Artifact-based domain generalization of skin lesion models

Alceu Bissoto[1], Catarina Barata[2], Eduardo Valle[3], Sandra Avila[1]

[1]Institute of Computing [3]School of Electrical and Computing Engineering
Recod.ai, University of Campinas (UNICAMP), Brazil

[2]Institute for Systems and Robotics, Instituto Superior Técnico, Portugal

MIT Technology Review

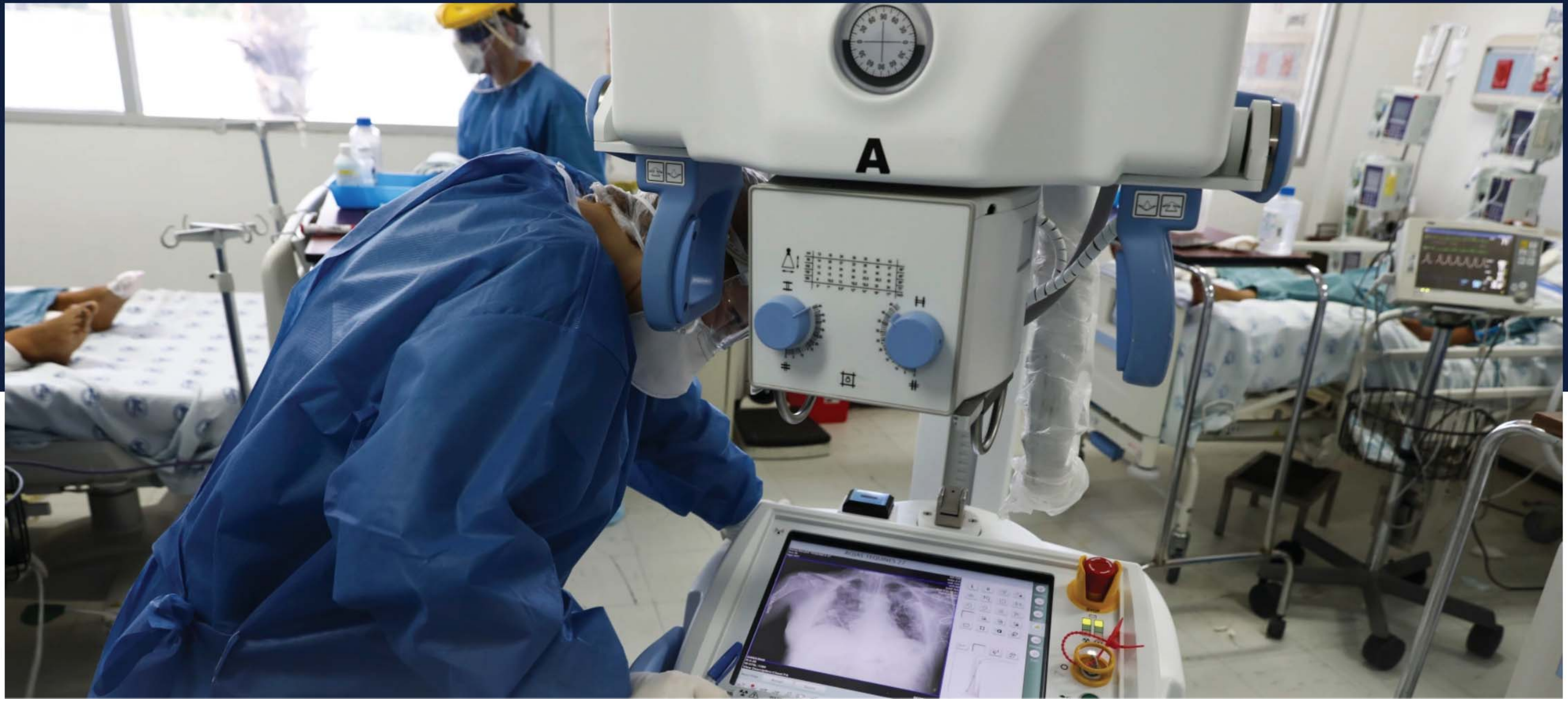
ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021



https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

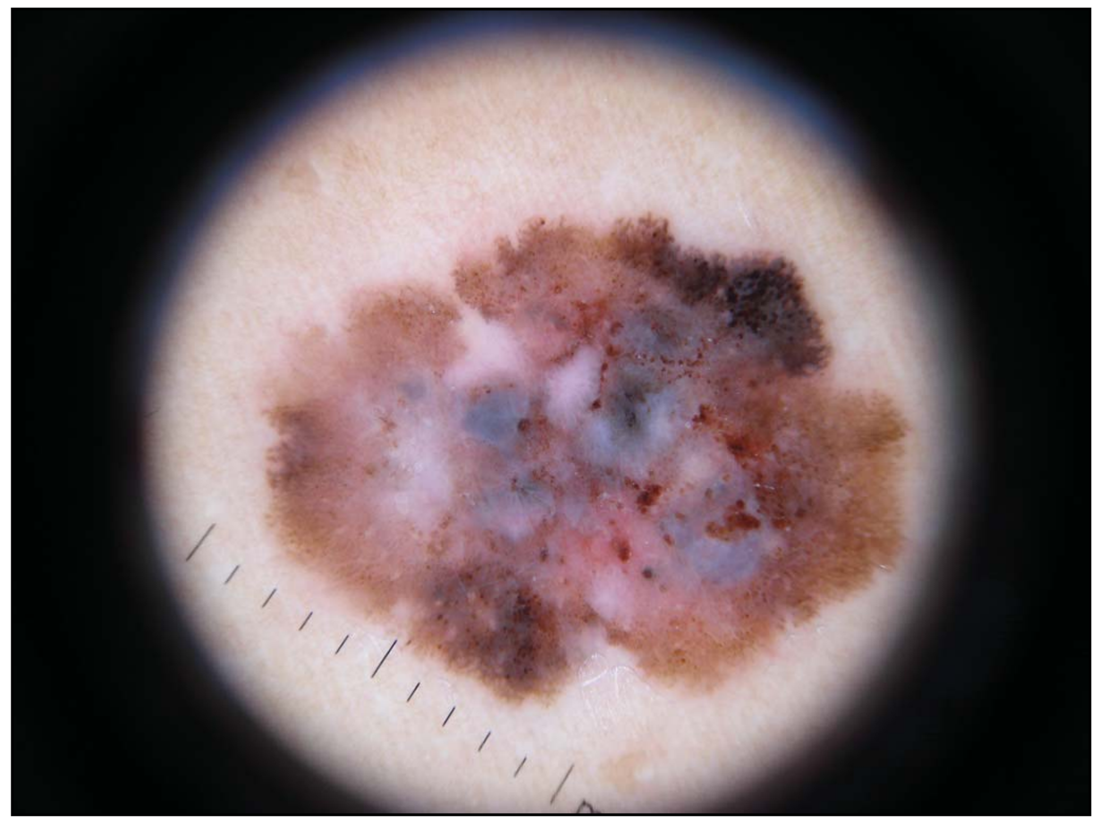
The problem of dataset bias

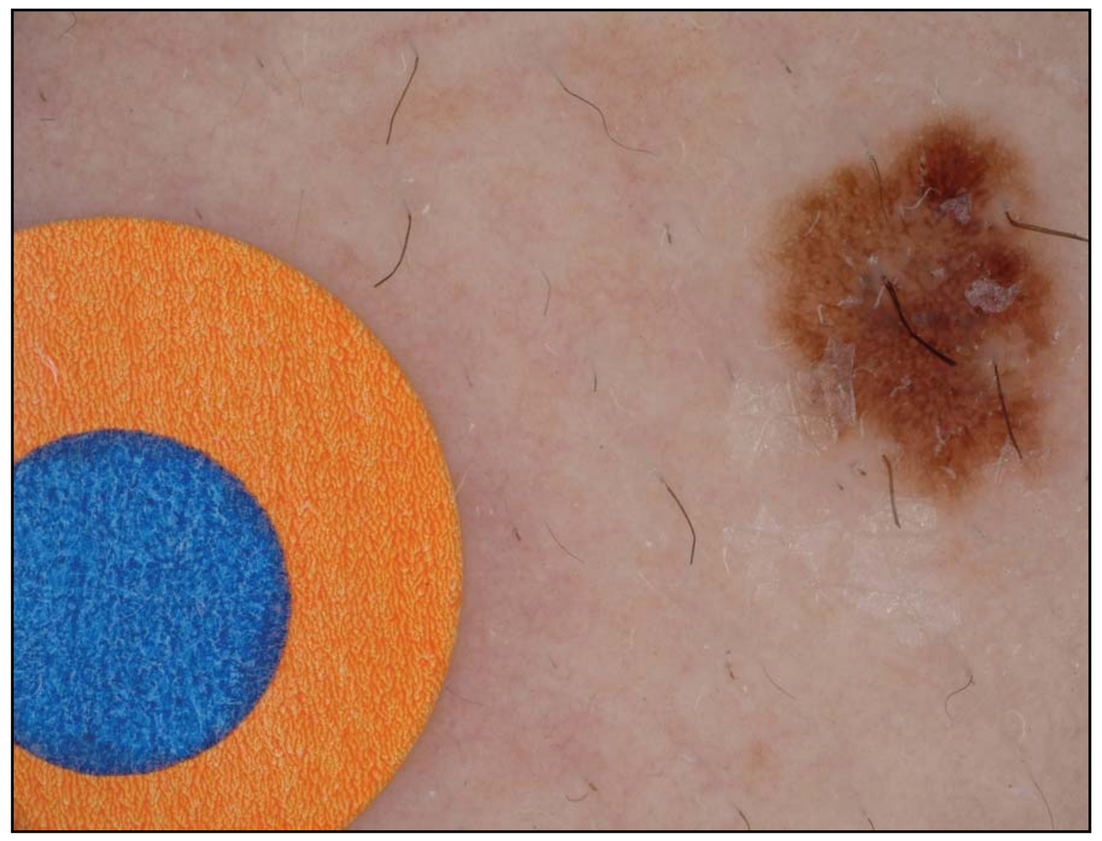
Diversity Shift

Clinical vs. Dermatoscopical

Correlation Shift

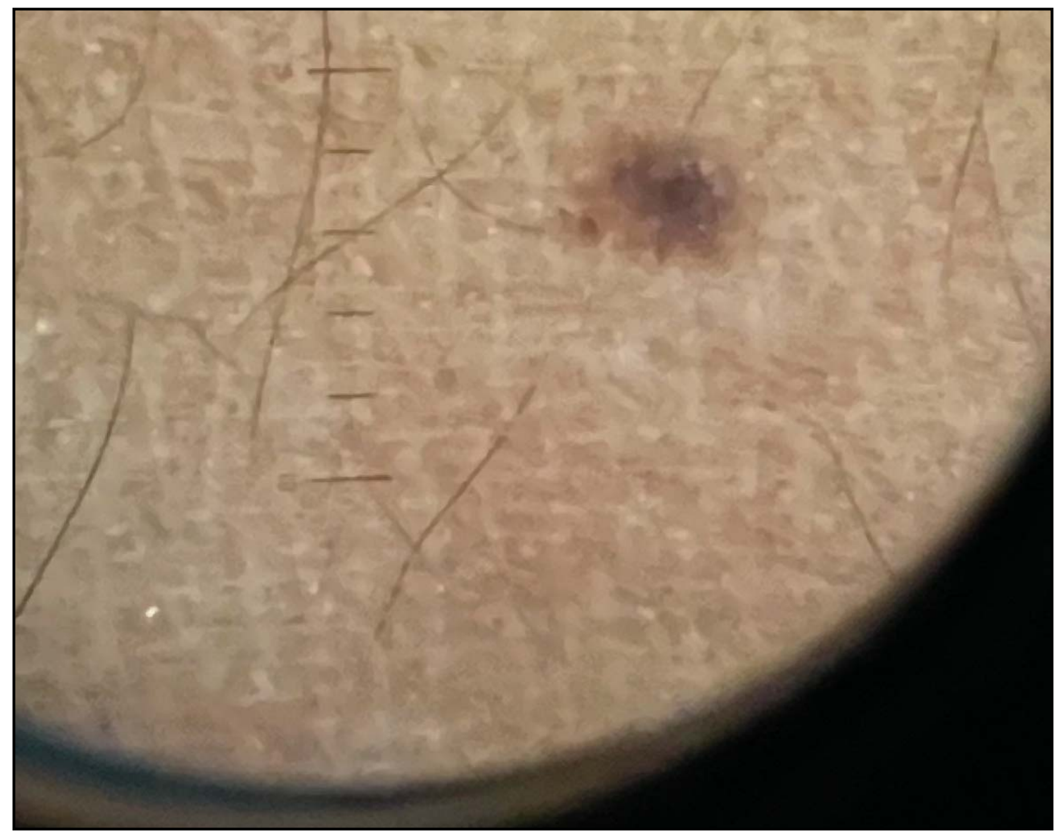
Artifacts





Subpopulation Shift

Underrepresented Skin Colors



De(Constructing) Bias

ISIC Workshop @ CVPR 2019

(De)Constructing Bias on Skin Lesion Datasets

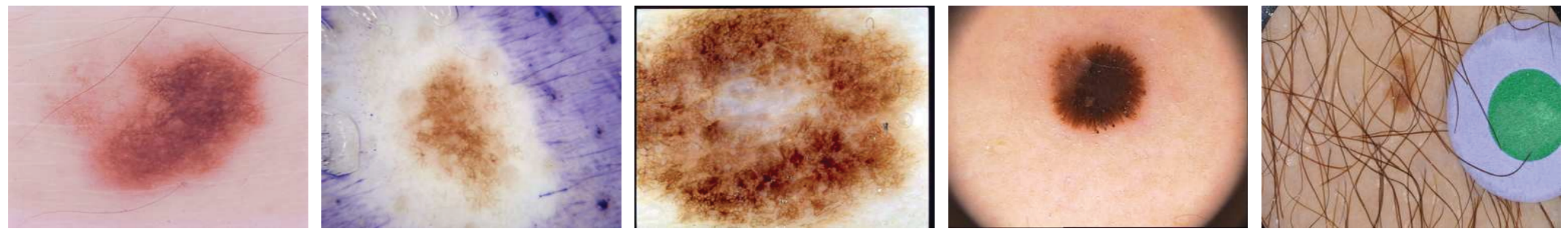
Alceu Bissoto[1] Michel Fornaciali[2] Eduardo Valle[2] Sandra Avila[1]
[1]Institute of Computing (IC) [2]School of Electrical and Computing Engineering (FEEC)
RECOD Lab., University of Campinas (UNICAMP), Brazil

Abstract

Melanoma is the deadliest form of skin cancer. Automated skin lesion analysis plays an important role for early detection. Nowadays, the ISIC Archive and the Atlas of Dermoscopy dataset are the most employed skin lesion sources to benchmark deep-learning based tools. However, all datasets contain biases, often unintentional, due to how they were acquired and annotated. Those biases distort the performance of machine-learning models, creating spurious correlations that the models can unfairly exploit, or, contrarily destroying cogent correlations that the models could learn. In this paper, we propose a set of experiments that reveal both types of biases, positive and negative, in existing skin lesion datasets. Our results show

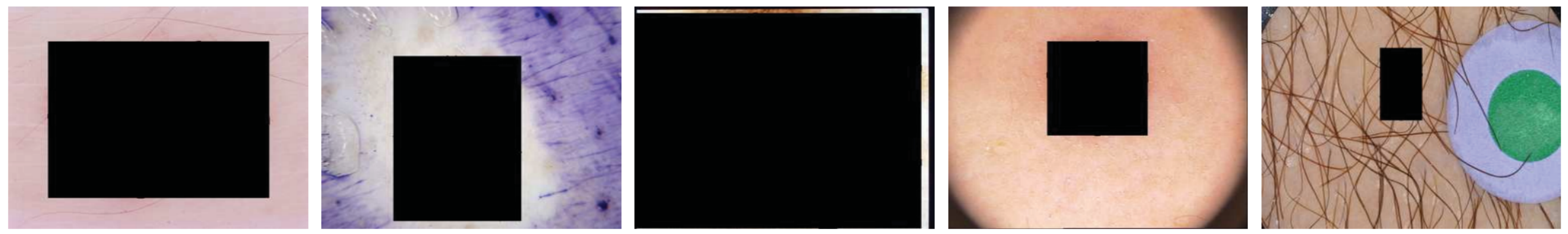
Deep learning methods are the state-of-the-art
cancer classification [11, 13]. That task is challengi
to the vast visual variability of skin lesions, and the s
of the cues that differentiate benign and malignant
To compound the difficulty, datasets to train the data-
models are small, when compared with general-purp
age datasets (*e.g.*, ImageNet, MSCOCO, LabelMe).

Due to the scarcity of good-quality, annotated skin images, two datasets dominate research on automated skin lesion analysis: the Interactive Atlas of Dermoscopy [5] 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.

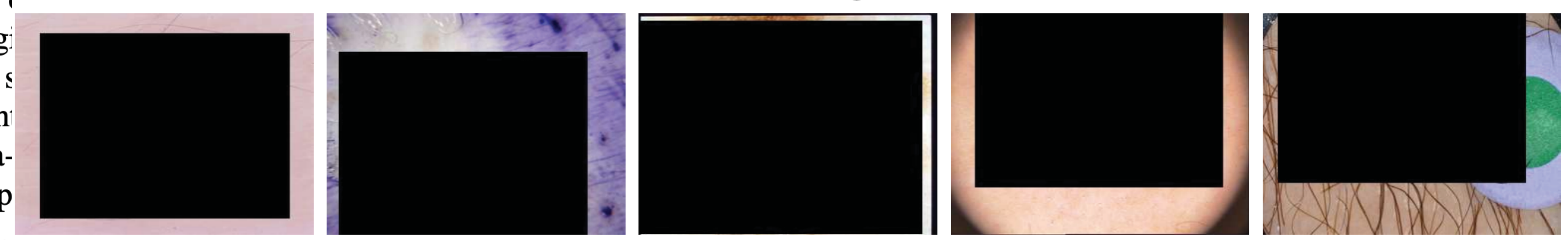


(a) Traditional images

(b) Only Skin images



(c) Bbox images



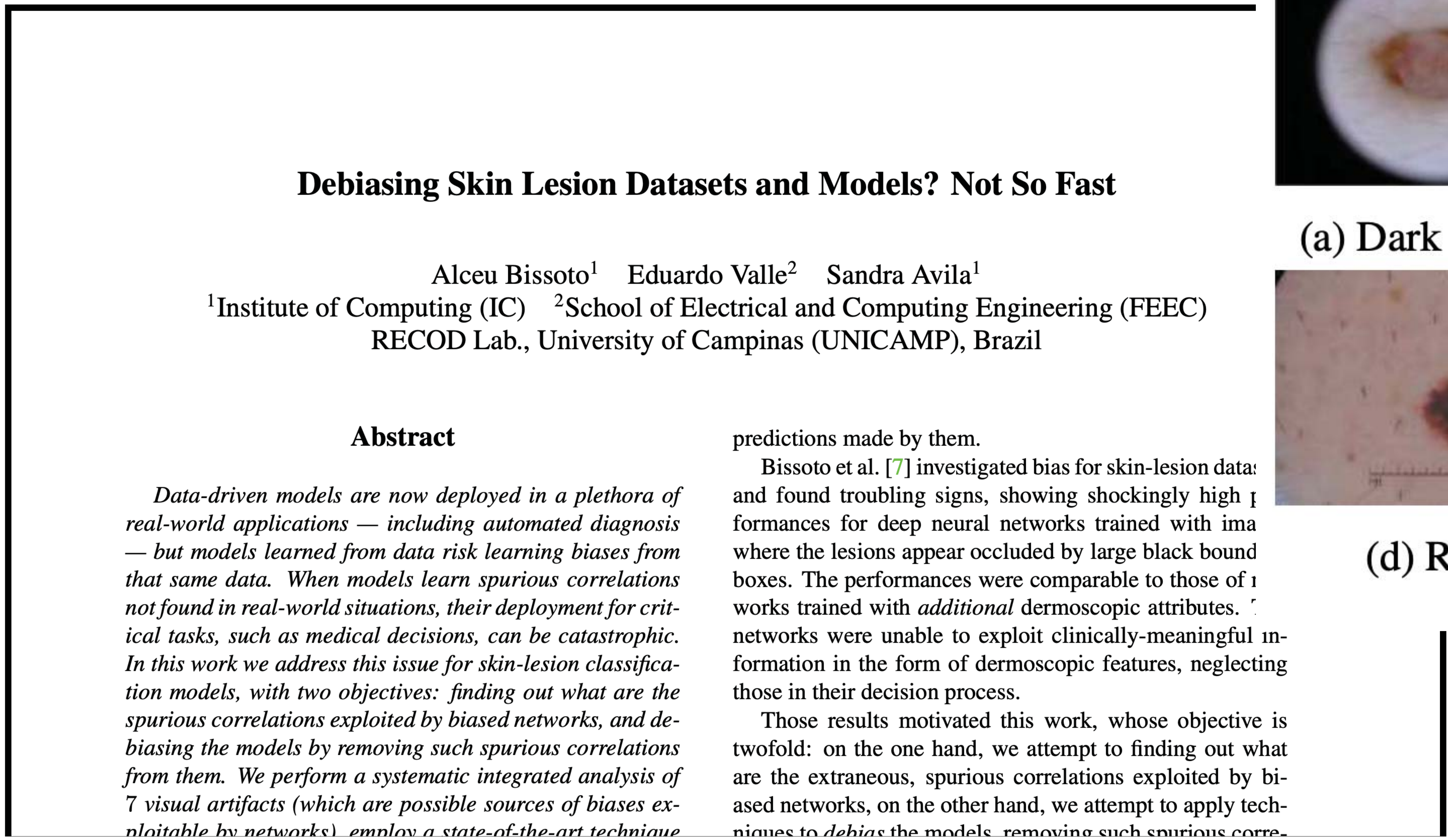
(d) Bbox70 images

Benchmarks

Robustness

Debiasing on Skin Lesion Analysis Models

ISIC Workshop @ CVPR 2020



Debiasing Skin Lesion Datasets and Models? Not So Fast

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

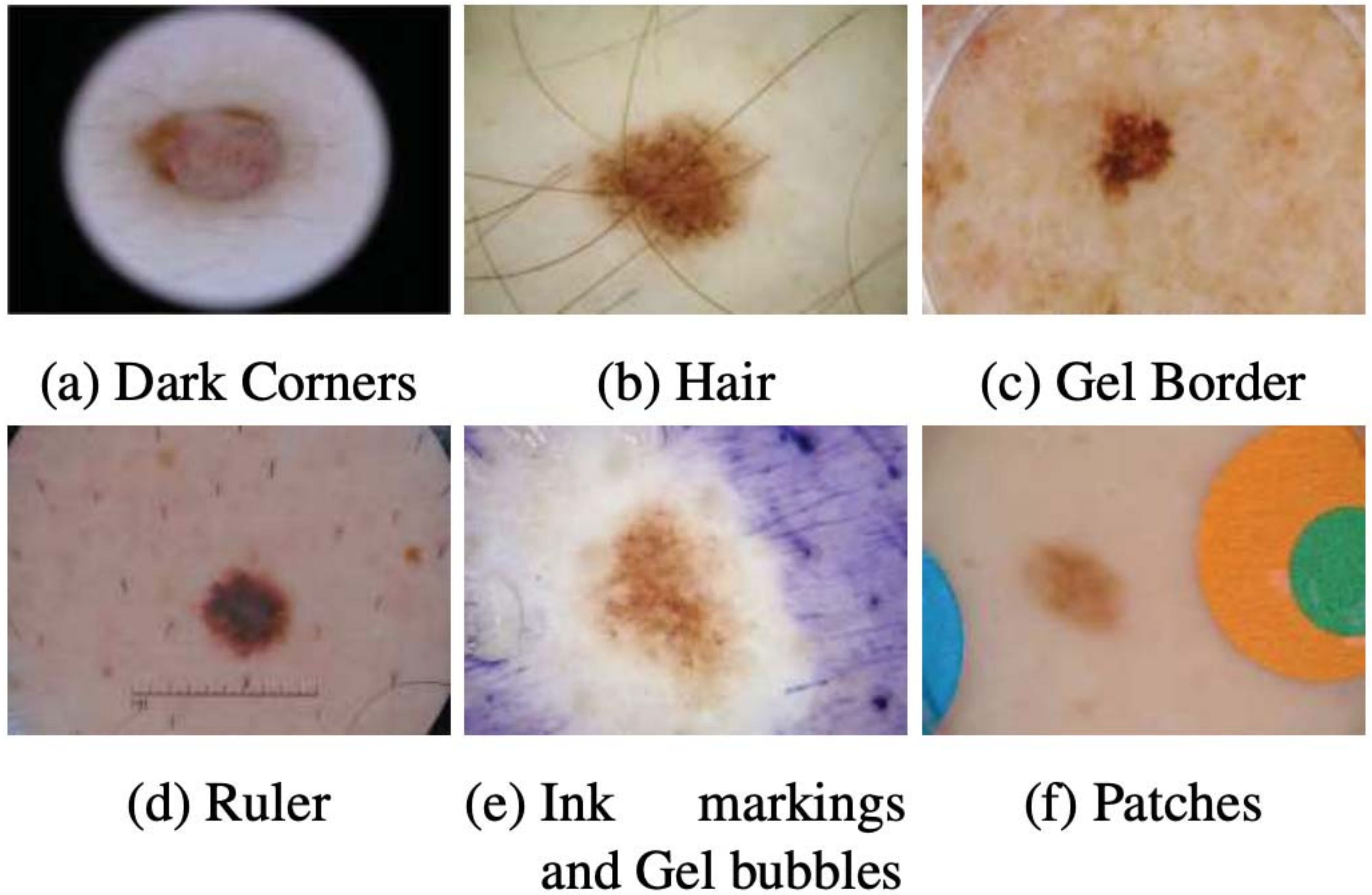
Data-driven models are now deployed in a plethora of real-world applications — including automated diagnosis — but models learned from data risk learning biases from that same data. When models learn spurious correlations not found in real-world situations, their deployment for critical tasks, such as medical decisions, can be catastrophic. In this work we address this issue for skin-lesion classification models, with two objectives: finding out what are the spurious correlations exploited by biased networks, and debiasing the models by removing such spurious correlations from them. We perform a systematic integrated analysis of 7 visual artifacts (which are possible sources of biases exploitable by networks), employ a state-of-the-art technique

predictions made by them.

Bissoto et al. [7] investigated bias for skin-lesion data and found troubling signs, showing shockingly high p formances for deep neural networks trained with ima where the lesions appear occluded by large black bound boxes. The performances were comparable to those of works trained with *additional* dermoscopic attributes. networks were unable to exploit clinically-meaningful in formation in the form of dermoscopic features, neglecting those in their decision process.

Those results motivated this work, whose objective is twofold: on the one hand, we attempt to finding out what are the extraneous, spurious correlations exploited by biased networks, on the other hand, we attempt to apply techniques to *debias* the models, removing such spurious corre-

Manually annotated ISIC 2018 and Derm7Pt



Debiasing on Skin Lesion Analysis Models

ISIC Workshop @ CVPR 2020

Domain Generalization

Debiasing Skin Lesion Datasets and Models? Not So Fast

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

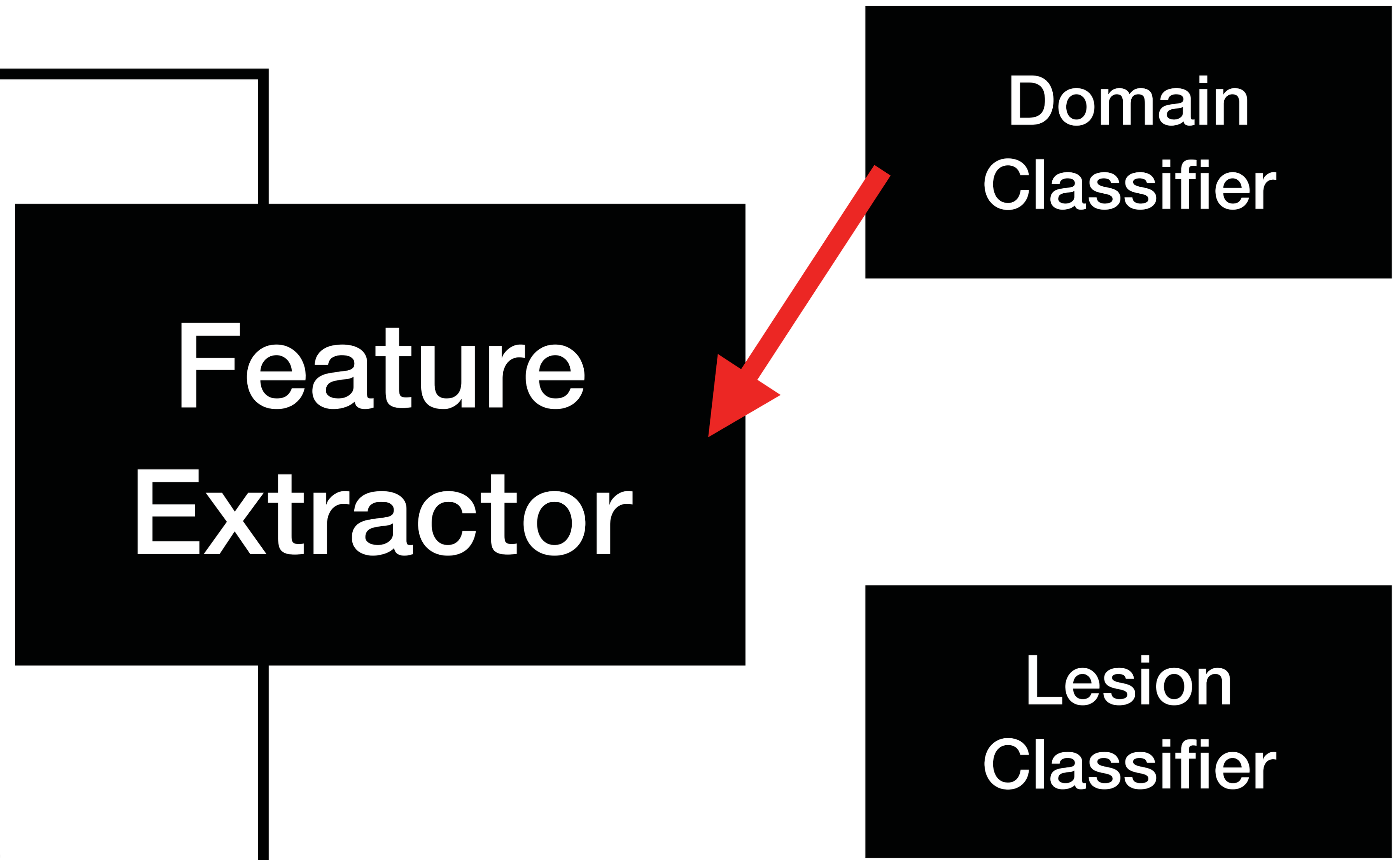
Abstract

Data-driven models are now deployed in a plethora of real-world applications — including automated diagnosis — but models learned from data risk learning biases from that same data. When models learn spurious correlations not found in real-world situations, their deployment for critical tasks, such as medical decisions, can be catastrophic. In this work we address this issue for skin-lesion classification models, with two objectives: finding out what are the spurious correlations exploited by biased networks, and debiasing the models by removing such spurious correlations from them. We perform a systematic integrated analysis of 7 visual artifacts (which are possible sources of biases exploitable by networks), employ a state-of-the-art technique

predictions made by them.

Bissoto et al. [7] investigated bias for skin-lesion datasets and found troubling signs, showing shockingly high performances for deep neural networks trained with images where the lesions appear occluded by large black bounding boxes. The performances were comparable to those of networks trained with *additional* dermoscopic attributes. The networks were unable to exploit clinically-meaningful information in the form of dermoscopic features, neglecting those in their decision process.

Those results motivated this work, whose objective is twofold: on the one hand, we attempt to finding out what are the extraneous, spurious correlations exploited by biased networks, on the other hand, we attempt to apply techniques to *debias* the models, removing such spurious corre-



Artifact-based Domain Generalization

ISIC Workshop @ ECCV 2022

15 p.p. improvement

in biased scenarios

Artifact-based Domain Generalization
of Skin Lesion Models

Alceu Bissoto[0000-0003-2293-6160][1,4], Catarina Barata[0000-0002-2852-7723][2], Eduardo Valle[0000-0001-5396-9868][3,4], and Sandra Avila[0000-0001-9068-938X][1,4]

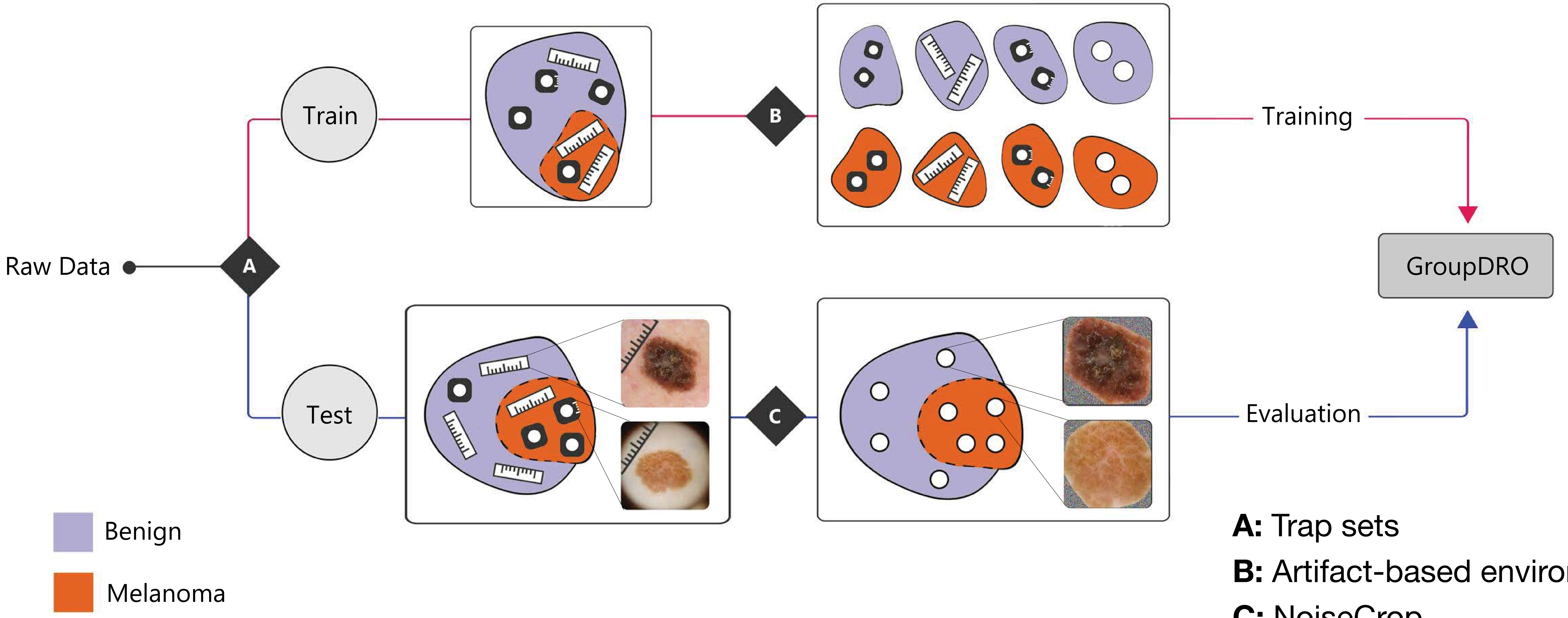
[1] Institute of Computing, University of Campinas, Brazil
{alceubissoto, sandra}@ic.unicamp.br
[2] Institute for Systems and Robotics, Instituto Superior Técnico, Portugal
ana.c.fidalgo.barata@tecnico.ulisboa.pt
[3] School of Electrical and Computing Engineering, University of Campinas, Brazil
dovalle@dca.fee.unicamp.br
[4] Recod.ai Lab, University of Campinas, Brazil

Abstract. Deep Learning failure cases are abundant, particularly in the medical area. Recent studies in out-of-distribution generalization have advanced considerably on well-controlled synthetic datasets, but they do not represent medical imaging contexts. We propose a pipeline that relies

Methodology

Debiasing Pipeline

Overview

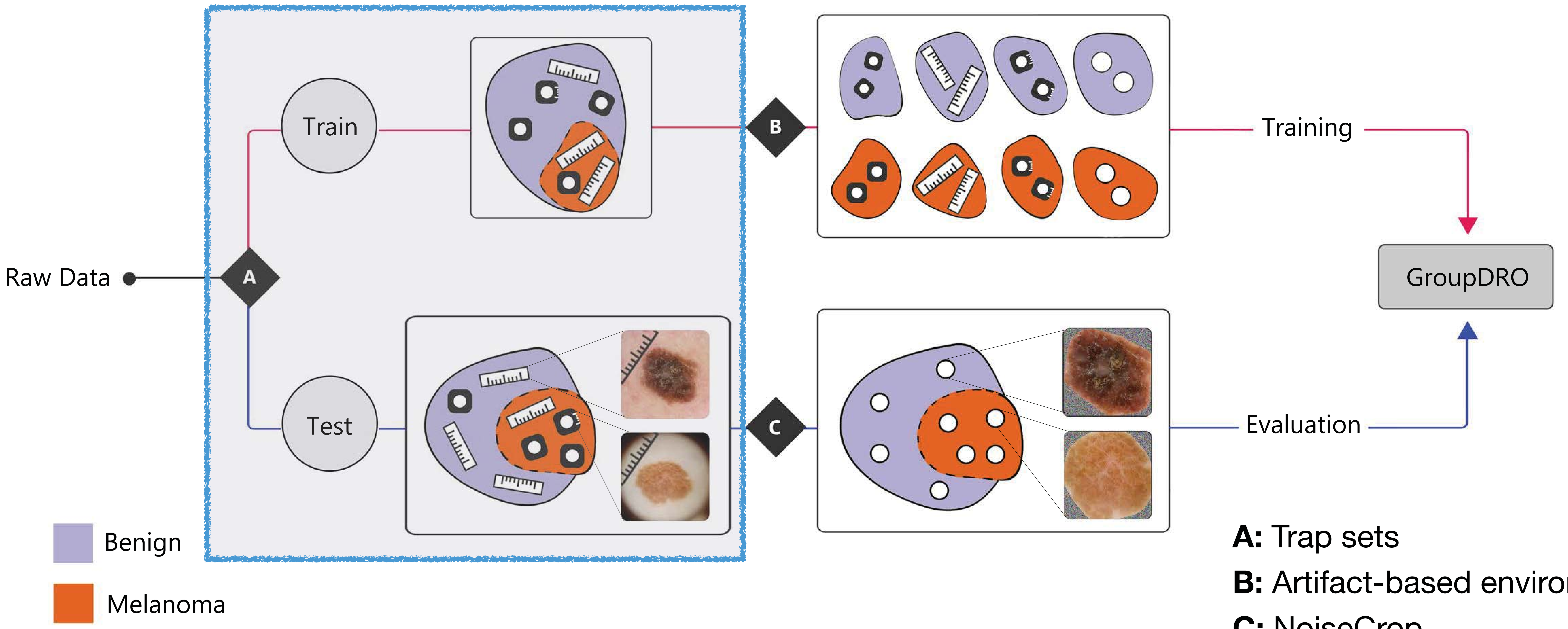


A: Trap sets
B: Artifact-based environments
C: NoiseCrop

Bissoto et al., "Artifact-based domain generalization of skin lesion models", ISIC Workshop @ ECCV 2022

Debiasing Pipeline

Overview



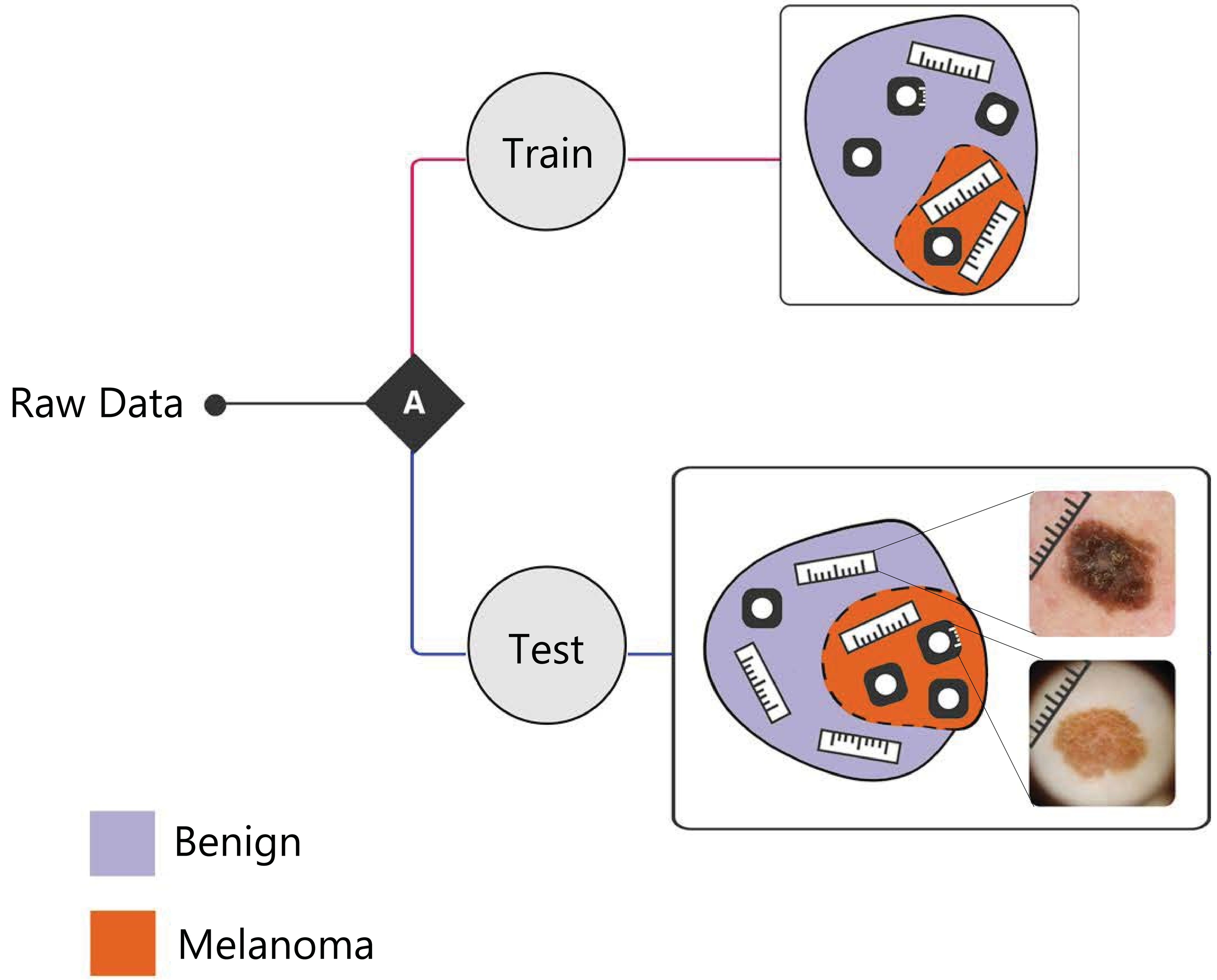
A: Trap sets
B: Artifact-based environments
C: NoiseCrop

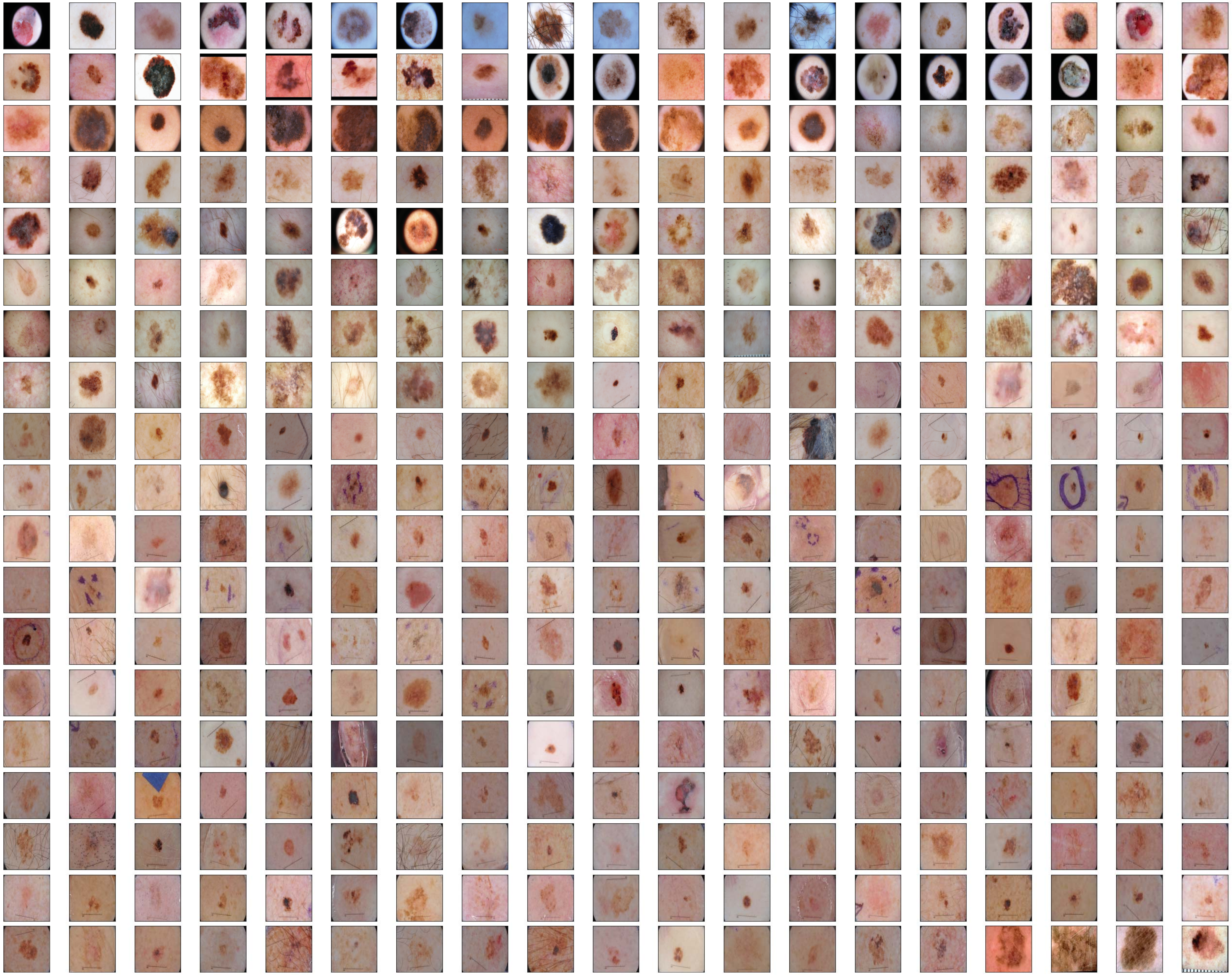
Bissoto et al., "Artifact-based domain generalization of skin lesion models", ISIC Workshop @ ECCV 2022

Trap Sets

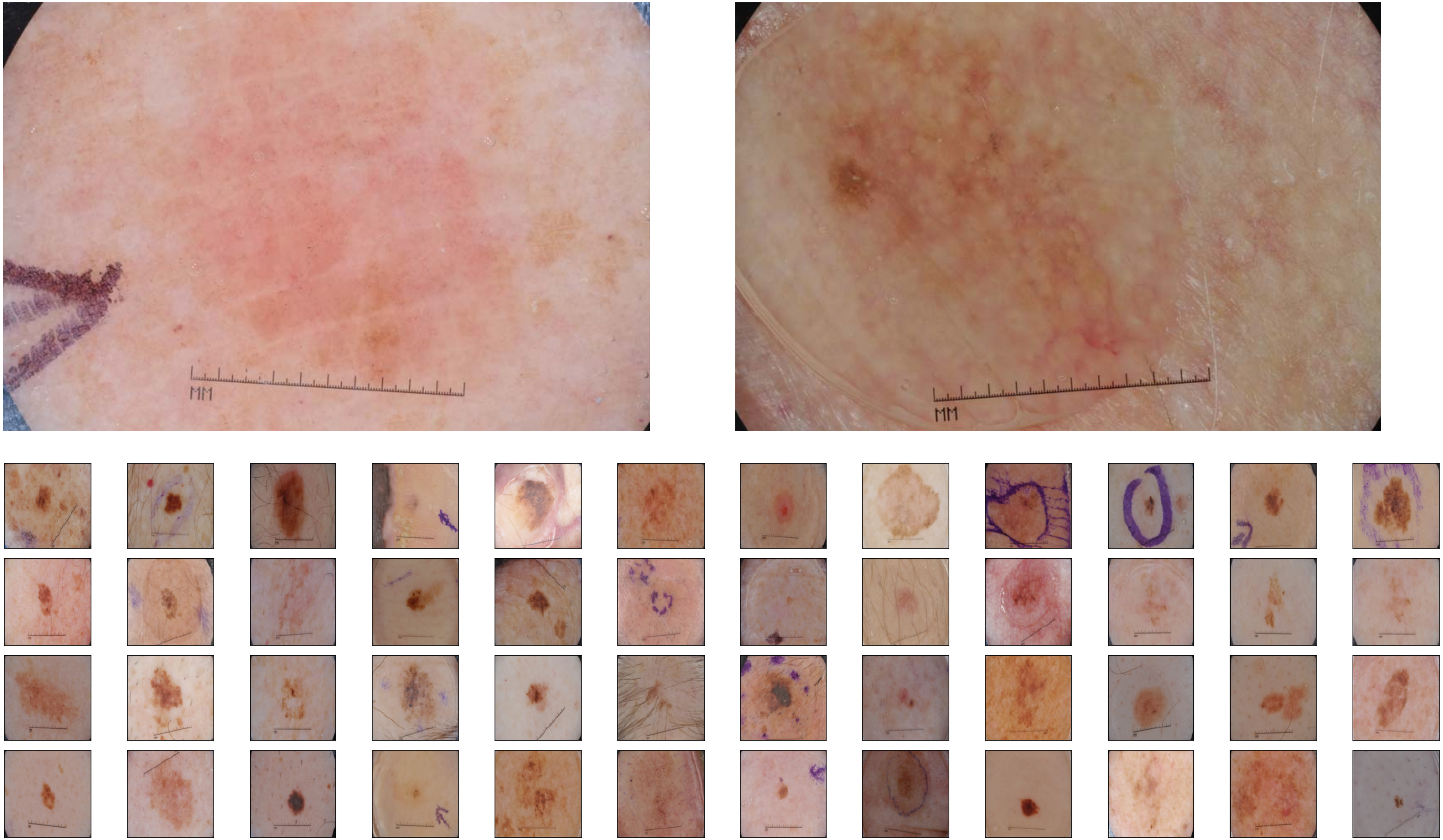
Debiasing Pipeline

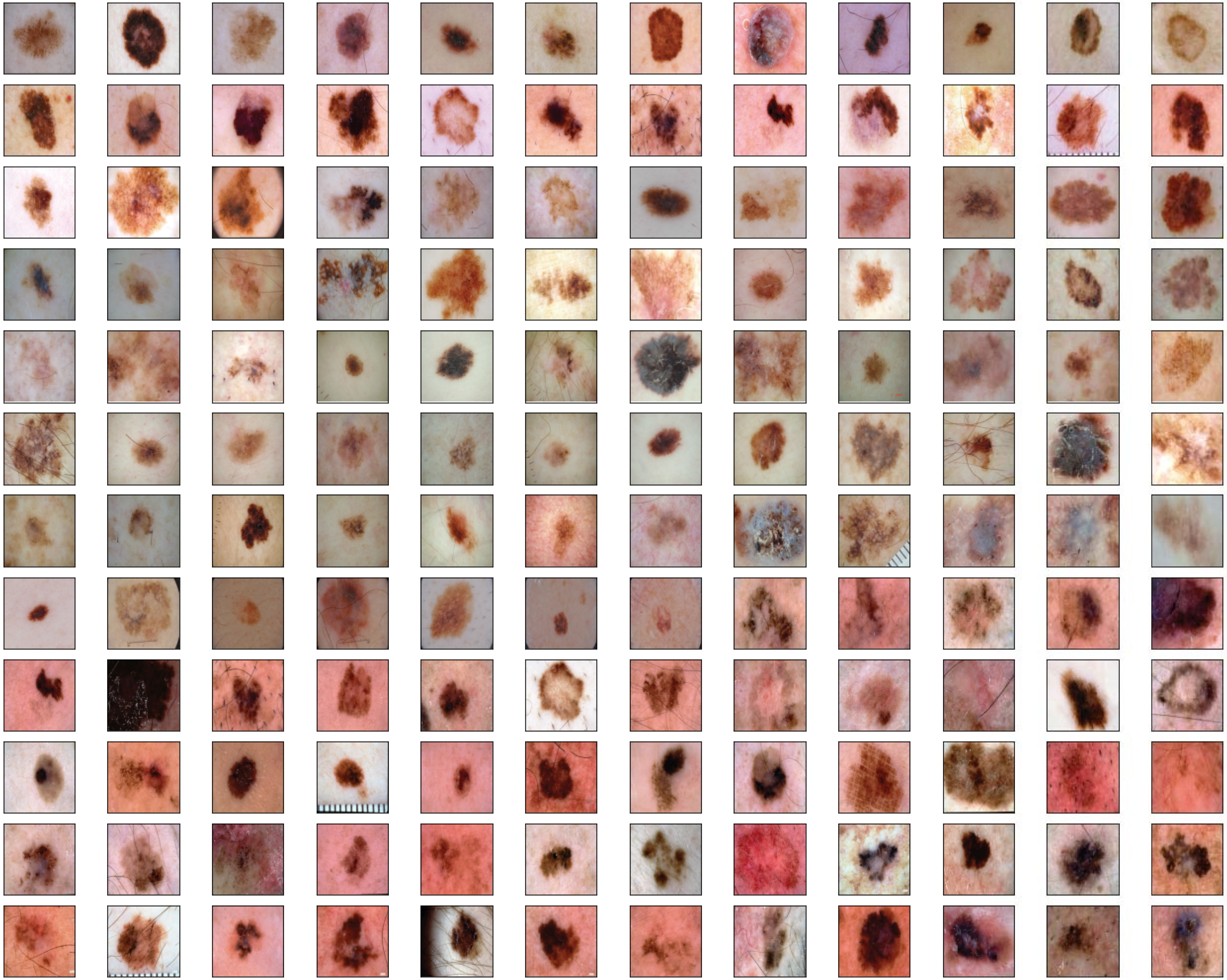
- **Control and amplify** the level of bias during training.
- Creating challenging test sets with **opposite correlations.**





Trap Train


MM
MM

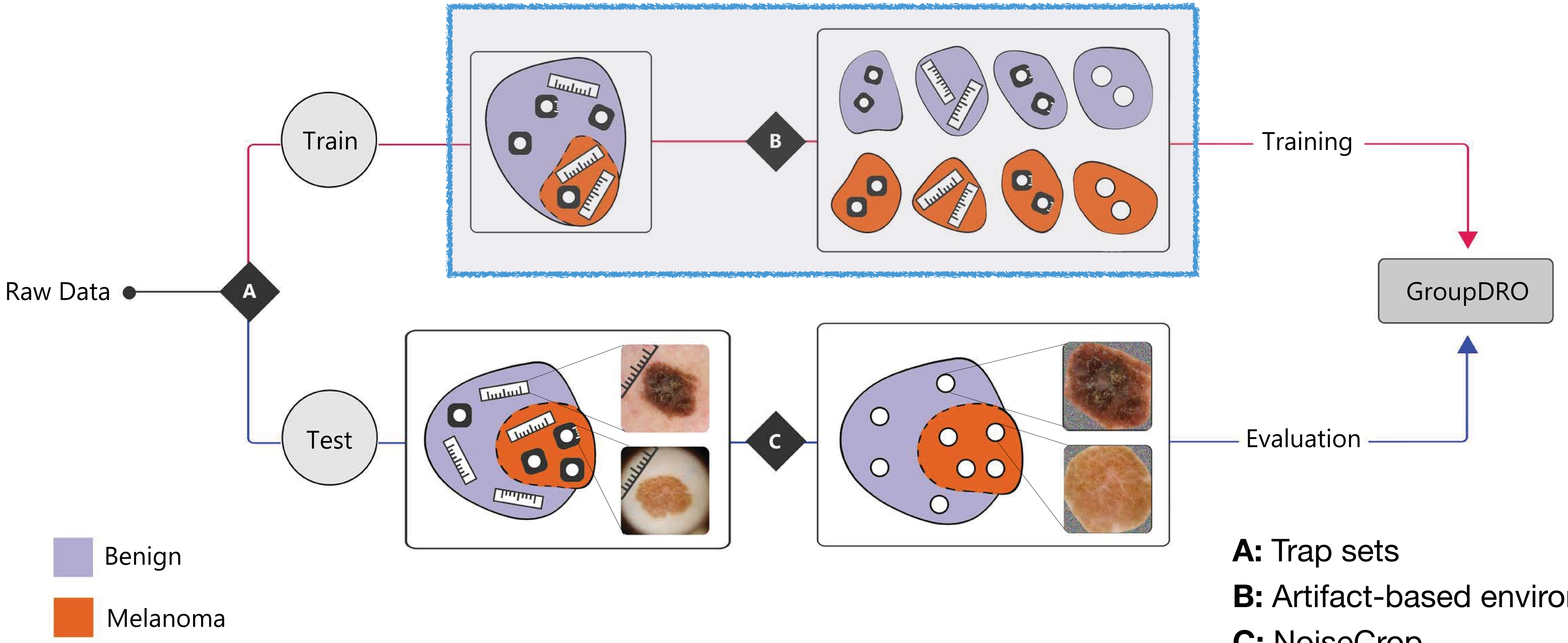


Trap Test

- ✗ No dark corners
- ✗ Few rulers
- ✗ No ink markings

Debiasing Pipeline

Overview

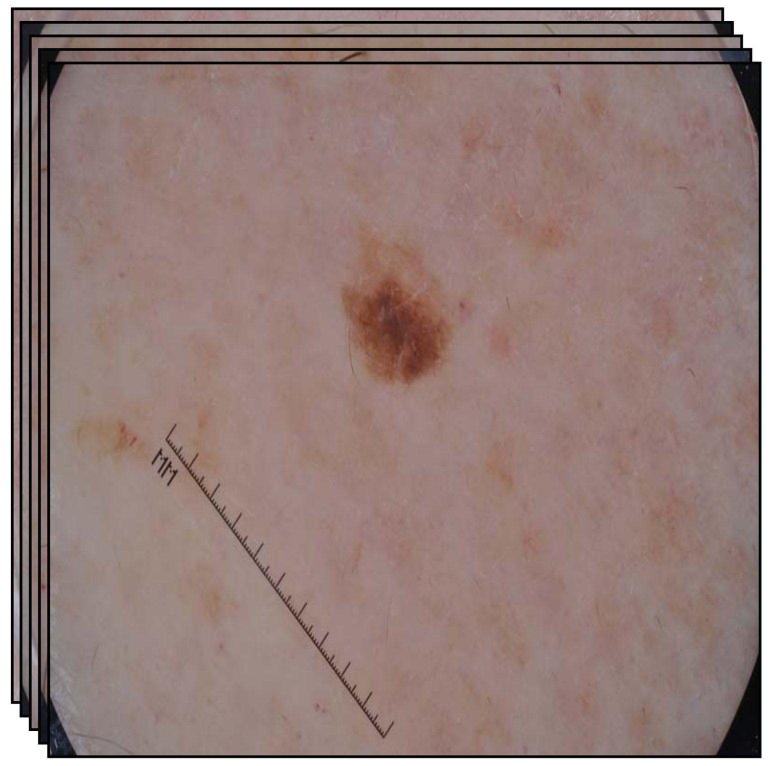


A: Trap sets
B: Artifact-based environments
C: NoiseCrop

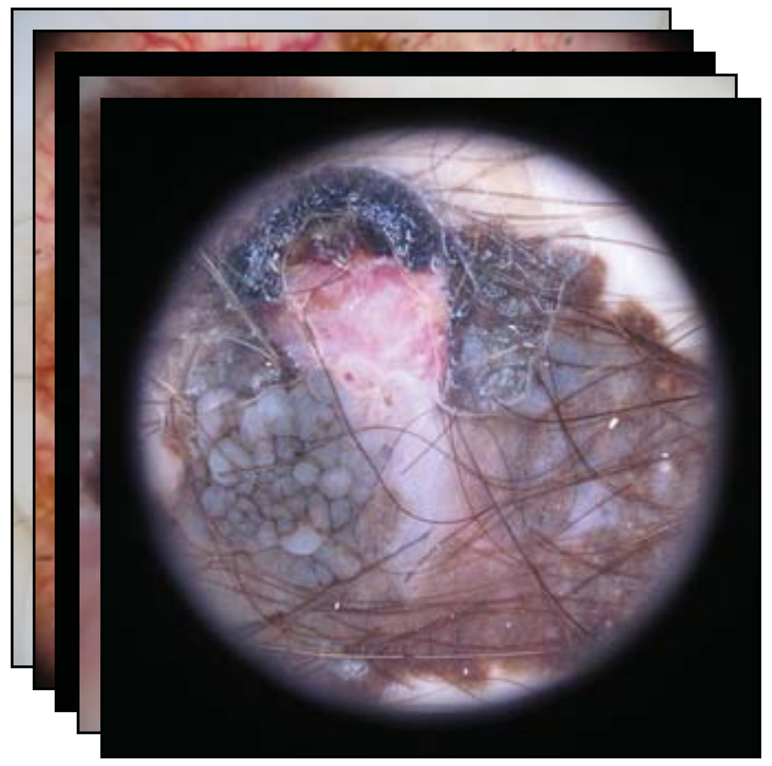
Artifact-based environments

Debiasing pipeline

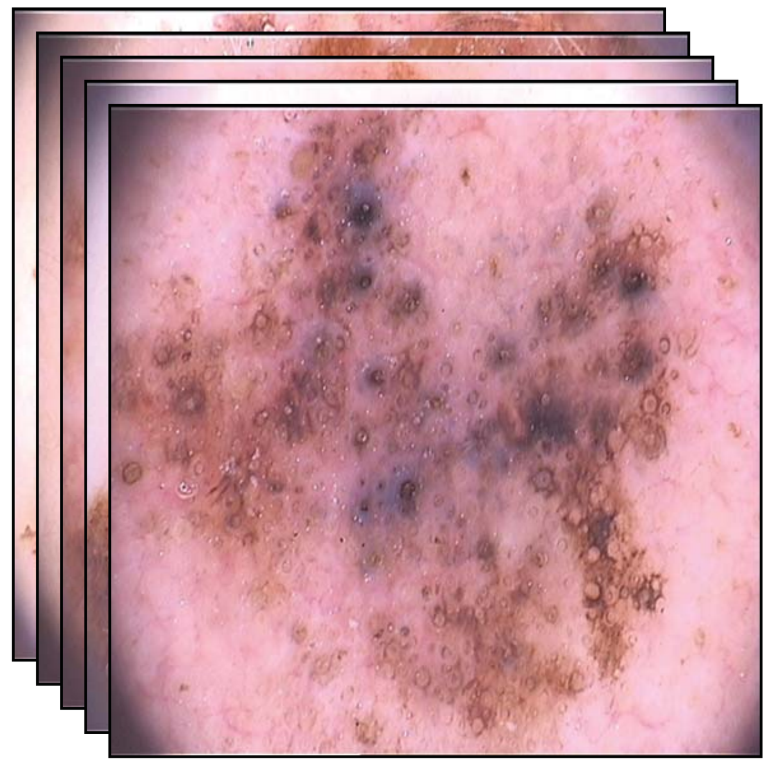
HAM10000



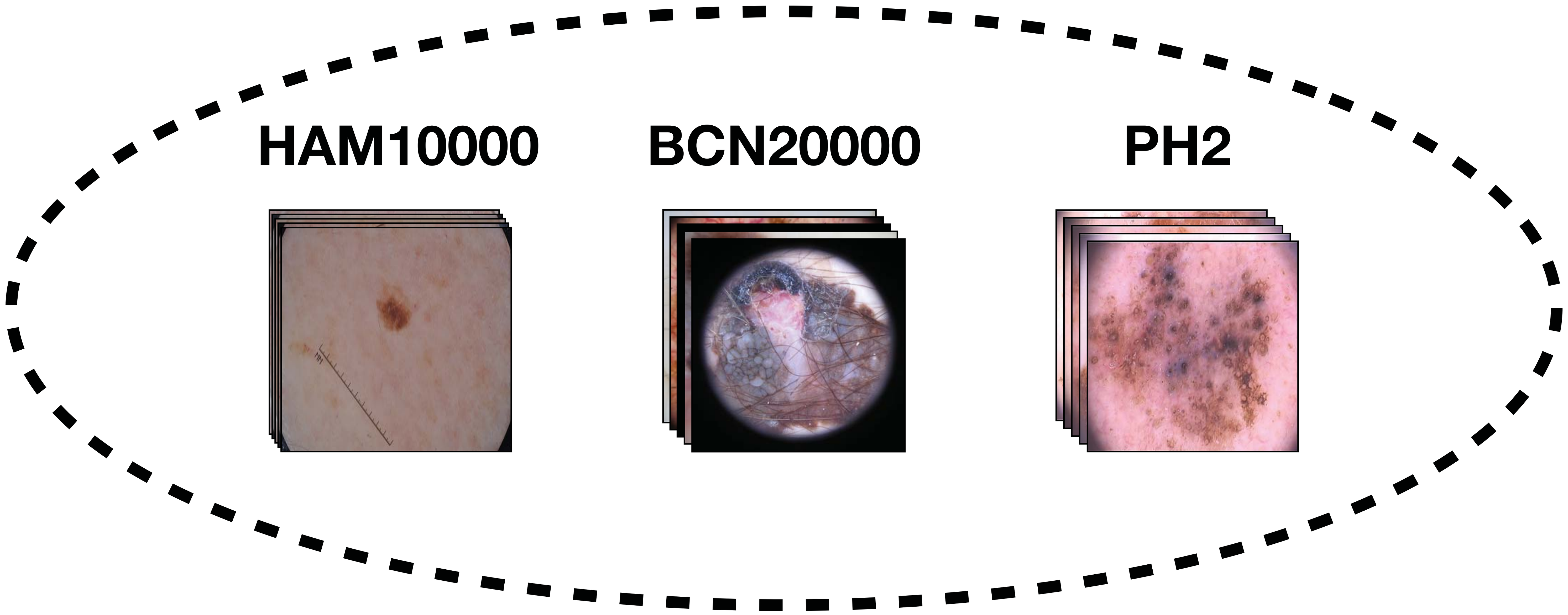
BCN20000



PH2



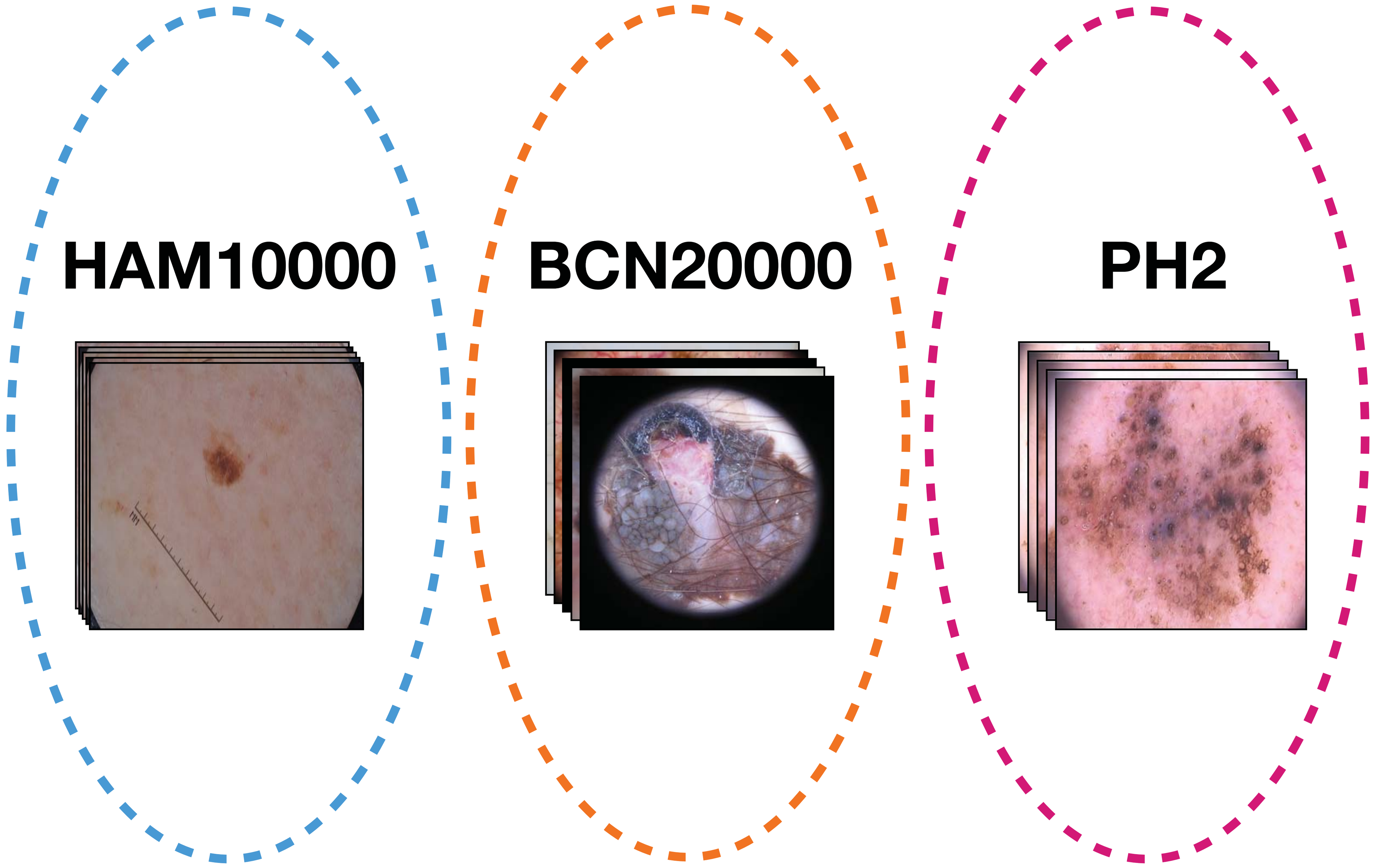
Classical Learning Method



Empirical Risk Minimization (ERM): Minimize the empirical risk among all samples (classical learning method)

"Principles of risk minimization for learning theory", Vapnik et al. NeurIPS 1991

Robust Learning

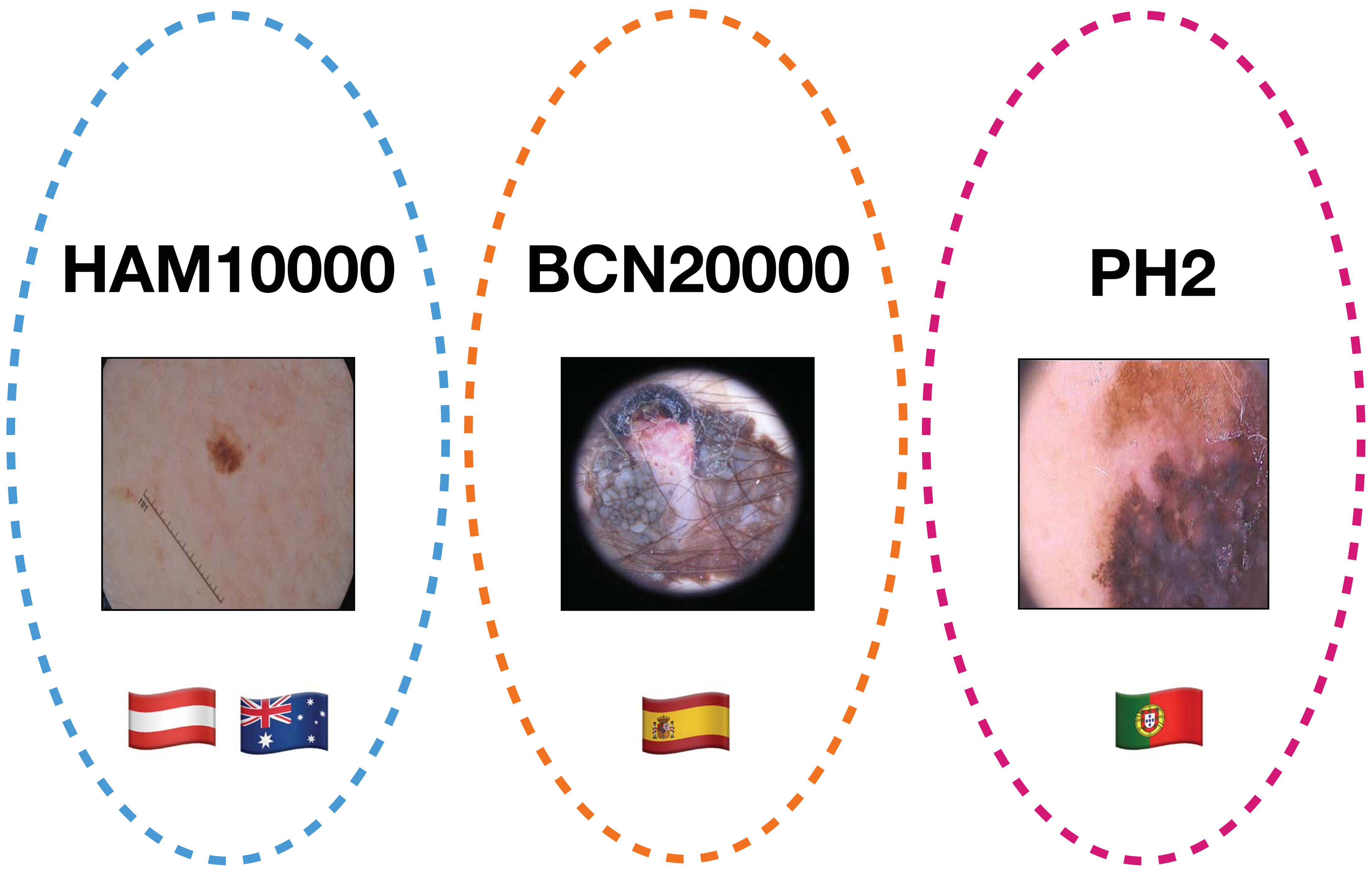


Distributionally Robust Optimization (DRO): Minimize the maximum risk across environments

"Distributionally Robust Neural Networks", Sagawa et al. ICLR 2020

Ideal Environments

Environments should differ in **single or few** aspects.



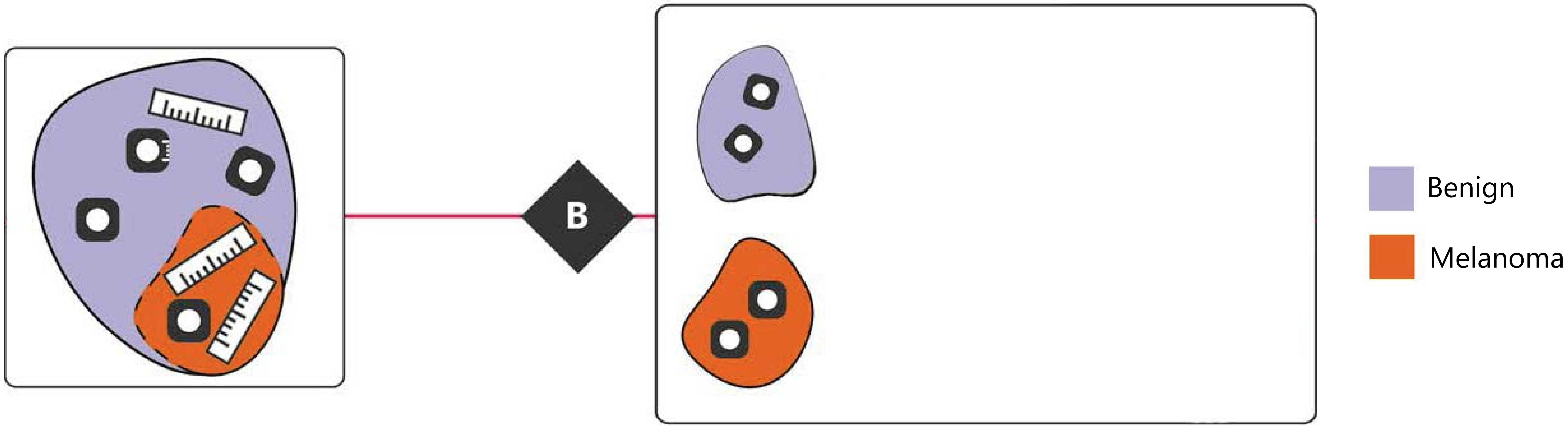
- Demographics
- Image acquisition devices
- Artifact distribution
- Artifact characteristics
- Class distribution

Bissoto et al., "Artifact-based domain generalization of skin lesion models", ISIC Workshop @ ECCV 2022

Artifact-based environments

Debiasing pipeline

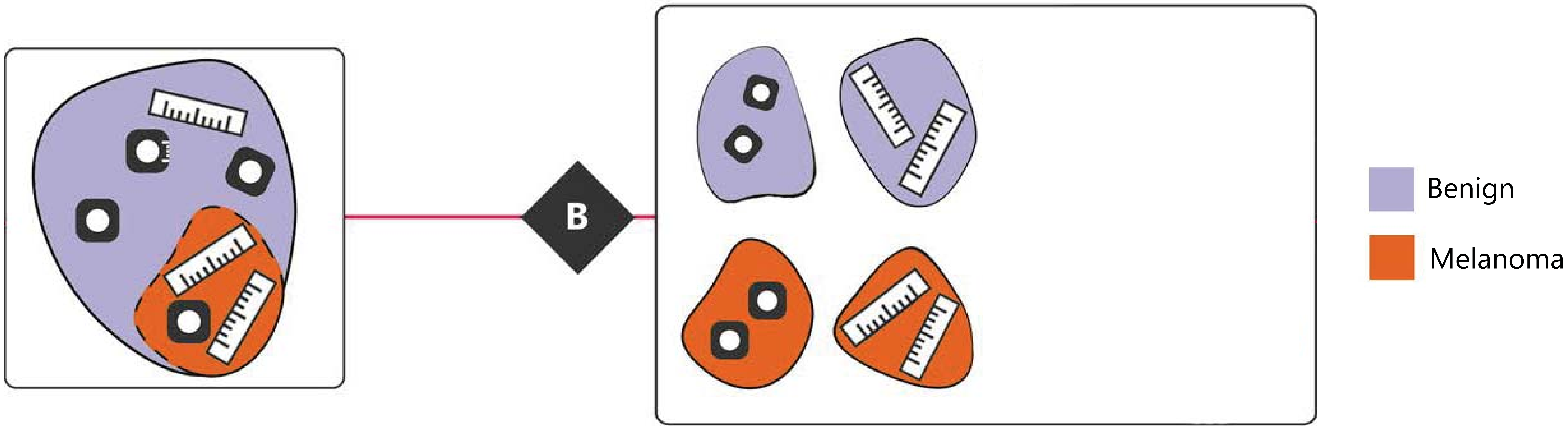
"**Separate** data into groups according to the presence of artifacts and its labels"



Artifact-based environments

Debiasing pipeline

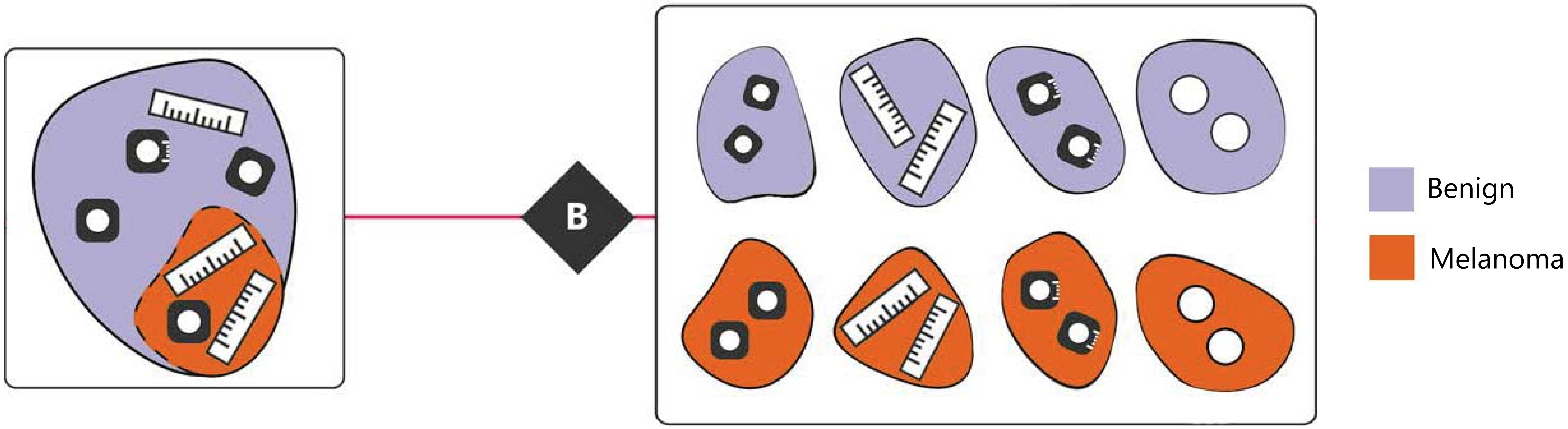
"**Separate** data into groups according to the presence of artifacts and its labels"



Artifact-based environments

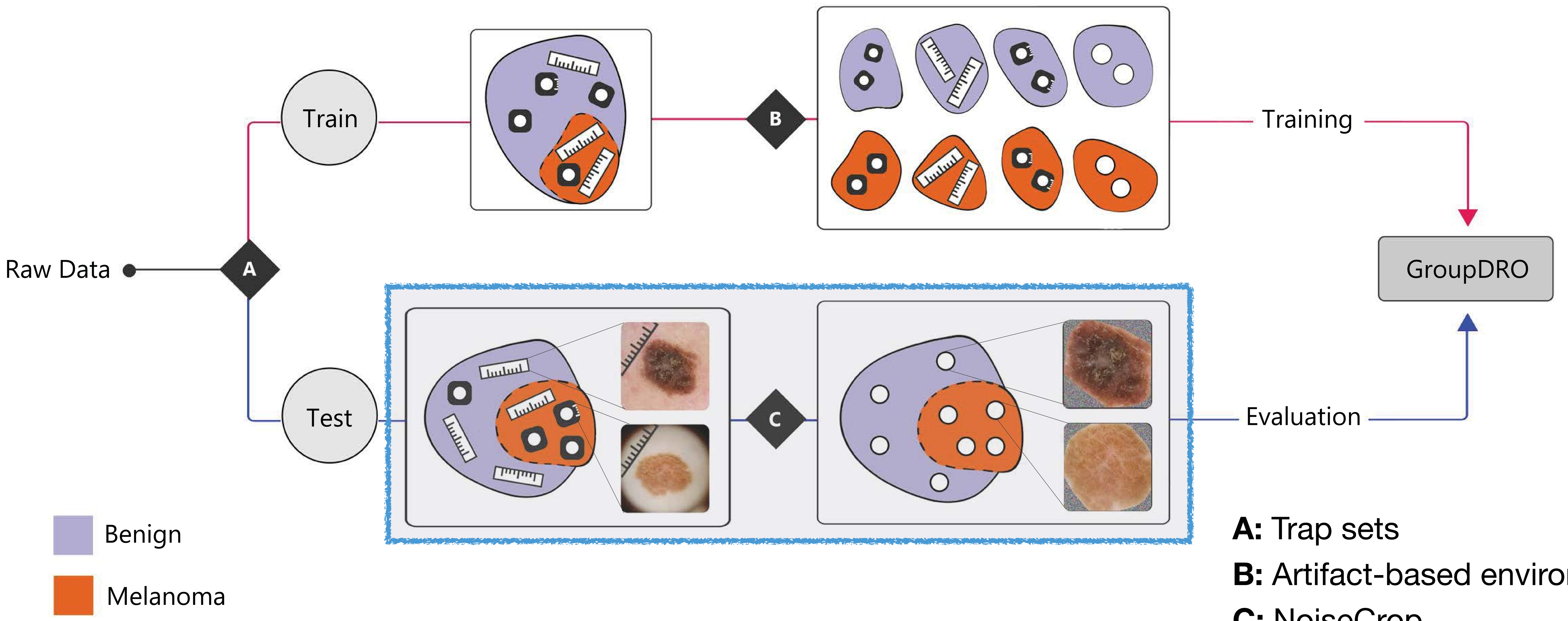
Debiasing pipeline

"**Separate** data into groups according to the presence of artifacts and its labels"



Debiasing Pipeline

Overview

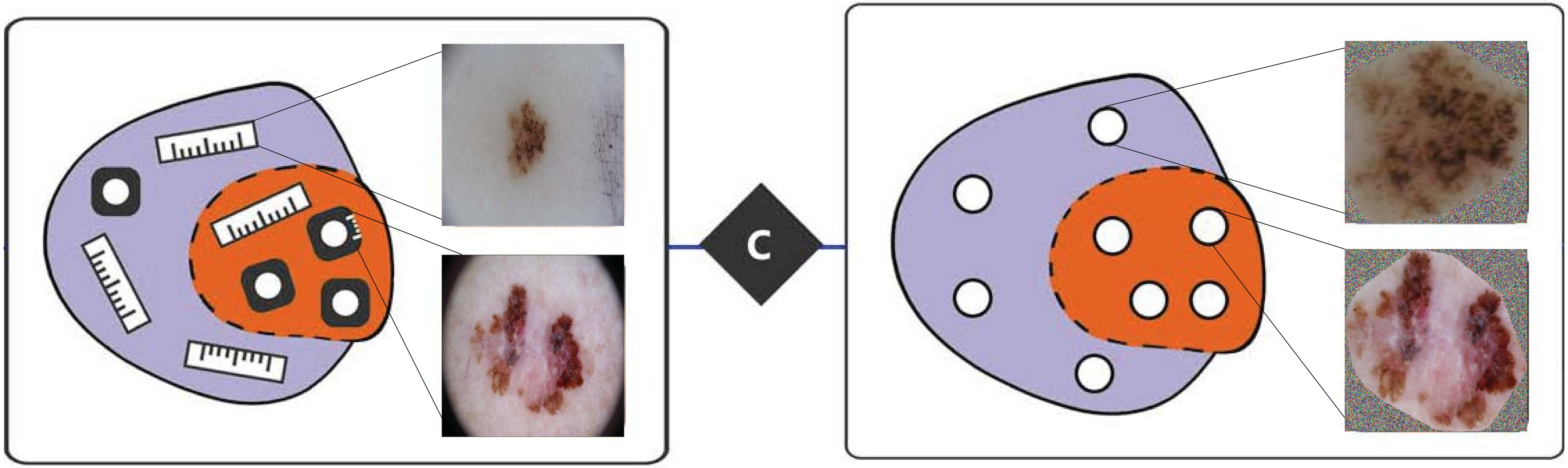


A: Trap sets
B: Artifact-based environments
C: NoiseCrop

NoiseCrop

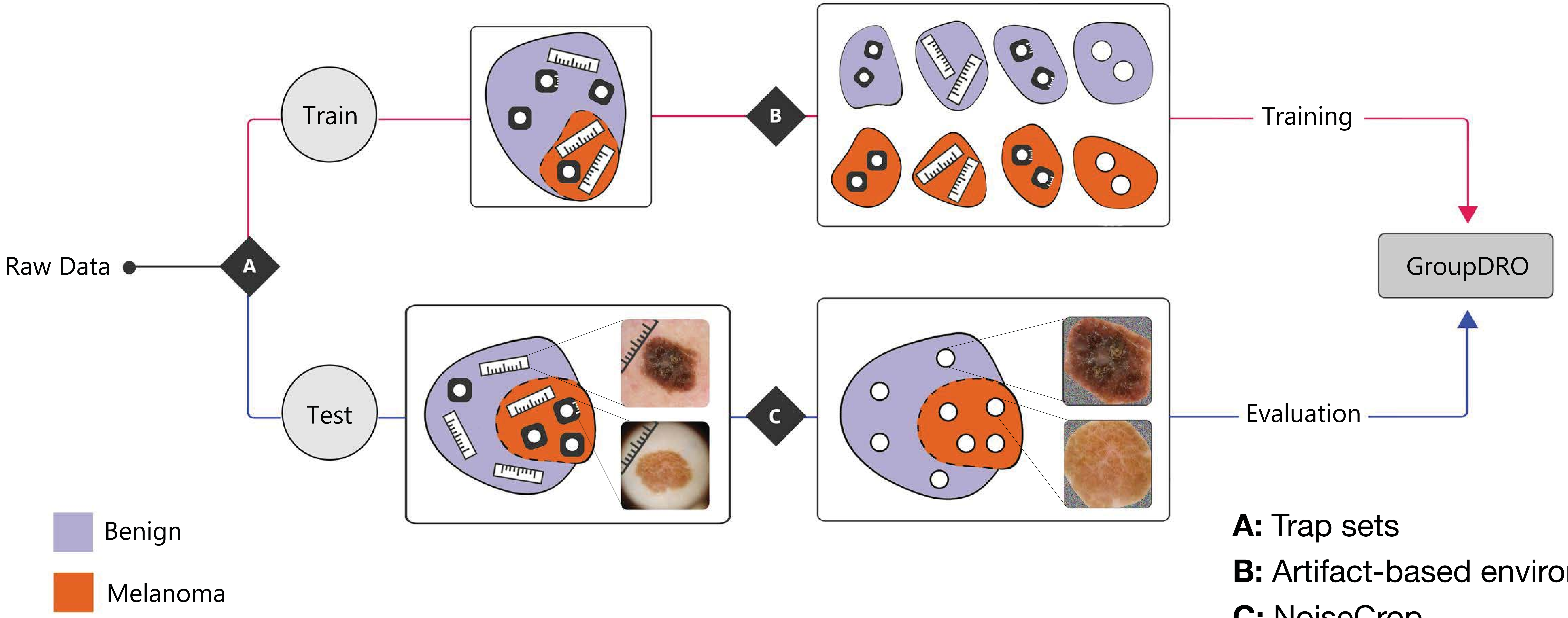
Debiasing pipeline

"Remove confounders from **test** samples"



Debiasing Pipeline

Overview



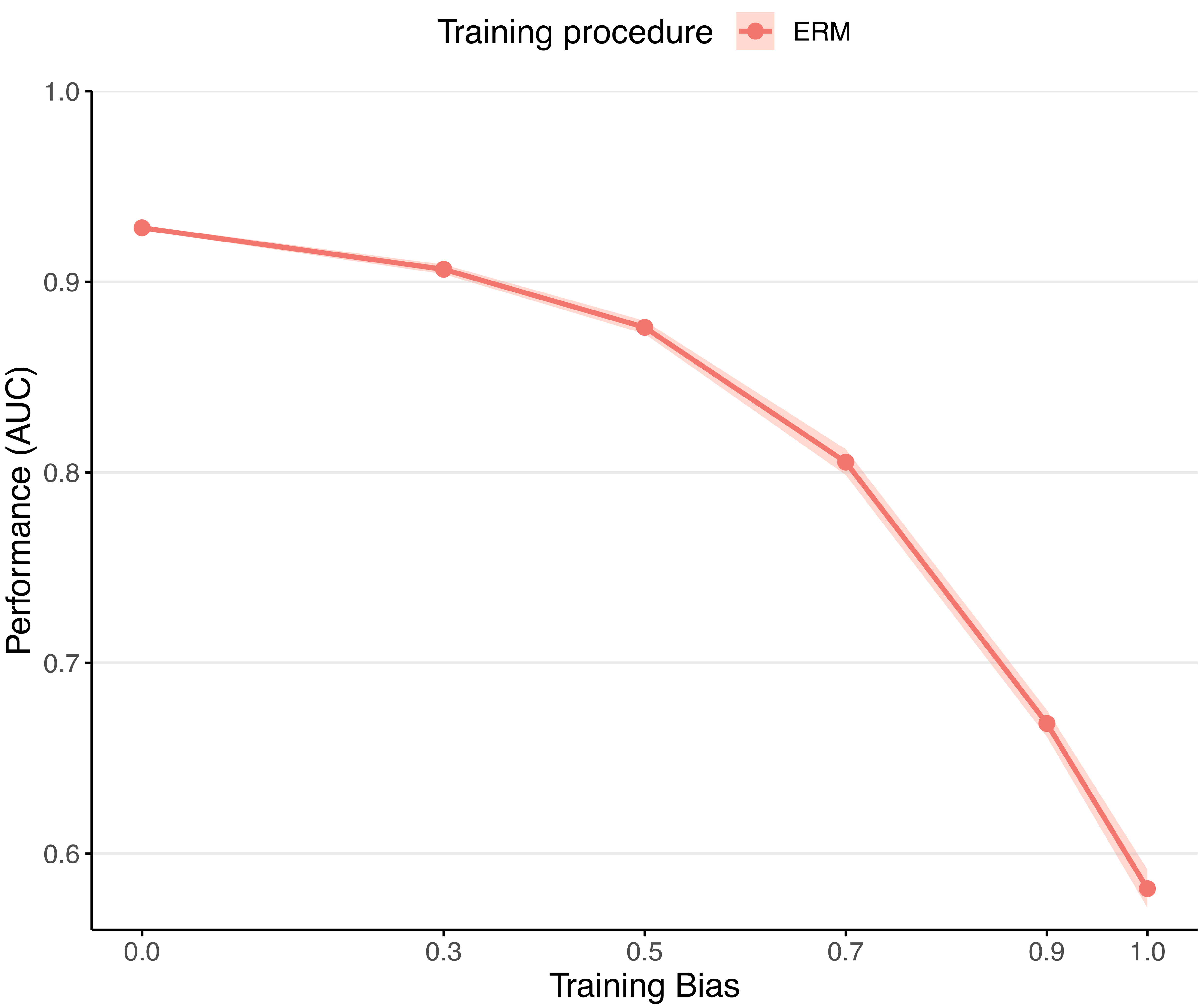
A: Trap sets
B: Artifact-based environments
C: NoiseCrop

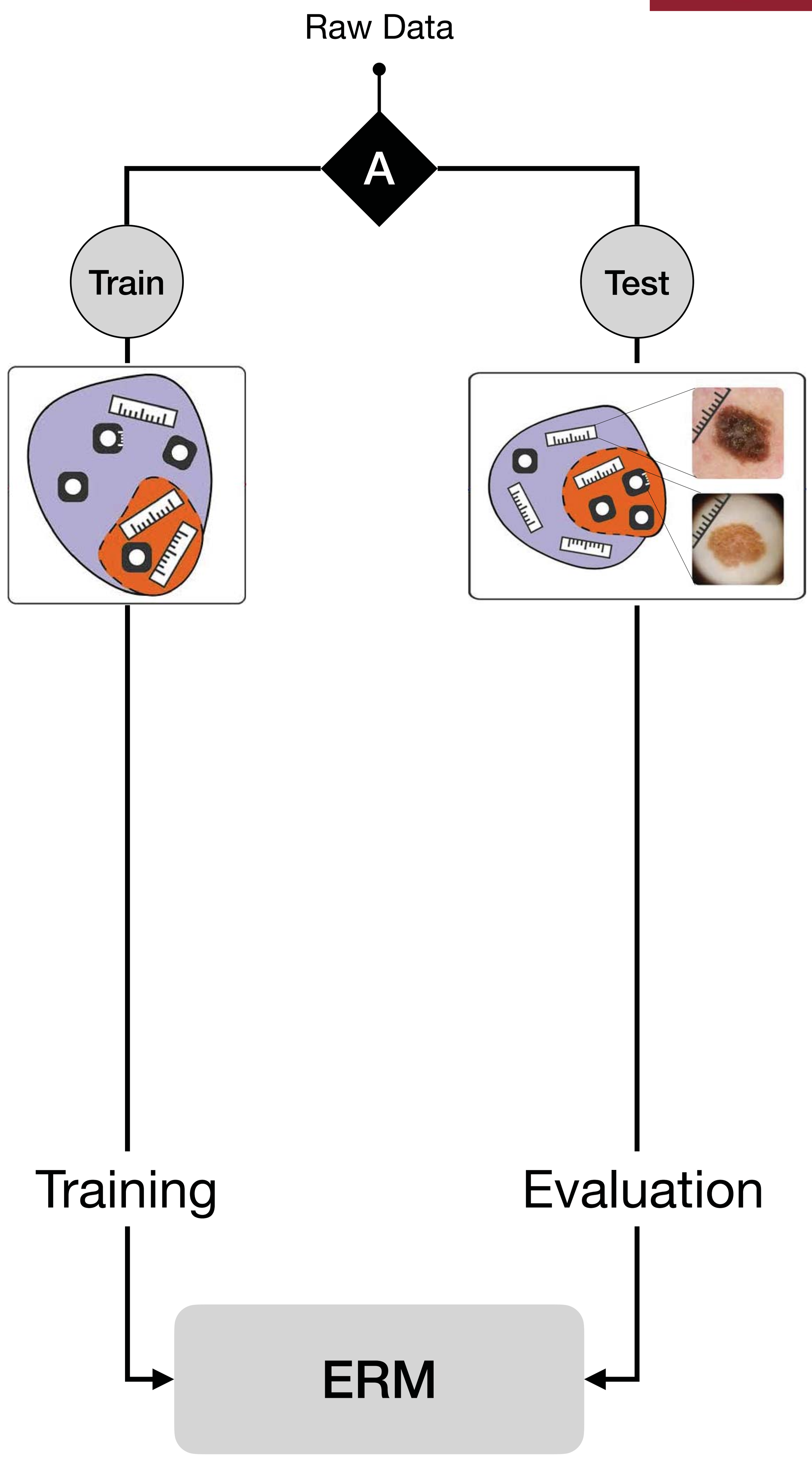
Bissoto et al., "Artifact-based domain generalization of skin lesion models", ISIC Workshop @ ECCV 2022

Results

Results

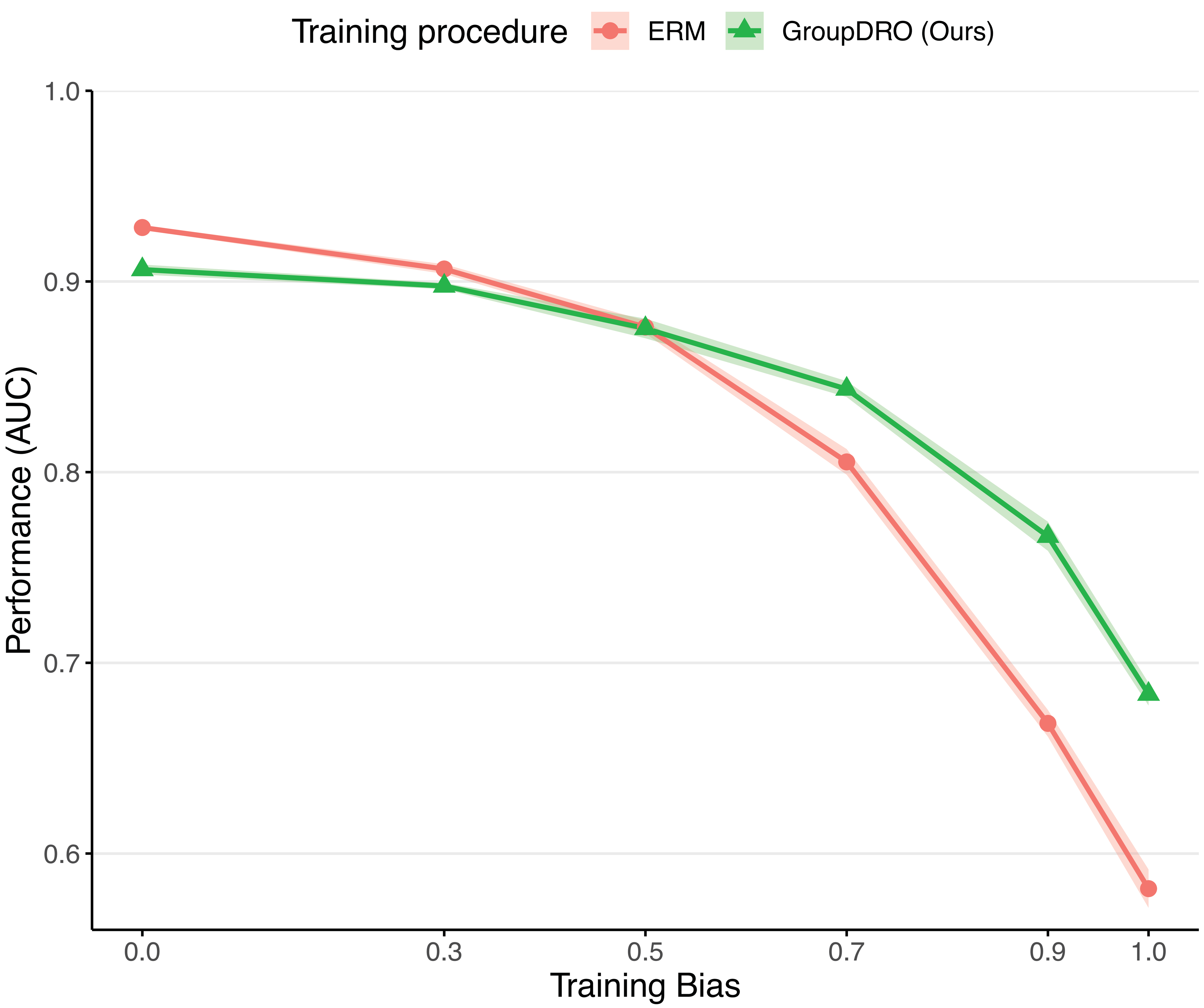
Trap Sets on ISIC 2019

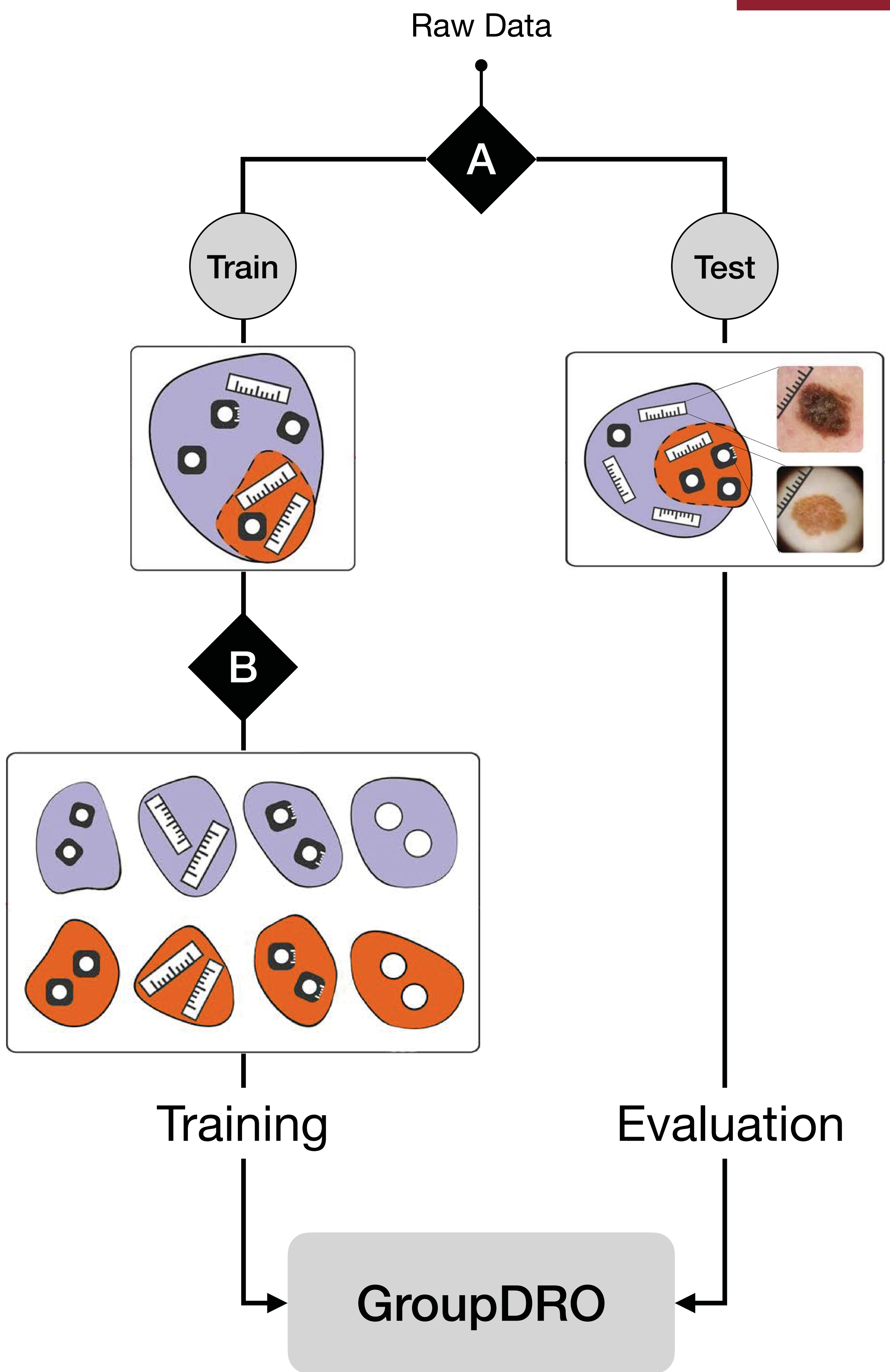




Results

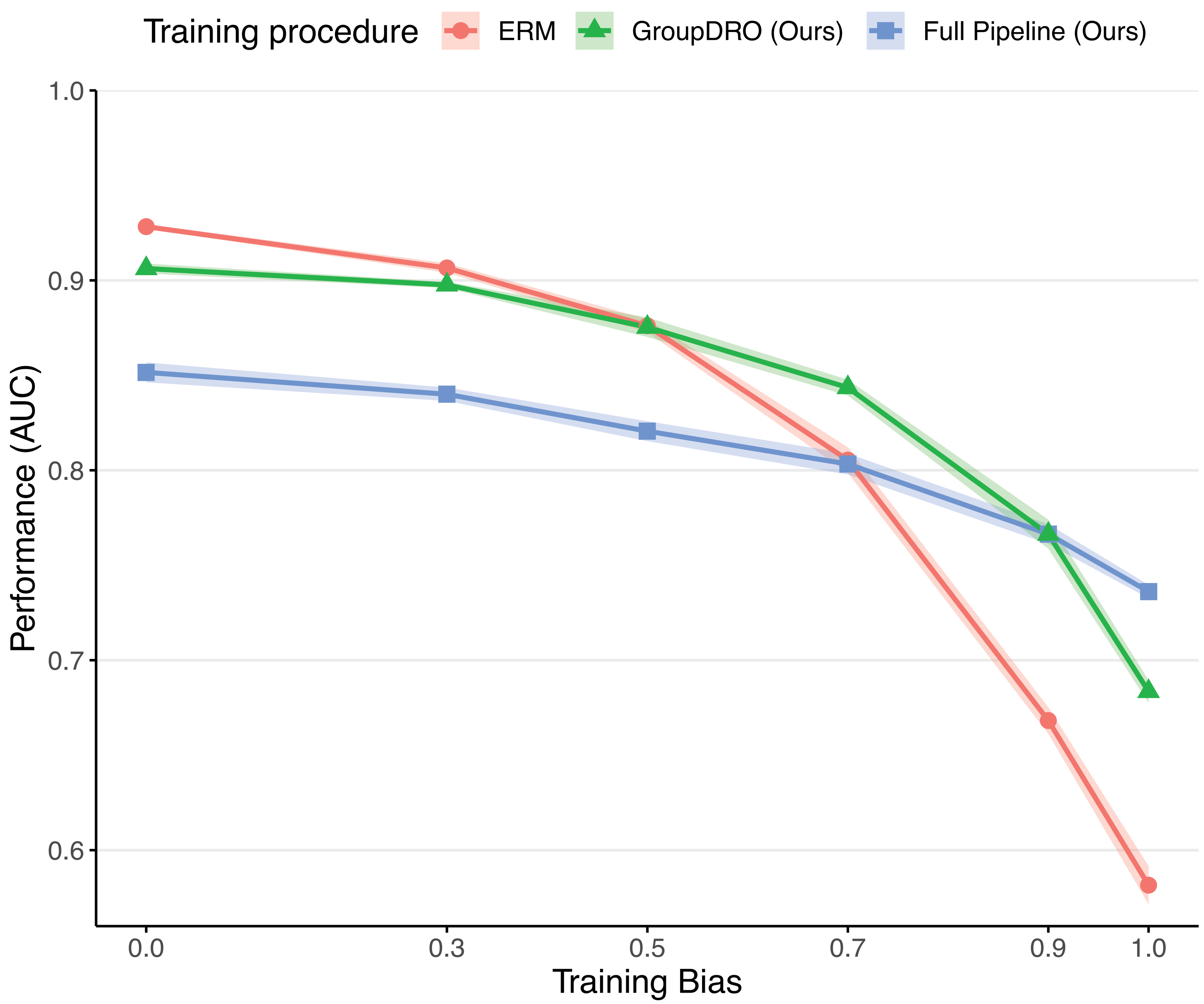
Trap Sets on ISIC 2019

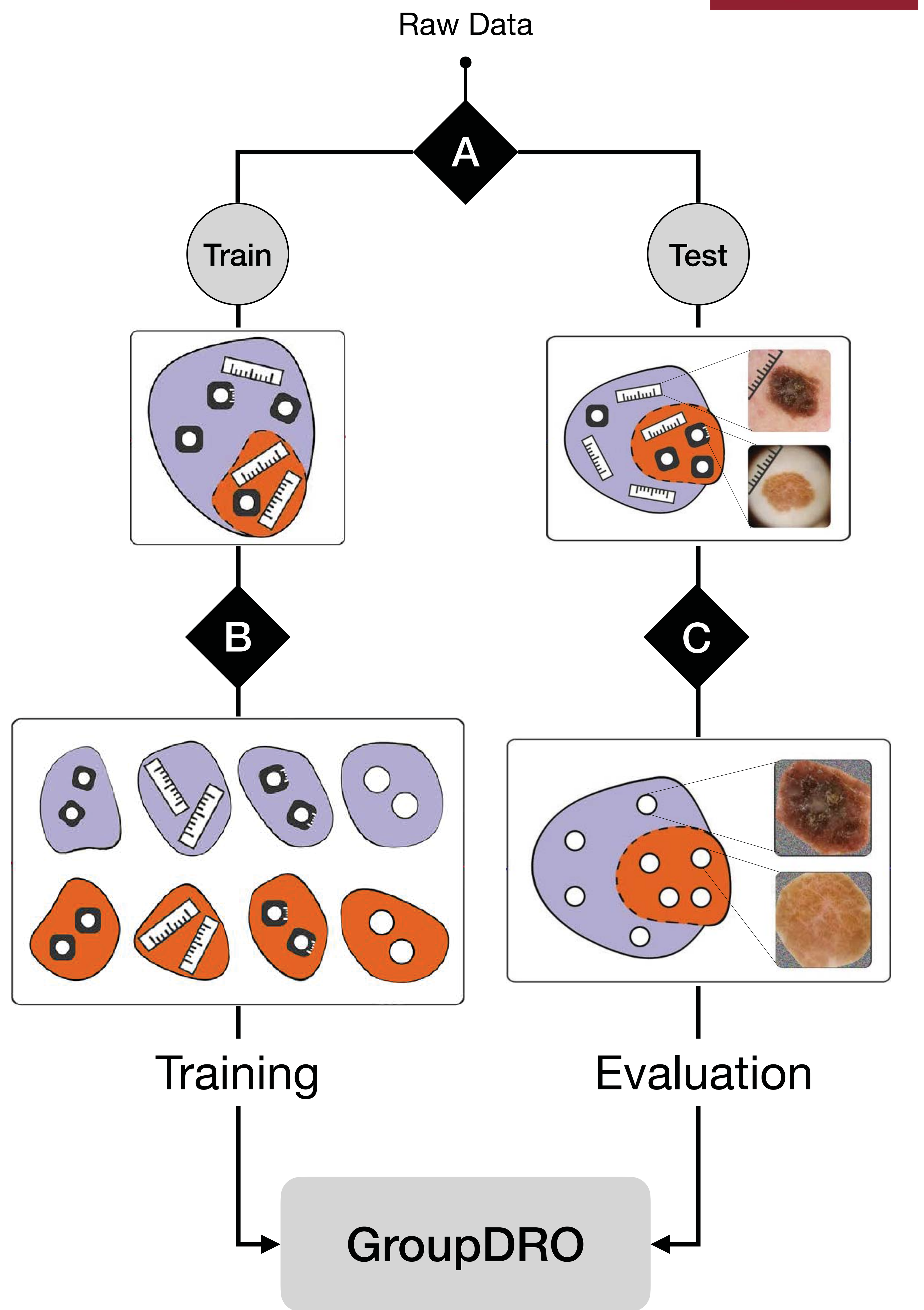




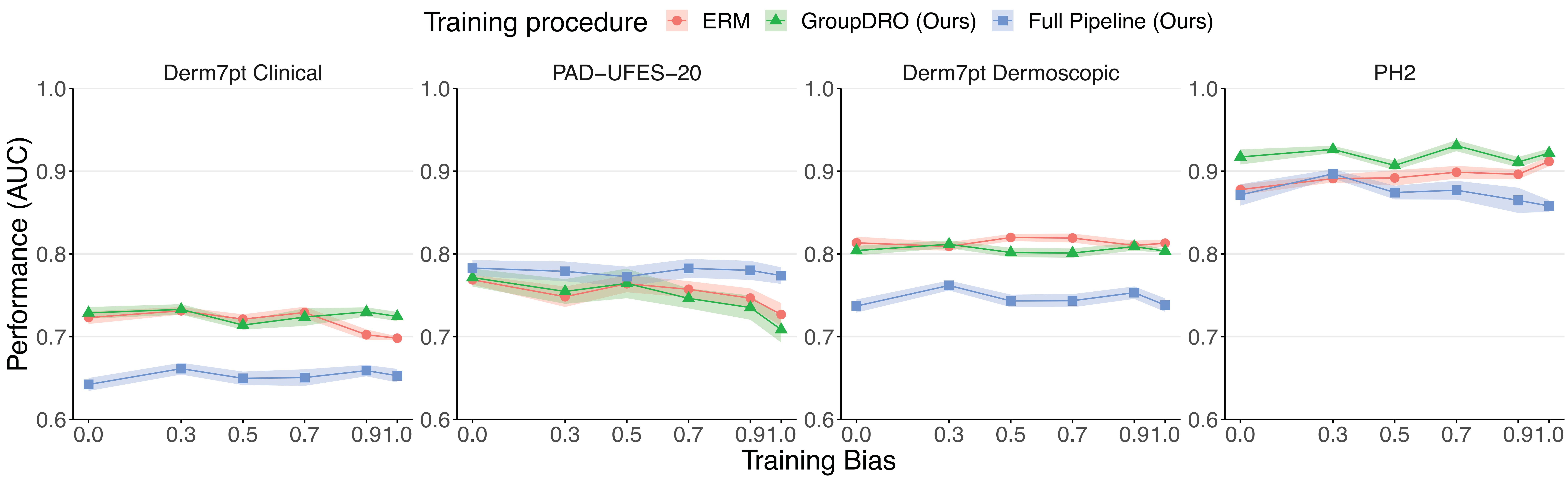
Results

Trap Sets on ISIC 2019

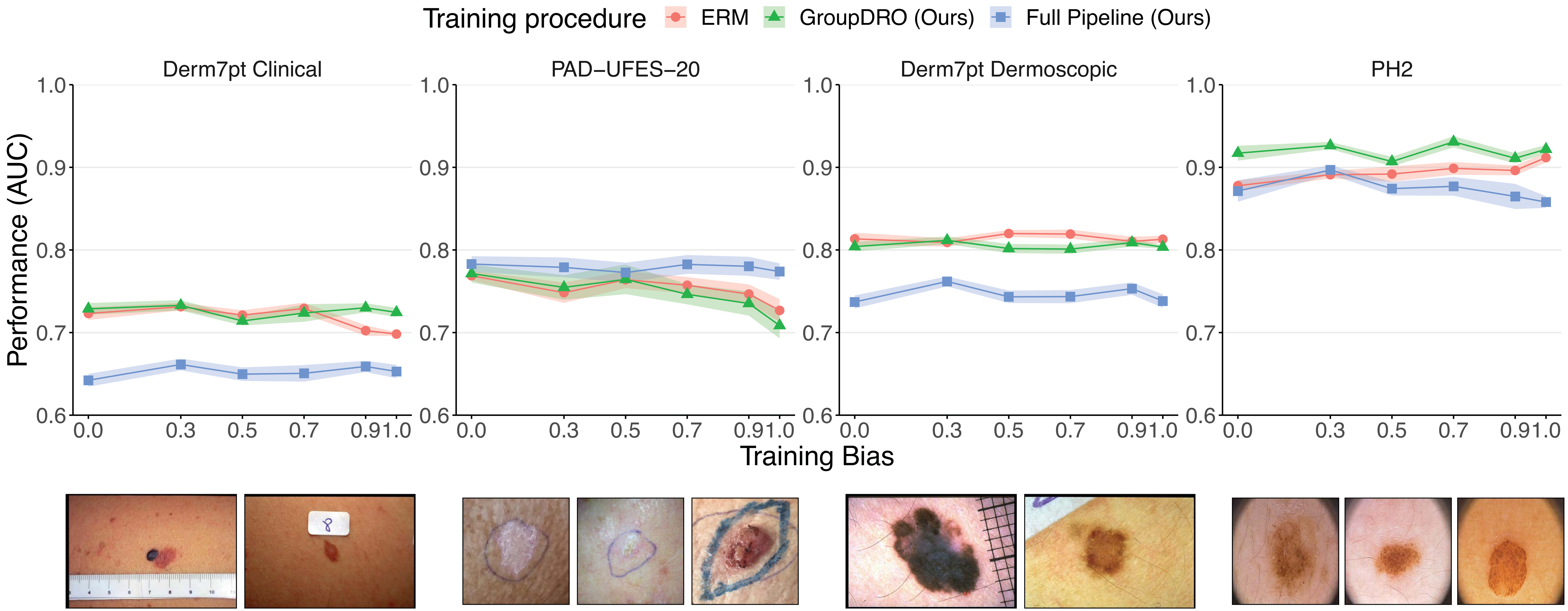




Out-of-Distribution Results



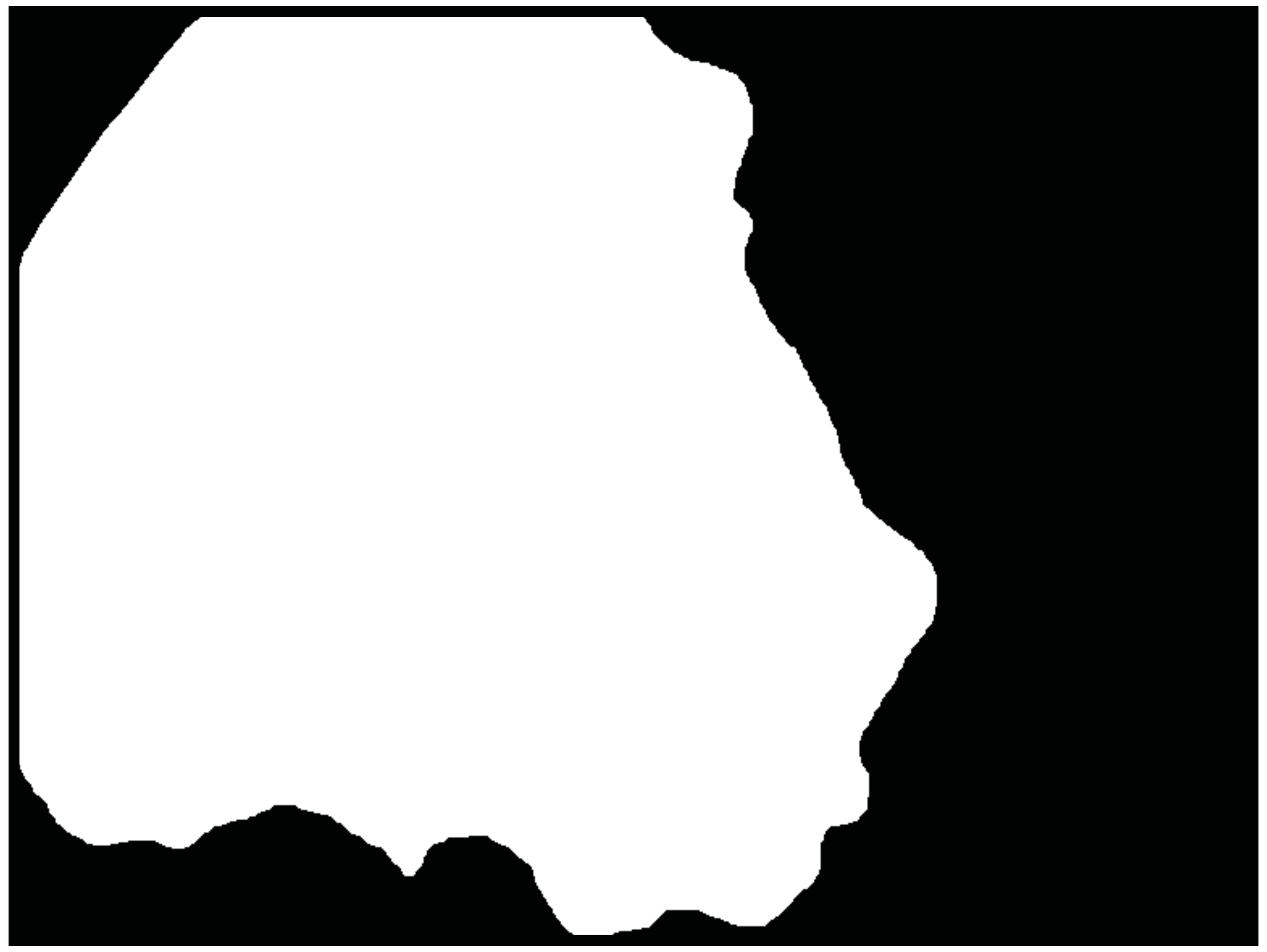
Out-of-Distribution Results



Limitations

- We still need extra annotations (in form of artifacts annotations and segmentation masks) to perform our debiasing pipeline.

	Dark corner	Ruler	Ink Markings
ISIC_0000001	✗	✗	✓
ISIC_0000002	✓	✗	✗
ISIC_0000003	✗	✓	✗
ISIC_0000004	✓	✓	✗



Limitations

- We still need extra annotations (in form of artifacts annotations and segmentation masks) to perform our debiasing pipeline
- Debiasing with respect to artifacts may not translate to out-of-distribution performance
  - Performance in out-of-distribution depends on the confounders available on test

Takeaways

- Is debiasing research useful only when biases on train are very high?

Takeaways

- Is debiasing research useful only when biases on train are very high?

*"Broadly, our analysis indicates that internet-trained models have **internet-scale biases**."*

Brown et al., "Language Models are Few-Shot Learners", NeurIPS 2020

Bissoto et al., "Artifact-based domain generalization of skin lesion models", ISIC Workshop @ ECCV 2022

Takeaways

- Is debiasing research useful only when biases on train are very high?
  - No! Even colossal models trained with billions of data such as GPT-3 reproduce mild biases. For medical data, the problem is compounded
- We can improve robustness to KNOWN biases through both training and test debiasing
  - We must continue handling different bias problems that may arise in the clinical scenario

Bissoto et al., "Artifact-based domain generalization of skin lesion models", ISIC Workshop @ ECCV 2022

Code, Data & Paper:
`https://github.com/alceubissoto/artifact-generalization-skin`

Thank you!

Alceu Bissoto alceubissoto@ic.unicamp.br
Catarina Barata ana.c.fidalgo.barata@tecnico.ulisboa.pt
Eduardo Valle dovalle@dca.fee.unicamp.br
Sandra Avila sandra@ic.unicamp.br

UNICAMP

recod

LARA

FAPESP