GAN-Based Data Augmentation and Anonymization for Skin-Lesion Analysis: A Critical Review

Alceu Bissoto[1]   Eduardo Valle[2]   Sandra Avila[1]

[1]Institute of Computing (IC)   [2]School of Electrical and Computing Engineering (FEEC)

RECOD Lab., University of Campinas (UNICAMP), Brazil

Abstract

Despite the growing availability of high-quality public datasets, the lack of training samples is still one of the main challenges of deep-learning for skin lesion analysis. Generative Adversarial Networks (GANs) appear as an enticing alternative to alleviate the issue, by synthesizing samples indistinguishable from real images, with a plethora of works employing them for medical applications. Nevertheless, carefully designed experiments for skin-lesion diagnosis with GAN-based data augmentation show favorable results only on out-of-distribution test sets. For GAN-based data anonymization — where the synthetic images replace the real ones — favorable results also only appear for out-of-distribution test sets. Because of the costs and risks associated with GAN usage, those results suggest caution in their adoption for medical applications.

1. Introduction

The lack of training images is perhaps the main challenge faced by medical deep learning, deep skin-analysis being no exception. Although the availability of high-quality public datasets has mitigated the issue [4, 10, 22, 26, 33, 41], the total number of annotated skin-lesion images available to researchers is still 1–2 orders of magnitude smaller than the size of general-purpose computer vision datasets [34, 39].

More training images translate to better results, especially for deeper network architectures [27, 42]. However, annotating skin-lesion images to increase datasets process is very costly, depending on the scarce time of medical specialists. Generative Adversarial Networks (GANs) [15] appear as an alternative to increase the amount of training data without incurring those costs.

GANs aim to artificially synthesize samples that are indistinguishable from real images. They may be employed as a complement to traditional data augmentation [30], artificially increasing the amount of training samples. A plethora of existing works, which we discuss in our literature sur-

vey, suggest applying them for that purpose, but — as we will show — obtaining reliable improvements from GAN-based data augmentation is *far* from obvious: in our carefully designed experiments, GAN-based augmentation failed to *reliably* improve the classification performance of skin-lesion diagnosis, although we obtained good results for selected special cases.

Another potential application of GANs is data anonymization, where the synthetic images are used to *replace*, instead of augment, the original training set. Here again, our results recommend caution, although they were promising for out-of-distribution tests.

The main contribution of this work is a detailed study of the factors that can impact GAN-based augmentation, including GAN architectures, amount of real images used, proportion of real and synthetic images, and method for sampling synthetic images. The procedure adopted, from GAN checkpoint selection until classification network evaluation, can serve as a guideline to increase the reliability of future works using GAN-based augmentation. In addition, we carry out a systematic literature review, where we summarize the techniques used for GAN-based augmentation. In that review, we list issues in experimental design that may lead to overoptimistic results.

The text follows the usual organization, with the literature review next, proposed approach in Section 3, followed by results in Section 4. We close the paper with a discussion of the main findings, the risks of using GAN-based augmentation in medical applications, as well as cautious avenues for continuing their use in this context.

2. Literature review

The review in this section started to seed an attempt to obtain reliable performance improvements from GAN-based data augmentation. Since several existing works reported measurable gains, we explored literature to understand which factors previous authors had tested in their experiments. For more details over different GAN methods, we direct the reader to surveys in the area [37, 44].

	Ref$_{Year}$	GAN	Target Model	Domain	Dataset	Real:Synth.	Sampling	Gain/Metric†
Classification	[7]$_{2018}$	pix2pixHD, PGAN	Inception-v4	skin lesions	ISIC 2018 Tasks 1–2	1:1, 1:2	random	1 / AUC
	[12]$_{2018}$	DCGAN	custom CNN	liver CT	private (182 samples)	1:0.6	random	7.1 / acc
	[45]$_{2018}$	translation-based	ResNet-50	mammography	DDSM	1:0.111 to 1:1	fixed	0.9 / AUC
	[5]$_{2019}$	DCGAN	ResNet-50	skin lesions	ISIC 2017	1:0.26	random	4 / AUC
	[9]$_{2019}$	translation-based	VGG-16, ResNet-18, DenseNet	chest x-ray	ChestXRay14	learned	Bayesian NN	—
	[16]$_{2019}$	PGAN, MUNIT, SimGAN	ResNet-50	brain MRI	BraTS 2016	1:0.96	—	1.7 / acc
	[11]$_{2020}$	DCGAN	LeNet-5, AlexNet	esophagus	MICCAI 2016 EndoVis Challenge	1:1, 1:5	random	2 / acc
	[14]$_{2020}$	pix2pix-based	MobileNet	skin lesions	private (49920 clinical images)	1:0.4	balancing	-11.3–13.4 / F1-score
	[23]$_{2020}$	PGAN	VGG19	histopathology	TCGA, OVCARE	1:22.8	random	4 / balanced acc
	[46]$_{2020}$	—	ResNet-34	histopathology	PatchCamelyon, private (2647 patches)	RL-based	RL-based	4.6 / AUC
Segmentation	[38]$_{2018}$	pix2pix	GAN-based	brain MRI	BraTS 2015	1:1*	fixed	1 / Dice
	[3]$_{2019}$	pix2pixHD	FCN-8s	red blood cells	private (100 images)	1:1*	fixed	0.3 / Dice
	[36]$_{2019}$	CycleGAN	U-Net	abdomen CT	Kidney NIH, (Liver, Spleen) DataDecathlon	1:0.06 to 1:2.6	random	1–60 / ood Dice
	[1]$_{2020}$	translation-based	modified U-Net	cardiac MRI	ACDC, SCD, York Cardiac MRI, private test set	1:0.66	fixed	1–11 / Dice
	[24]$_{2020}$	translation-based	cascaded net, U-Net, DeepLab-v3	brain MRI	BraTS 2017	1:1	fixed	3 / Dice
	[25]$_{2020}$	translation-based	U-Net, DenseUNet	opt. coher. tomogr.	RETOUCH	0:1*	fixed	2 / Dice
	[31]$_{2020}$	DCGAN, LAPGAN	custom CNN	skin lesions	ISIC 2017	—	random	2 / Jaccard
	[19]$_{2021}$	translation-based	U-Net	covid lung CT	Radiopaedia	1:0.1 to 1:.0.5	fixed	2.5 / Dice

Table 1: The 18 works selected for our analysis of existing art. —: unclear or missing. *: missing details. †All gains in percentage points over the metric, min–max: ranges of gains found in experiments, ood:out of distribution, acc: accuracy.

However, as our experiments progressed without revealing reliable improvements, we returned to literature to subsidize a larger-scale experiment. Our review grew, in scope and formality, proportionally to our experimental ambitions. Although this review does not intend to be a formal meta-analysis, we took inspiration from Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [28] to gather a representative sample of existing art.

Our starting point was all GAN-related works published in the past ISIC Workshops. To that we added a database search, in Google Scholar, with the query "GAN generative_adversarial_networks medical_image synthetic_image data_augmentation classification OR segmentation -NLP -temporal -tabular", which gave 251 results. Notice that we did not restrict our query to skin-lesion analysis but to all medical-image applications. We excluded from the sample all works outside our scope (*i.e.*, no GAN data augmentation, no test on a medical dataset, no classification or segmentation), as well as works without experimental results. We excluded surveys and reviews from our sample as well. To get a manageable sample of papers to study, we kept our sample only works from top conferences (CVPR, NeurIPS, MICCAI), and their respective workshops, or published in journals of impact factor 3.0 or higher. For the same reason, we did not include unpublished preprints.

The resulting collection of 18 works appears in Table 1. The table details which GANs each work evaluated, which deep-learning model was the target of the data augmentation, what was the medical application domain, which datasets were used on the evaluation, and which improvements the authors reported over which metrics.

Depending on the domain, dataset, and task at hand, different families of GANs may be more appropriate, or conversely, completely unfeasible. Translation-based GANs, which include pix2pix [18], pix2pixHD [43], and SPADE [29], learn to translate between different types of images, *e.g.*, from a segmentation mask into a new synthesized input, or from a non-contrast to a contrast CT-scan. Their main advantage is that adherence to the mask tends to improve their biological/medical coherence.

In contrast, noise-based generation models, like DCGAN [32] and PGAN [20] offer flexibility, the latter being able to generate high-resolution images. Noise-based generation may suffer, however, if the training sets are too small, requiring mitigating techniques such as patch extractions and traditional data augmentation.

During the review, we took freehand notes about the experimental protocol of each paper, to understand which factors literature considered the most important to evaluate, and which varied the most among works. We found the proportion of real training images to images synthesized by the GAN to be one of the foremost factors, as well as the technique used to sample the images from the GAN.

Application	GANs	Real:Synthetic Ratios (Real:Synthetic Benign:Synthetic Malignant)	Sampling Technique
Augmentation	pix2pixHD, PGAN, SPADE, StyleGAN2	$1{:}0$, $1{:}^1\!/_4$, $1{:}^1\!/_2$, $1{:}1$, $(1{:}^1\!/_2{:}^1\!/_2)$, $(1{:}^1\!/_2{:}^3\!/_4)$, $(1{:}^1\!/_2{:}1)$	random, diverse, best, worst
Anonymization	StyleGAN2	$^1\!/_{16}{:}0$, $^1\!/_{16}{:}^1\!/_{16}$, $^1\!/_{16}{:}^{15}\!/_{16}$, $^1\!/_8{:}0$, $^1\!/_8{:}^1\!/_8$, $^1\!/_8{:}^7\!/_8$, $^1\!/_4{:}0$, $^1\!/_4{:}^1\!/_4$, $^1\!/_4{:}^3\!/_4$, $^1\!/_2{:}0$, $^1\!/_2{:}^1\!/_2$, $1{:}1$, $1{:}0$	random

Table 2: Summary of the two proposed experiments, with the main factors and their levels.

In addition to requiring an input to serve as a guide for generation, translation-based GANs tend to have a limited range of outputs, each input being able to create a single synthetic image (which is why most of them are marked as *fixed* for sampling in Table 1). For that same reason, works using them tend to employ a 1:1 proportion of real:synthesized images, although 1: < 1 ratios are also possible. Works using noise-based GANs tend to display more diverse choices of sampling and image ratios.

In our freehand comments, we also noticed possible issues with experimental protocols that we wished to avoid in our large-scale experiments. The main issues we found amidst existing art were: (1) giving to the GAN-augmented models better data-access than to the baseline model, especially by choosing hyperparameters directly on the test set; (2) (hyper)optimizing the GAN-augmented models more thoroughly than the baseline models; (3) failing to use best current training practices on the baseline model, *e.g.* best available (conventional) data-augmentation, learning-rate choice, normalization, *etc.*; (4) ignoring performance fluctuations, *e.g.*, by performing a single run, or by failing to report the deviation statistics. Not all works suffer from all those issues, of course, but we believe they may explain the discrepancy between the results we report next, and those found in current art.

We decided to limit the scope of this review to works that apply GANs to medical applications, and not to the GANs themselves. We remark, however, that choosing the GAN model for medical-image augmentation puts stringent requirements on the model. First, the model must be able to generate high-resolution images, to accommodate the visual patterns that characterize medical images. For skin-lesion analyses, the patterns that differentiate benign or malignant skin lesions are rather fine-grained, and state-of-the-art classification networks have inputs from 224×224 to 1024×1024 pixels. Second, the model must able to generate class-conditional samples, *i.e.*, to create synthetic samples which convincingly belong for each of the dataset classes, so those may join the supervised training dataset coherently.

3. Experiment design

When we started this study, the initial goal was to maximize the performance of GAN data-augmentation for skin-lesion analysis. As our preliminary experiments progressed, we found our results to be extremely noisy. Performance improvement, when it happened at all, was completely random: the choice of GAN model or other factors had no explanatory power. We changed our research question to a more fundamental one: *can GAN data-augmentation actually improve the performance of skin-lesion analysis?* Failing that, *can GANs be used to anonymize the training data?* The latter application — using synthetic samples *instead of* the actual real data — would be profitable even with a small, tolerable drop in accuracy, since it could make feasible for different countries and institutions to share knowledge in situations where direct patient images could not be exchanged (*e.g.*, due to incompatible privacy laws). We design two experiments (Table 2) to answer those questions, which we detail below.

GAN-based data augmentation. As seen in the previous section, literature shows no consensus on how to perform GAN-based data augmentation. Our experimental design attempted to reflect the diversity of approaches found in existing art, contemplating a diversity of GANs, real:synthetic image ratios, real training dataset sizes, and synthetic sampling techniques. We present those choices next.

The GAN models investigated were pix2pixHD [43], PGAN [20], SPADE [29], and StyleGAN2 [21]. We chose pix2pixHD and PGAN because they are known to work on skin-lesion data augmentation [7], while SPADE and Style-GAN2 are considered the state of the art on image generation. Pix2pixHD and SPADE are translation-based, while PGAN and StyleGAN2 are noise-based techniques. While the former tend to generate very high-quality images, they have stringent limitations on the amount of images they can generate, due to the requirement of using segmentation masks or different image modalities as inputs. Noise-based techniques have no such limitations, but tend to generate images with lower visual quality, and risk reproducing artifacts (*e.g.*, vignettes, rulers) that may reinforce biases in the dataset [6]. For pix2pixHD and SPADE, we use the whole training set's mask to generate our synthetic set. For PGAN and Style-GAN2, we sample enough images to keep a $1{:}^1\!/_2$ ratio. In Figure 1, we show lesions generated by the GANs used in our investigation. For pix2pixHD and SPADE, we use a mask from the training set as input, causing the synthetic

Test Dataset	Size	Imaging Tech.	Diagnostic Classes	Notes
isic19 [10]	3,863	Dermoscopic	melanocytic nevus, melanoma, benign keratosis, actinic keratosis, dermatofibroma, vascular lesion	in-distribution, same classes as train
isic20 [33]	1,743	Dermoscopic	melanocytic nevus, melanoma, benign keratosis, actinic keratosis, lentigo, benign unknown	mainly in-distribution, potential out-of-distribution samples among the 'benign unknown' samples
derm7pt–derm [22]	872	Dermoscopic	melanocytic nevus, melanoma, seborrhoeic keratosis	out-of-distribution, fewer classes than train
derm7pt–clinic [22]	839	Clinical	melanocytic nevus, melanoma, seborrhoeic keratosis	out-of-distribution, fewer classes than train
dermofit [4]	973	Clinical	melanocytic nevus, melanoma, seborrhoeic keratosis, actinic keratosis, pyogenic granuloma, haemangioma, dermatofibroma	out-of-distribution benign classes

Table 3: Description of the test sets used in the evaluation of the classification networks trained with the augmented training set.

images to be almost identical to the real ones for the human eye (although features from deep networks may be still considerably different between the two kinds of images).

For real:synthetic ratios, we considered the ratios $1{:}^1/_4$, $1{:}^1/_2$, and $1{:}1$, where 1 is the size of our original real-image training set ($14,805$ samples). We considered additional experiments varying the proportion of benign and malignant synthetic images proportion, to evaluate the opportunity of using GAN-based data augmentation to correct class-imbalance [14]. Those experiments are notated (real:synthetic benign:synthetic malignant), and we considered the ratios $(1{:}^1/_2{:}^1/_2)$ (which is the same as $1{:}^1/_2$ and has no balancing effect), $(1{:}^1/_2{:}^3/_4)$, and $(1{:}^1/_2{:}1)$. The baseline for all the experiments in this group is the ratio $1{:}0$, *i.e.*, the experiment with the entire real training set and no synthetic data augmentation.

We varied the amount of *real* images employed in training mainly as an evaluation of GAN-based anonymization (see below). Those experiments may also be interpreted as the impact of GAN-based data-augmentation for different training-set sizes. We used real-image training set sizes with fractions of $^1/_{16}$, $^1/_8$, $^1/_4$, $^1/_2$, and $^1/_1$ (the baseline) of the whole dataset.

Finally, we investigate different ways of sampling the synthetic images from the generated pool. That only applies to the noise-based GANs (PGAN and StyleGAN2), which can generate a limitless amount of images. We generated $100,000$ images for each of the 2 classes (benign and melanoma), and evaluated different methods to select the ones to compose the training set: choosing them at *random*, choosing them at random but with a criterion of *diversity* inspired on the perceptual-sensitive hash (pHash) to exclude near-duplicates, or choosing the ones *best*-classified (lowest error) or *worst*-classified (highest error) by an ancillary skin-lesion classification model.

GAN-based anonymization. In anonymization, instead of using GANs to augment the training set, synthetic images *replace* real images. That application has received less attention in literature [25], but could be invaluable for researchers and institutions wishing to share knowledge while having to navigate issues of patient confidentiality.

For that experimental design, we evaluate a single GAN (StyleGAN2) and a single sampling technique (random), explained above. In contrast, we evaluate many more real:synthetic ratios in this experiment, varying the amount of both kinds of images, in order to evaluate the situation where an institution was training a model with its own real images, adding synthetic images from a GAN provided by another institution. For each group of experiments with a fraction of the initial real training-set $^1/_x$, we evaluated experiments with ratios $^1/_x{:}0$ (the baseline for the group, with no synthetic images), $^1/_x{:}^1/_x$ (doubling the training set), and $^1/_x{:}^{x-1}/_x$ (topping up the training set). We included the $1{:}0$ ratio as a reference for the expected upper bound on the accuracy for those experiments.

3.1. Datasets

For all experiments, the reference training set of real images was based on the training set of the ISIC 2019 challenge [10, 41]. We split that dataset into a training set ($14,805$ samples) and a validation set ($1,931$ samples) used in all our experiments, and a test set, with $3,863$ samples, added to our collection of test sets (*isic19* on Table 3).

We trained the noise-based PGAN and StyleGAN2 with the entire training set, but pix2pixHD and SPADE require semantic segmentation masks to guide the generation. We employ the clinical attribute semantic masks of the ISIC Challenge 2018 Task 2, which are available for only $2,594$ images, or about $^1/_6$ of the training set.

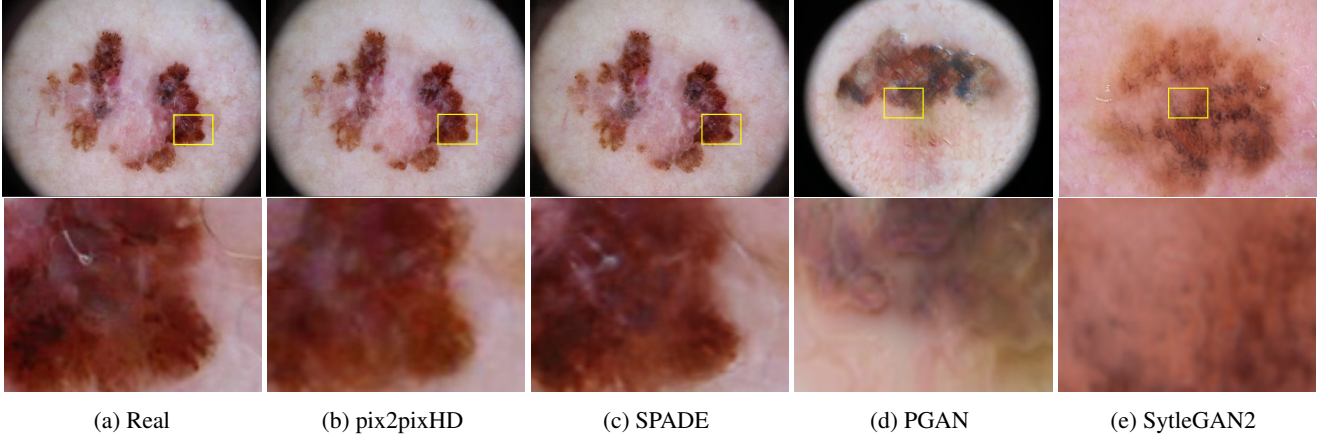|				
(a) Real	(b) pix2pixHD	(c) SPADE	(d) PGAN	(e) SytleGAN2

Figure 1: Synthetic samples for different GAN-based approaches: (a) Real, (b) pix2pixHD, (c) SPADE, (d) PGAN, (e) StyleGAN2. In the first row, we present the full image, while in the second, we zoom-in to focus on the details.

When fine-tuning the target models, fractions of real images refer to selections of that same training dataset (*e.g.*, $^1/_4$:0 would refer to a training set of $3,701$ randomly selected samples from the $14,805$ real ones). Fractions of the synthetic dataset refer to the same *size* of the real dataset, for a selection on the synthetic generated images (*e.g.*, 0:$^1/_4$ would refer to a training set of $3,701$ randomly selected while keeping the real dataset class ratio, sampled from the $200,000$ synthetic images, half benign, and half malignant).

We perform our tests in five gold-standard datasets (Table 3), selecting the classes to always perform a melanoma-*vs.*-benign task (carcinomas, if present in the dataset, are discarded from both training and testing). Having an array of test sets, both similar to our training set ("in-distribution") and very different ("out-of-distribution") is an attempt to mitigate the effect of dataset bias [6, 8, 13] and measure the models' generalization ability. For derm7pt, we remove near duplicates, and keep only classes present in the ISIC 2018 Challenge Task 2 dataset (melanoma, sebhorreic keratosis, and nevus). Those alterations result in 872 samples for derm7pt-derm and 839 samples for derm7pt-clinic. For dermofit, we remove the carcinomas, leaving 973 samples.

Analyses. We replicated each experiment ten times, varying the selection of the real images from the training set and repeating the fine-tuning of the target model. In all experiments, the metric was the area under the ROC-curve of the target model (AUC) for melanoma-*vs.*-benign classification. A visual analysis of the results is given by blotting the individual data-points superimposed with a box-plot that, as usual, reveals the medians, quartiles, and range. We also plot the arithmetic means (red dots). There are two sets of plots, one for each application (augmentation and anonymization), plots were separated per test dataset, and within each plot, experiments were grouped (blue and

black colors) to facilitate comparison. The labels of the experiments reveal the proportion of real:synthetic — or (real:synthetic benign:synthetic malignant) — images used in the training set used to fine-tune the target model, with additional information for the choice of GAN (p2p: pix2pixHD, spd: SPADE, pgn: PGAN2, sgn: StyleGAN2, all: samples from all GANs together), and the choice of sampling method (wst: worst, bst: best, div: diverse). When omitted, the GAN is StyleGAN2 and the sampling method is random.

3.2. Implementation details

For PGAN and pix2pixHD, we follow Bissoto *et al.*'s [7] implementation, modifications, and hyperparameters. We also adopt their pix2pixHD generation procedure to SPADE. For StyleGAN2, we do not need to adapt the original GAN model implementation[1], as it could generate class-conditioned samples from the start.

Evaluating the quality of synthetic images is an open problem. Available metrics often fail to evaluate diversity, and both broad- and fine-grained details. Most popular metrics (*e.g.*, FID [17], Inception Score [35]) rely on ImageNet-trained networks to extract characteristics, and are, thus, questionable for applications where classes are very different, such as medical applications. We address that issue using a broad analysis, comprising traditional GAN metrics, metrics on classifier models, and visualization techniques of the trained classifier.

To select the best training checkpoint for the GAN model, we considered both the time spent on training, and the FID metric [17] preferring later checkpoints (longer training) for similar FID (< 0.3 difference). The exact training lengths, and corresponding FIDs, appear in Table 4.

---

[1]https://github.com/NVlabs/stylegan2

GAN Architecture	Epochs	FID
SPADE	300	16.62
pix2pixHD	400	19.27
PGAN	890	39.57
StyleGAN2	565	15.98

Table 4: Amount of epochs and FIDs for each of the generative models used in this research. To select the checkpoint used to generate the samples that compose our classification model's training dataset, we consider both FID, and time spent training the GAN.

The target model for both the data-augmentation and anonymization experiments is a skin-lesion classification Inception-v4 [40], pre-trained on ImageNet. We chose this model based on its well-known performance on skin-lesion analysis, including its ability to profit from larger training sets [42]. The full set of weights of the target model is fine-tuned with stochastic gradient descent with momentum of 0.9, weight decay of 0.001 and learning rate of 0.001, which we reduce by 10 on each validation plateau after a 10-epoch patience until a minimum of 10^{-5}. We use a batch size of 32, shuffling the data before each epoch. We resize input images to 299×299 pixels, and z-normalize the input with ImageNet's training set mean and standard deviation. We train for a maximum of 100 epochs, and an early stopping with a patience of 22 epochs. We apply (conventional) data augmentation to *all* experiments both during train *and test*, with random horizontal and vertical flips, resized crops containing 75–100% of the original image area, rotations from -45–45° degrees, and hue changes of -10–10%. We apply the same augmentations on both train and test. For test augmentation we average the predictions of 50 augmented replicas of each sample [30]. We run each experiment 10 times, and in each one, we vary the real images, but keep the synthetics the same (following the sampling criteria).

Full-reproducible source code is available in our repository https://github.com/alceubissoto/gan-aug-analysis.

4. Results

In this section, we evaluate GAN-based augmentation procedures for skin lesion analysis. In all the following experiments, we want to make the comparisons as fair as possible, giving equal opportunity to all models to be at their peak performance.

4.1. Augmentation

The experiments for GAN-based data augmentation appear on the top row of Figure 2. Comments on how to read and interpret those plots appear in Section 3, under the heading "Analyses".

The leftmost experiment in each plot, with label 1:0, is the baseline with no synthetic-data augmentation. The plots reveal that, for dermoscopic test images, augmented-train sets are unable to confer a significant advantage, with augmented models showing lower — or at best similar — to the baseline. The experiments suggest that the more synthetic images we add to the training set, the worse the results are. The performance of different GANs fluctuated across datasets, but translation-based GANs tended to work better than noise-based GANS — but please notice that this factor is slightly confounded with image proportion in our tests due to the very limited generation ability of translation-based GANs. Of the two noise-based GANs, StyleGAN2 performed better (or at least, less worse).

The scenario was less clear for the derm7pt-clinic dataset, where most experiments significantly improved the results. Those results departed from the other datasets also by showing StyleGAN2 ahead of all other GANs. However, the results on the dermofit dataset, also with clinical images, were more similar to the results on the dermoscopic datasets than to the ones on derm7pt-clinic.

We remind that the experiments on anonymization (explained next), may also be interpreted as experiments on data-augmentation for small training datasets, *i.e.*, an anonymization experiment with ratio $1/x$:y can be reinterpreted as a data augmentation experiment with ratio 1:x/y for a test dataset with $1/x$ of the samples of our main one. As we will show, those experiments failed to reveal the advantage of synthetic augmentation, even for small datasets.

4.2. Anonymization

The experiments for GAN-based anonymization appear on the bottom row of Figure 2. Comments on how to read and interpret those plots appear in Section 3, under the heading "Analyses".

Those experiments are organized in groups of alternating blue and black colors, the rightmost experiment in each group, with label $1/x$:0, being the baseline for the group. GAN-based anonymization only has interest if the performance of the experiment is significantly above that baseline — otherwise, it can be trivially obtained by simply discarding the problematic samples. The rightmost experiment shows the model trained with all real samples, and gives an expected upper bound of accuracy for the plot.

Unfortunately, for dermoscopic test images, no experiment appeared above that baseline. The derm7pt-clinic was again an exception, showing three experiments above the baseline, and one even slightly above the expected upper bound. In this set of experiments, the results on the (also clinical-image) dermofit dataset were also positive, with several configurations above the baseline.
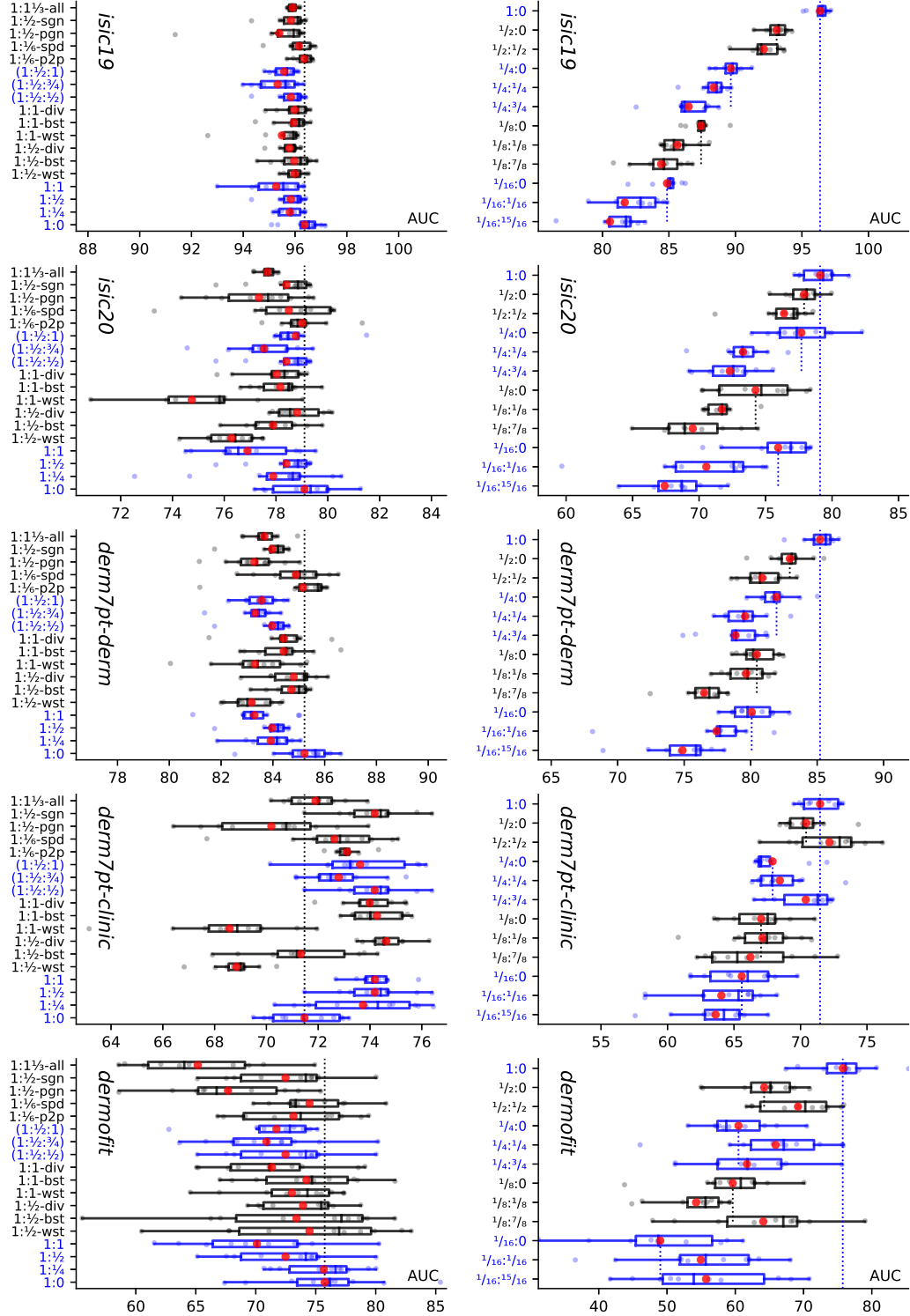
Figure 2: Results for GAN-based data augmentation (left) and GAN-based anonymization (right), separated by test dataset. Box-plots display medians, quartiles, and range, as usual; red dots show arithmetic means. Y-axis labels show real:synthetic or (real:synthetic benign:synthetic malignant) training image proportion, with additional information for the choice of GAN (p2p: pix2pixHD, spd: SPADE, pgn: PGAN2, sgn: StyleGAN2, all: samples from all GANs together), and the choice of sampling method (wst: worst, bst: best, div: diverse). When omitted, the GAN is StyleGAN2 and the sampling method is random. Experiments showcasing a factor are grouped together in alternating blue/black color.

5. Discussion

Our results suggest future authors interested in GAN-based data augmentation to be conservative about expected results, and cautious about evaluation protocols.

GAN-based augmentation is a technique extensively explored in the literature for mitigating the scarcity of training images, being particularly interesting for the medical images community. However, we found our preliminary experiments to be excessively noisy, and noticed flaws in some experimental protocols during our literature review. Reliably transforming synthetic images into reliable performance gains is far from obvious.

We will *not* go as far as *condemning* GAN-based data augmentation. Our experiments suggest that for some specific out-of-distribution scenarios (*e.g.*, training on isic19 and testing on derm7pt-clinic) the technique may provide reliable improvements. Characterizing exactly which scenarios are those is, however, still an open question, as experiments on the also clinical-image dermofit dataset did not confirm those findings.

Because training GANs requires a huge computation investment — a single training of StyleGAN2 takes up to a month of GPU time — researchers and practitioners should carefully evaluate whether their application to data-augmentation is justifiable, considering, among other things, their energetic footprint [47]. The finicky nature of GAN training also brings other risks: missing modes [2] may aggravate dataset biases, reinforcing the disparity of over/underrepresented groups instead of correcting them. GANs may also fixate on artifacts (such as vignettes, rulers, gel bubbles, *etc.*) and introduce or reinforce spurious correlations on the data [6].

Our results for GAN-based anonymization show modest results, but here, at least, there seems to be a trend, with results for out-of-distribution data being generally favorable, and results for in-distribution data being generally unfavorable. Those data suggest that using GANs may be possible at least as an ancillary method for sharing knowledge while preserving patient privacy. For making that application safe, however, we need further studies on how much the GAN "remembers" each original training sample and on its ability to (purposefully or accidentally) reconstructing original samples.

Possible avenues for unblocking GAN-based data augmentation point towards attempting to conciliate the advantages of translation-based and noise-based techniques — obtaining the high-quality of the former and the limitless sampling availability of the latter — but such conciliation is a hard open problem. More achievable may be enhanced sampling methods, able to select the highest-quality, or better yet most relevant for decision, samples from the limitless sample of a noise-based technique. Even if future works fail to improve the ability of GANs to reliably provide data-augmentation or anonymization, a better characterization of the cases they are able to improve may provide interesting insights on the fundamental workings of deep learning.

Acknowledgements

References

[1] S. Amirrajab, S. Abbasi-Sureshjani, Y. Al Khalil, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer. Xcat-gan for synthesizing 3d consistent labeled cardiac mr images on anatomically variable xcat phantoms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 128–137. Springer, 2020. 2

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

[3] O. Bailo, D. Ham, and Y. Min Shin. Red blood cell image generation for data augmentation using conditional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019. 2

[4] 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. 1, 4

[5] D. Bisla, A. Choromanska, R. S. Berman, J. A. Stein, and D. Polsky. Towards automated melanoma detection with deep learning: Data purification and augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019. 2

[6] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila. (De)constructing bias on skin lesion datasets. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 3, 5, 8

[7] 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. 2, 3, 5

[8] A. Bissoto, E. Valle, and S. Avila. Debiasing skin lesion datasets and models? Not so fast. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 5

[9] B. Bozorgtabar, D. Mahapatra, H. von Tengg-Kobligk, A. Poellinger, L. Ebner, J. P. Thiran, and M. Reyes. Informative sample generation using class aware generative adversarial networks for classification of chest xrays. *Computer Vision and Image Understanding*, 184:57–65, 2019. 2

[10] M. Combalia, N. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. Halpern, S. Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 1, 4

[11] L. A. de Souza Jr, L. A. Passos, R. Mendel, A. Ebigbo, A. Probst, H. Messmann, C. Palm, and J. P. Papa. Assisting barrett's esophagus identification using endoscopic data augmentation based on generative adversarial networks. *Computers in Biology and Medicine*, 126:104029, 2020. 2

[12] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. 2

[13] R. Geirhos, J. H Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 5

[14] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu. Dermgan: synthetic generation of clinical skin images with pathology. In *Machine Learning for Health Workshop (NeurIPS ML4H)*, pages 155–170. PMLR, 2020. 2, 4

[15] 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. 1

[16] C. Han, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, and H. Hayashi. Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection. *IEEE Access*, 7:156966–156977, 2019. 2

[17] 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. 5

[18] 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. 2

[19] Y. Jiang, H. Chen, M. H. Loew, and H. Ko. Covid-19 ct image synthesis with a conditional generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 2021. 2

[20] 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. 2, 3

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and Timo. Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 3

[22] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, mar 2019. 1, 4

[23] A. B. Levine, J. Peng, D. Farnell, M. Nursey, Y. Wang, J. R. Naso, H. Ren, H. Farahani, C. Chen, D. Chiu, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of pathology*, 252(2):178–188, 2020. 2

[24] Q. Li, Z. Yu, Y. Wang, and H. Zheng. Tumorgan: A multimodal data augmentation framework for brain tumor segmentation. *Sensors*, 20(15):4203, 2020. 2

[25] D. Mahapatra, B. Bozorgtabar, and L. Shao. Pathological retinal region segmentation from oct images using geometric relation based augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9611–9620, 2020. 2, 4

[26] T. Mendonça, P. Ferreira, J. Marques, A. Marcal, and J. Rozeira. PH2: A dermoscopic image database for research and benchmarking. In *IEEE EMBS*, 2013. 1

[27] 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. 1

[28] D. Moher, A. Liberati, J. Tetzlaff, and D. Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *BMJ*, 339, 2009. 2

[29] 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. 2, 3

[30] 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. 1, 6

[31] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana. Augmenting data with gans to segment melanoma skin lesions. *Multimedia Tools and Applications*, 79(21):15575–15592, 2020. 2

[32] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[33] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021. 1, 4

[34] 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. 1

[35] 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. 5

[36] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):1–9, 2019. 2

[37] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, E. Celebi, and J. Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, 2021. 1

[38] H. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI MICCAI)*, pages 1–11. Springer, 2018. 2

[39] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 1

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 6

[41] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1, 4

[42] 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*, 2020. 1, 6

[43] 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. 2, 3

[44] Z. Wang, Q. She, and T. Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 1

[45] E. Wu, K. Wu, D. Cox, and W. Lotter. Conditional infilling gans for data augmentation in mammogram classification. In *Breast Image Analysis Workshop (BIA MICCAI)*, pages 98–106. Springer, 2018. 2

[46] J. Ye, Y. Xue, L. R. Long, S. Antani, Z. Xue, K. C. Cheng, and X. Huang. Synthetic sample selection via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 53–63. Springer, 2020. 2

[47] N. Yu, L. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *IEEE International Conference on Computer Vision (CVPR)*, pages 7556–7566, 2019. 8