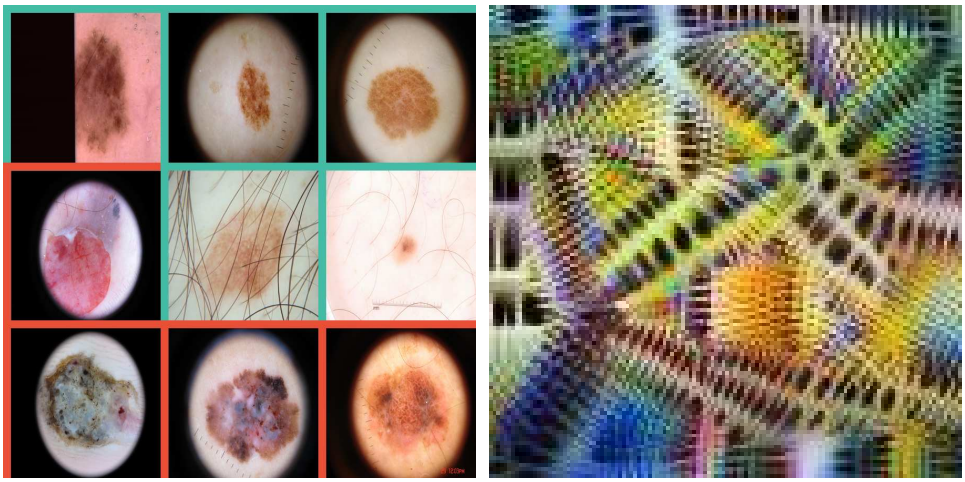
(a) Real/Synthetic Image (b) GradCAM (c) Occlusion

Figure 3.10: Saliency maps results from (b) GradCAM [86] and (c) Occlusion [29] 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. Although synthetic images are almost identical (at least to our eyes) to real ones, the network perceives differences between them, causing the saliency maps to output differently between real and synthetic. 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.

In Figure 3.11 we show that the filters that compose our classification network can encapsulate concepts for analyzing skin lesion images. Although those filters content are very similar to the ones present in ImageNet pre-trained networks, these filters can possibly identify meaningful patterns to base their decisions, once they are concerned about simple structures and objects. However, using feature visualization is very inefficient to verify this property because of the vast amounts of filters present in state-of-the-art classification models. Also, as we visualize filters closer to the top of the network, where more complex concepts are learned (and where the more interesting conclusions are situated), the visualizations' information decrease greatly.



(a) 1st neuron from layer mixed4a. It is apparently fired by thin hairs and rulers.



(b) 60th neuron from layer mixed4a. The neuron is apparently attracted to black corners and its curved lines.

Figure 3.11: Our preliminary results when visualizing features on a skin lesion classification network. On the left we show images of the training set that maximally activate the target neuron. On the right, we show an artificially created image that maximally activate the same neuron. The green canvas display benign lesions, while red canvas display malignant ones.

It is necessary to evaluate our network to help us understand its predictions. In the

medical context, it is especially critical since the decisions made can directly influence the treatment of a patient. Also, interpretability is critical to gain the trust of specialists, that can take advantage of this technology to aid difficult decisions.

The interpretation of these methods can sometimes be subjective, and we use to project our expectations to the visualizations. The differences between the saliency methods for real and synthetic images help us understand the performance improvement for classification, but also raises several questions: by augmenting our training set with synthetic images that do not exist, are we inserting biologically correct correlations? Does it matter as long as it improves the models' generalization? Are our test datasets, in their current state, sufficient to measure generalization?

The same questions are present even disregarding synthetic images. Adversarial examples [12, 92] are noise-looking inputs that when added to an image, do not alter them visually (to our eye), but are enough to change the model's prediction completely. To the network, there are no patterns that are "correct" or "incorrect" to be exploited. Recent work [44] investigated the adversarial examples and advocated that they are actually features that contain highly-predictive power (and therefore are learned by the models), but since we can not make sense of them they seem incorrect to exist. Since our models learn from these (for now) incoherent patterns, they must also be present in our visualizations' saliency maps, failing our expectations that our models are fully human-meaningful.

We believe that although we can not fully understand and make sense of the network's decision process, we need to continue investigating and applying visualization methods. By carefully designing experiments, we can verify other dangerous aspects of classification networks, such as bias (see Chapter 4), that may be prejudicial to real-world solutions.

3.6 Conclusion

In this chapter, we show GAN-based methods to generate realistic synthetic skin lesion images. We visually compare the results, showing high-resolution samples (up to 1024×512) that contain fine-grained details. Malignancy markers are present with coherent placement and sharpness, which result in visually-appealing images. We evaluate the synthetic images using visualization techniques, dermoscopic attribute segmentation, and classification networks.

Despite the qualities presented, we point out some aspects to improve. Attributes without much representatives in the dataset are harder for the GAN to learn, causing mode collapse. The clear examples are negative networks, streaks, and milia-like cysts. This deficiency was detected when evaluating the synthetic images using the semantic segmentation network.

The synthetic images also present low variation concerning their real counterparts. This limitation comes with our design choices, and the current state-of-the-art of image-to-image translation in the middle of 2018 (the period where we produced this work). First, state-of-the-art image-to-image translation models were able to create a single synthetic image given an input (segmentation mask, for example). At that time, literature was still researching a way to make good use of noise to add variation over the synthetic images.

Note that if the noise is not carefully incorporated into the generation procedure, the network can learn to ignore it. Recent works [51, 75] introduced new architectures that incorporate class information more efficiently, enabling sampled noise to influence the final generated image.

Also, despite greatly helping to improve the overall quality of the synthetic samples, the use of perceptual loss to guide the generation to match synthetic and real images (which is common even in current state-of-the-art solutions [75]) causes lack of diversity between real and synthetic images. More recent works attempted to approach this problem, suggesting the use of normalized feature loss [43].

Even if we can generate multiple outputs given a single input, these outputs are naturally similar to each other, since they respect the same input mask. To create a more diverse dataset to train our classification models, we still need more lesions to provide the semantic and instance maps. A possible solution to this problem is incorporating a semantic segmentation network to the GAN pipeline that is trained with the generator and discriminator. During training, it provides meaningful gradients generated in the dermoscopic attribute segmentation process, that guide generation towards adding clinical-relevant patterns and improving quality.

Finally, we need to investigate how our classification models are receiving synthetic information and make sure they are providing correct correlations to improve generalization. Nevertheless, our results when augmenting our training datasets with synthetic images show that this technique can significantly aid classification for small datasets.

Chapter 4

Bias in Skin Lesion Datasets

Despite our success when generating high-resolution images containing clinically-meaningful information, our synthetic samples are deterministic, with a single lesion for a given set of semantic and instance maps (see Chapter 3). This is an undesired model characteristic, since ideally, we want to increase the variance over our training examples to lead the classification network to better generalization.

Image-to-image translation models in the literature suffer from the same problem, and creating a stochastic process of translation that can generate varied samples from a single set of inputs is still an open and challenging problem.

On traditional GANs, the noise that feeds the generator enables to generate varied samples. For image-to-image translation methods, naively including sources of noises in the architecture failed to add more variation in the generation process, since the network learns to ignore it [47].

Manipulating semantic maps, which simply maps the location of the different dermoscopic attributes is easy. We could generate an infinite number of those (*e.g.*, with a GAN, or even with a simpler method, with enough manual effort). However, we can not say the same for instance maps. They are extracted from real images and depend on the image's superpixels, which group nearby pixels concerning color and the formed shapes. Simple transformations or manipulations are not enough to create a new instance map that does not resemble a previous example.

Thus, our solution to the lack of variation in our synthetic samples (with respect to the training set) was to combine different lesions' instance and semantic maps, creating new, unseen skin lesion images (see Figure 4.1). Since skin lesion combination is a phenomenon that occurs in nature, if we can generate those combinations, they can be appropriate to compose a skin lesion classification network. If we succeed in this procedure, we can create infinite variations, without using any source of noise or architectural modification by manipulating the semantic map while using other lesions' instance maps.

There are noticeable variations in the sampled distribution. Despite keeping the overall lesion structure, we can visually see differences over the attributes they contain. The synthetic lesion's overall structure (*e.g.*, border shape, and size) is defined by the original lesion used to evaluate the model's perceptual loss, while instance maps controlled fine-grained details content information, and semantic maps controlled the displayed dermoscopic attributes. Since lesions and attributes in those maps can have different sizes

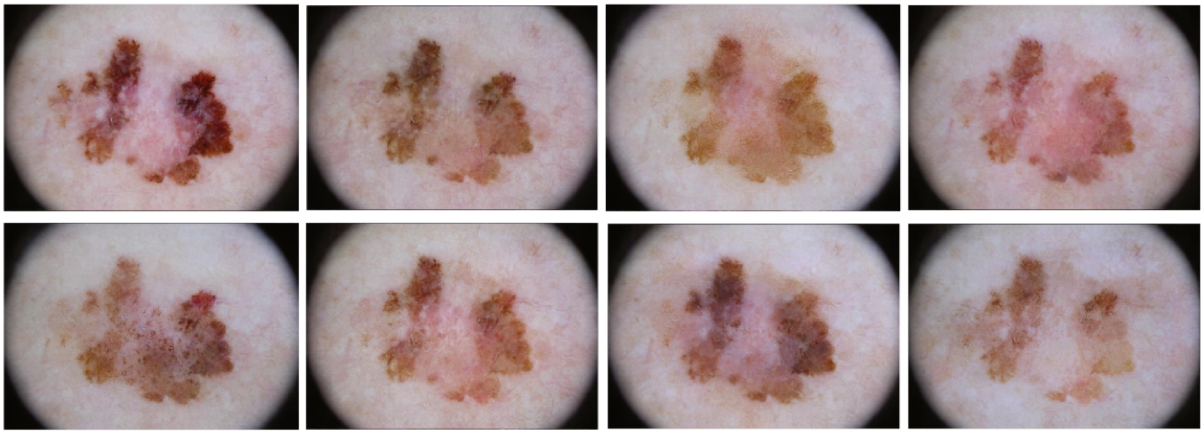


Figure 4.1: Different examples of new images created keeping the instance map the same, while using different lesions' semantic maps. By combining different lesions' maps, we create variation between samples. However we do not know the new image's diagnostic, and therefore they are meaningless for classification purposes.

and occupy different locations, the position of the dermoscopic attributes in the resultant synthetic lesion is uncertain.

Despite the apparent success, a new problem arises when inflating a classification network training dataset with those new synthetic images. What are their labels? During generation we are dealing with labels — the dermoscopic attributes — but they are not the ones used for classification: we want the diagnostic of the images (*e.g.*, nevus, melanoma, seborrheic keratosis; or malignant and benign). From a data augmentation point of view (which is the purpose of generating skin lesion images), we want the transformation learned (from map to image) to carry the original label, creating a link between the different lesions' source maps and the result combination of them. We know that the attributes within each lesion are the main information used by dermatologists for diagnostic. It is also core for medical algorithms, such as the 7-points [5], which evaluate a score according to the presence or absence of dermoscopic attributes (see Section 3.1.1).

Since dermoscopic attributes are so important for human specialists, they should be also important for the neural network. To investigate our assumption, we designed an experiment to verify the importance of these attributes for classification. Finally, if by controlling the present attributes in the synthetic images, we are also controlling the diagnostic, then we solved our problem of infinite skin lesion image synthesis.

In this chapter, we detail our experiments and results that were strongly based on our published work "(De)Constructing Skin Lesion Dataset Bias" [13], presented at the ISIC Skin Image Analysis Workshop at the Conference on Computer Vision and Pattern Recognition (CVPR 2019). In that opportunity, this work received the Best Paper Award. All our source code is readily available on `https://github.com/alceubissoto/deconstructing-bias-skin-lesion`.

4.1 Construction Experiments

Specialists can infer the diagnosis of a skin lesion analyzing its dermoscopic attributes. We want to know if our classification models can do the same. That is, can the neural network exploit the same correlation exploited by specialists to infer diagnosis with dermoscopic attributes information?

To answer this question, we design a construction experiment, where we train and evaluate a classification network with different variations of the same dataset. We call it "construction experiment" because we are increasingly feeding the network with more information, expecting it to have better performance in each step.

These variations were designed to guide the network's learning to exploit dermoscopic attributes. Despite the same motivation, within each set, we feed the network presenting dermoscopic attributes in a different way, exploring to find the best way to display this crucial information.

4.1.1 Constructing Data

To determine how important dermoscopic attributes are for automated skin lesion analysis, we perform constructive actions in the dataset, building from clinically-meaningful information (dermoscopic attributes) to guide the network's learning (see Figure 4.2).

We introduce modifications that are only possible with the dermoscopic attributes masks available on the ISIC dataset. The **ISIC Archive (ISIC)** dataset [1] is a large and generic dataset, composed of more than $13,000$ images collected from different leading clinical centers internationally, using a variety of devices for acquisition. Since the first ISIC Challenge in 2016 [62], this dataset is increasing in size and in the amount of information available for each lesion.

Segmentation masks and maps over five dermoscopic attributes (pigment network, negative network, streaks, globules, and milia-like cysts[1]) are available for smaller subsets of the dataset. This is crucial for our experiment and the main reason we choose to use it. These images and their respective maps are also used to train our GAN for skin lesion synthesis [15] (Chapter 3). We build our dataset modifications using the $2,594$ images that contain dermoscopic attributes maps.

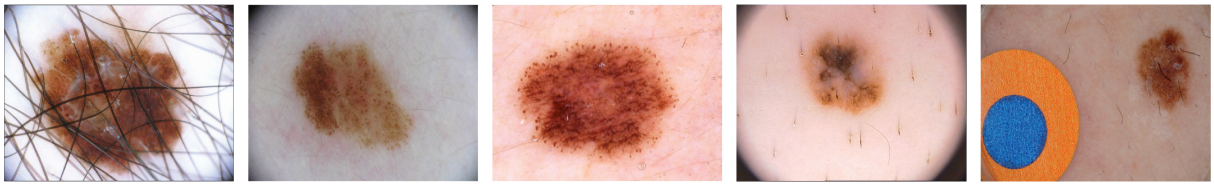
We summarize the datasets used below. For our classification experiments, we always consider a melanoma versus benign (others) scenario:

- ISIC 2018 Challenge[2] — Task 2 (Lesion Attribute Detection), composed of 2,594 images from the ISIC Archive (nevus, melanoma, and seborrheic keratosis). All images contain masks for each of five dermoscopic attributes (pigment network, negative network, streaks, milia-like cysts, and globules). This annotation, that maps the presence of each attribute to a location in the lesion, is unique to this set.

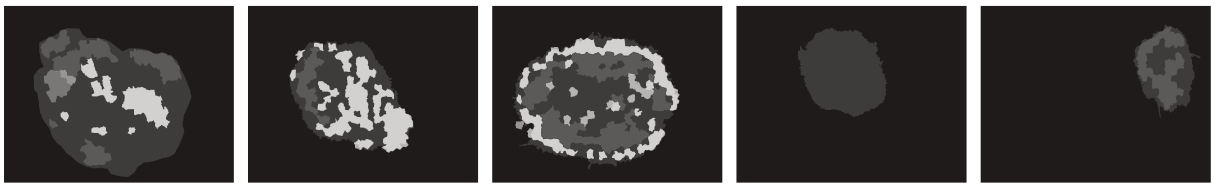
- ISIC 2018 Challenge — Task 1 (Lesion Boundary Segmentation), composed of the same images from Task 2, but containing each lesion's segmentation mask.

[1] Please refer to Section 3.1.1 for more details.

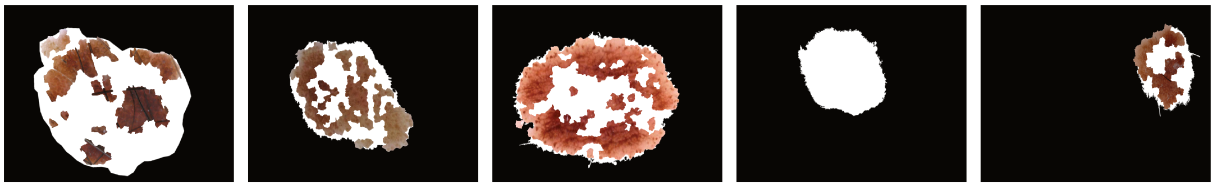
[2] `https://challenge2018.isic-archive.com`



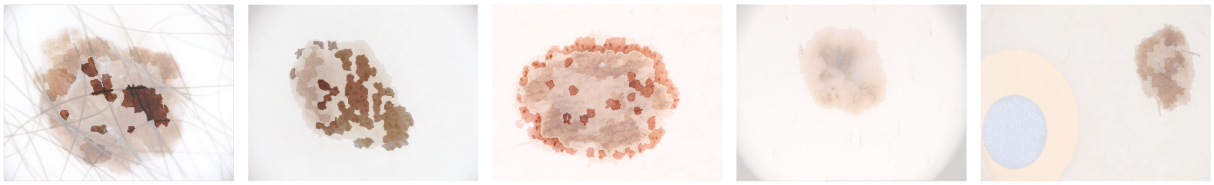
(a) Traditional images



(b) Grayscale Attributes images



(c) RGB Attributes images



(d) Traditional+Grayscale Attributes images

Figure 4.2: Samples from each of the variations created for the information construction experiment. We build from the dermoscopic attribute and segmentation information, gradually adding information until the samples contain all image's pixels plus an additional channel containing extra dermoscopic attribute and segmentation information.

Next, we detail our modified datasets and the intuition used for their design. Please refer to examples in Figure 4.2.

Traditional: This dataset contains the usual information used for training and evaluating skin lesion analysis networks. The images contain all pixels' information, and we expect it to have the highest scores in our tests, being our upper bound baseline.

Grayscale Attributes: To compose each image in this set, we use a lesion's masks from ISIC that show the location of five dermoscopic attributes and the same lesions' segmentation masks. The skin without lesion, the lesion without markers, and each dermoscopic attribute are assigned a different value, equally spaced from each other. Dermatologists look for this information to diagnose skin lesions, and it is the basis for different medical algorithms, therefore being one of the most critical parts of the image.

RGB Attributes: This dataset only shows the RGB values of the regions of the image that belongs to an annotated dermoscopic attribute, and mask the others. This way, the network does not know in principle what are the skin patterns in the image or how many of them are present, but it gains access to their RGB values. We keep the segmentation mask information from *Grayscale Attributes* in this set to display some

information for cases that do not present any skin patterns. ISIC's annotation over the dermoscopic attributes is not as detailed as Atlas'. By letting the network analyze the RGB pixels that belong to a dermoscopic attribute, we are forcing the network to focus on the attributes, to discover more details about them (*e.g.*, typical or atypical, regular or irregular, etc.), and to rely the classification on this information.

Traditional+Grayscale Attributes: Here, we aim to guide the learning process by giving to the network extra information that is very relevant to dermatologists. We concatenate a fourth channel to the *Traditional* image, containing the information described in the *Grayscale Attributes*. We need to adapt the network to receive the extra channel in the input. We add an extra convolutional layer at the beginning of the network, initialized to prioritize receiving information from the RGB channels, and progressively learn to make use of the mask provided. We expected the results to be better than *Traditional* since we are adding clinically-meaningful information to guide the network to a better understanding of the process according to human knowledge.

4.1.2 Training and Evaluation Setup

For every experiment, we use 10 splits that we keep the same throughout all sets of images (*Traditional, Grayscale Attributes, RGB Attributes, Traditional+Grayscale Attributes*) to make comparisons fair.

We use the same network architecture and hyperparameters for all experiments. We employ an Inception-v4 network [91], widely used for computer vision, and well-established for skin lesion analysis. To train each network, we use Stochastic Gradient Descent (SGD) with momentum 0.9, weight decay 0.001 and learning rate 10^{-3}, which we reduce to 10^{-4} after epoch 25. We use a batch size of 32, shuffling the data before each epoch.

We fine-tune the ImageNet [83] pre-trained network to the target dataset. We resize the input images to 299×299 to fit the input size of Inception-v4. To augment the dataset [78], we apply random horizontal and vertical flips, random resized crops that contain from 75% to 100% of the original image, random rotations between -45 and 45 degrees, and random hue changes between -20% to 20% We apply the same augmentations on both train and test. For the evaluation, we average the predictions over 50 augmented versions of each image. We normalize the input using the z-score, computed on ImageNet's training set mean and standard deviation. For all experiments, we report the Area Under the ROC Curve (AUC).

Since our datasets are relatively small, we choose not to use a validation set, using the weights after the 60^{th} epoch for test evaluation.

4.1.3 Results and Discussion

We show in Figure 4.3 our results evaluating all different sets on the ISIC dataset.

Our attempt to guide the network's learning process, verifying if the network's behavior mimics human specialists when diagnosing skin lesions revealed important insights that we discuss next.

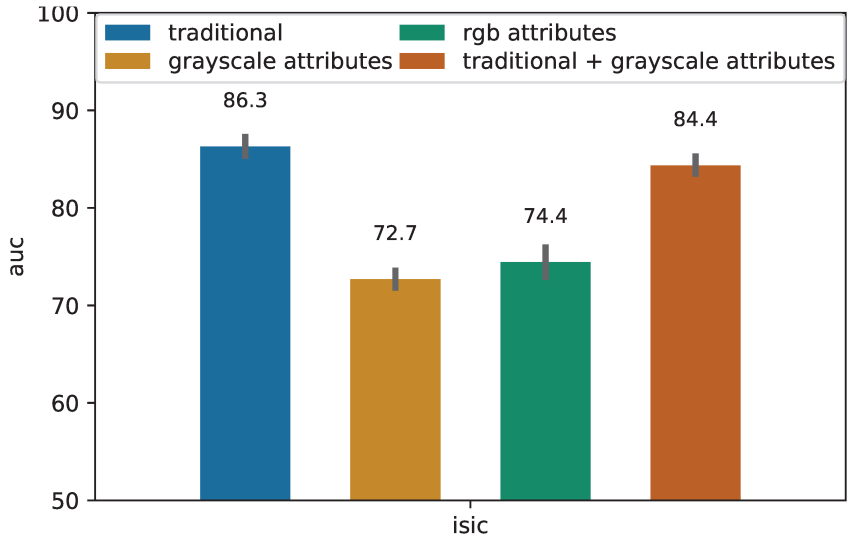


Figure 4.3: Performance comparison of the different sets of images with the ISIC dataset. Surprisingly, when we try to simplify the learning process, feeding the network with dermoscopic attributes that are clinically-meaningful, the result does not improve.

Starting from *Grayscale Attributes*, we are feeding the network with enough information to verify global patterns present in the lesion, and location of some local features (pigment network, globules, streaks, negative network, and milia-like cysts). We note that the dermoscopic attributes information is not as detailed as the ones used in medical algorithms such as the 7-points [5], and this may affect the capability of the network to make correct predictions exploring clinically-meaningful correlations.

In *RGB Attributes*, we add pixel information to the images. That enables the network to learn details about each different dermoscopic attribute, reaching the same annotation detail used in 7-points [5], and improving classification. However, we did not observe that behavior. The extra information did not help the network to improve its understanding of the problem.

In *Traditional+Grayscale Attributes*, where we are adding clinical relevant information to the usual classification procedure to guide the learning process, the result did not improve as well in comparison to the *Traditional* baseline.

The way dermatologists learn to interpret dermoscopic attributes, analyzing global and local patterns, and even the way that medical algorithms are designed [5], simply evaluating the presence of different dermoscopic attributes and assigning a score to each one, suggests that this task is suitable for a convolutional neural network. Not only that, but also that it should perform well, learning these clinically-meaningful correlations. However, our experiments showed the contrary behavior. We believe that this result shows that our classification networks are not exploiting correct, meaningful correlations, and that studying dataset bias over our skin lesion datasets are crucial for the deployment of those solutions in the real world.

In the next section, we discuss dataset bias and investigate its presence in the two most used datasets for skin lesion analysis. Then, we perform destruction experiments to investigate the information the network is exploiting when classifying skin lesion images.

4.2 The Problem of Bias in Skin Lesion Datasets

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 [6] and the ISIC Archive [1]. 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 refer to these datasets for training, evaluating or comparing its models [15, 17, 19, 95], and the ISIC Archive deserves special mention as the source of the images used in the ISIC Challenge [24, 25, 62], an annual event where different teams compare the performance of their algorithms under the controlled supervision of the organizers.

The problem of having so few, relatively small datasets dominating much of research in automated skin analysis, is the risk of datasets biases. Indeed, the (re)use of relatively small datasets by a research community poses particular risks for research on Machine Learning [76]. Dataset biases may both 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, and 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 [94]. If bias is present even in bigger and more diverse datasets [94] like ImageNet [83], 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) [68] that could inflate models performances due to spurious correlations.

Despite being impossible to eliminate wholly, it is important to understand bias and its sources to improve our image acquisition processes and deep learning models further. A useful way to measure a possible effect of a dataset bias (undue inflation of performances due to spurious correlations in the dataset), is a counterfactual experiment, which destroys the cogent information in the data, and measures how much the performance of models drops. Therefore, our destructive 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.

Finally, we contrast the results from both destruction (where we study the model's exploitation of spurious correlations) and construction experiments (where we evaluate the model's capabilities of exploiting correct clinically-meaningful correlations) to help us understand how our classification models learn with current skin lesions datasets.

4.3 Destruction Experiments

In this section, we detail our information destruction experiments. In our construction experiments, we could verify that the network is not learning from correct correlations, which is also a characterization of bias. In this section, we investigate the presence of dataset bias by gradually removing cogent information. To make the results comparable between these two experiments, the network architecture, augmentation strategies, and model selection, are the same used in the previous design.

First, we introduce the disrupted datasets used and proceed to show and discuss our results.

4.3.1 Destructing Data

To evaluate the presence and effect of dataset bias in Atlas and ISIC, we propose to:

- Perform destructive actions (see Figure 4.4) in the dataset to analyze if the network can still learn patterns to correctly classify skin lesions, even without clinically-meaningful information available.

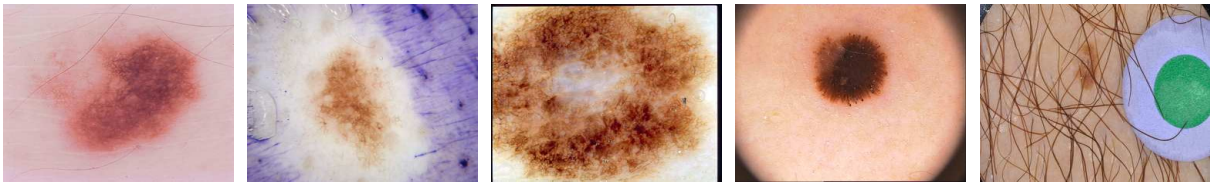
- Apply the 7-point checklist algorithm [5] to the Atlas dataset, and analyze the result comparing it with the recent melanoma classification benchmark for AI [17] to verify how biased it is due to its educational purposes and acquisition methods.

To compose our modified sets, we use two different data sources: a subset of the ISIC Archive [1], and the Interactive Atlas of Dermoscopy [6]. To keep constructive (Section 4.1) and destructive experiments comparable, we use the same ISIC dataset of our construction experiments (2,594 images from the 2018 ISIC Challenge — Task 1, which provides segmentation masks for every sample).

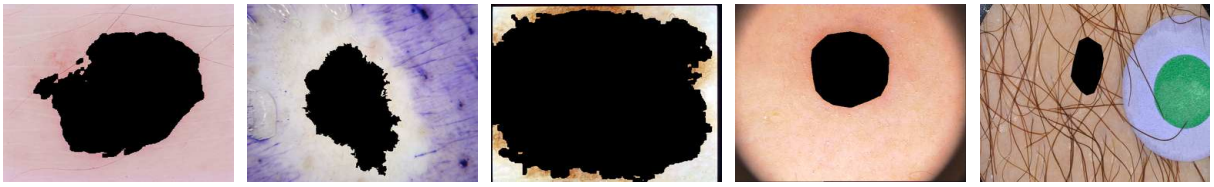
The **Atlas** [6] is a medical educational dataset composed of +1,000 cases of pigmented skin lesions. Each case is associated with clinical and dermoscopic images. Each skin lesion has clinical data (*e.g.*, location, diameter, elevation), histopathological results, diagnosis, and the presence or absence of dermoscopic attributes. The presence of those rich metadata corresponds to the pedagogical objectives of the Atlas of teaching dermoscopy through reliable and understandable medical algorithms (*e.g.*, the 7-point checklist). The Atlas also groups the lesions according to their level of diagnostic difficulty (low, medium or high), which indicates how difficult it is to identify the medical attributes (*e.g.*, networks, dots-and-globules, etc.) in the lesions. The difficulty relies on the morphological variability of a given criterion, which explains the sometimes low intra- and interobserver agreement of such medical algorithm.

We are especially interested in the dermoscopic attributes annotation. Lesions' dermoscopic attributes analysis (through pattern-based medical algorithms) is crucial for dermatologists to diagnose skin cancer. This information enables us to verify bias by comparing the medical algorithm performance (7-points), the network performance, and an Artificial Intelligence benchmark for melanoma classification [17].

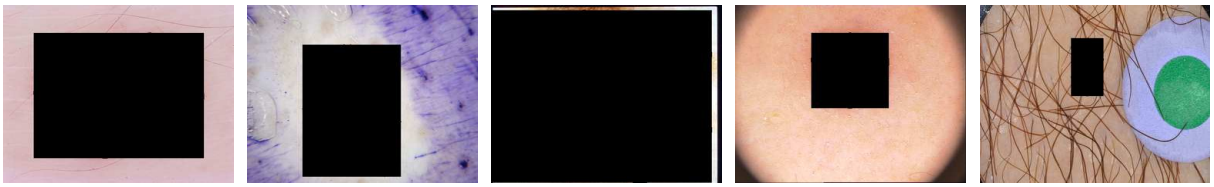
For our experiments, we select only the dermoscopic samples from the Atlas, remove "duplicates" (some medical cases have multiple images), and include only the classes



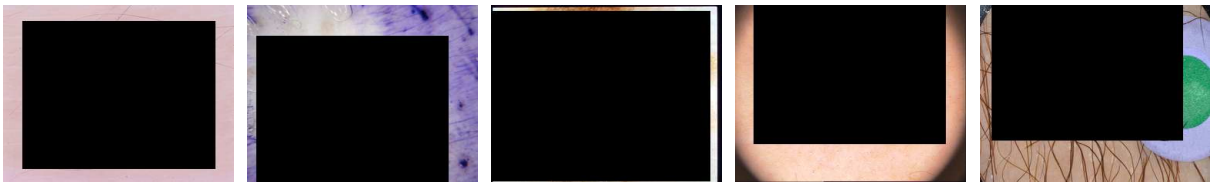
(a) Traditional images



(b) Only Skin images



(c) Bbox images



(d) Bbox70 images

Figure 4.4: Samples from each of our disrupted datasets. We gradually remove cogent information, until there is no information left to apply any aspect of medical score algorithms [5, 33]. Next, we use those sets to evaluate if the network can still learn patterns with the information left to correctly classify skin lesions. Best seen in digital format.

present in the dataset of Task 2 of ISIC 2018 Challenge (melanoma, nevus, and seborrheic keratosis). Those alterations result in a dataset containing 872 images. Unlike the ISIC dataset, the Atlas dataset does not provide the lesions' ground truth segmentation masks. To obtain them, we choose to use the SeGAN model [98], which placed 4^{th} on the segmentation task at the ISIC 2018 Challenge making use of a generative approach for skin lesion segmentation.

It is import to note that the dermoscopic attributes annotations in ISIC and Atlas differ in two ways. First, in ISIC the annotation is a mask that maps the dermoscopic attributes in the original images. In the Atlas dataset, we only have the information about the presence or absence of each dermoscopic attribute. Second, the two datasets annotated information about different dermoscopic attributes, with different levels of detail. Unfortunately, only the patterns present in the Atlas dataset allow to apply (and evaluate) the medical pattern-based algorithms.

Next, we present the different datasets modifications made for our first experiments and our motivations behind each one. In Figure 4.4 we show examples of each variation. We point out that we keep the same modifications for both training and testing our

networks.

Traditional: This dataset contains the usual information used for training and evaluating skin lesion analysis networks. The images contain all pixels' information, and we expect it to have the highest scores in our tests, being our upper bound baseline.

Only Skin: To create this dataset, we take advantage of segmentation masks. We apply the mask in the samples from the *Traditional* dataset, removing the pixels' information (they turn black) inside the actual lesion. We keep only the silhouette of the lesion and the skin of the image. Our intention when creating this dataset is to destroy the lesion information while verifying if the network could still make sense of the remained pixels to classify the samples correctly.

Bounding Box (Bbox): The lesion border is an essential feature to diagnose skin lesions. The classic ABCD medical algorithm [71] consider this feature, which accounts border symmetry and border regularity. To destroy this information from the dataset, we cover the silhouette of the lesion with a black bounding box. At this point, we already removed the lesion and its borders information. Only healthy skin and artifacts reminiscent from the image acquisition process are available for the network to learn.

Bounding Box 70**% (Bbox**70**)**: The diameter (size) of the lesion is considered by dermatologists to diagnose skin lesions since melanomas are usually bigger (start with a diameter of more than 6mm [33] than benign lesions. The diameter is the last clinical feature we attempt to remove from the network's learning possibilities. For this purpose, we define that every bounding box must at least have the size of a 250×250 square (note that images are 299×299). We keep intact bounding boxes that need to be bigger to cover the lesion. The 250×250 square is sufficient to cover 70% of the pixels. We place this square at the center of the lesion. If the lesion is not in the center of the image, part of the box is not visible. In these specific cases, the bounding boxes may cover less than 70% of the pixels. At this point, there is no information left to apply any of the factors from the ABCD [33, 71], ABCDE [2] or any pattern-based algorithm [5].

4.3.2 Training and Evaluation Setup

Since both construction and destruction experiments are useful for detecting and understanding bias on skin lesion datasets, we make them comparable. For both designs, we use the same network architecture, hyperparameters, augmentation strategies, and data splits (each adapted with its respective modification). For more details of our network setup, please refer to Section 4.1.2.

Since we do not require dermoscopic attributes maps (differently from our construction experiments), we have the opportunity to perform a cross-dataset design employing both ISIC Archive and Atlas datasets. However, since we want this destructive experiment to be comparable with the construction one, we keep ISIC with the same set of $2,594$ images used in our construction experiments.

Next, we introduce our ideas to exploit deep neural network learning capabilities.

Destructing Atlas-dataset: We employ the Atlas dataset with our disruptive actions for both training and testing the network in the destruction of information approach. We use the same 10 splits used in the construction experiments, keeping it the

same throughout all sets of images (*Traditional, Only Skin, Bounding Box, Bounding Box 70%*) to make comparisons fair. To compose each training split, we randomly select 70% of the images of each diagnostic difficulty present in the Atlas dataset (low, medium, and high). We compose the corresponding test split using the 30% that is left. Following this procedure, we reduce the possibility of biasing our results with a split that is especially good for a given set of images. Since the training and test sets come from the same data distribution (same dataset), we expect these results to be optimistic, and that motivates our three next designs.

Destructing ISIC-dataset: We also apply the destruction of information approach to the ISIC dataset. We do that to confirm the behavior verified in Atlas in a more generic dataset, with fewer effects of human bias. We apply the same 10 split generation procedure we described for this experiment, except for the diagnostic difficulty stratification (the information is not present for the ISIC dataset).

Destructing Cross-dataset: We increase the difficulty by experimenting with a cross-dataset fashion. We train with all 2,594 samples from the ISIC dataset and evaluate on the complete 872 images set from Atlas. The differences between the statistics between those two datasets make this task harder, and better reflect a real-world setting [94]. We repeat that experiment 10 times, for statistical significance.

4.3.3 Results and Discussion

We employ the melanoma classification benchmark [17] to measure the expected performance for dermatologists, in an unbiased scenario. This benchmark is the result of a study with 157 German dermatologists to be a reliable benchmark for artificial intelligence algorithms. Brinker et al.'s [17] procedure was to send an electronic questionnaire to dermatologists containing 100 dermoscopic images (80 nevi and 20 biopsy-verified melanoma) randomly chosen from the ISIC Archive, asking for their evaluation. The AUC achieved by dermatologists for dermoscopic images (which is the case for our Atlas set) is 67%.

We employ 7-point checklist [5], a score-based medical algorithm, to verify bias in the Atlas dataset. This way we can isolate the neural network's learning capabilities. Dermatologists use attribution pattern analysis to diagnose malignant cases. The 7-point medical algorithm assigns a score to each of the dermoscopic attributes. The medical practitioner needs to accumulate the scores over the detected present attributes. If this score surpasses a threshold, the lesion is assigned as a melanoma. Dermatologists use this information in addition to clinical information (if the lesion is growing, if it itches, if it bleeds, if it hurts, its location and patient's age and sex), to diagnose skin lesions. We use the 7-points checklist score available as metadata of the Atlas dataset[3]. It achieves 91.7% AUC over all selected Atlas samples (see Figure 4.5).

The huge gap between the 7-point checklist performance with the melanoma classification benchmark reveals it is biased due to the characteristics and educational objectives of the Atlas dataset. Low and medium difficulty cases selected to compose the dataset are probably hand-picked to be good examples to teach new medical practitioners to identify

[3] `http://derm.cs.sfu.ca`

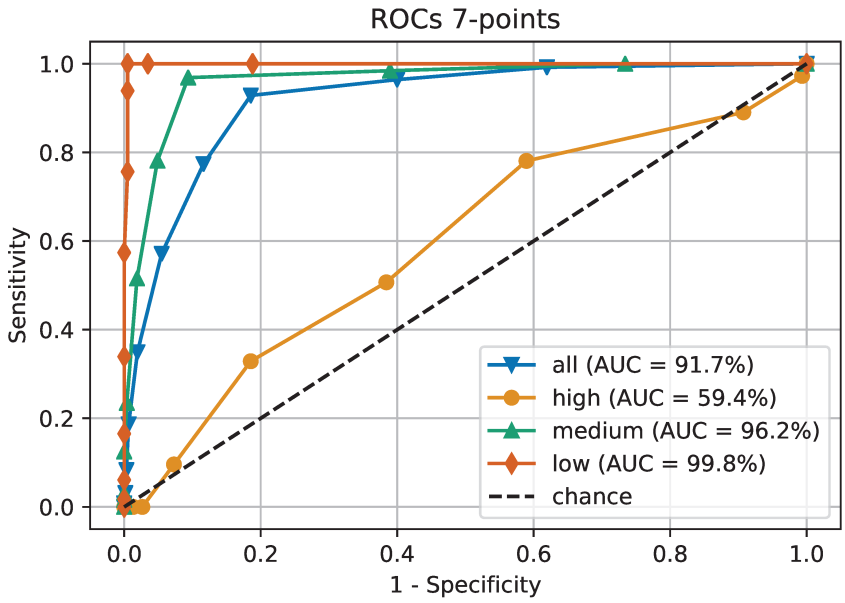


Figure 4.5: Performance of the 7-point checklist algorithm on the Atlas dataset. It shows a huge gap to the performance of dermatologists evaluated in 100 random dermoscopic samples from the ISIC Archive, which is 67% [17]. The results for 7-point checklist applied on Atlas is optimistic considering the dataset's bias towards its educational aspects.

and classify dermoscopic attributes, while hard cases are exceptions to the pattern-based analysis.

Next, we try to find the source of bias, by gradually destructing clinical-meaningful information from the images, and assessing the network's performance on them. Figures 4.6 and 4.7 show the network's performance for the different sets in the Atlas, Cross-dataset, and ISIC experiments respectively.

High difficulty lesions classification seem to be a very hard and specific task to the network, as it is for dermatologists. It could not learn clinical patterns properly with the training set, and destroying information do not influence the results. We understand that the network is probably exploiting image acquisition artifacts and dataset bias.

When experimenting in a cross-dataset fashion, the performance drops as expected, because of the differences between the statistics of Atlas and ISIC. The behavior of the network is similar in all experiments, and the following analysis can be generalized.

Traditional has the best overall performance, as expected. The network results follow the annotation of difficulty to diagnose by dermatologists. The results start to drop in *Only Skin*, where we start to deconstruct the information. When we remove the pixel information inside the lesion, we are removing all the information about dermoscopic attributes. The only clinically-meaningful information present is the border of the lesion, that could be used to verify its symmetry and irregularity, and skin features, such as vascularization.

When we remove the information of the borders, on *Bbox*, the performance lower, even more, revealing that we removed an essential feature for classification. An explanation, referring to medical algorithms like ABCD [71], is that the diameter of the box contains the information on the size of the lesion, which is also relevant information when diagnosing

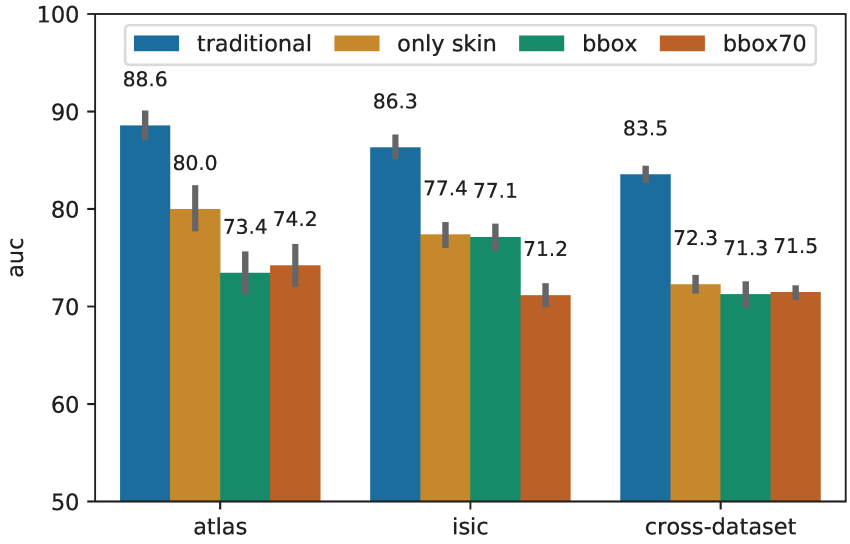


Figure 4.6: Models' performance over the disturbed datasets. We first remove all the pixel colors inside the lesion (*only skin*), proceeding to remove border information (*bbox*), and finally, removing the size (diameter) of the lesion (*bbox70*). Surprisingly, even when we destruct all clinical-meaningful information, the network finds a way to learn to classify skin lesion images much better than chance.

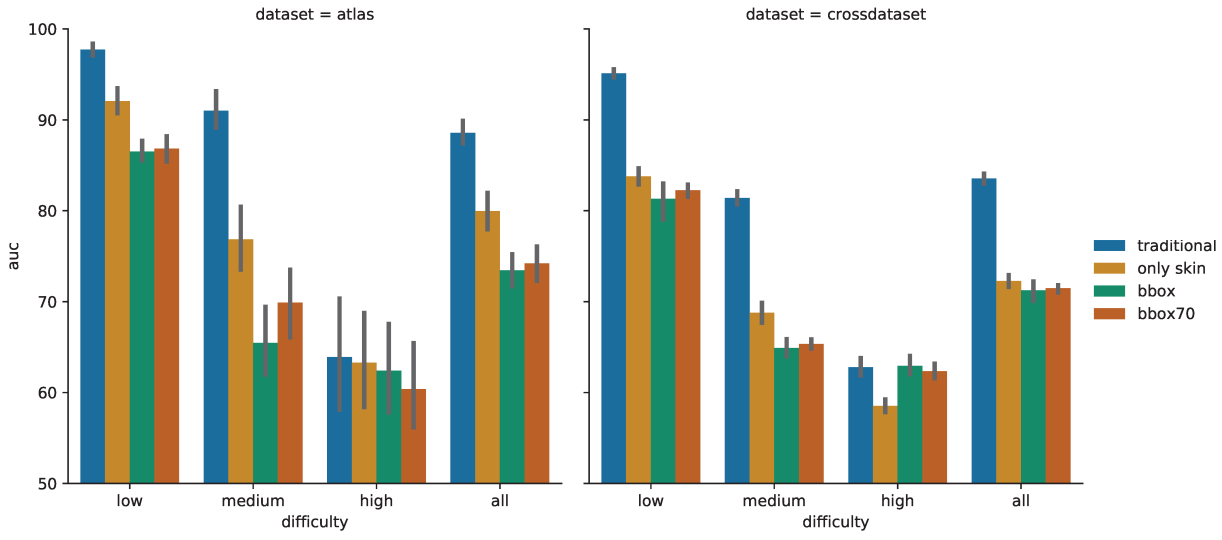


Figure 4.7: The differences over the disturbed datasets, stratifying the performance into the different diagnostic difficulties. High difficulty diagnostic present resilience to the removal of cogent information. Despite not presenting as high numbers as the other difficulties, they are still much better than chance, revealing the patterns learned are not clinical. Other difficulties are more affected by the disturbances, but the overall result for *bbox*, and even *bbox70*, shockingly surpasses melanoma classification benchmark [17] of 67%. This result suggests that dataset bias inflates our model's results.

skin lesions.

At *Bbox*70, we remove 70% of all pixels in the image and all medical relevant features that could aid the classification. Still, surprisingly, the network can make sense of visual features to make decisions that are much better than chance. There is a pattern within the available pixels that contain information that leads to the correct label. This is shocking. The numbers achieved by the network at this point even surpass the AUC achieved by dermatologists on the melanoma classification benchmark. As a sanity check, we performed an experiment hiding all image information, feeding the network (for training and testing) only zero-filled images. We achieved an AUC of 50%, which is expected since AUC is insensitive to class balance.

We believe that dataset bias is the culprit for inflating the network's performance in our destructive experiments, introducing artifacts [68] that undesirably can deviate the network's attention from more critical features. We also verify that bias is not only present in the smaller educational purpose Atlas dataset, but also the most diversified ISIC dataset. Even performing the experiments in a cross-dataset fashion (the network is trained on ISIC, and tested on Atlas), the unnatural behavior persists, attesting to the fact that these two datasets may also share the same bias. We will address the exact causes and artifacts in future works.

Another possibility is that there is meaningful information at the borders of the images (parts that were not affected by the destruction procedures). This is unlikely because according to medical algorithms [2, 5, 71], there is no information left to account.

4.4 Conclusion

If we hide the same lesion information from the networks, can it still learn patterns that help 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.

Surprisingly, the result when feeding the network with clinically-meaningful information from the dermoscopic attribute maps (*Grayscale Attributes* and *RGB Attributes* sets) is worse than feeding it only with healthy skin information (*Only Skin* and *Bounding Box* sets). That leads us to believe that also our networks' results towards both datasets is optimistic, not only the performance of 7-points over Atlas (which is expected).

That problem is critical for deploying automated skin lesion analysis. When performing in the real world, we want the network to 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.

Chapter 5

Conclusion

In this chapter, we review our findings, covering the major topics discussed in this Master thesis. Also, we discuss future directions to improve our results and the skin lesion analysis area as a whole.

5.1 Contributions

- **GAN literature review:** The GAN literature is fast-paced, with the number of works in the topic raising sharply after every year since 2014 when it was proposed. To enable the reader to have a grasp of it, we provide, in Chapter 2, a comprehensive GAN state-of-the-art, splitting it into six topics, showing how works influenced each other until we arrive at the stage we are today.

- **Skin lesion image synthesis:** We proposed, in Chapter 3, a method for skin lesion synthesis that generates high-definition clinically-meaningful synthetic skin lesions. The incorporation of dermoscopic attributes to compose the maps that are translated to the new image is the main factor that aided generation. Not only the quality of the images increased, since we are feeding the model with more information to guide the generation, but also it forced the presence of dermoscopic attributes in the final output. The presence of those attributes is crucial because they are used by dermatologists to diagnose melanoma, and can contribute to the classification models by delivering correct correlations. Although our method still is up today the current state-of-the-art for skin lesion synthesis, we have many paths to investigate to increase the quality and variability of our synthetic samples.

- **Interpretability:** We studied, in Section 3.5, and applied methods for visualization and interpretability of our classification models, analyzing the networks' behavior when feeding them with both real and synthetic data. Due to the fine-grained requirements of our problem, where the lack of detail on the visualizations can lead to a very different analysis, or by the high complexity of our skin lesion classifiers, the results are far from what we expect from a system that can be used to aid specialists when diagnosing melanoma.

- **Bias on skin lesion datasets:** Finally, in Chapter 4, we investigated the data used for both synthesis and classification models. The performance of the model even when no clinically-meaningful information is presented to it (according to medical algorithms [5, 66, 71]) is shockingly high, surpassing benchmarks that quantify a specialist's performance [17]. This is not a good sign for AI research. We think that the network learned to exploit artifacts that are introduced during the images acquisition, which caused inflated performances. Our work in this matter raised awareness in the community to this matter, and we hope it can be approached soon.

5.2 Limitations and Future Works

We believe our work contributed to the skin lesion analysis scenario with innovative techniques to deal with data. Data augmentation using synthetic skin lesions generated by a GAN has much room for improvement, pushed forward by a rapidly evolving literature. Our solution's main drawback is the lack of variety of the generated images, which look too similar to the training dataset. This happened because GANs at the time were unable to insert noise into the generation process in a way to positively affect it. Very recently, SPADE [75] managed to change this scenario by combining image-to-image translation with a noise component to increase variability. However, the performance of these models in general-purpose datasets, which are often vast and diversified, is usually better than the performance of the same models in limited, specific, unbalanced data (*e.g.*, medical context).

Another limitation is related to the dermoscopic attributes annotation, which the inclusion granted a higher level of details and a clinical meaning to our solution. This annotation is rare in skin lesion datasets, being present for a tiny subset of our data (less than 10%). Skin lesion synthesis could benefit from having more images annotated (concerning dermoscopic attributes), and also from the stratification of the current labels (*e.g.*, regular, irregular). GANs benefit from having a higher number of labels available during training [59, 73], and it would also enable classification networks to compare their performance on large datasets with medical algorithms. Applying recent GANs' strengths by experimenting with recent models, and exploring new ones dedicated to deal with limited and unbalanced data, can impact classification even more deeply.

Classification also needs to be studied so we can make sure we are moving towards a solution that can be used in the real world (generalization), instead of building a solution which is optimal only for our current datasets. The ISIC 2019 Challenge contributed for this purpose this year by including an "unknown class" to the task that was exclusively present in the closed test dataset. The new unknown class mimics a real-world scenario where the solution must be robust to work even with previously unseen labels, without assigning an incorrect diagnostic to it.

Since efforts for a complete AI solution for skin lesion analysis are increasing in the last few years, we need to make sure the datasets we use are delivering correct correlations to our models, as seen in our dataset bias investigation (Chapter 4). When evaluating our models, we need to worry about the challenges those solutions will face when being

applied in the real world and try to mitigate the data limitations we face. For this, we need to understand our models, investigating the artifacts that are being exploited. This knowledge can guide us to avoid our overfitting to too narrow data distributions (from which currently, most train and test datasets are coming from), until the point where datasets are vast and diverse.

Interpretability also is beneficial to create trust with dermatologists, which is crucial to take full advantage of the technology's potential. Interpretability would also enable AI solutions for skin cancer analysis to assist specialists in hard cases. However, current methods for visualizing our classification models are not robust enough to present good performance on complex networks and contexts [29, 48], the output saliency maps are not precise enough to be useful to understand the predictions [86], or their result is subjective to human interpretation [74]. Better visualization methods can improve our understanding of the problems we are facing, and maybe guide future modifications on our solutions.

Finally, we witnessed the difference that the incorporation of dermoscopic attributes had in our skin lesion synthesis solution, and we think that working with different meta-data can highly benefit future results. However, it is very challenging to acquire, organize, and annotate this data. For the machine learning point of view, we also need to learn how to extract the most of each information while combining different domains (text from medical records, dermoscopic and histopathologic images, and genomics), but we believe it can be the next revolution for skin cancer analysis.

Bibliography

[1] International Skin Imaging Collaboration: Melanoma Project. `https://isic-archive.com`. 48, 57, 61, 62

[2] N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, and D. Polsky. Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. *Jama*, 292(22):2771–2776, 2004. 64, 68

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels. Technical Report No. EPFL-REPORT-149300, 2010. 41, 42

[4] American Cancer Society. Survival rates for melanoma skin cancer, by stage, 2019. `www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging`. 13

[5] 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*, 134(12):1563–1570, 1998. 15, 38, 46, 56, 60, 62, 63, 64, 65, 68, 70

[6] G. Argenziano, H. P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, et al. Dermoscopy: a tutorial. *EDRA, Medical Publishing & New Media*, page 16, 2002. 13, 37, 43, 44, 48, 61, 62

[7] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017. 21, 27

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017. 27, 35

[9] L. Ballerini, R. Fisher, B. Aldridge, and J. Rees. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*. 2013. 43, 44, 48

[10] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549, 2017. 50

[11] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The cramer distance as a solution to biased wasserstein gradients. *arxiv:1705.10743*, 2017. 27

[12] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013. 53

[13] 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*, 2019. 16, 56

[14] 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:1808.08480*, 2018. 16, 48

[15] A. Bissoto, F. Perez, E. Valle, and S. Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 294–302, 2018. 14, 15, 16, 34, 40, 57, 61

[16] A. Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 34

[17] 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*, 111:30–37, 2019. 61, 62, 65, 66, 67, 70

[18] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 21

[19] 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*, pages 97–129, 2015. 61

[20] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016. 23

[21] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018. 29

[22] L. Chongxuan, T. Xu, J. Zhu, and B. Zhang. Triple generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4088–4098, 2017. 24, 25, 39

[23] N. Codella, D. Gutman, M. Celebi, B. Helba, M. Marchetti, 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). *arXiv:1710.05006*, 2017. 43, 44

[24] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. 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 *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 168–172, 2018. 14, 61

[25] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arxiv:1902.03368*, 2019. 14, 16, 48, 61

[26] E. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1486–1494, 2015. 22

[27] M. Dietz. On the intuition behind deep learning gans - towards a fundamental understanding, February 2017. `https://blog.waya.ai/introduction-to-gans-a-boxing-match-b-w-neural-nets-b4e5319cc935`. 20

[28] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 14

[29] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3429–3437, 2017. 50, 51, 71

[30] R. Fong and A. Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8730–8738, 2018. 50

[31] M. Fornaciali. Towards robust melanoma screening: A case for enhanced mid-level features. Master's thesis, University of Campinas, 2015. 13

[32] M. Fornaciali, M. Carvalho, F.. Bittencourt, S. Avila, and E. Valle. Towards automated melanoma screening: Proper computer vision & reliable results. *arXiv:1604.04024*, 2016. 14

[33] R. J. Friedman, D. S. Rigel, and A. W. Kopf. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians*, 35(3):130–151, 1985. 63, 64

[34] I. González-Díaz. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE Journal of Biomedical and Health Informatics (JBHI)*, 23(2):547–559, 2018. 14

[35] 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 (NeurIPS)*, pages 2672–2680, 2014. 15, 18, 20

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017. 26

[37] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 26

[38] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 40, 48

[39] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 34

[40] R. D. Hjelm, A. P. Jacob, T. Che, K. Cho, and Y. Bengio. Boundary-seeking generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 27

[41] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 48

[42] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 23, 25, 31

[43] X. Huang, M. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 25, 31, 32, 54

[44] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019. 53

[45] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 26

[46] 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. 15, 40

[47] 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)*, pages 1125–1134, 2017. 24, 28, 29, 55

[48] Springenberg J., Dosovitskiy A., Brox T., and Riedmiller M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2015. 71

[49] 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)*, pages 694–711. Springer, 2016. 27, 29, 30

[50] 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 (ICLR)*, 2018. 15, 22, 23, 24, 26, 34, 35, 39, 43, 45

[51] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 18, 23, 24, 25, 32, 35, 54

[52] D. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 24

[53] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 39

[54] Chen L., Zhu Y., Papandreou G., Schroff F., and Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 833–851. Springer, 2018. 46, 47

[55] A. Larsen, S. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning (ICML)*, pages 1558–1566, 2016. 24

[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 22

[57] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 30, 41

[58] Z. Lin, A. Khetan, G. Fanti, and S. Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1498–1507, 2018. 21, 27

[59] M. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz. Few-shot unsupervised image-to-image translation. *arxiv:1905.01723*, 2019. 25, 32, 70

[60] M. Lucic, K. Kurach, M.n Michalski, S. Gelly, and O. Bousquet. Are gans created equal? a large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 700–709, 2018. 21, 28

[61] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017. 27, 35

[62] M. A Marchetti, N. C. F. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology*, 78(2):270–277, 2018. 14, 57, 61

[63] T. Mendonça, P. Ferreira, J. Marques, A. Marcal, and J. Rozeira. PH2: A dermoscopic image database for research and benchmarking. In *IEEE EMBS*, 2013. 43, 44, 48

[64] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle. Knowledge transfer for melanoma screening with deep learning. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 297–300, 2017. 14

[65] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle. RECOD Titans at ISIC Challenge 2017. *arxiv:1703.04819*, 2017. 44

[66] Ingvar C. Crotty K. A. McCarthy W. H. Menzies, S. W. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. *Archives of Dermatology*, pages 1178–1182, 1996. 15, 46, 70

[67] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arxiv:1411.1784*, 2014. 24

[68] N. K. Mishra and M. E. Celebi. An overview of melanoma detection in dermoscopy images using image processing and machine learning. *arXiv preprint arXiv:1601.07843*, 2016. 61, 68

[69] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 22, 26

[70] T. Miyato and M. Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018. 22, 25

[71] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994. 15, 16, 36, 64, 66, 68, 70

[72] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4467–4477, 2017. 24

[73] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning (ICML)*, pages 2642–2651. JMLR. org, 2017. 25, 28, 39, 70

[74] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 50, 71

[75] T. Park, M. Liu, T. Wang, and J. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 25, 26, 32, 33, 35, 54, 70

[76] R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011. 61

[77] F. Perez, S. Avila, and E. Valle. Solo or ensemble? choosing a CNN architecture for melanoma classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 14

[78] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311, 2018. 14, 49, 59

[79] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011. 34

[80] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. 15, 22, 26, 35, 43, 45

[81] D. Rigel, J. Russak, and R. Friedman. The evolution of melanoma diagnosis: 25 years beyond the abcds. *CA: A Cancer Journal for Clinicians*, 60(5):301–316, 2010. 13

[82] 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)*, pages 234–241. Springer, 2015. 28

[83] 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*, 115(3):211–252, 2015. 14, 59, 61

[84] A. Sáez, C. Serrano, and B. Acha. Model-based classification methods of global patterns in dermoscopic images. *Medical Imaging, IEEE Transactions on*, 33(5):1137–1147, 2014. 37

[85] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2234–2242, 2016. 15, 21, 25, 27, 33, 40

[86] 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 (ICCV)*, pages 618–626, 2017. 50, 51, 71

[87] K. Shmelkov, C. Schmid, and K. Alahari. How good is my GAN? In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2018. 34, 45

[88] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 28

[89] J. Susskind, A. Anderson, and G. Hinton. The toronto face dataset. Technical report, University of Toronto, 2010. 22

[90] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 44, 48

[91] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 59

[92] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 53

[93] L. Theis, A. Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016. 34

[94] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011. 61, 65

[95] 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*, 2019. 14, 61

[96] 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 International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 15, 28, 35, 40, 41, 43, 45

[97] 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)*, pages 8798–8807, 2018. 29, 30, 32

[98] Y. Xue, T. Xu, and X. Huang. Adversarial learning with multi-scale loss for skin lesion segmentation. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 859–863, 2018. 63

[99] D. Yoo, N. Kim, S. Park, A. Paek, and I. Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision (ECCV)*, pages 517–532. Springer, 2016. 28

[100] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014. 50

[101] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2019. 26

[102] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 28, 29