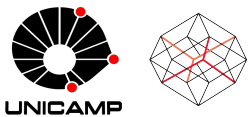
GAN-Based Data Augmentation and Anonymization for Skin-Lesion Analysis:

A Critical Review

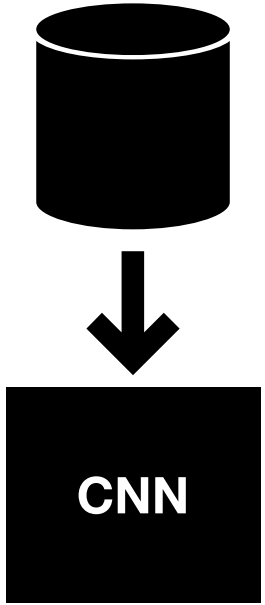


Alceu Bissoto[1], Eduardo Valle[2], Sandra Avila[1]

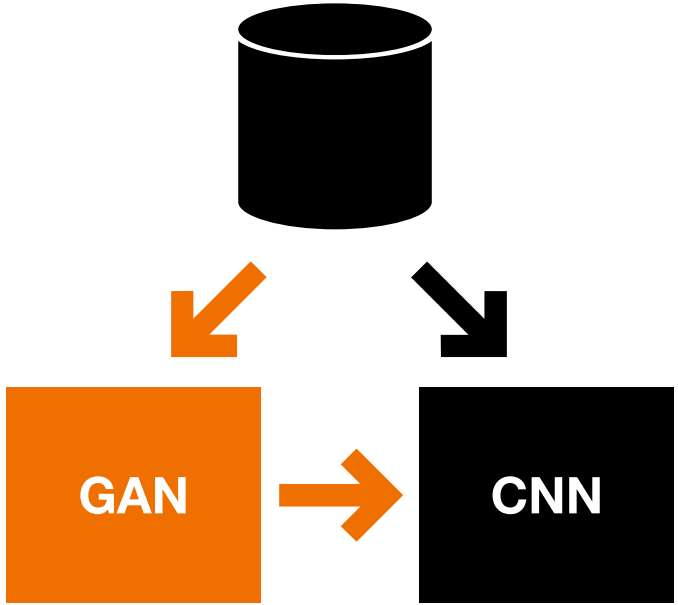
[1]Institute of Computing [2]School of Electrical and Computing Engineering

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

GAN-based augmentation is a method to mitigate the lack of data



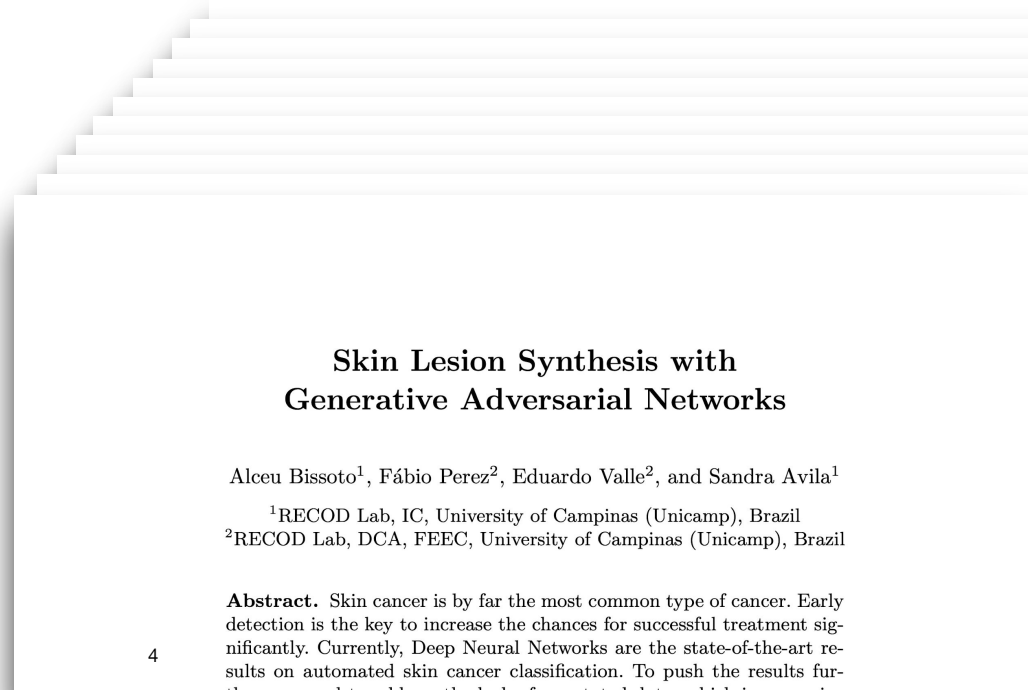
GAN-based augmentation is a method to mitigate the lack of data



Preliminary experiments did not reliably improve performance

What are we doing wrong?

- Systematic Literature Review on GAN-based augmentation in the medical context.
- What did we learn?

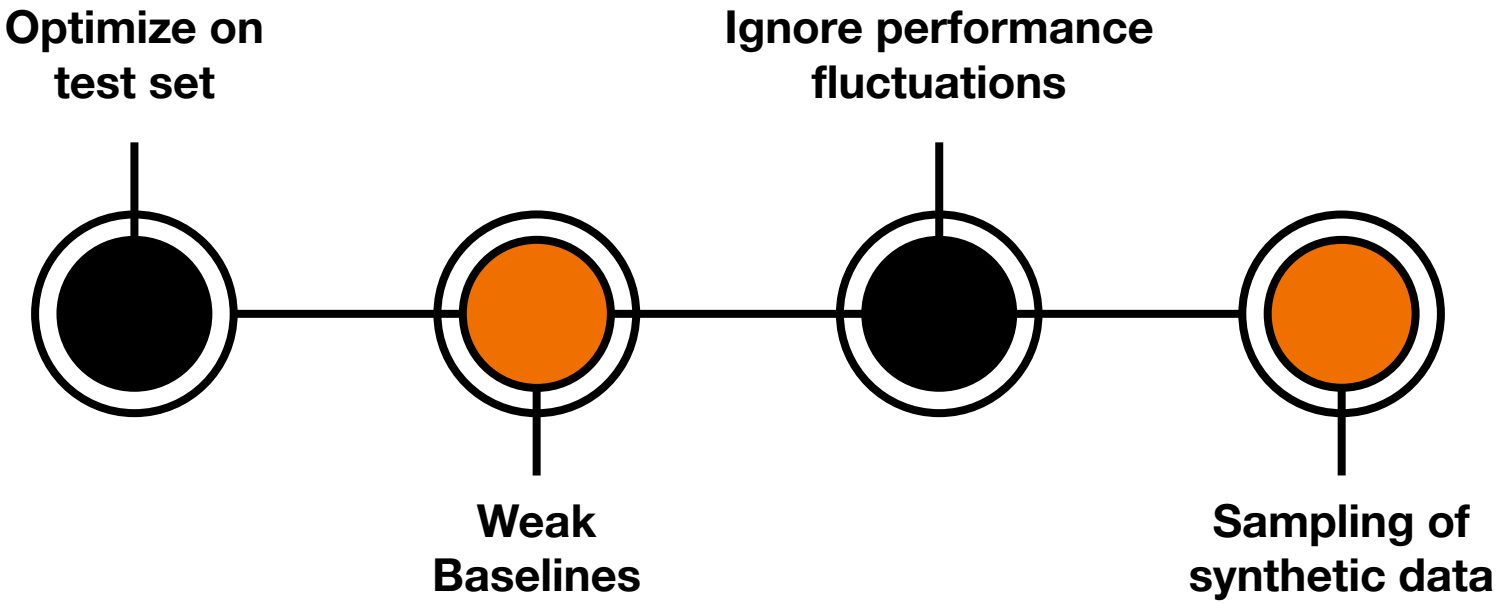


Skin Lesion Synthesis with Generative Adversarial Networks

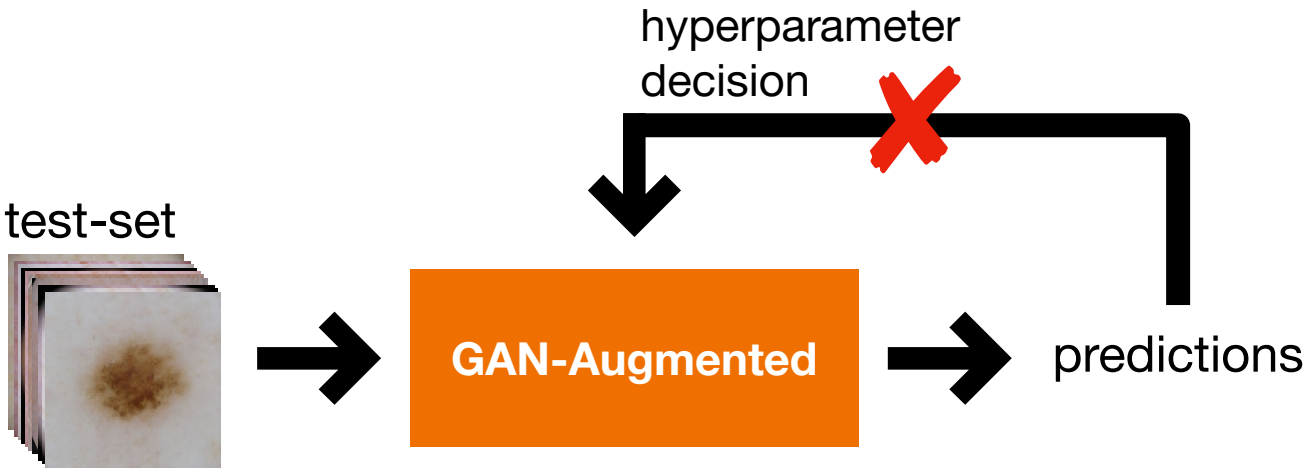
Alceu Bissoto[1], Fábio Perez[2], Eduardo Valle[2], and Sandra Avila[1]

[1]RECOD Lab, IC, University of Campinas (Unicamp), Brazil
[2]RECOD Lab, DCA, FEEC, University of Campinas (Unicamp), Brazil

Abstract. Skin cancer is by far the most common type of cancer. Early detection is the key to increase the chances for successful treatment significantly. Currently, Deep Neural Networks are the state-of-the-art results on automated skin cancer classification. To push the results fur-

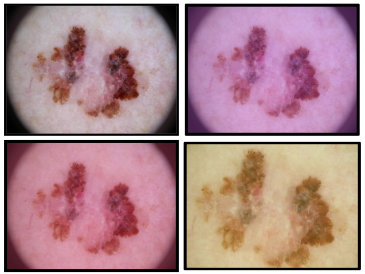

Optimize on test set
Ignore performance fluctuations
Weak Baselines
Sampling of synthetic data

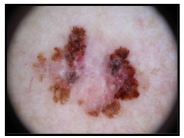
Choosing hyperparameters directly on the test-set ✗

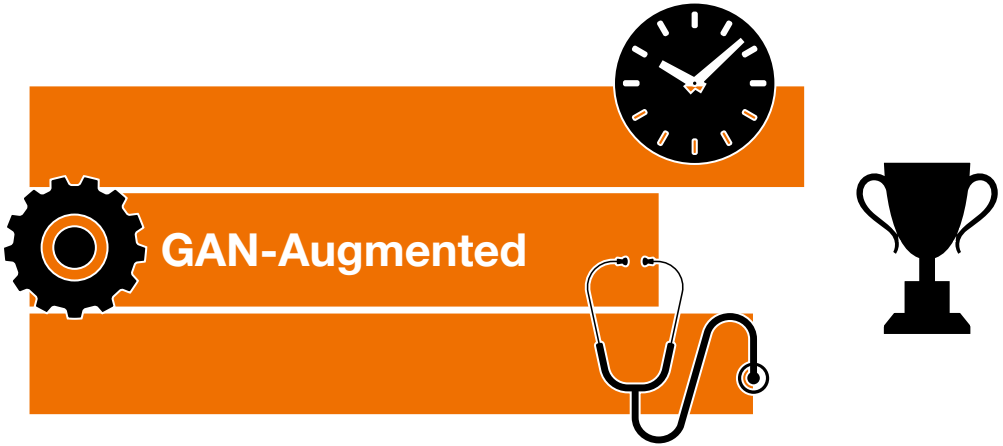


GAN-augmented models are more thoroughly optimized ✗

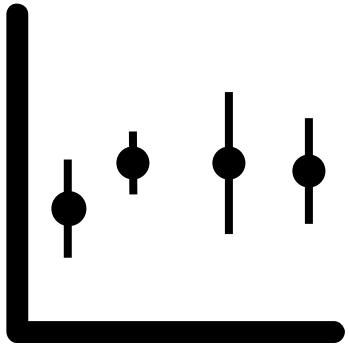
Weak Baselines





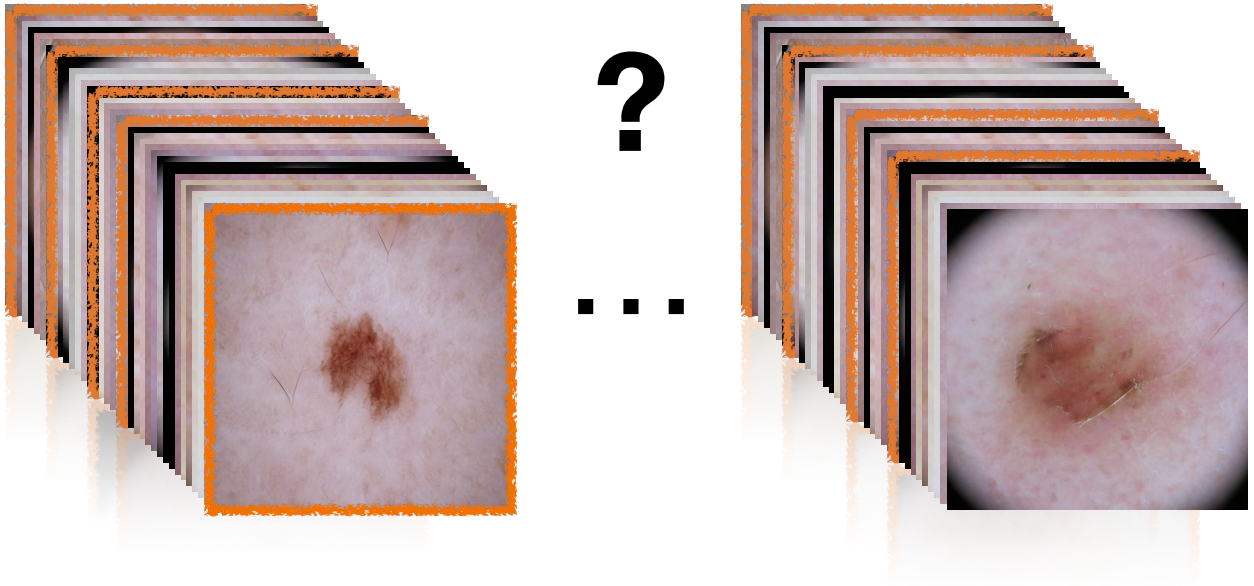


Ignoring performance fluctuations ✘



$$\mu \pm \sigma$$

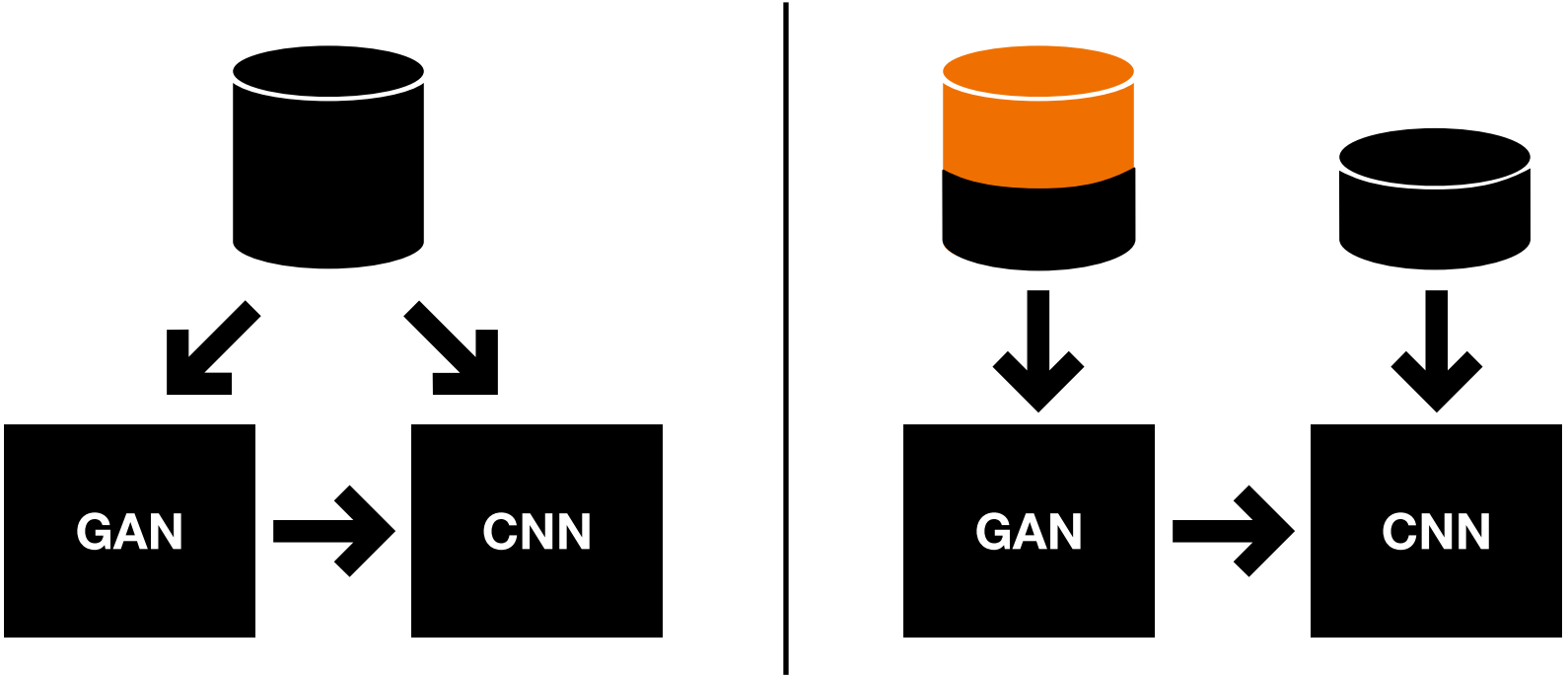
Are the sampling method and the amount of synthetic images key factors for GAN-based augmentation?



Methods

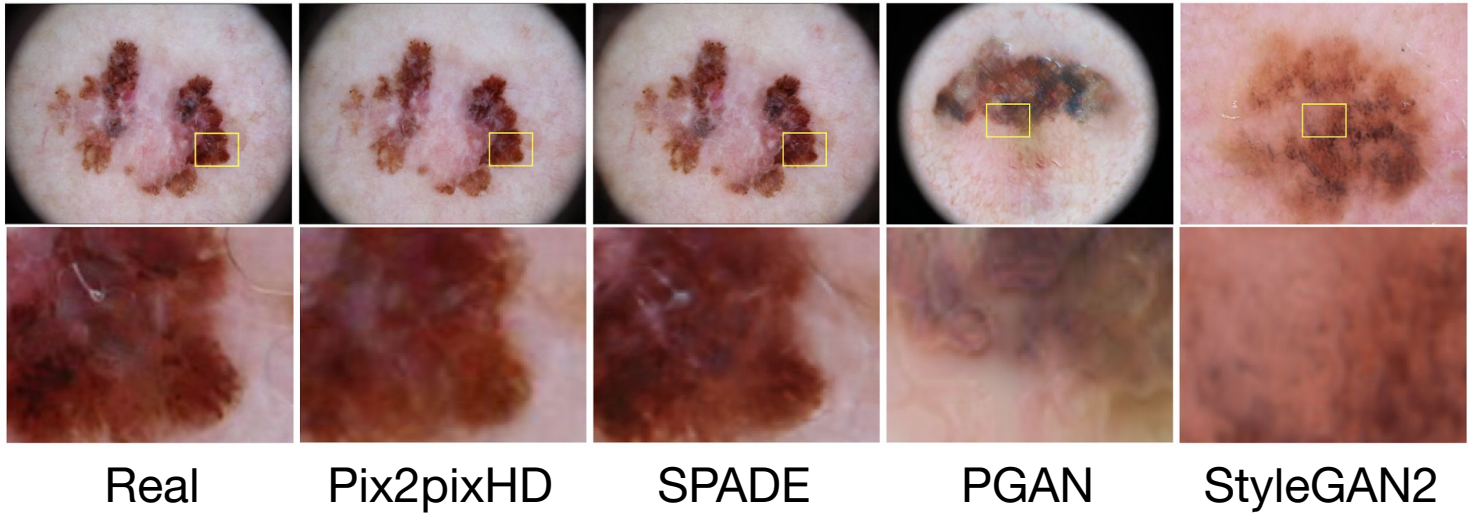
Our work evolved to a critical analysis of GAN-based augmentation

Augmentation vs. Anonymization



We consider different GANs

Both translation and noise-based

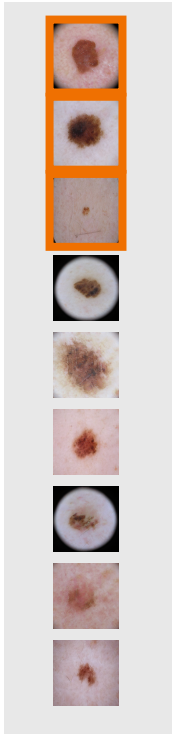


We evaluate different sampling methods

random

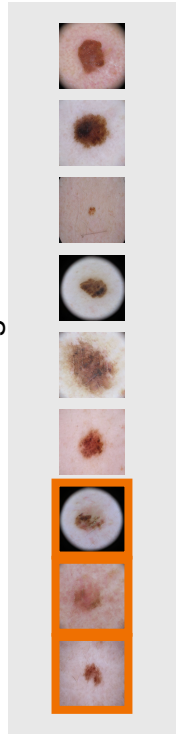
best

sorted according to CNN scores



worst

sorted according to CNN scores



diverse

pHash-based removal of near duplicates

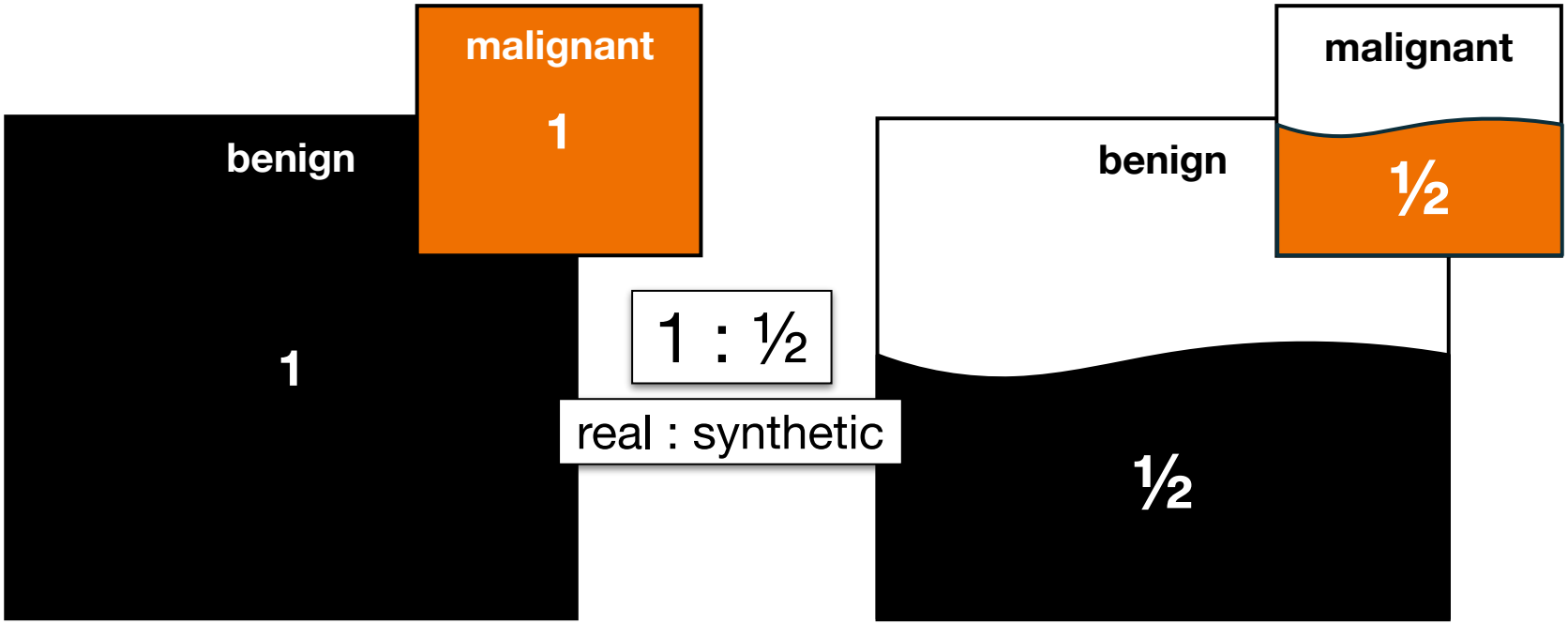
We sample different ratios of real and synthetic

real

14.805 images from ISIC 2019

synthetic

(14.805 / 2) images generated with a GAN

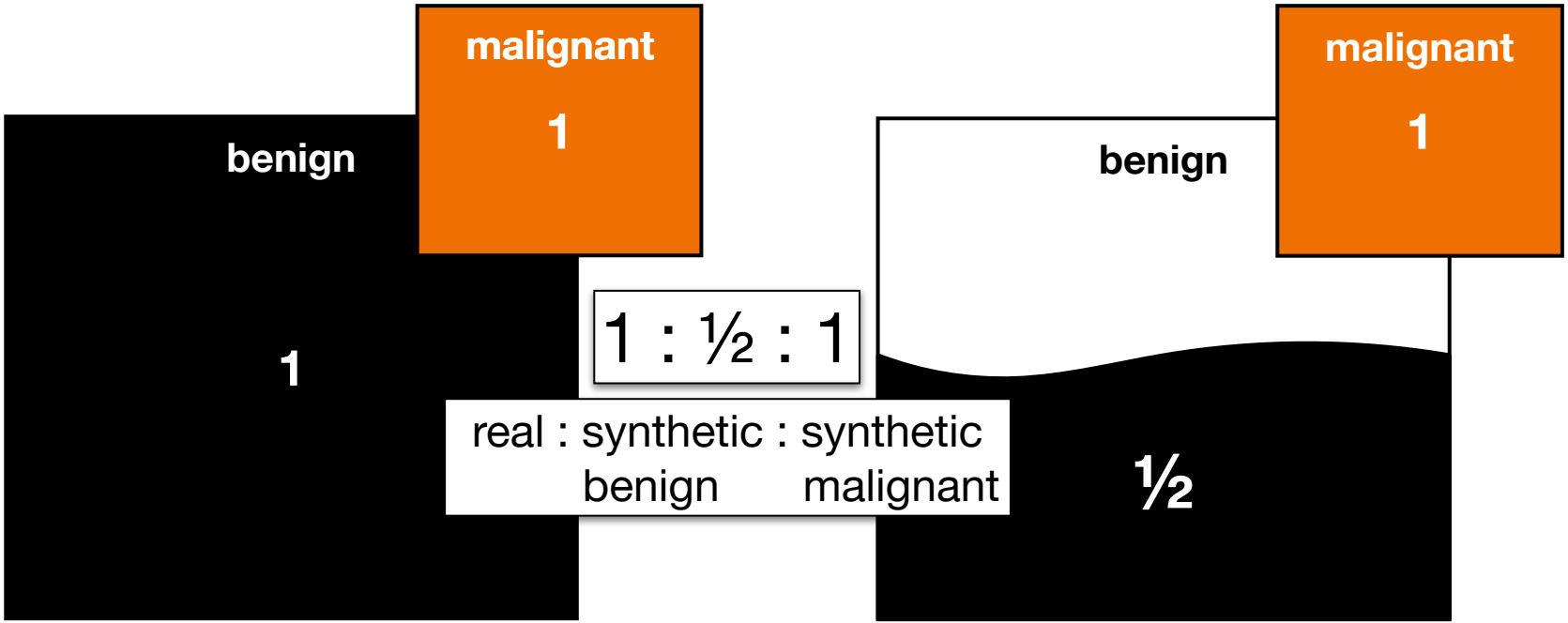


We sample different ratios of real and synthetic

real

14.805 images from ISIC 2019

synthetic



All models are trained under the same design

To select the best training checkpoint for the GAN, we consider both the **time spend on training** and the **FID score**.

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

All models are trained under the same design

To select the best training checkpoint for the GAN, we consider both the **time spend on training** and the **FID score.**

We perform **early stopping** based on the validation loss.

All models are trained under the same design

To select the best training checkpoint for the GAN, we consider both the **time spend on training** and the **FID score.**

We perform **early stopping** based on the validation loss.

We apply conventional data augmentation to **all experiments** (both during train and test).

All models are trained under the same design

To select the best training checkpoint for the GAN, we consider both the **time spend on training** and the **FID score.**

We perform **early stopping** based on the validation loss.

We apply conventional data augmentation to **all experiments** (both during train and test).

We evaluate our models in **5 different test sets.**

All models are trained under the same design

To select the best training checkpoint for the GAN, we consider both the **time spend on training** and the **FID score.**

We perform **early stopping** based on the validation loss.

We apply conventional data augmentation to **all experiments** (both during train and test).

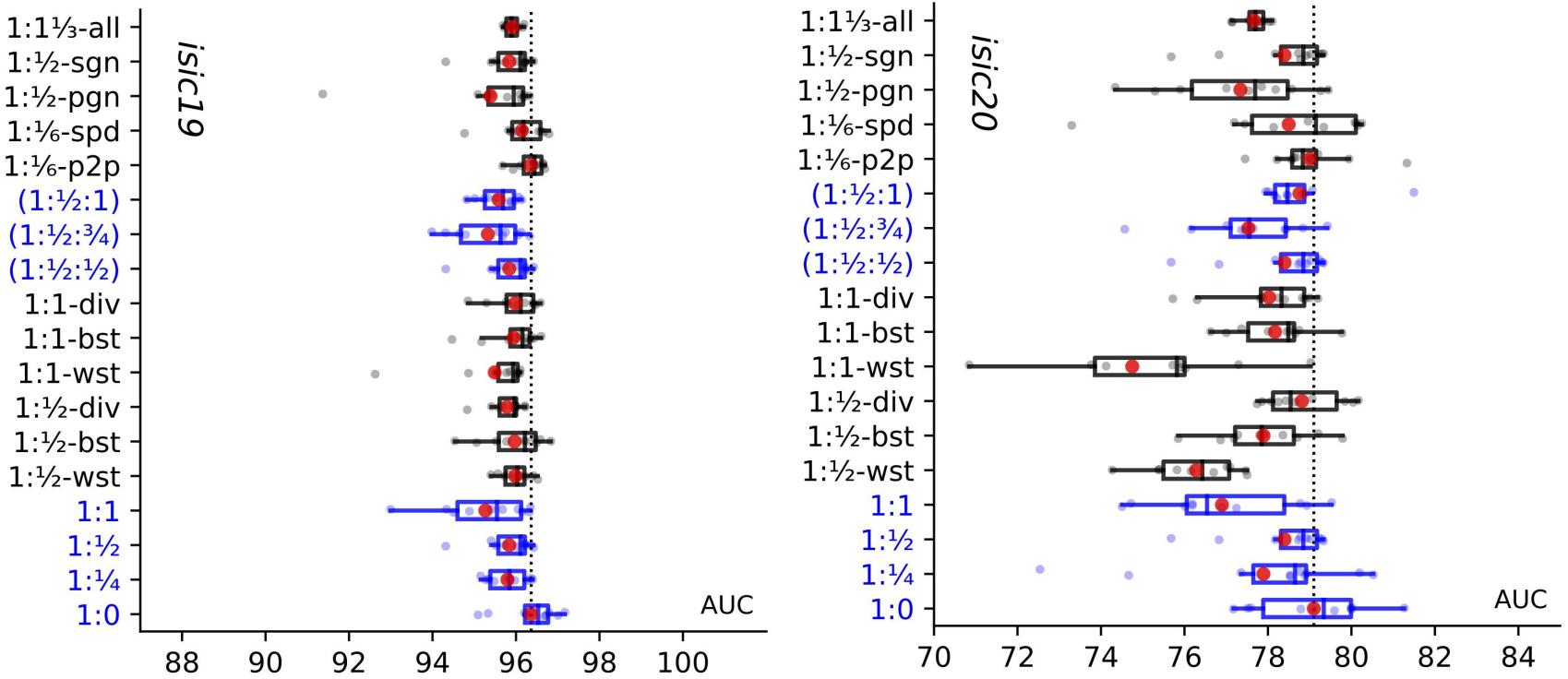
We evaluate our models in **5 different test sets.**

For statistical significance, we **run our experiments 10 times.**

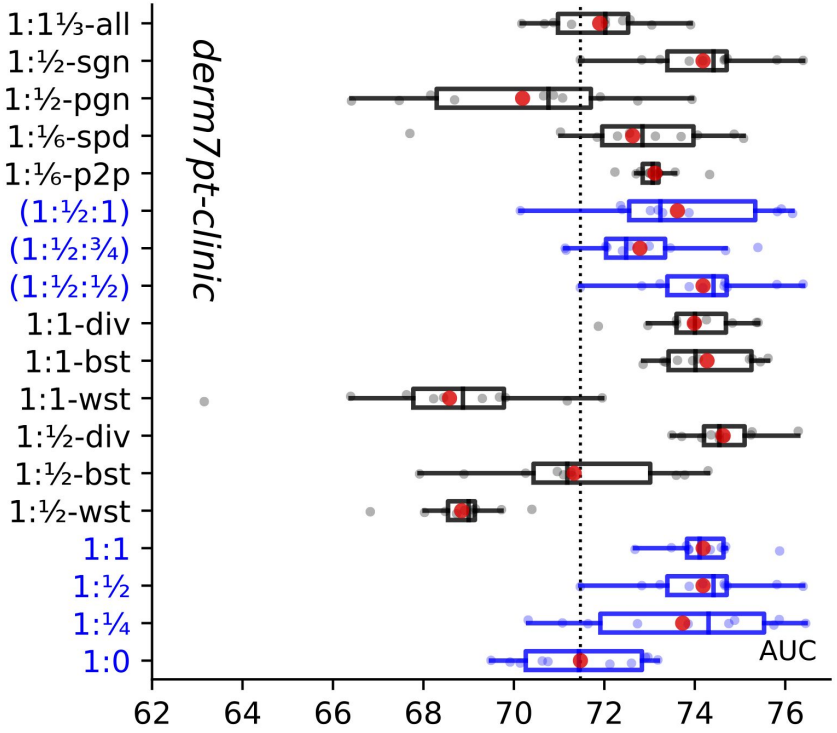
Results

GAN-based augmentation did not reliably improve performance (with some exceptions)

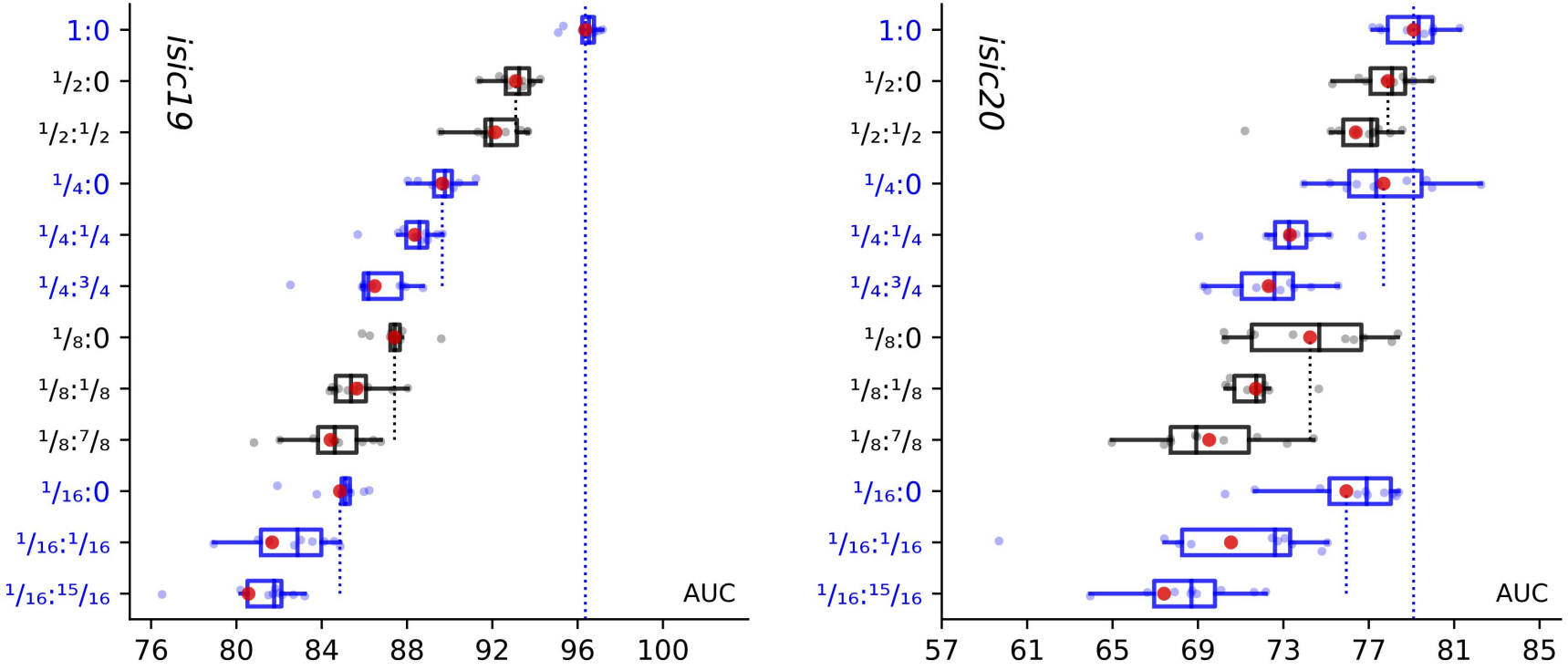
GAN-based augmentation on in-distribution test did not reliably improve performance



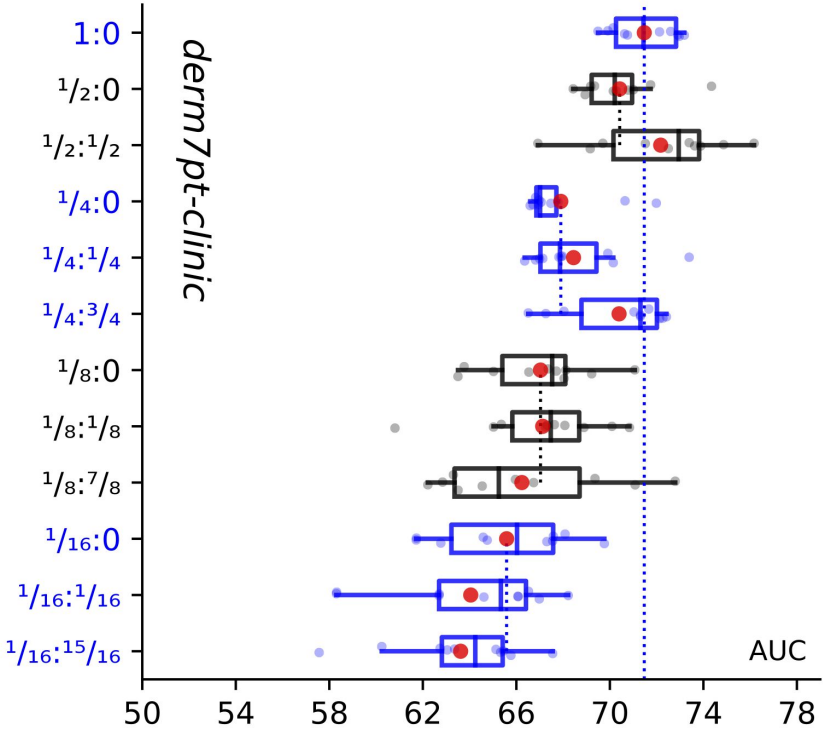
GAN-based augmentation on out-of-distribution test improved performance



GAN-based anonymization on in-distribution test did not reliably improve performance



GAN-based anonymization on out-of-distribution test improved performance



Takeaway:

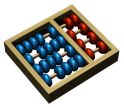
Be cautious about evaluation protocols

Code, Data & Paper:

`https://github.com/alceubissoto/gan-critical-review`

Thank you!

Alceu Bissoto alceubissoto@ic.unicamp.br
Eduardo Valle dovalle@dca.fee.unicamp.br
Sandra Avila sandra@ic.unicamp.br



ISIC Workshop @ **CVPR 2021**