Test-Time Selection for Robust Skin Lesion Analysis



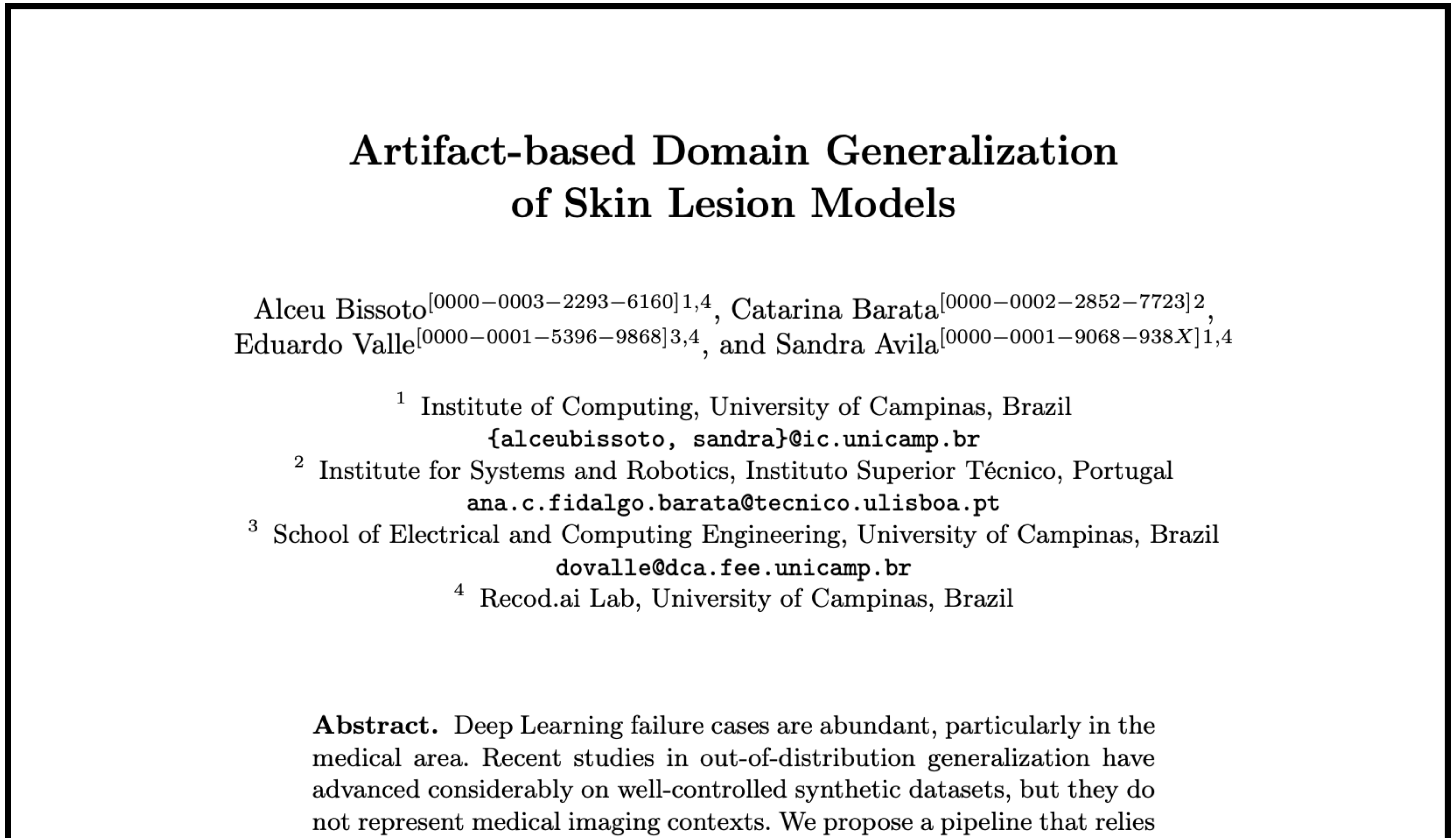

Alceu Bissoto[1], Catarina Barata[2], Eduardo Valle[3], Sandra Avila[1]

[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

Artifact-based Domain Generalization

ISIC Workshop @ ECCV 2022

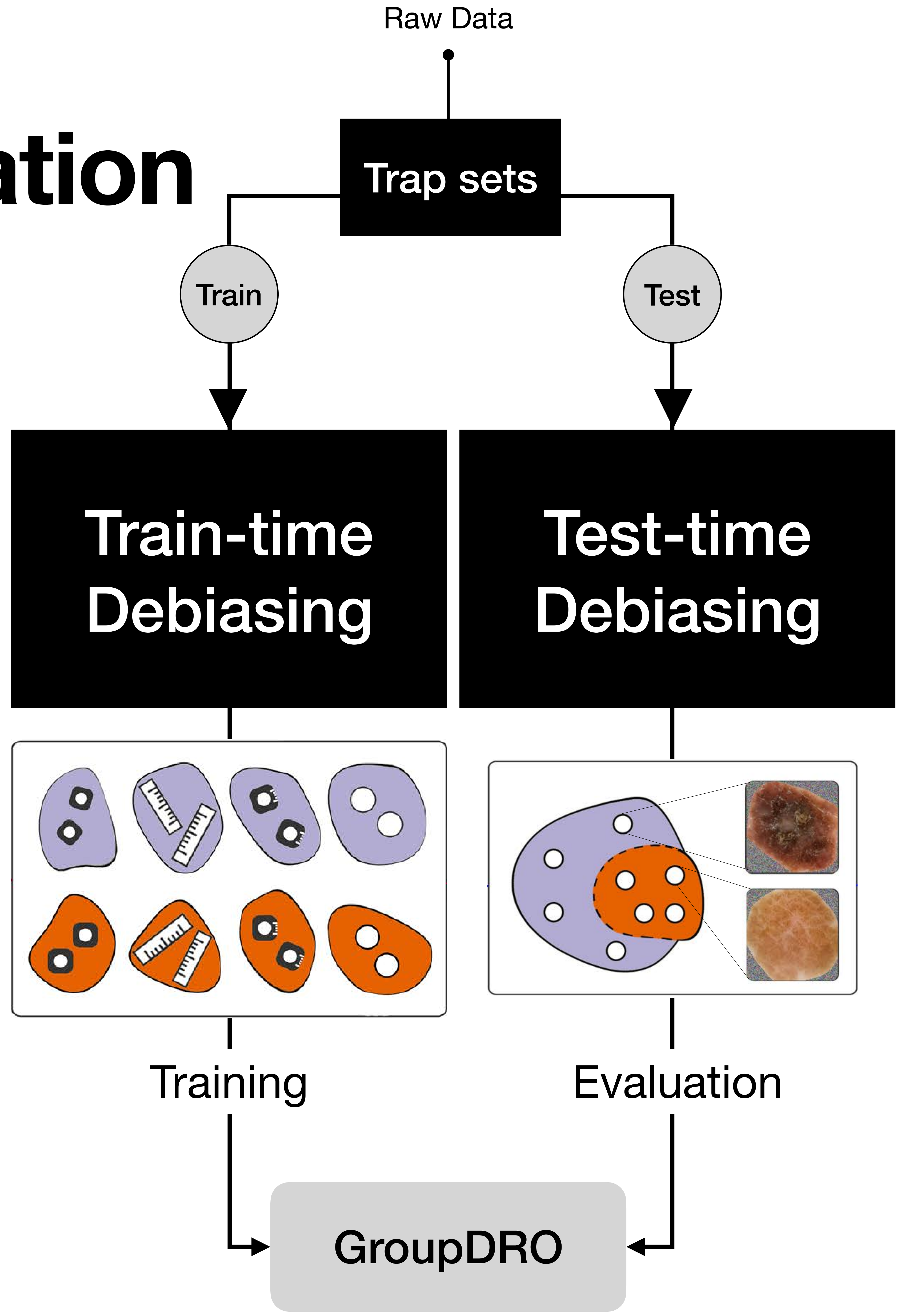


Artifact-based Domain Generalization of Skin Lesion Models

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]

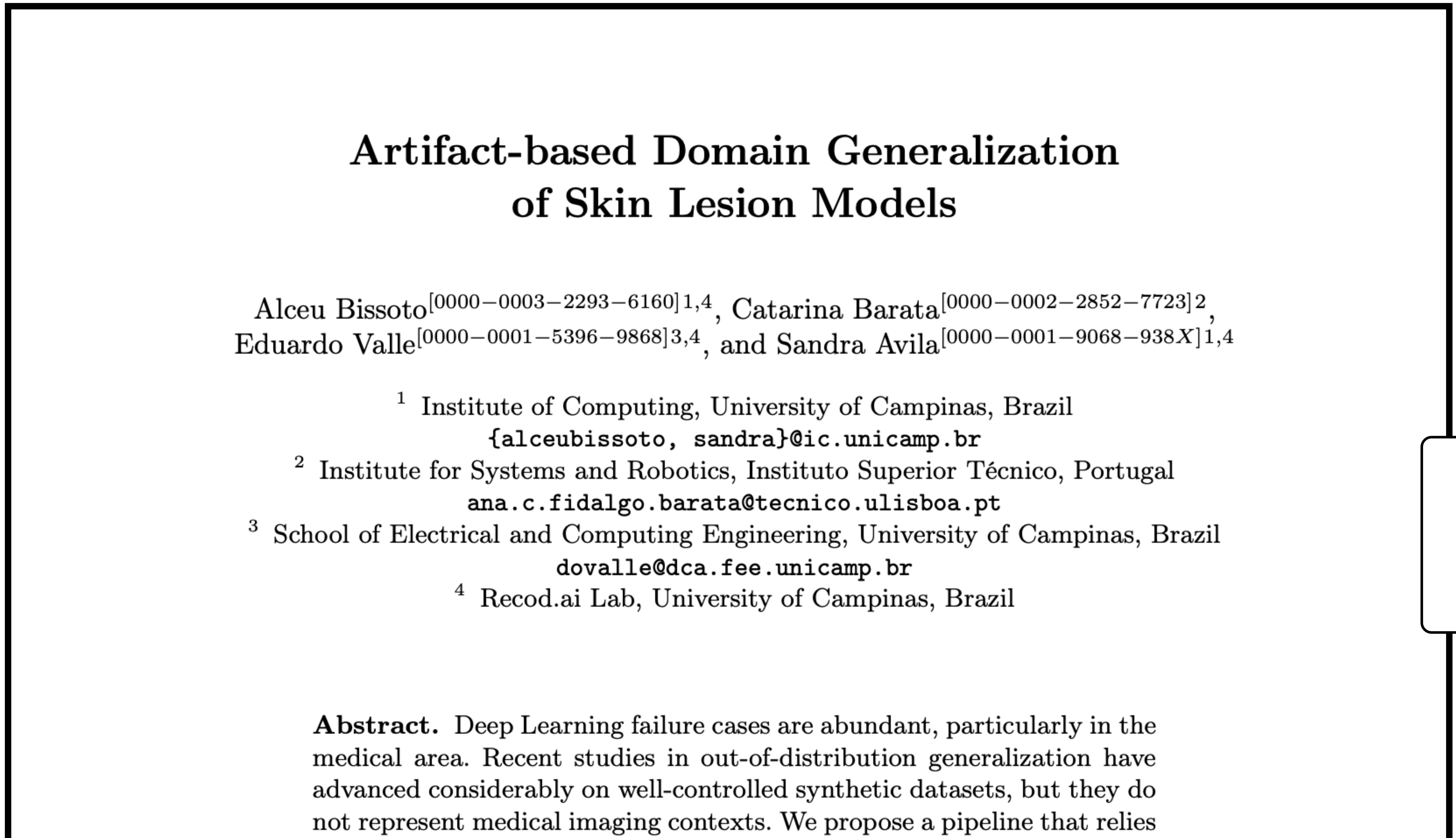
[1] Institute of Computing, University of Campinas, Brazil
{alceubissoto, sandra}@ic.unicamp.br
[2] Institute for Systems and Robotics, Instituto Superior Técnico, Portugal
ana.c.fidalgo.barata@tecnico.ulisboa.pt
[3] School of Electrical and Computing Engineering, University of Campinas, Brazil
dovalle@dca.fee.unicamp.br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



Artifact-based Domain Generalization

ISIC Workshop @ ECCV 2022

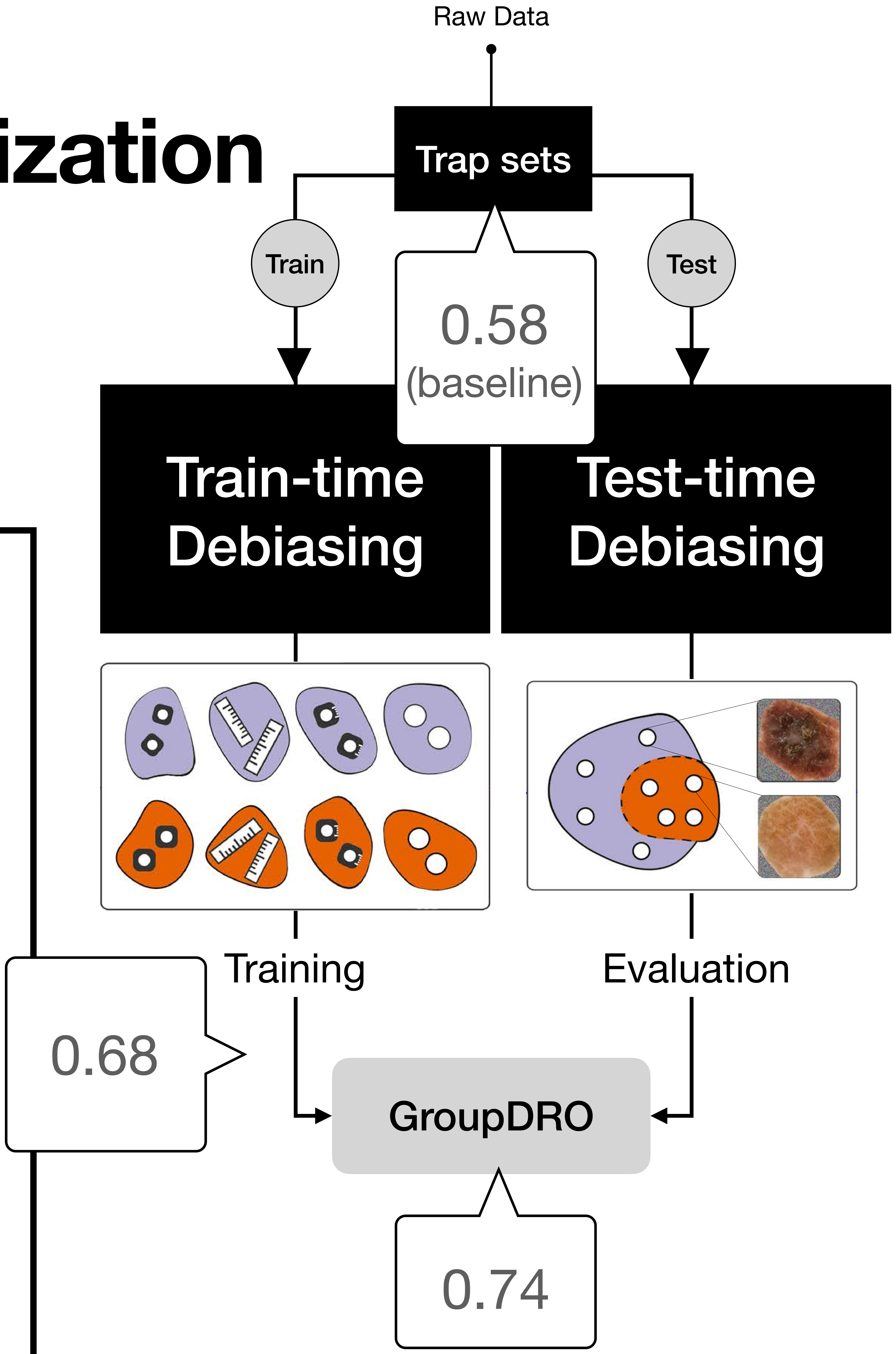


Artifact-based Domain Generalization
of Skin Lesion Models

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



Artifact-based Domain Generalization

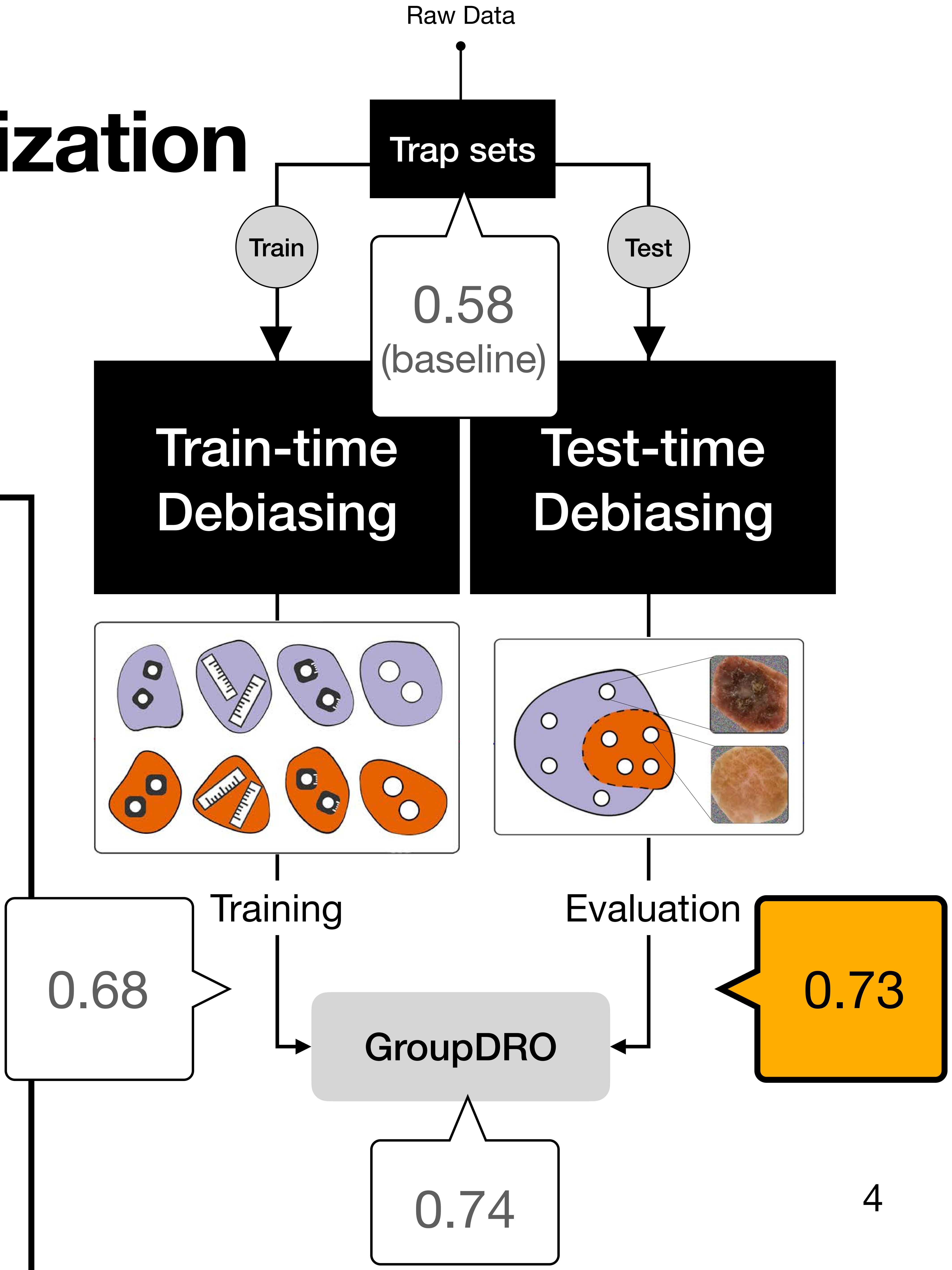
ISIC Workshop @ ECCV 2022

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



Test-time Debiasing Literature

Alignment with the Clinical Workflow

	Method	#Keypoints	AUC
Baseline	Test-time augmentation	-	58,4
Literature	T3A	-	56,7
Literature	Tent	-	54,1
Literature	NoiseCrop	50.176	72,7

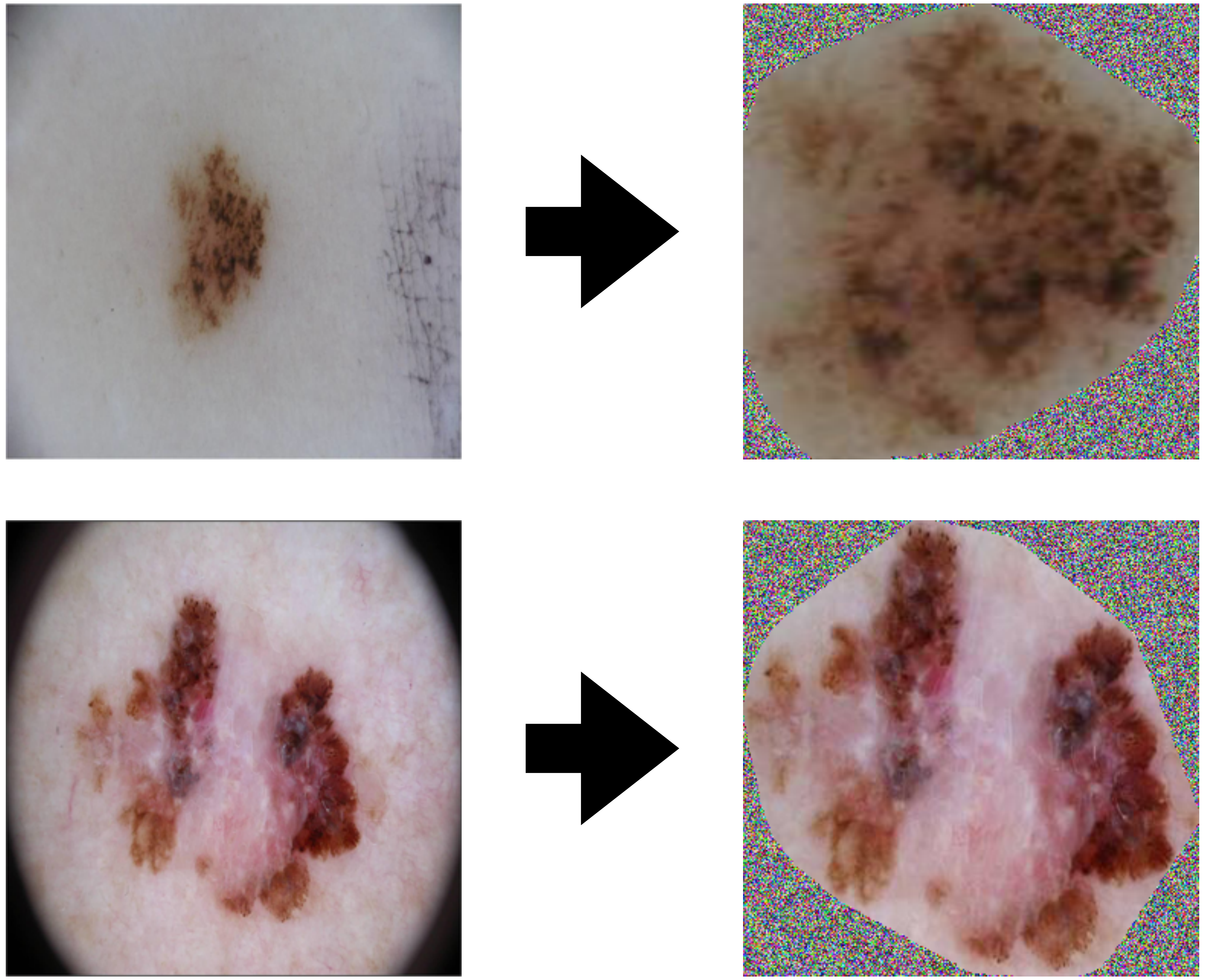
Test-time debiasing literature

- Relies on test batch statistics to update model weights.
- Fail when **only a single image is available.**
- Fail when **test distribution is heterogenous.**

Test-time Debiasing Literature

Alignment with the Clinical Workflow

NoiseCrop



- Relies on full segmentation masks, which are **<u>hard to annotate</u>**.
- Make modifications in the pixel-space, which **<u>might introduce unexpected features.</u>**

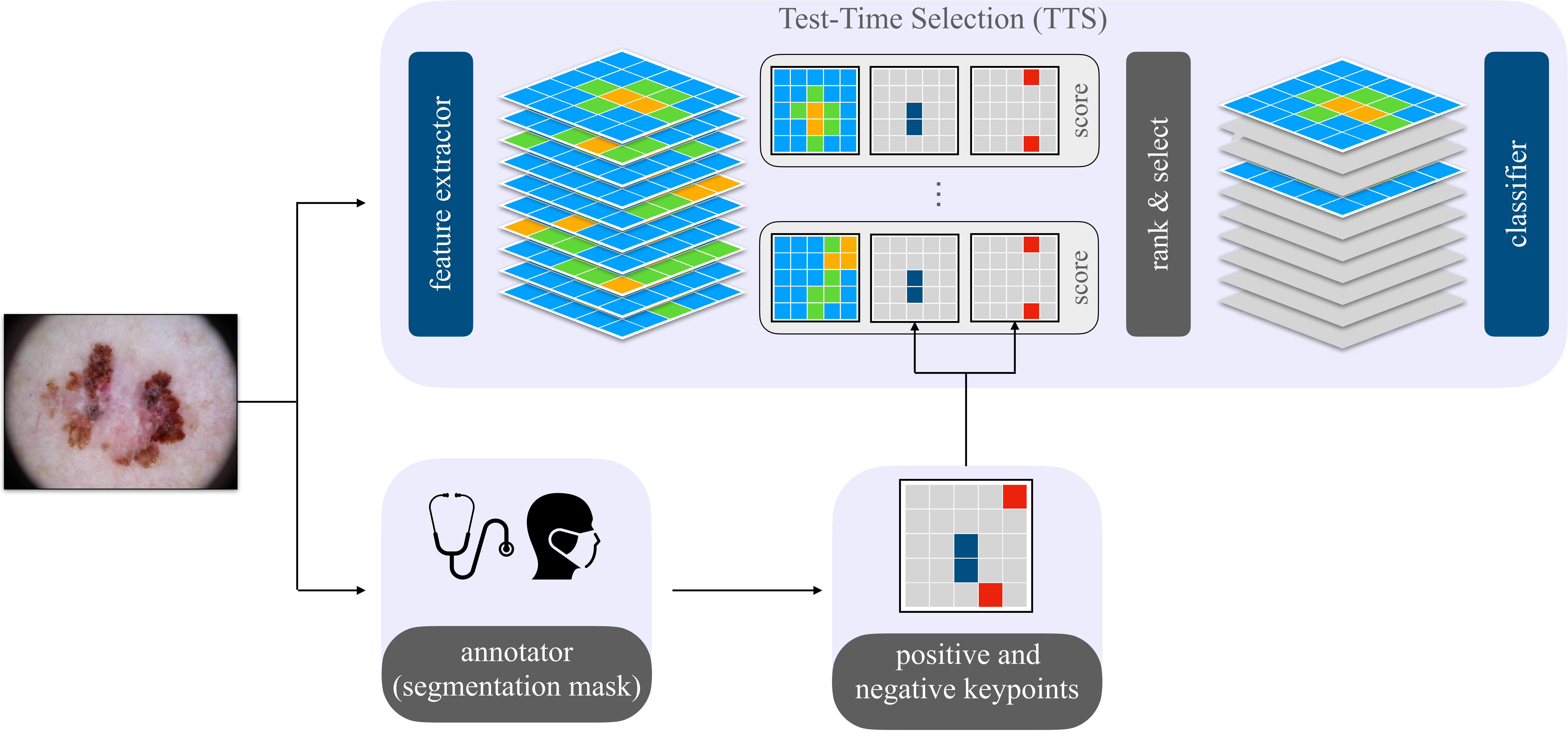
Test-time Selection (TTS)

- **Fast to annotate.**
- **Avoid introducing distribution shifts** by intervening on the feature space.
- **Does not rely** on test batch statistics.
- It's **cheap** as there are no model updates.

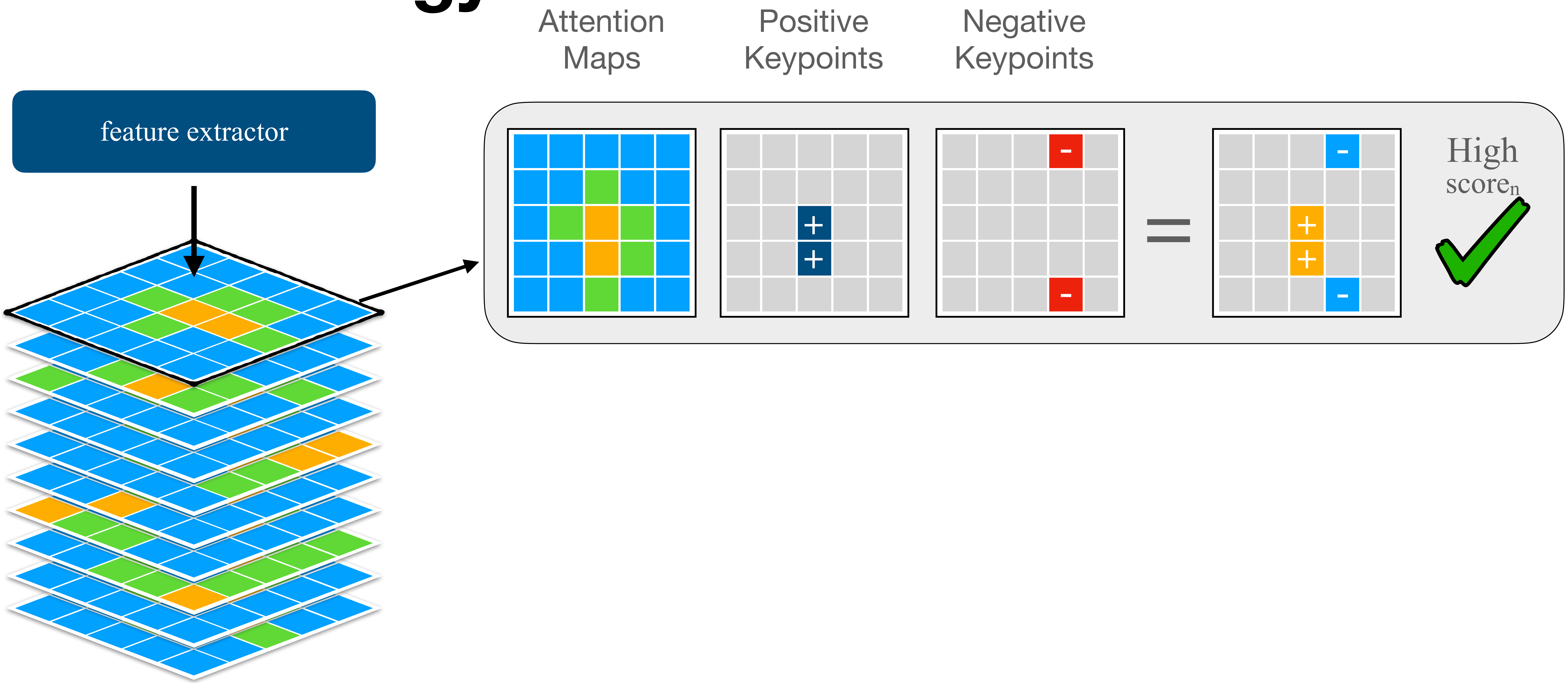
	Method	#Keypoints	AUC
Baseline	Test-time augmentation	-	58,4
Literature	T3A	-	56,7
Literature	Tent	-	54,1
Literature	NoiseCrop	50.176	72,7
Ours	TTS	40	75,0

Methodology

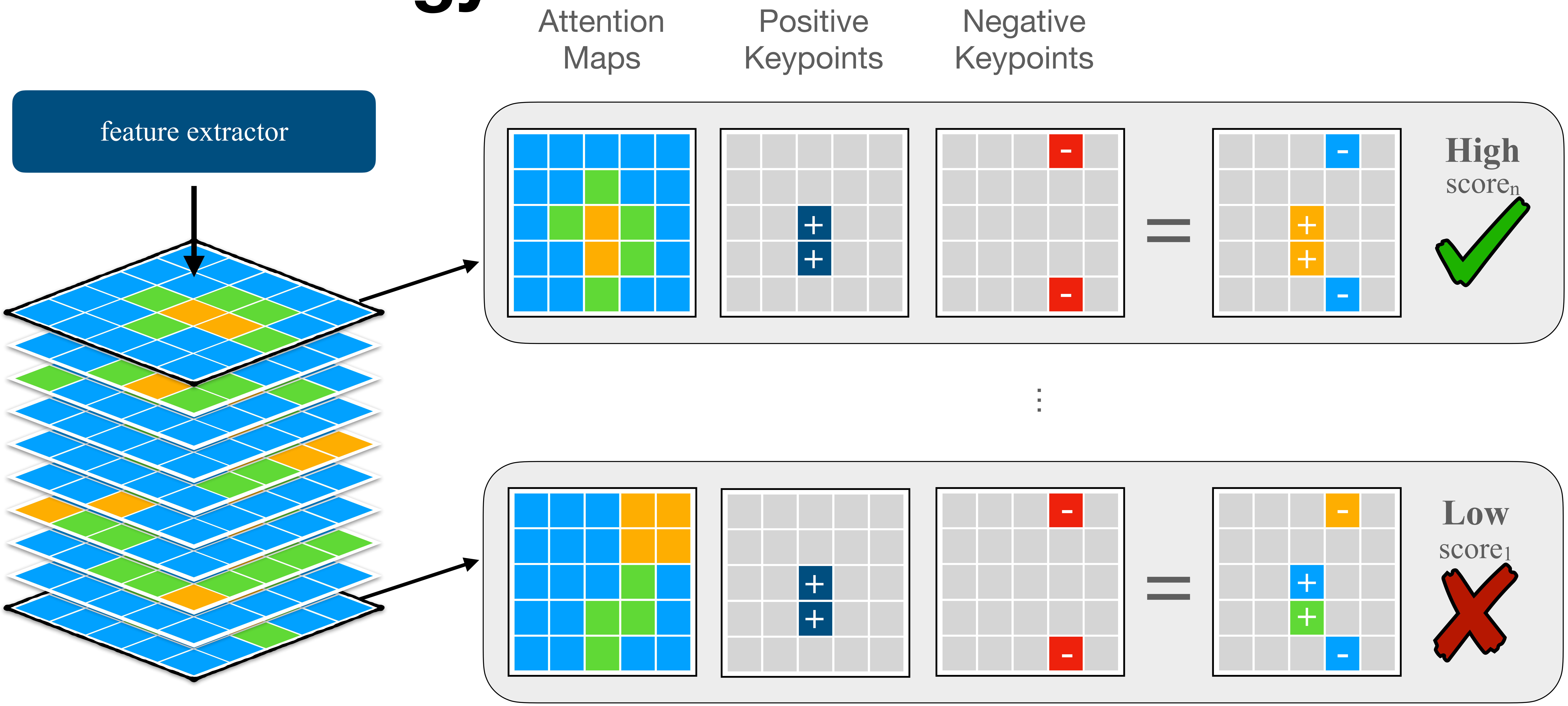
Methodology



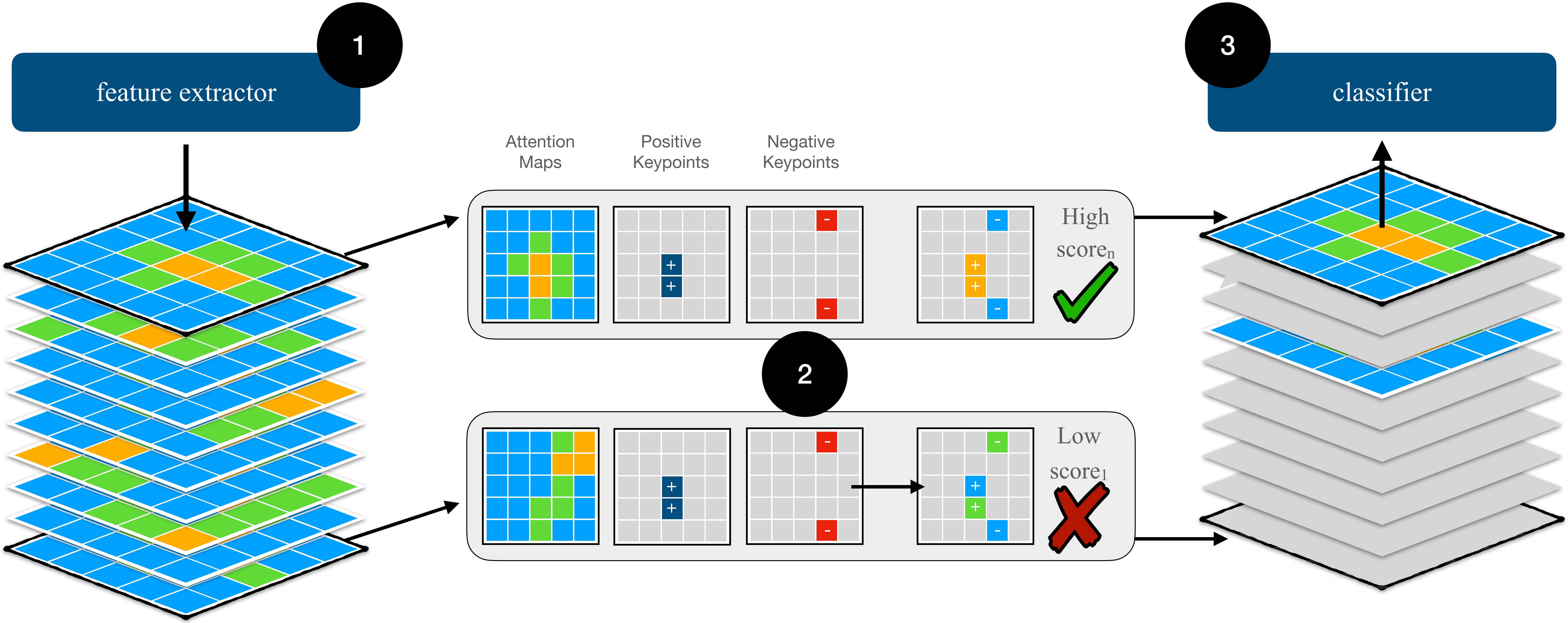
Methodology



Methodology

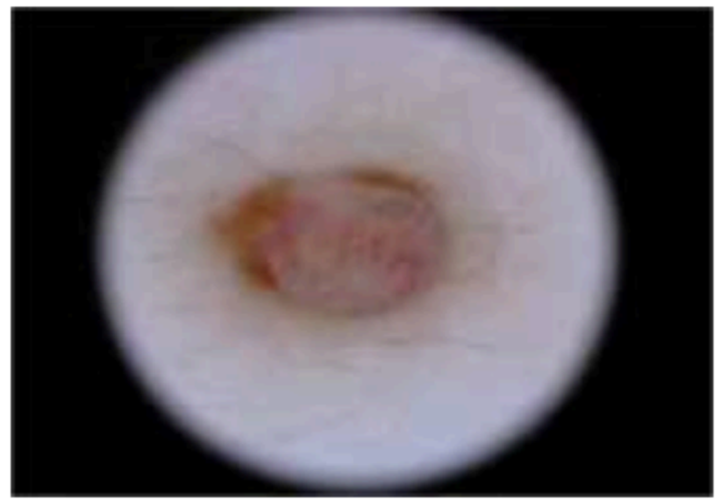


Methodology



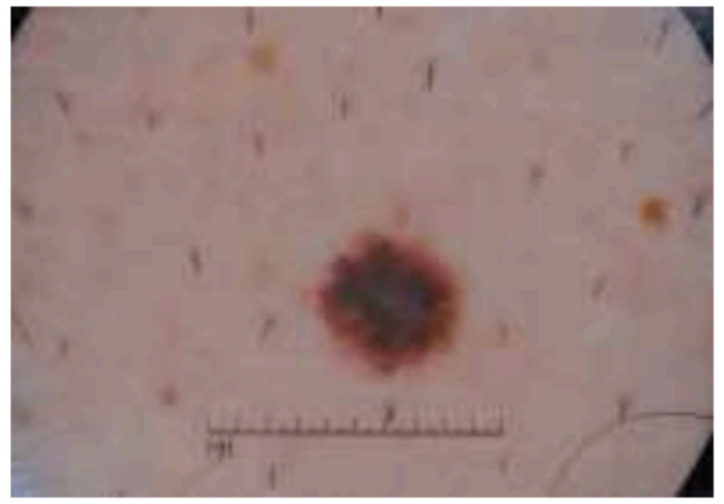
Evaluation Protocol

Artifacts providing spurious correlations

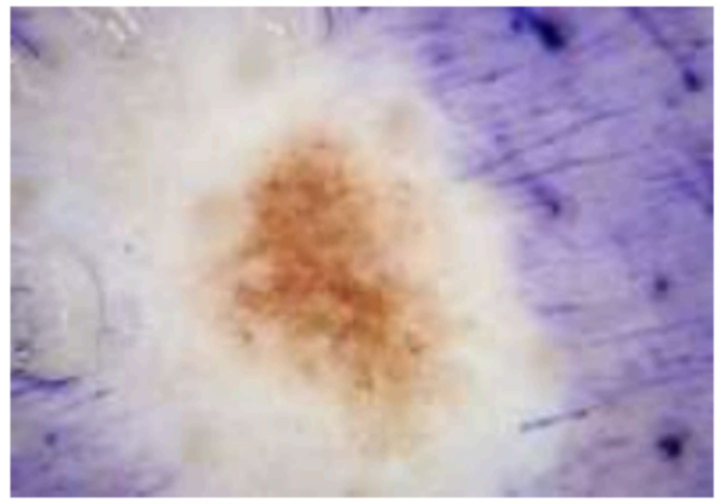


Dark Corners

Hair



Ruler



Ink markings

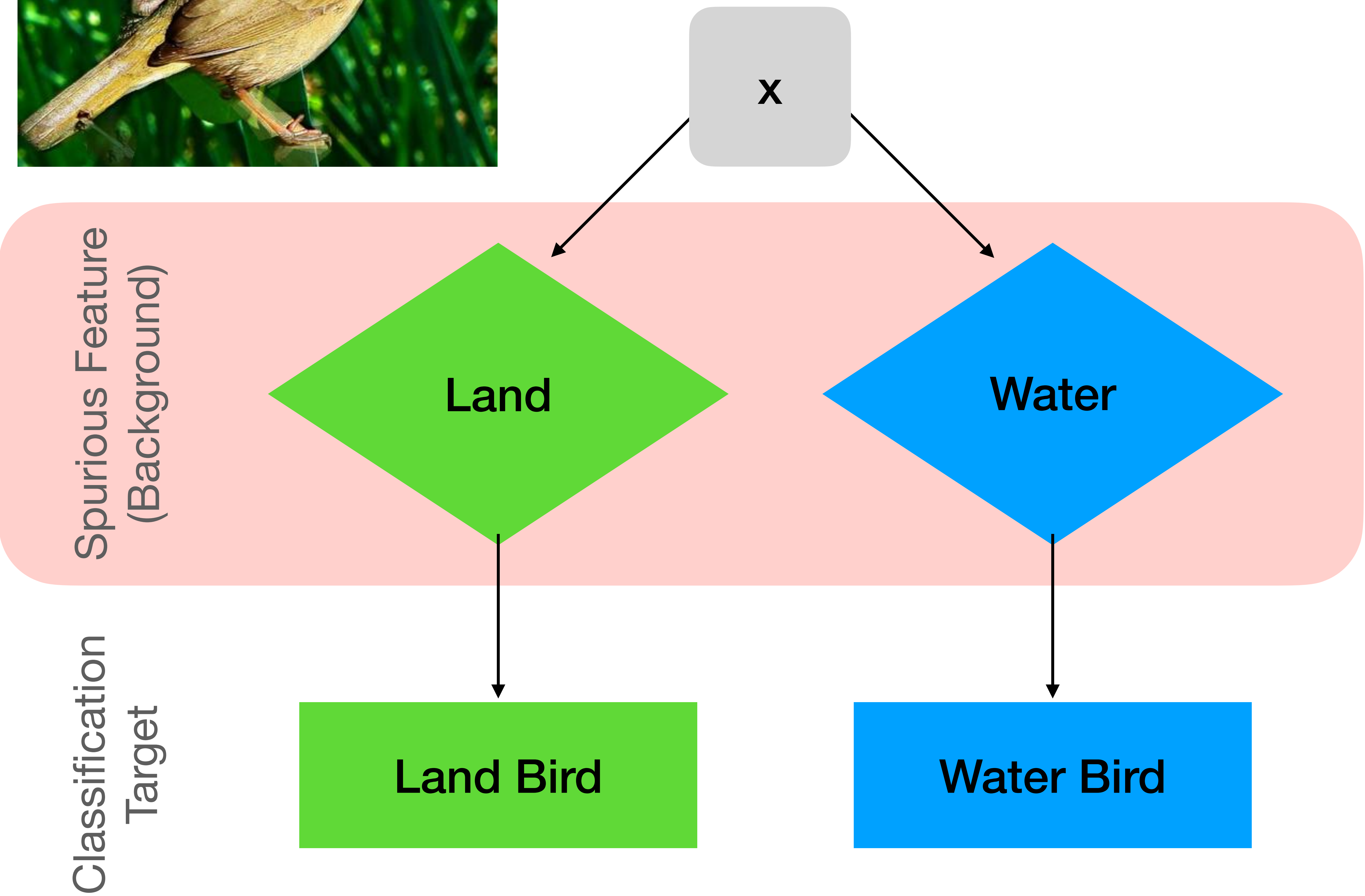
Patches

factor	set	dark corner	hair	ruler	ink	patches
0	train	0.119	-0.104	0.142	0.023	-0.138
	test	0.135	-0.112	0.162	0.030	-0.149

- Mild correlations.
- Gaining robustness to artifacts will hardly impact any metric.
- **<u>We need to control/amplify the correlations.</u>**

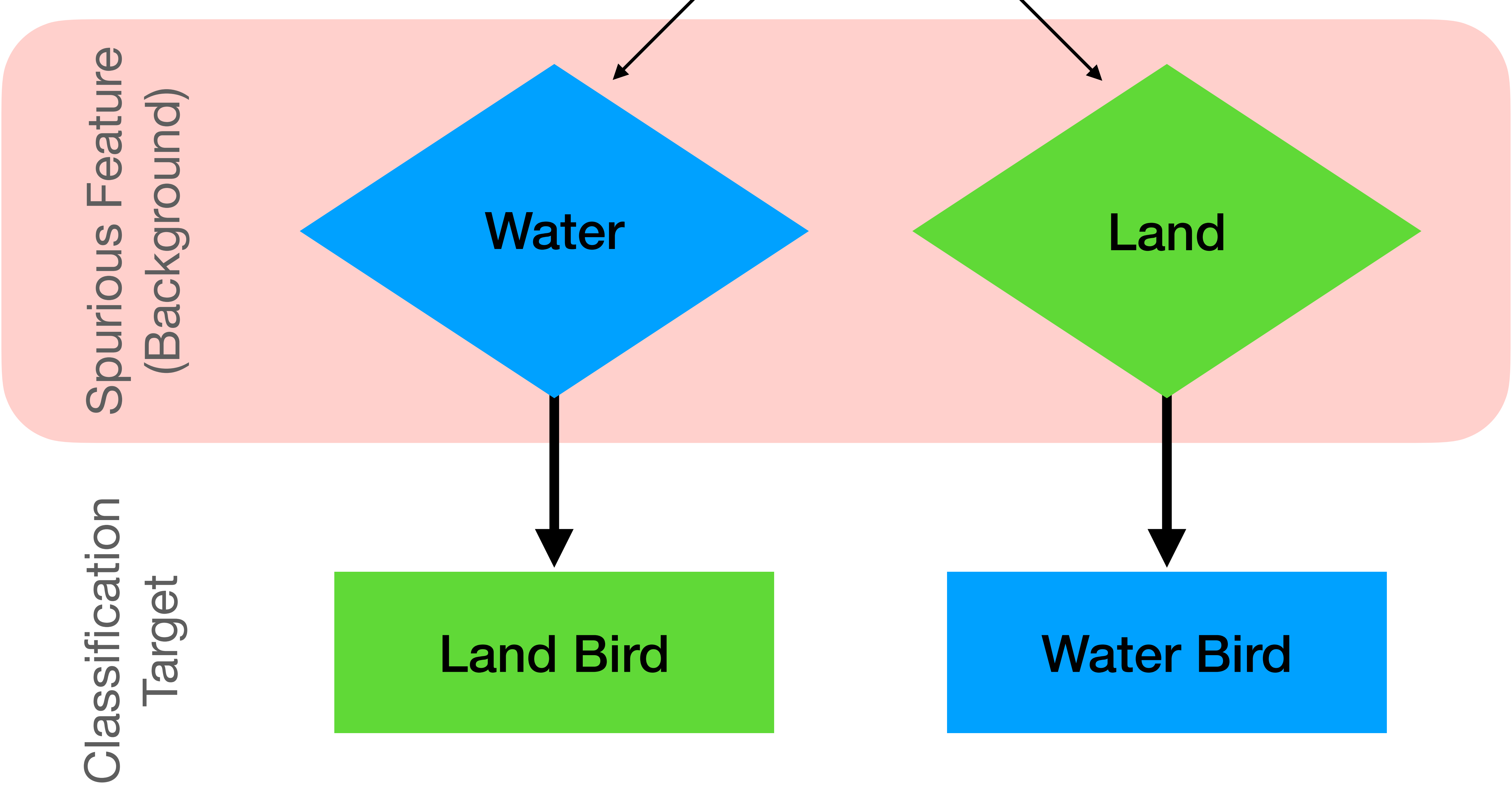
Spurious Features vs. Generalization

Train



Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.", ICLR 2020

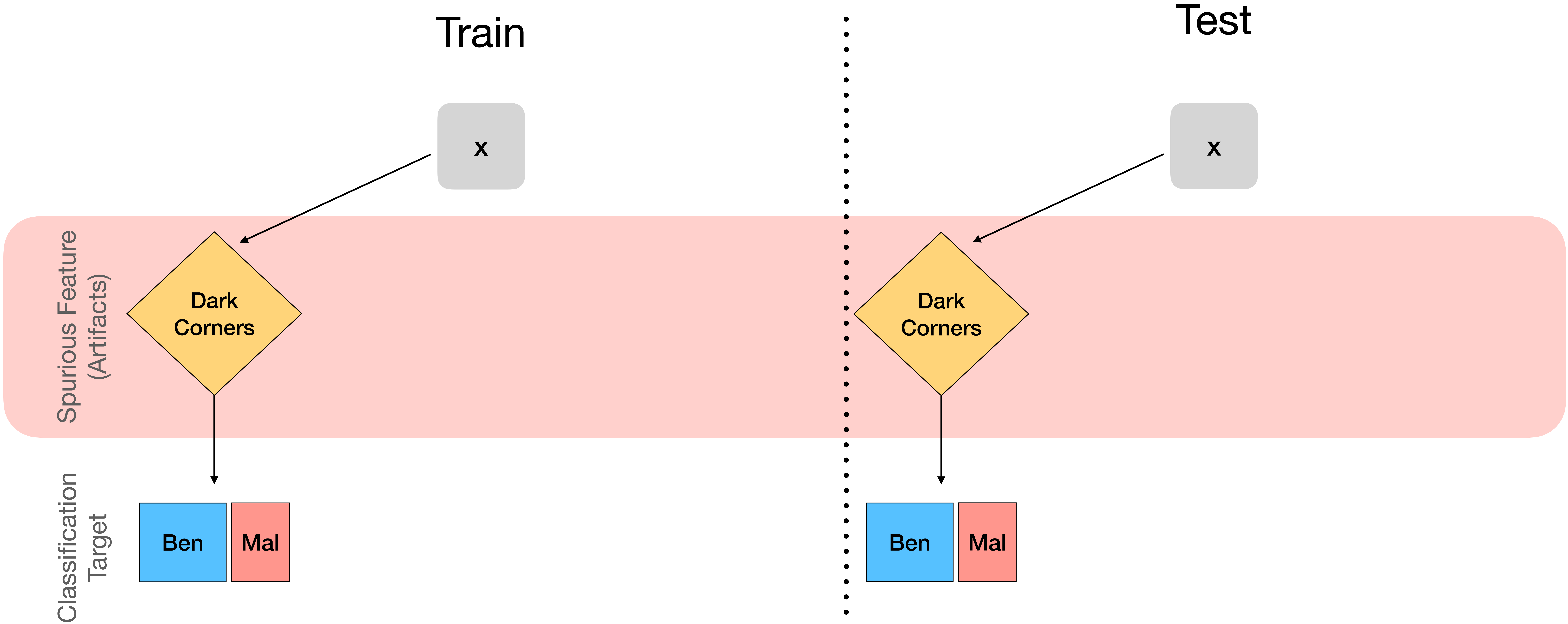
Spurious Features vs. Generalization



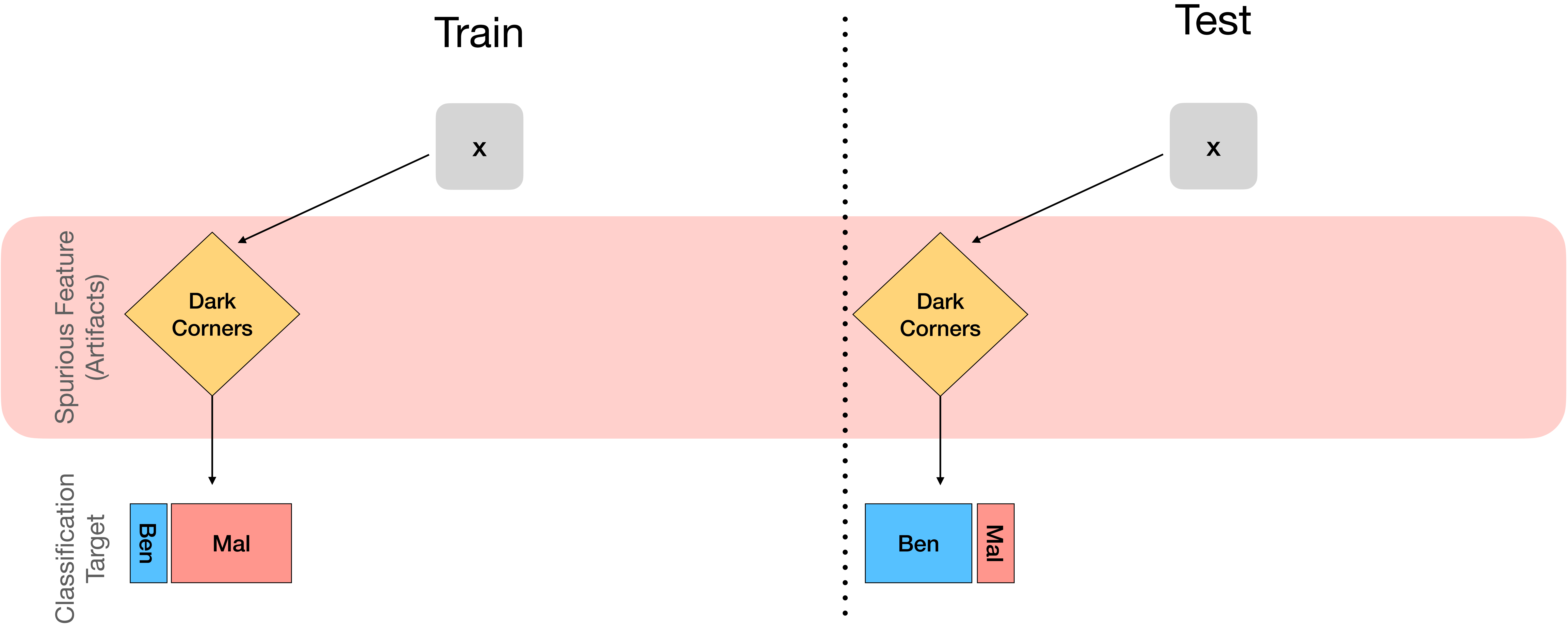
Train and Test sets present **<u>opposite</u>** correlations!

Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.", ICLR 2020

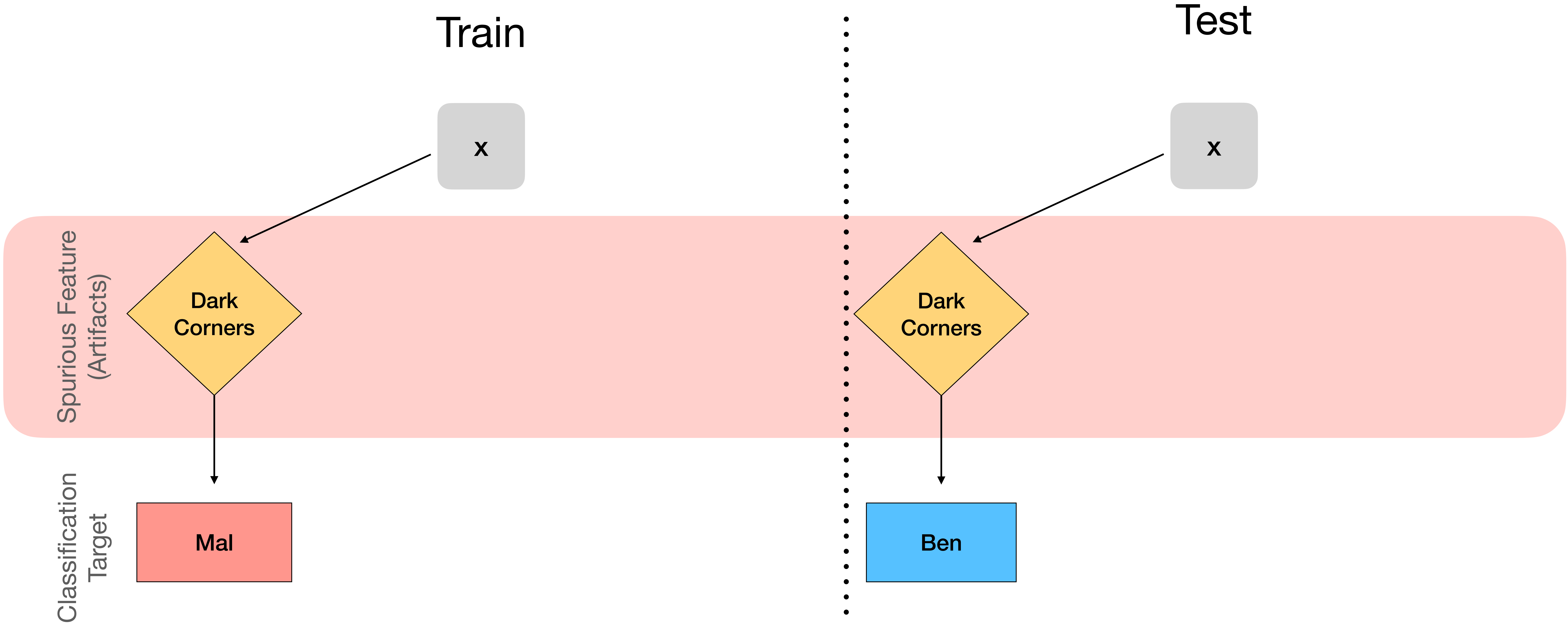
Trap Sets - Controllable known biases



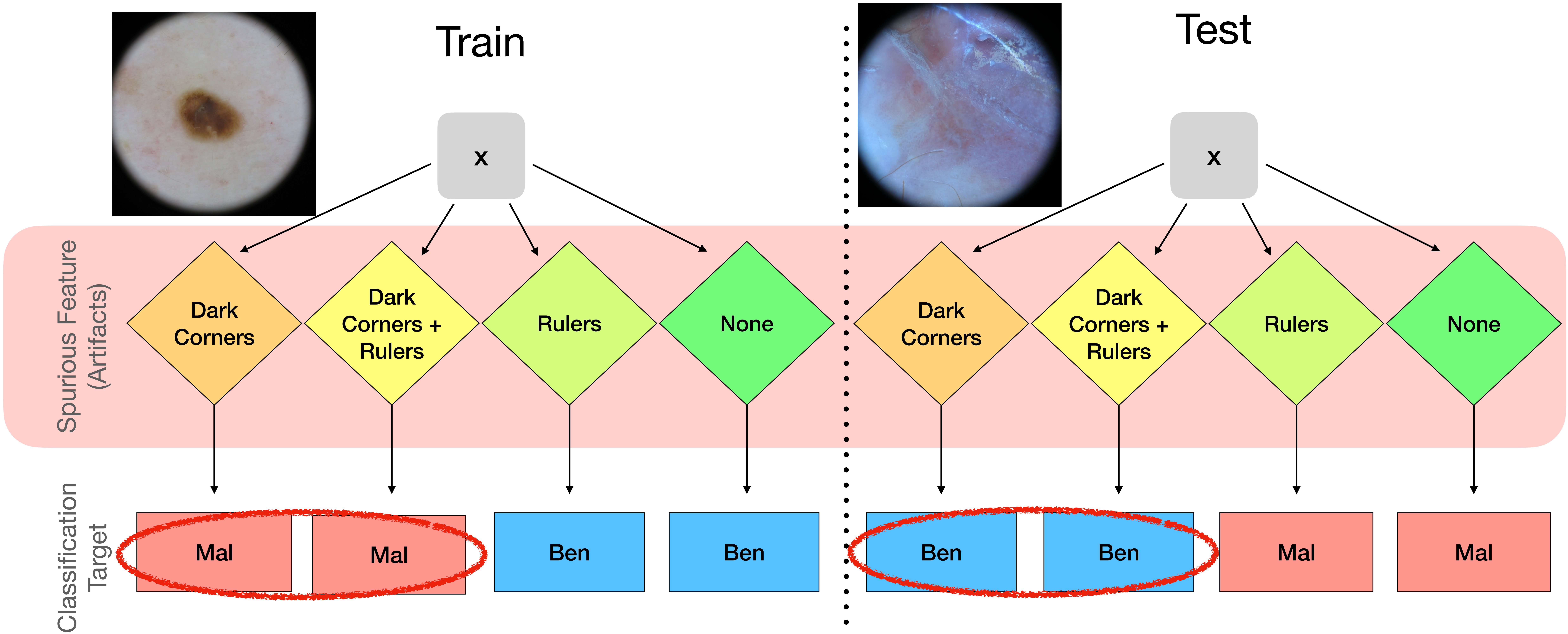
Trap Sets - Controllable known biases



Trap Sets - Controllable known biases

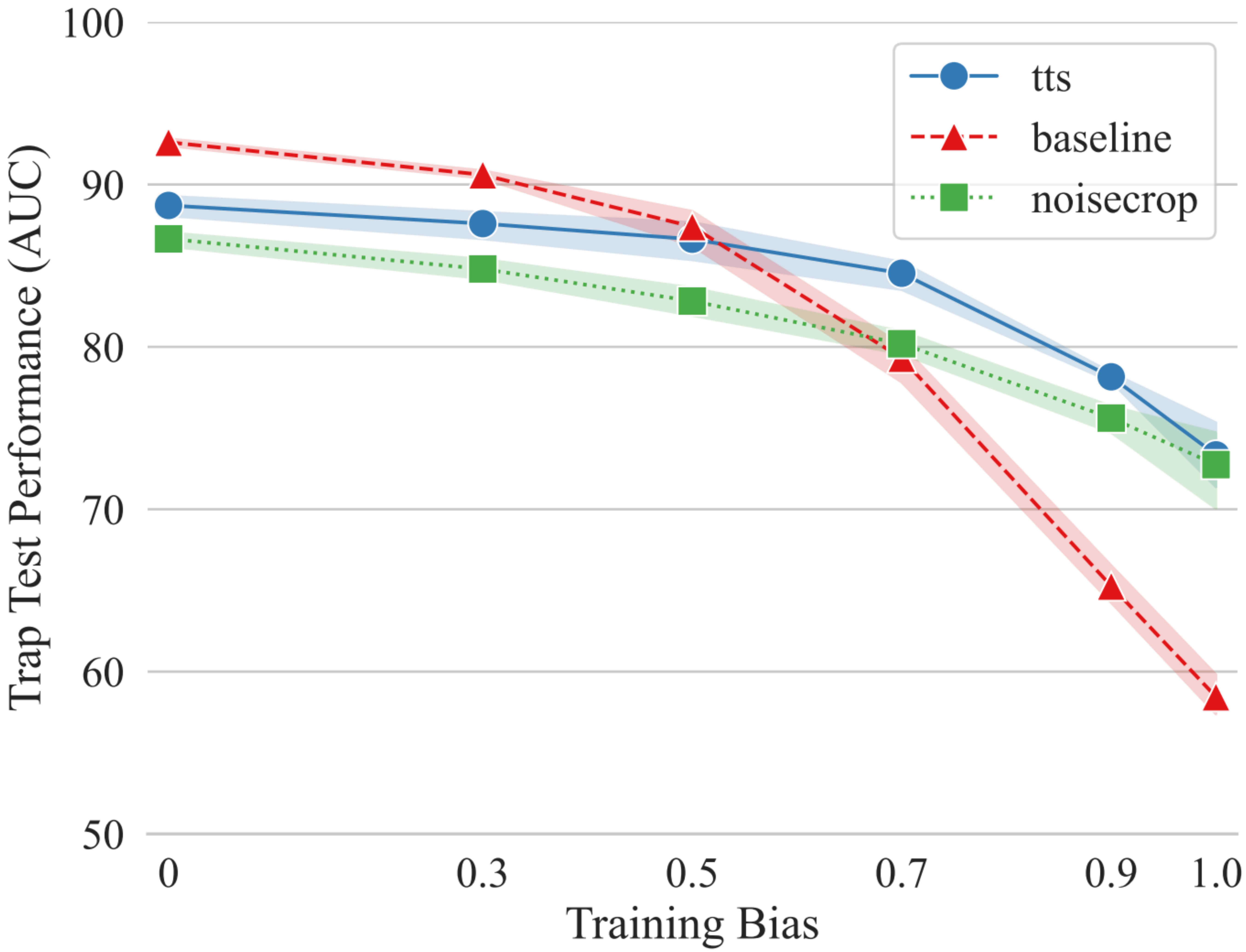


Trap Sets - Controllable known biases

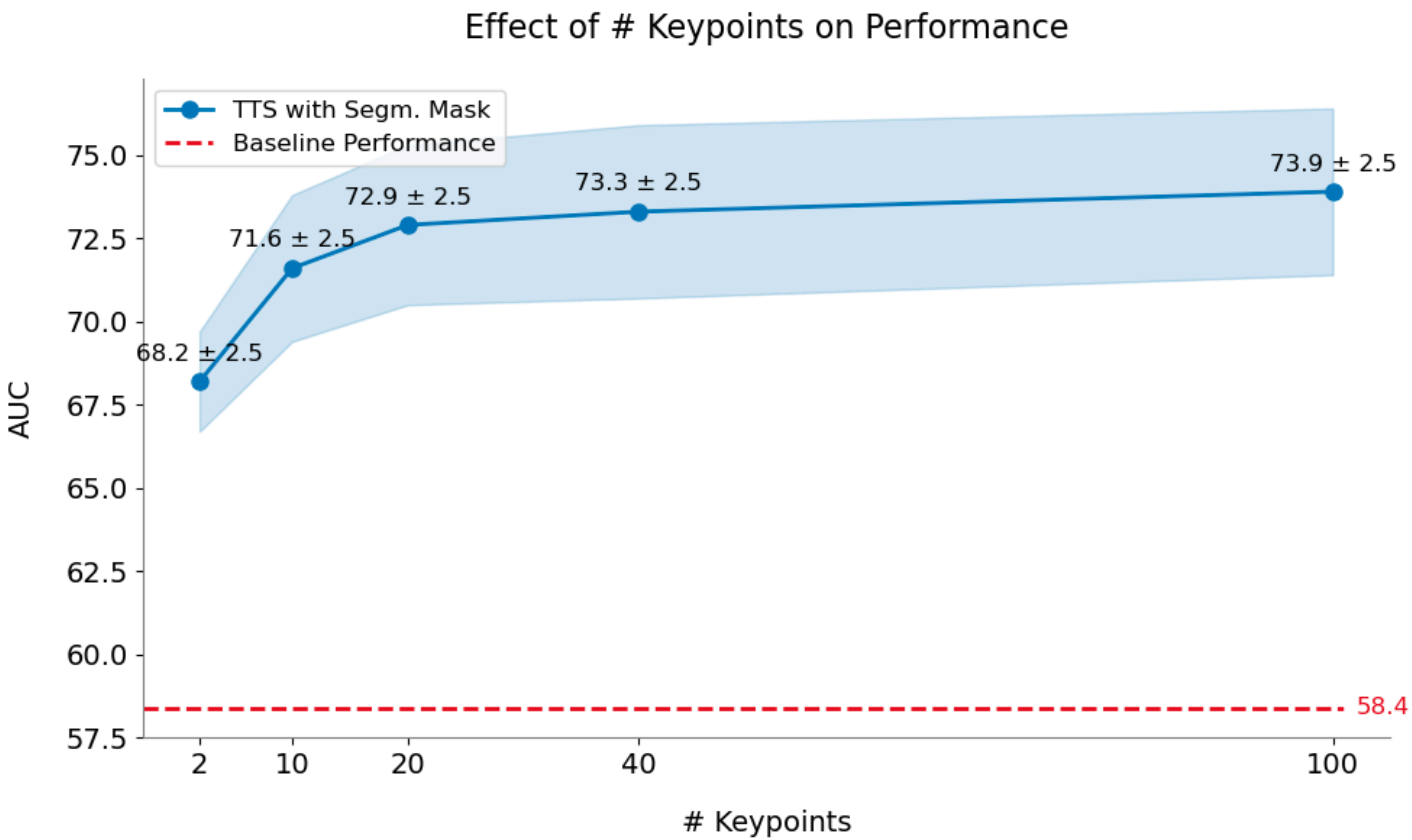


Results

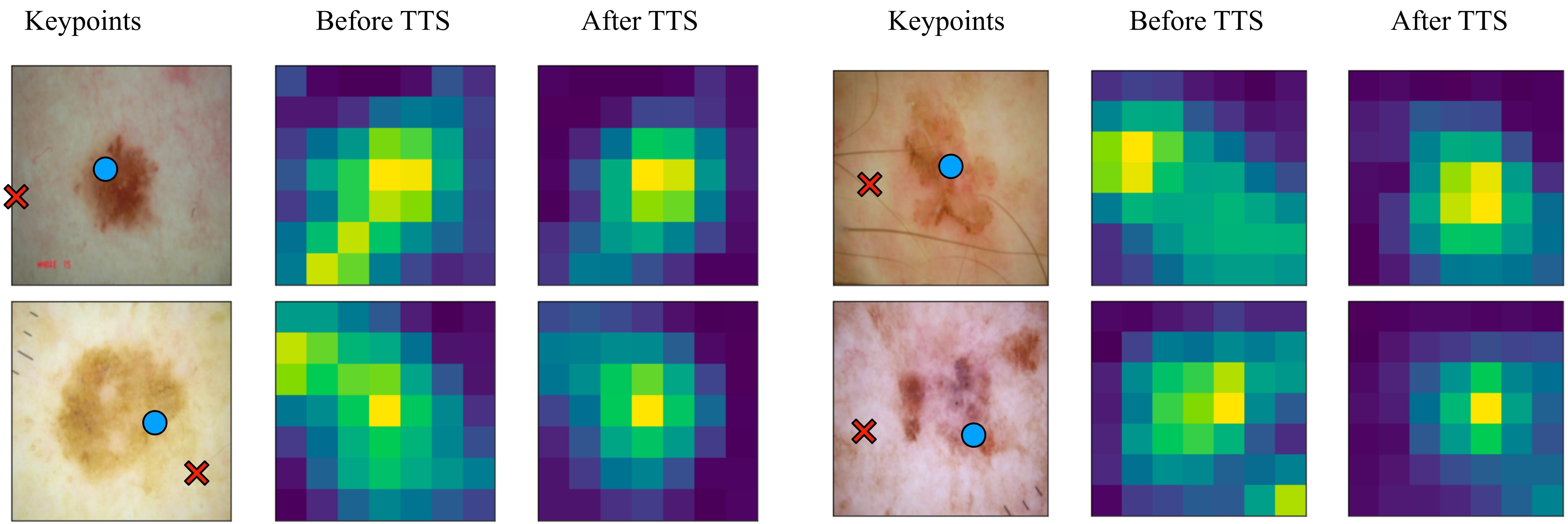
Effective throughout different training biases



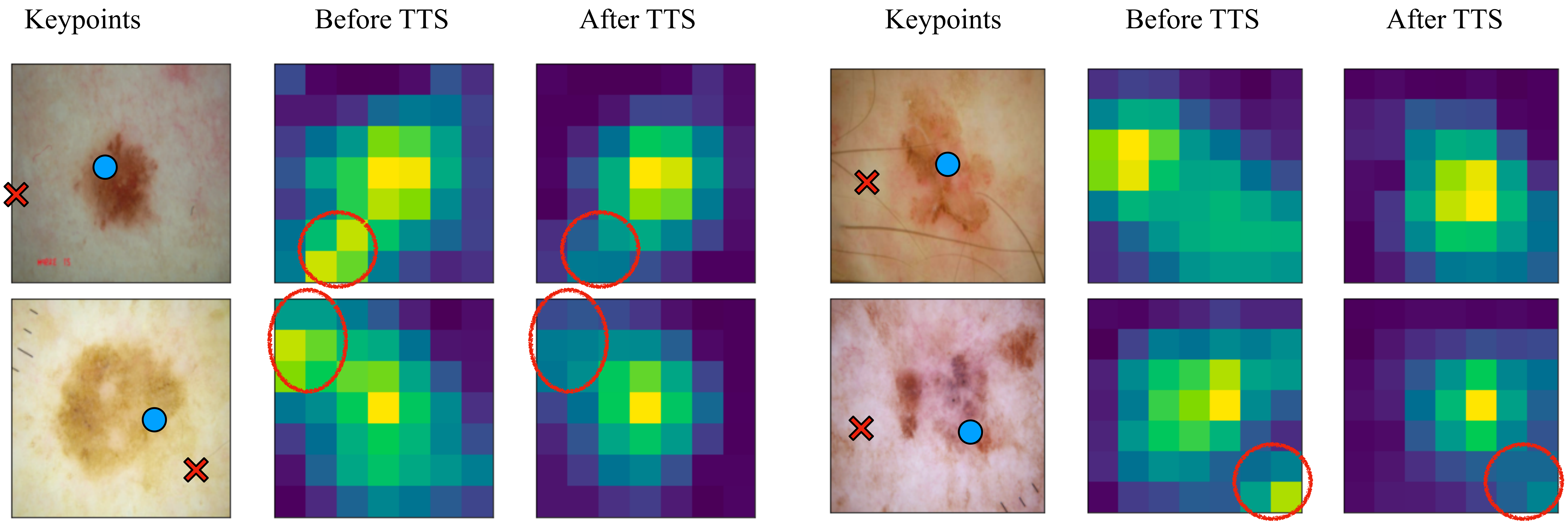
Effective even with a single pair of keypoints



Visualization

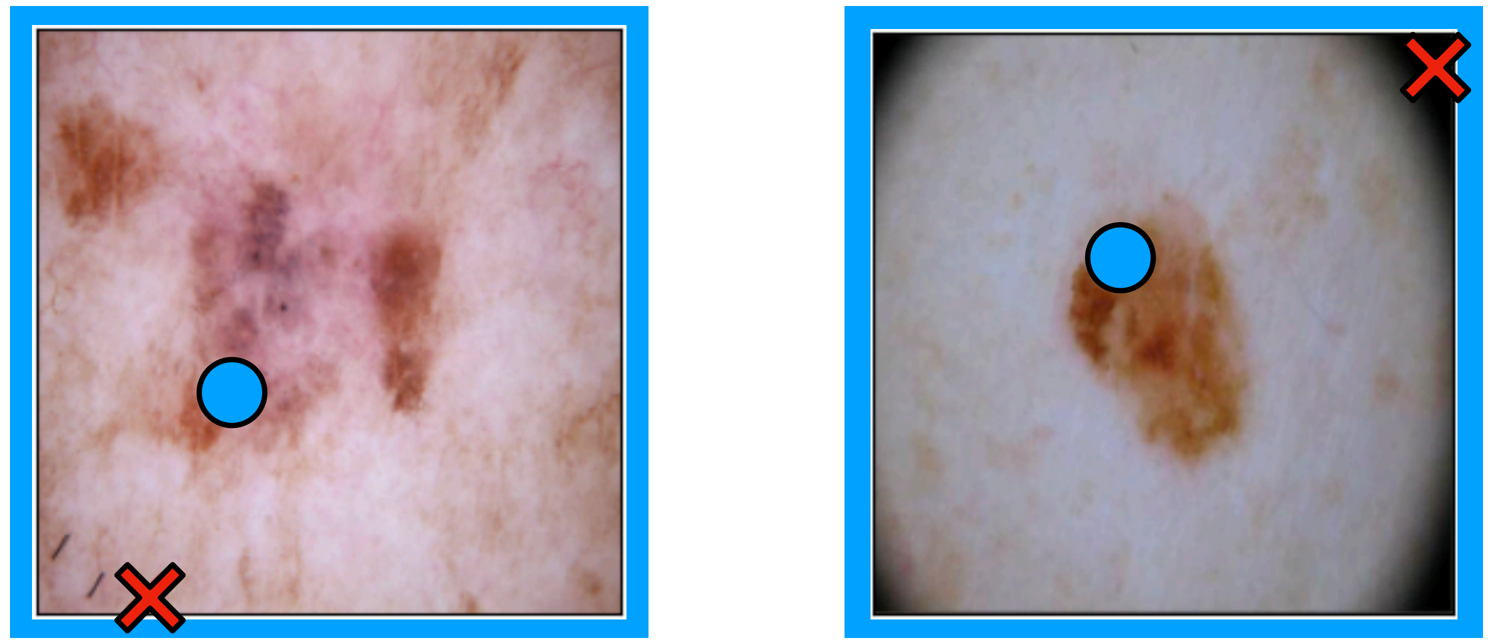


Visualization

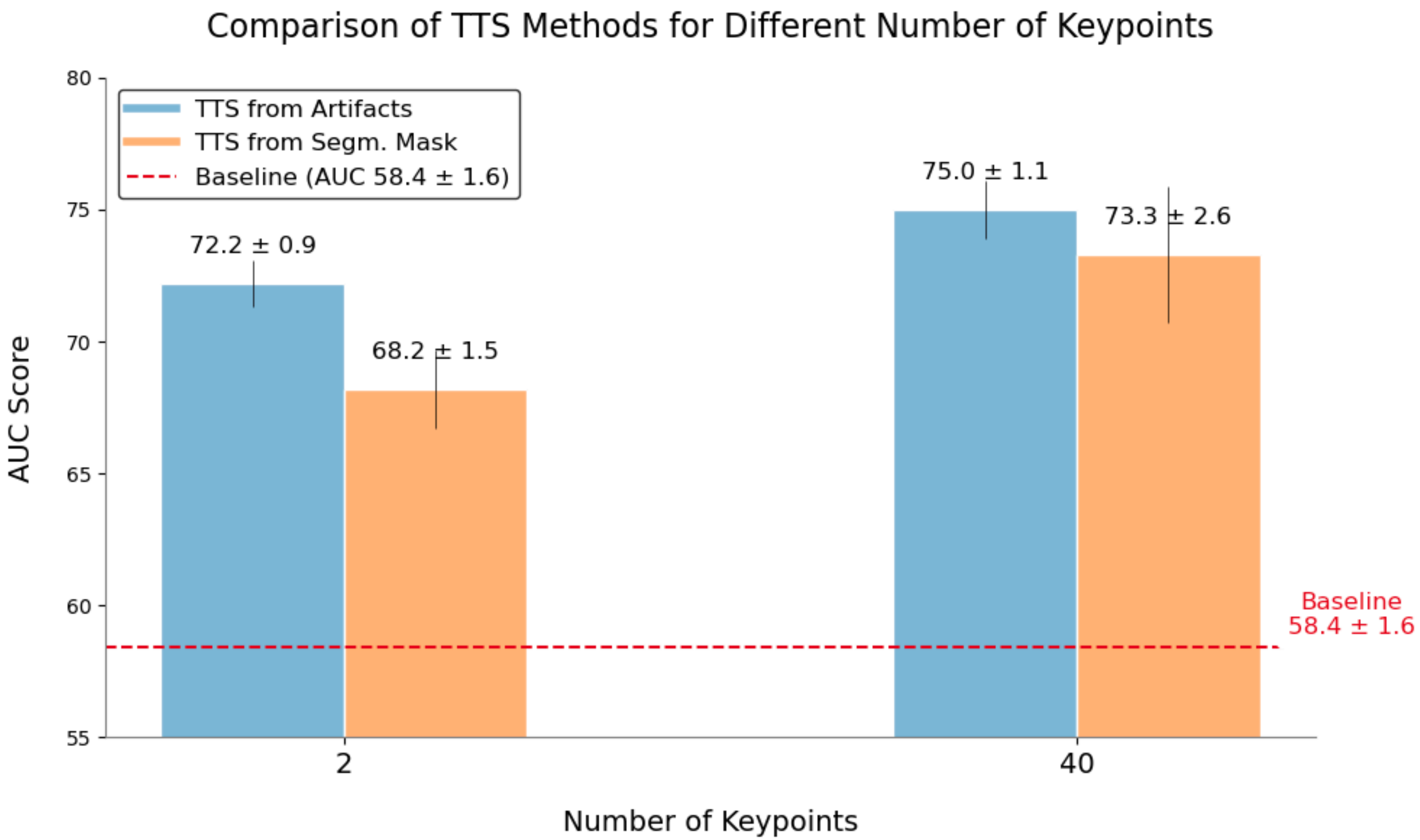


Flexible for different types of annotation

Keypoints on Artifacts



Keypoints from Segmentation Mask



Limitations

- How to adapt this solution to Vision Transformers?

- How to deal with biases uniformly spread across the image? (e.g., different acquisition devices.)

Takeaways

- Consider evaluating your models’ robustness on trap sets
- TTS improves robustness across different levels of bias
- TTS is effective even with a single pair of keypoints
- TTS is flexible to different types of annotations

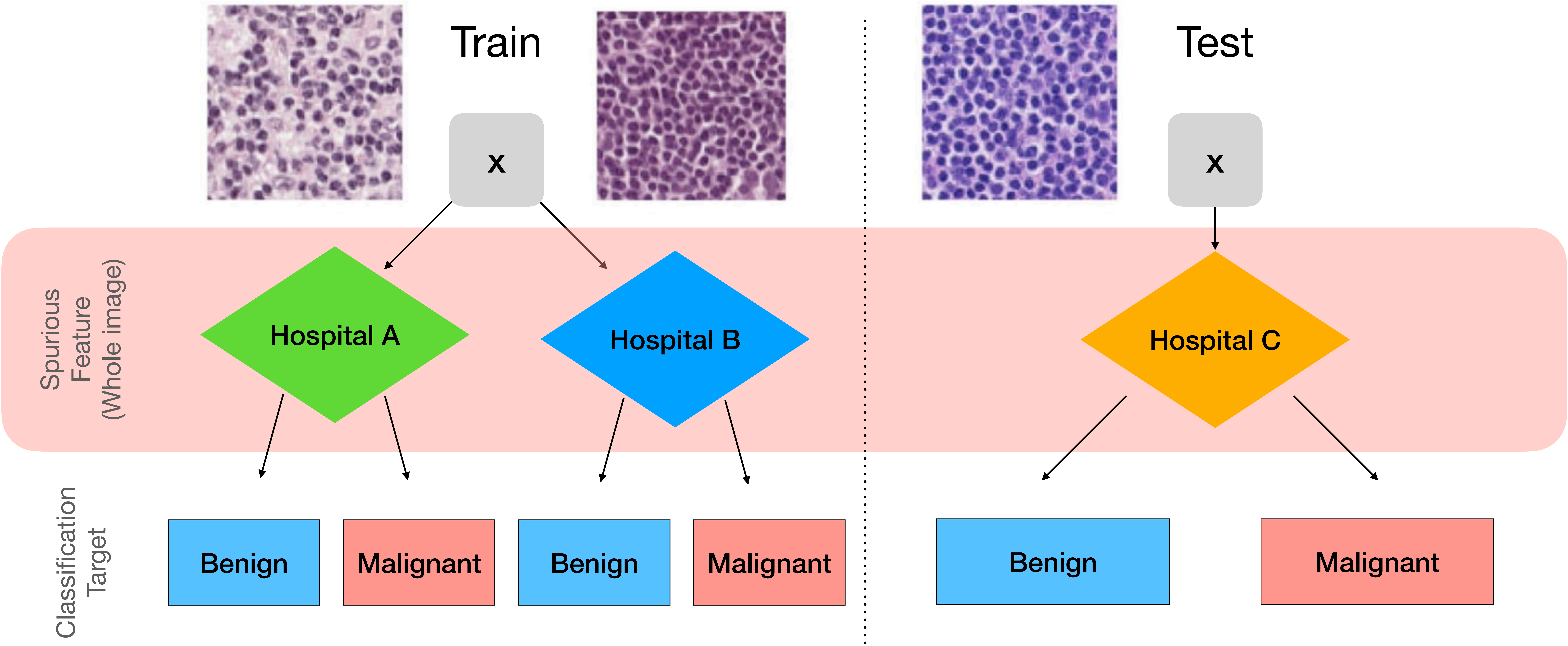
Code, Data & Paper:
https://github.com/alceubissoto/skin-tts

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

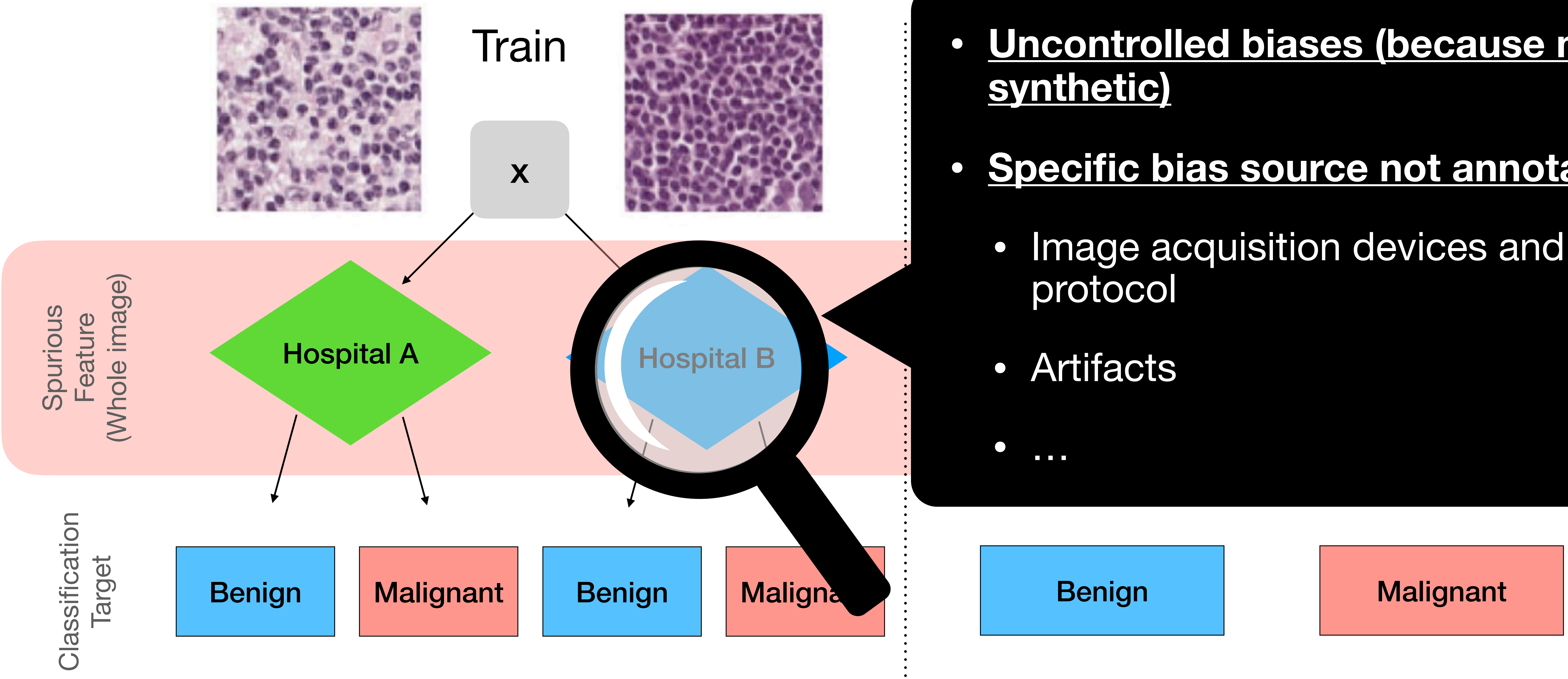
UNICAMP recod LARA FAPESP

Spurious Features vs. Generalization



X

Spurious Features vs. Generalization



- **<u>Uncontrolled biases (because not synthetic)</u>**
- **<u>Specific bias source not annotated:</u>**
  - Image acquisition devices and protocol
  - Artifacts
  - …

X

Evaluation - Generalization and Spurious Correlations

Subgroup Evaluation

Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge

Marc Combalia*, Noel Codella*, Veronica Rotemberg*, Cristina Carrera, Stephen Dusza, David Gutman, Brian Helba, Harald Kittler, Nicholas R Kurtansky, Konstantinos Liopyris, Michael A Marchetti, Sebastian Podlipnik, Susana Puig, Christoph Rinner, Philipp Tschandl, Jochen Weber, Allan Halpern*, Josep Malvehy*

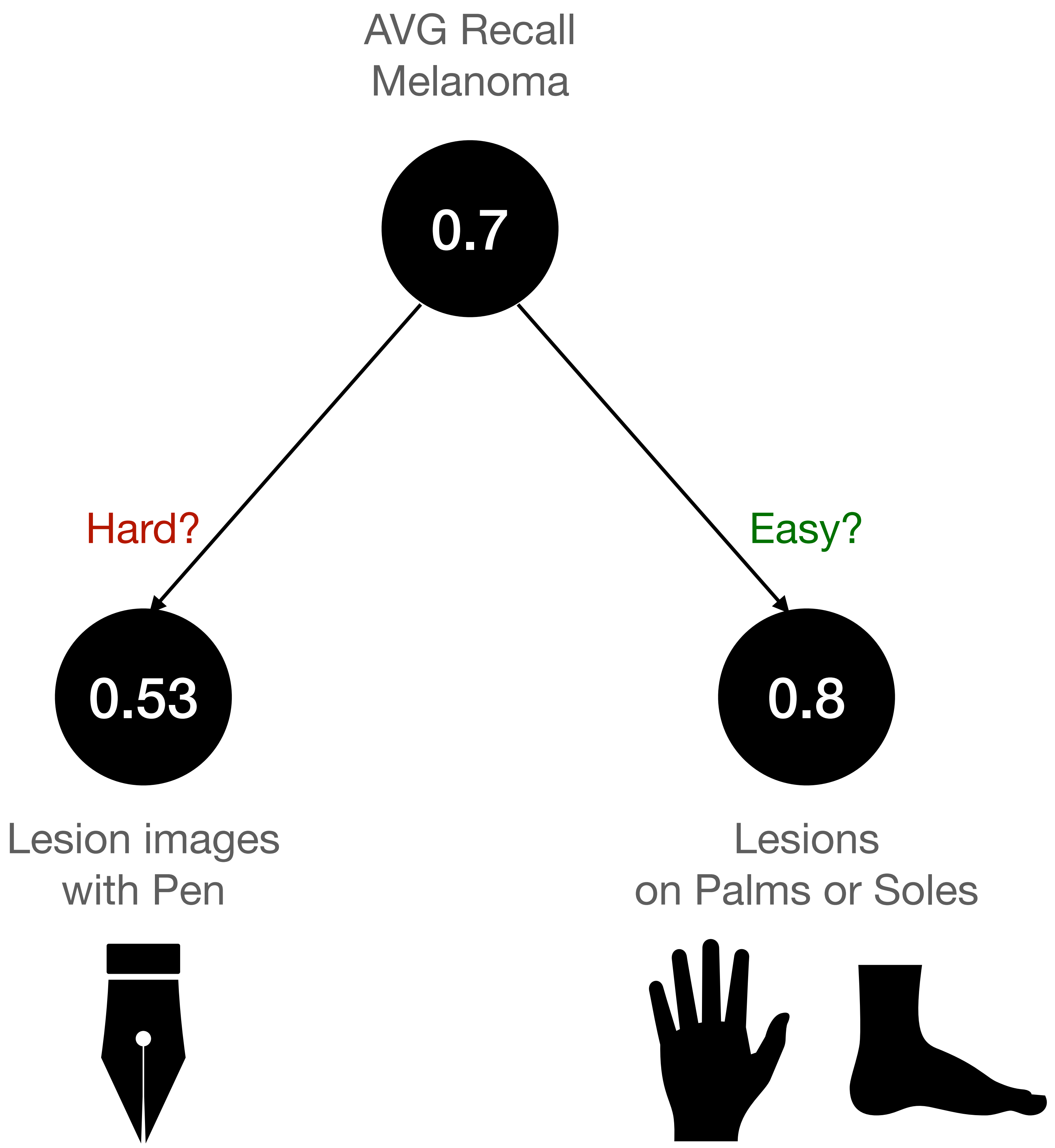
Summary

Background Previous studies of artificial intelligence (AI) applied to dermatology have shown AI to have higher diagnostic classification accuracy than expert dermatologists; however, these studies did not adequately assess clinically realistic scenarios, such as how AI systems behave when presented with images of disease categories that are not included in the training dataset or images drawn from statistical distributions with significant shifts from training distributions. We aimed to simulate these real-world scenarios and evaluate the effects of image source institution, diagnoses outside of the training set, and other image artifacts on classification accuracy, with the goal of informing clinicians and regulatory agencies about safety and real-world accuracy.

		MEL								
		MEL–NV	MEL–BKL	MEL–MEL	MEL–SCC	MEL–AK	MEL–BC	MELC–VASC	MEL–DF	MEL–NT
	Ref	0·14	0·04	0·7	0·0079	0·039	0·037	0·0076	0·0053	0·028
Artifacts	No crust	0·14	0·04	0·7	0·008	0·039	0·037	0·0077	0·0054	0·028
	Crust	0·016	0·019	0·83	0·0024	0·052	0·054	0	0	0·024
	No hair	0·14	0·038	0·7	0·0083	0·044	0·032	0·0064	0·0055	0·028
	Hair	0·15	0·052	0·67	0·0059	0·011	0·067	0·015	0·004	0·027
	No pen	0·13	0·039	0·71	0·0083	0·039	0·033	0·008	0·0054	0·026
	Pen	0·23	0·048	0·53	0·0024	0·037	0·098	0·002	0·0039	0·05
	No pigmentation	0·0075	0·018	0·46	0·045	0·046	0·27	0·075	0·0069	0·073
	Pigmentation	0·15	0·041	0·71	0·0056	0·039	0·022	0·0032	0·0052	0·025
	No ulceration	0·16	0·043	0·69	0·0053	0·042	0·024	0·005	0·0057	0·027
	Ulceration	0·0014	0·013	0·74	0·028	0·02	0·13	0·027	0·0022	0·037
Anatomical site	Torso	0·21	0·016	0·69	0·004	0·0076	0·037	0·012	0·0034	0·022
	Head or neck	0·021	0·089	0·64	0·0035	0·15	0·059	0·0002	0	0·032
	Lower extremity	0·17	0·022	0·73	0·011	0·0043	0·019	0·0021	0·016	0·019
	Upper extremity	0·13	0·057	0·67	0·015	0·022	0·051	0·01	0·0051	0·046
	Palms or soles	0·079	0·025	0·8	0·017	0·012	0·016	0·018	0·0003	0·032
	Oral or genital									

Evaluation - Generalization and Spurious Correlations

Subgroup Evaluation

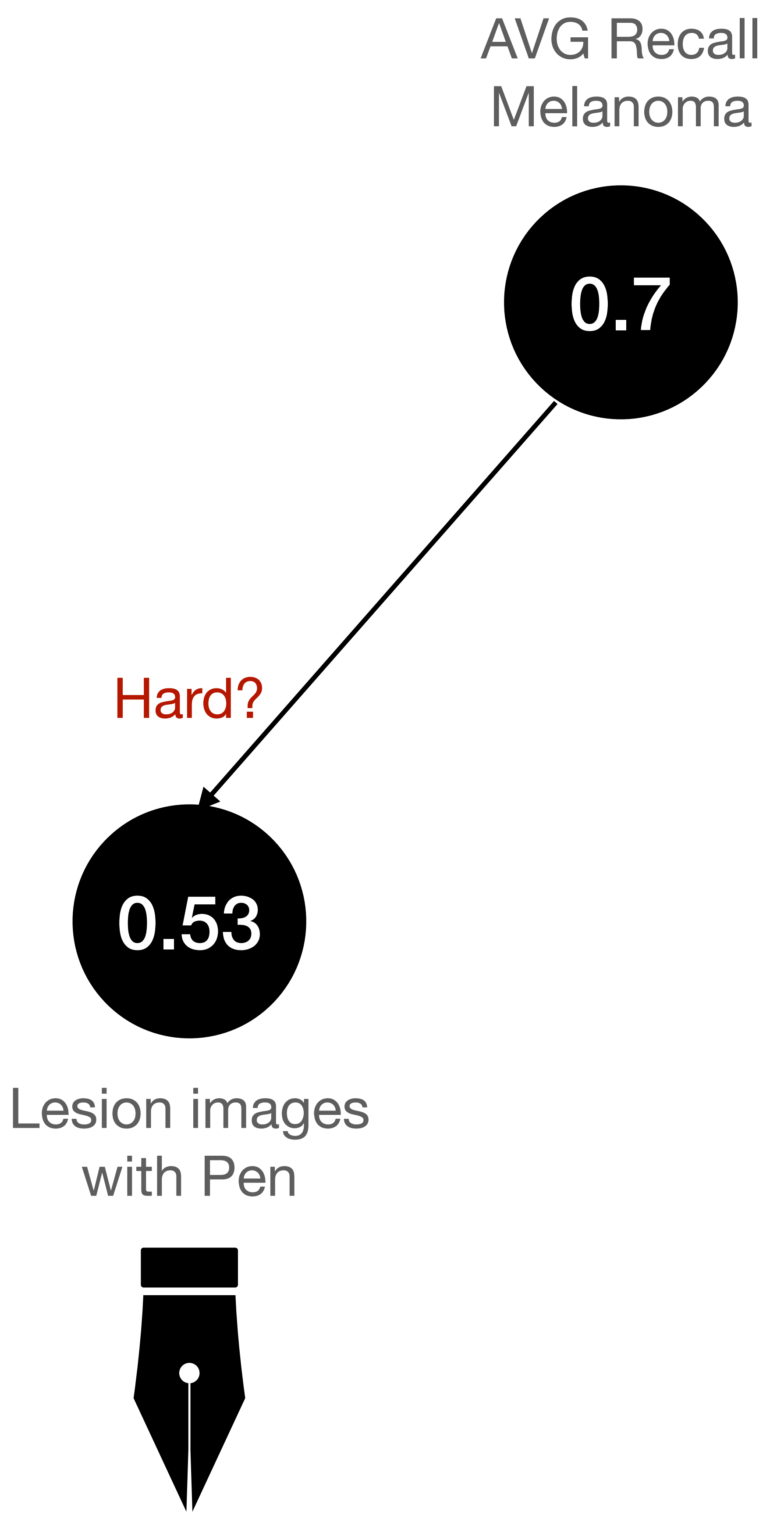


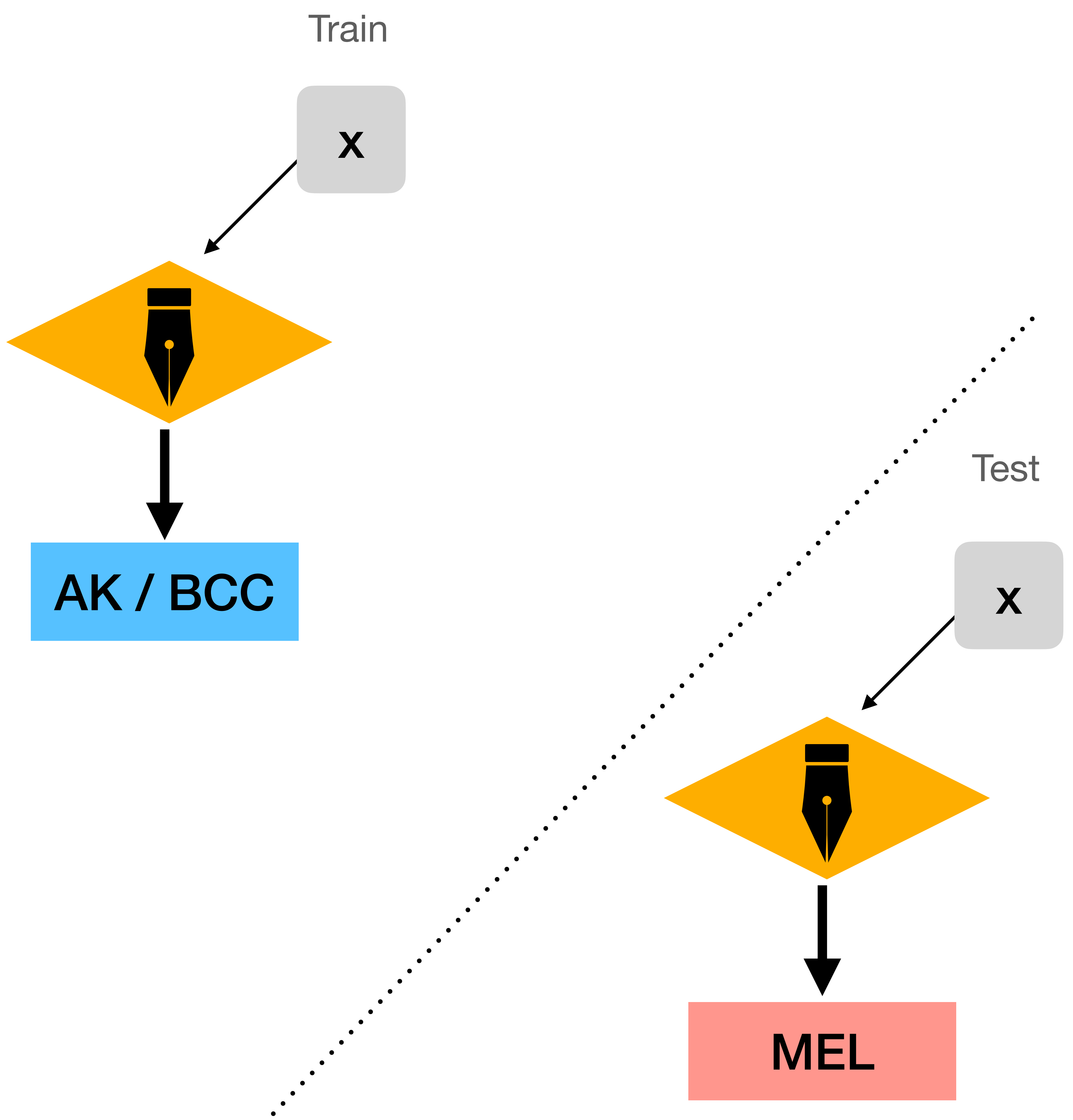
X

Combalia et al. "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022

Evaluation - Generalization and Spurious Correlations

Subgroup Evaluation



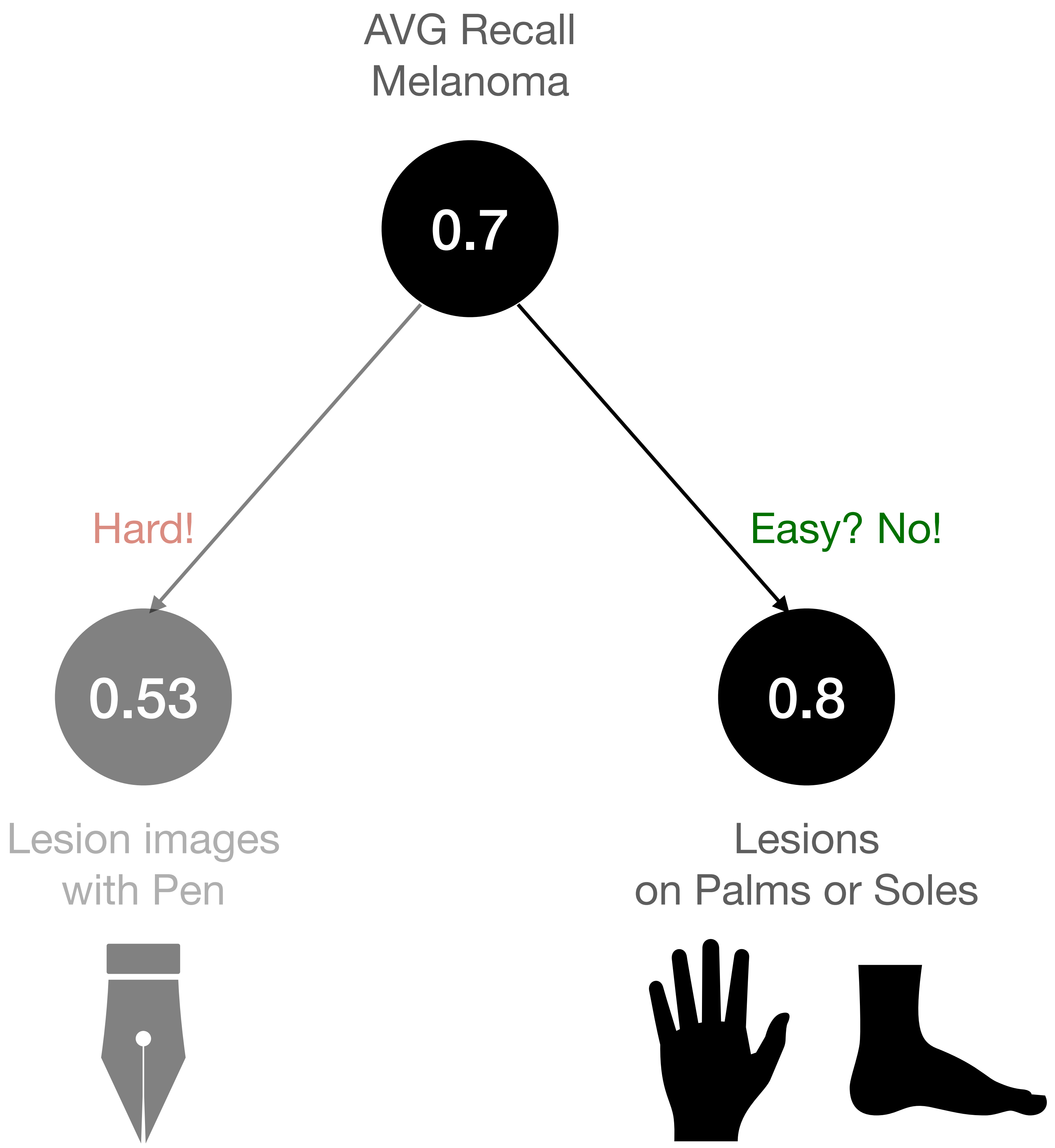


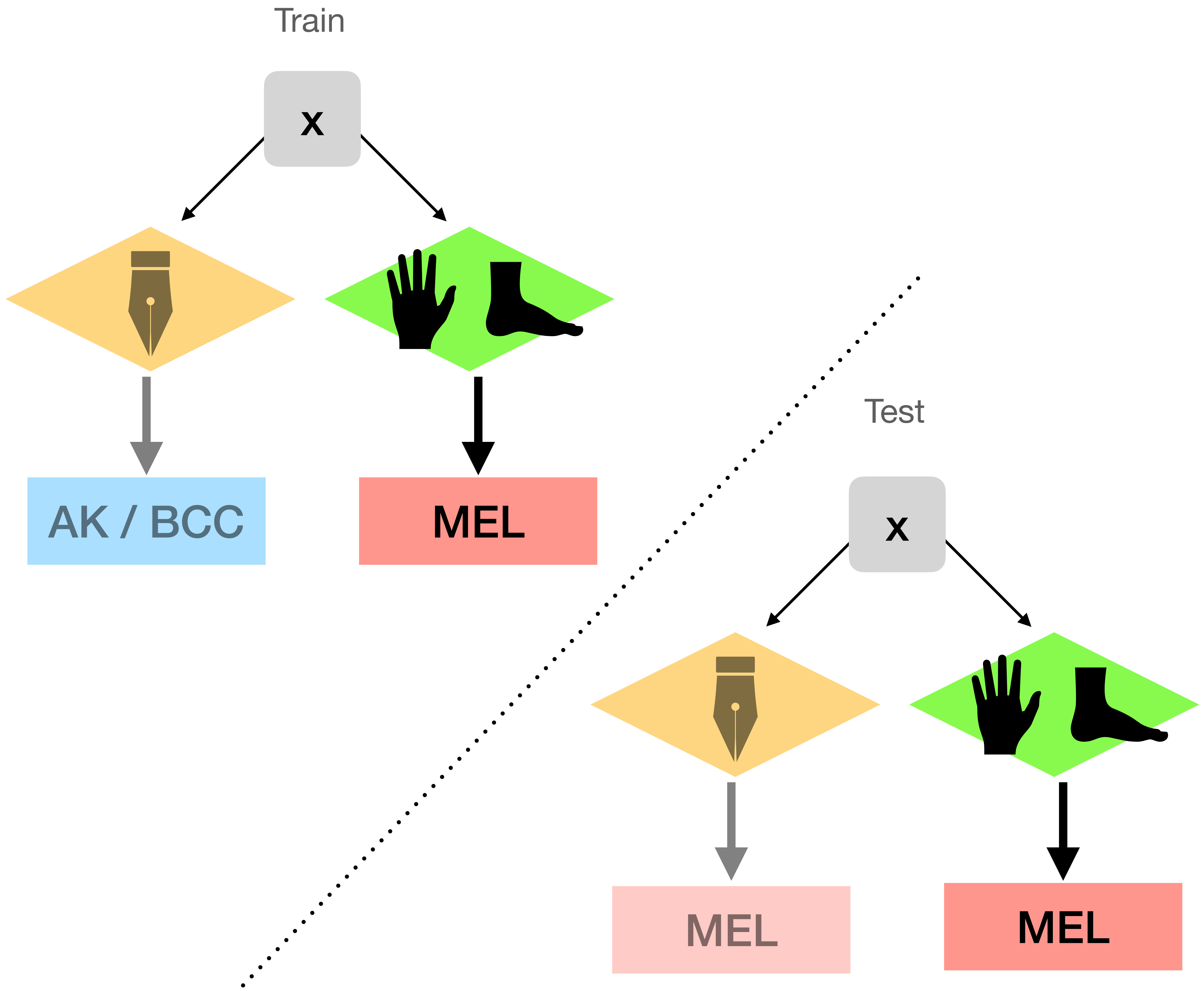
X

Combalia et al. "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022

Evaluation - Generalization and Spurious Correlations

Subgroup Evaluation





X

Combalia et al. "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge", The Lancet, 2022