CIRCLe: **C**olor **I**nvariant **R**epresentation Learning for Unbiased **C**lassification of Skin **Le**sions



Arezou Pakzad, Kumar Abhishek, Ghassan Hamarneh

Seventh ISIC Skin Image Analysis Workshop

In conjunction with **ECCV 2022**

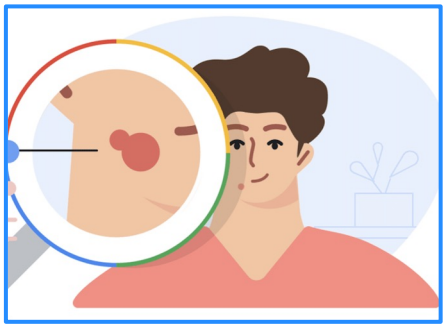
October 23, 2022

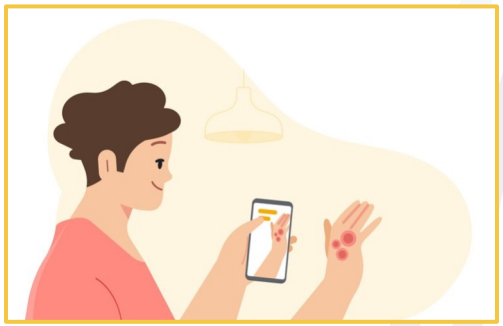
SFU SIMON FRASER UNIVERSITY

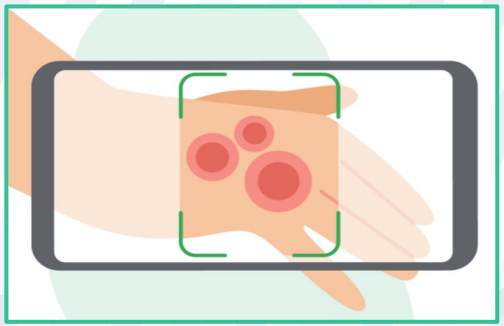
MIAL
MEDICAL IMAGE ANALYSIS LAB

Introduction

- **Convolutional neural networks (CNNs)** can be helpful decision support tools in healthcare.
- DL-based models can reach the dermatologist-level **classification** accuracies for skin diseases.



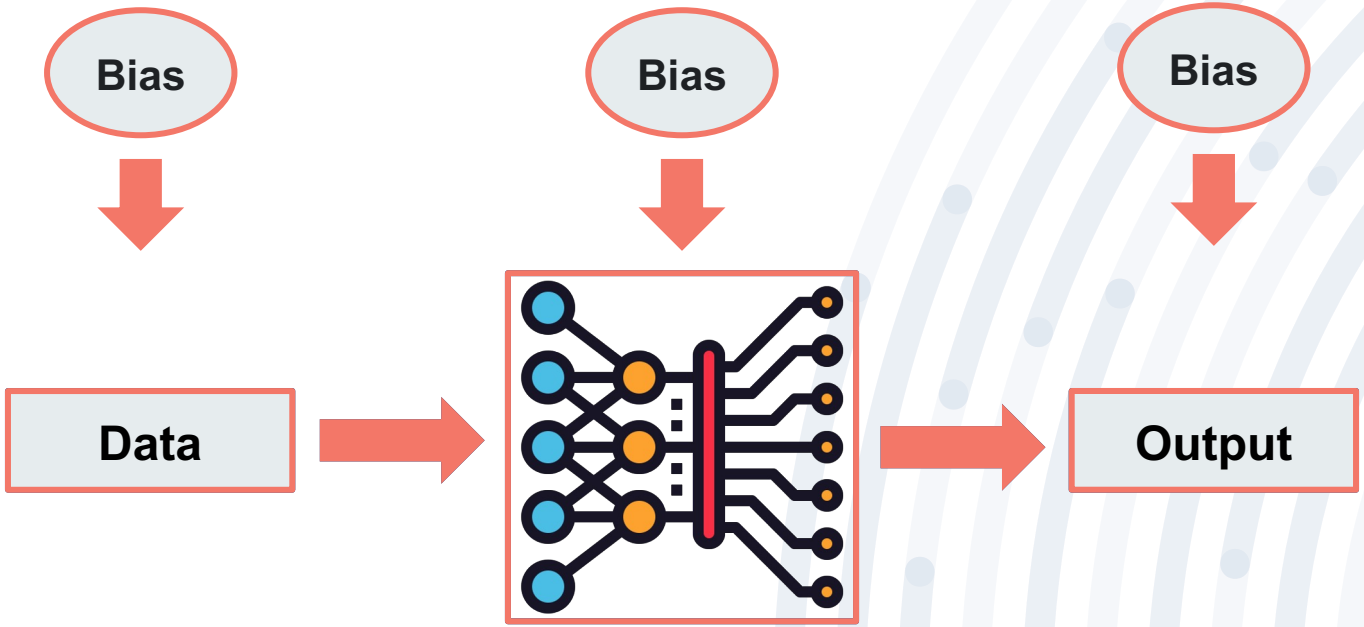




Images Source: https://blog.google/technology/health/ai-dermatology-preview-io-2021/

Bias In Predictions

- Data-driven learning paradigm



Fairness in Skin Image Analysis

- **Darker skin** is **under-represented** in most publicly available data sets.
- Skin conditions **appear differently** across different skin types.
- The **data imbalance** across different skin types → racial biases in **diagnosis**

Images Source: https://www.statnews.com/2020/07/21/dermatology-faces-reckoning-lack-of-darker-skin-in-textbooks-journals-harms-patients-of-color/

Images Source: https://www.verywellhealth.com/psoriasis-on-dark-skin-5218057

Contributions

- **C**olor **I**nvariant **R**epresentation learning for unbiased **C**lassification of skin **Le**sions (**CIRCLe**)
- **Skin color transformations** and skin **color-invariant** disease classification
- A new **fairness metric**: **Normalized Accuracy Range** → works with multiple protected groups
- Comprehensive evaluation of our proposed method

Problem

- Dataset $\mathcal{D} = \{X, Y, Z\}$

x: input

y: label

z: protected attribute

- M classes ($|Y| = M$)
- N protected attributes ($|Z| = N$)

Train a classification model that:

- Its prediction is invariant to the protected attribute z
- Model's classification performance is maximized.

Approach

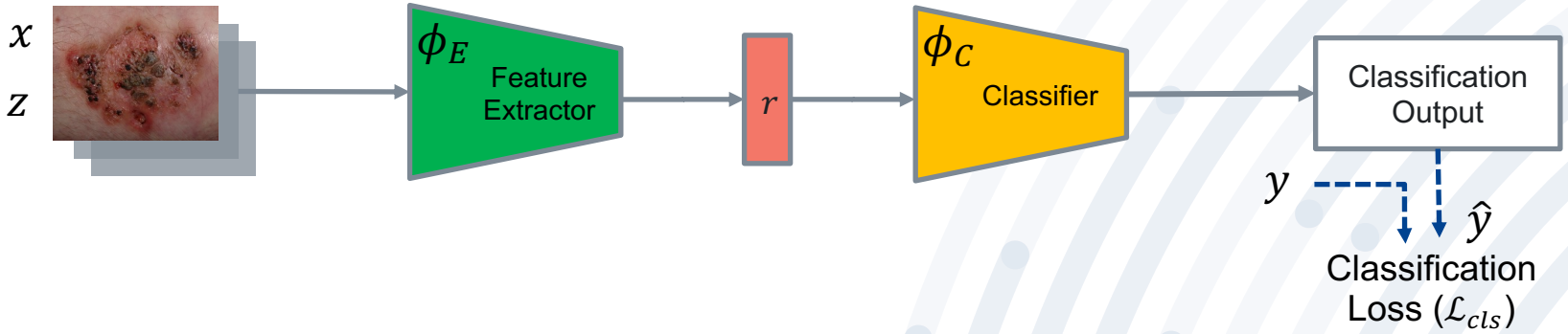
- **Domain Invariant Representation Learning**
  - **Fairness Definition**
    - **Statistical Parity:** independence between the model's prediction and the protected attribute

- Learn data distributions that are independent of the underlying skin types

Approach

1) **Feature Extractor and Classifier**
2) **Regularization Network**
   - **Skin Color Transformer**
     - To learn transformations between skin type domains
   - **Domain Regularization Loss**
     - To enforce the color invariant condition

Feature Extractor and Classifier



Prediction function $\hat{y} = f_\theta(x)$ $\hat{y} = \phi_C \circ \phi_E(x)$

Feature extractor $r = \phi_E(x)$

$r \in \mathbb{R}^p$

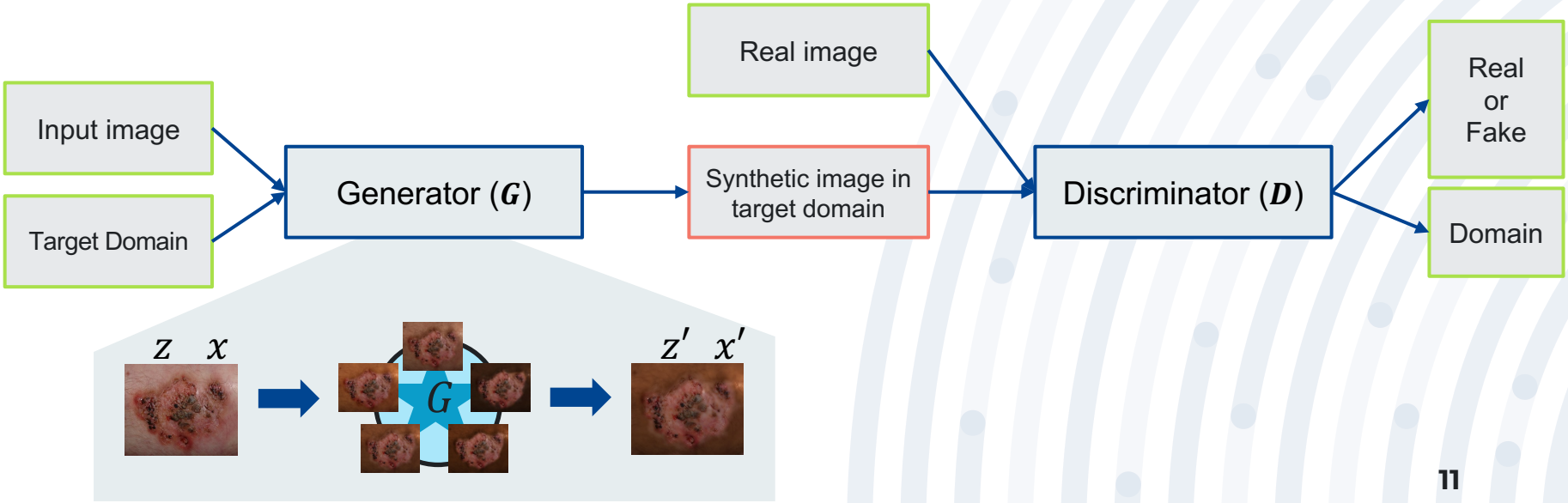
Classifier $\hat{y} = \phi_C(r)$

Skin Color Transformer

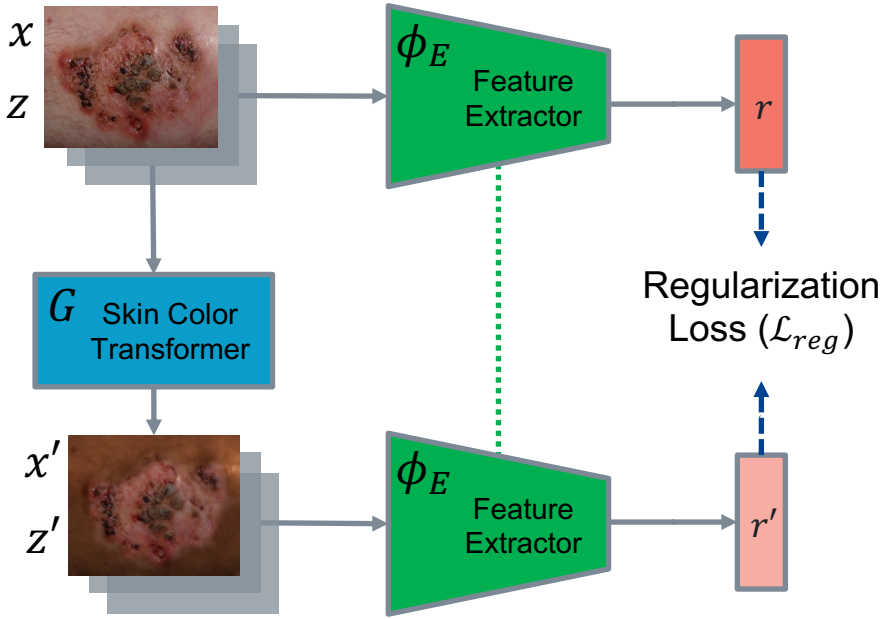
- Learn the function $G(x, z, z')$ that performs image-to-image transformations between skin type domains using **StarGAN.**

Skin Color Transformer

- Learn the function $G(x,\ z,\ z')$ that performs image-to-image transformations between skin type domains using **StarGAN.**



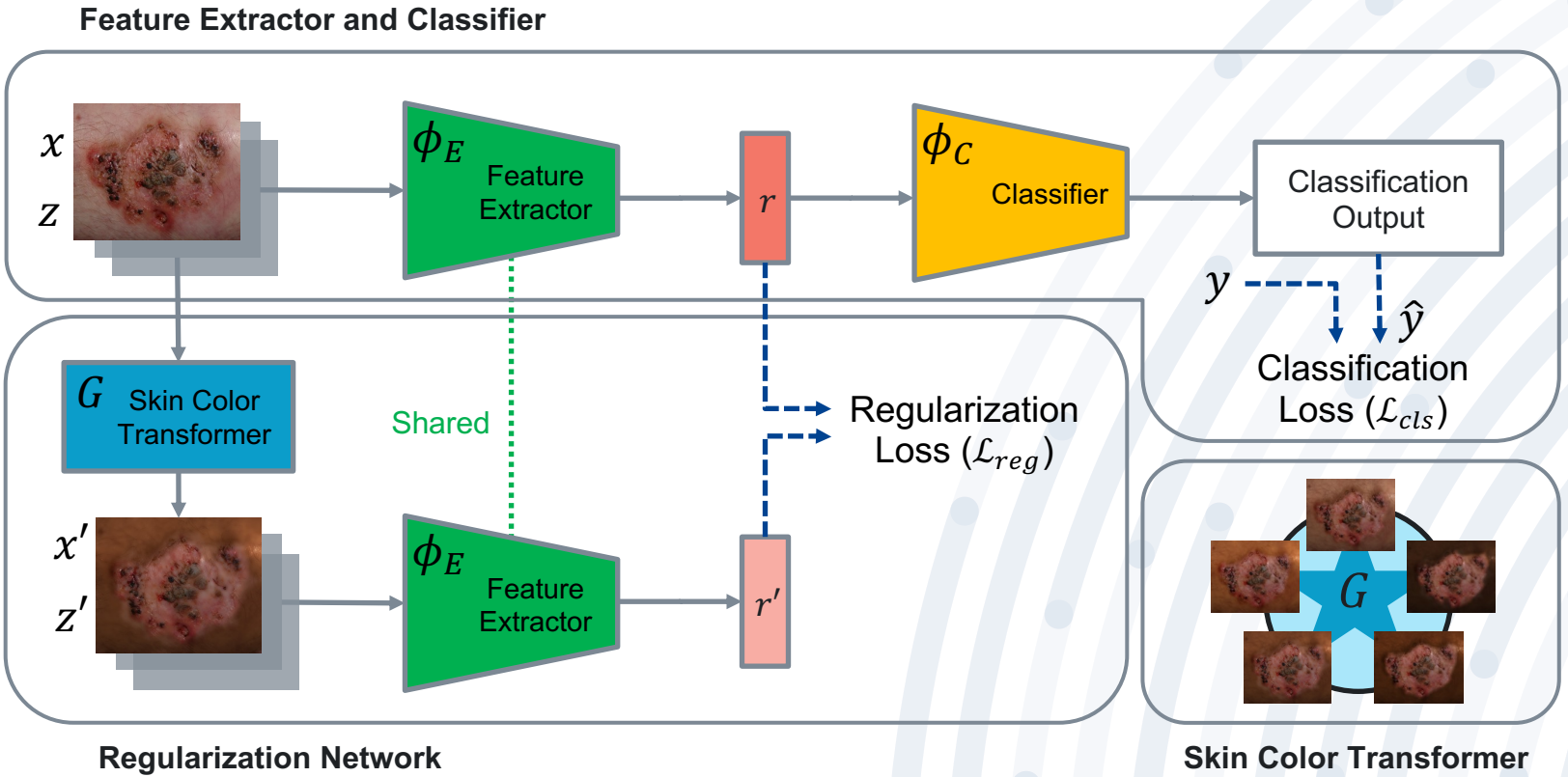
Domain Regularization Loss



- Enforce the model to learn similar representations for the original and the synthetic image
- $\mathcal{L}_{cls}$: **Classification loss**
  - Cross Entropy Loss
- $\mathcal{L}_{reg}$: **Regularization loss**
  - Squared Error Distance

$$\mathcal{L}_{reg} = \|r - r'\|_2^2$$

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg}$$


Feature Extractor and Classifier
x
z
ϕ_E
Feature Extractor
r
ϕ_C
Classifier
Classification Output
y
$\hat{y}$
Classification Loss ($\mathcal{L}_{cls}$)
G
Skin Color Transformer
Shared
Regularization Loss ($\mathcal{L}_{reg}$)
x'
z'
ϕ_E
Feature Extractor
r'
G
Regularization Network
Skin Color Transformer

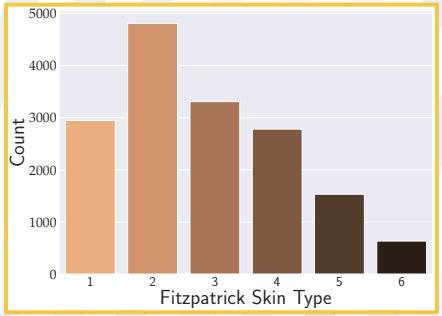
Dataset

- Fitzpatrick17K Dataset [1]
- 16,577 clinical images
- 114 skin conditions

[1] Groh et al., “Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset”, *CVPR* (2021).

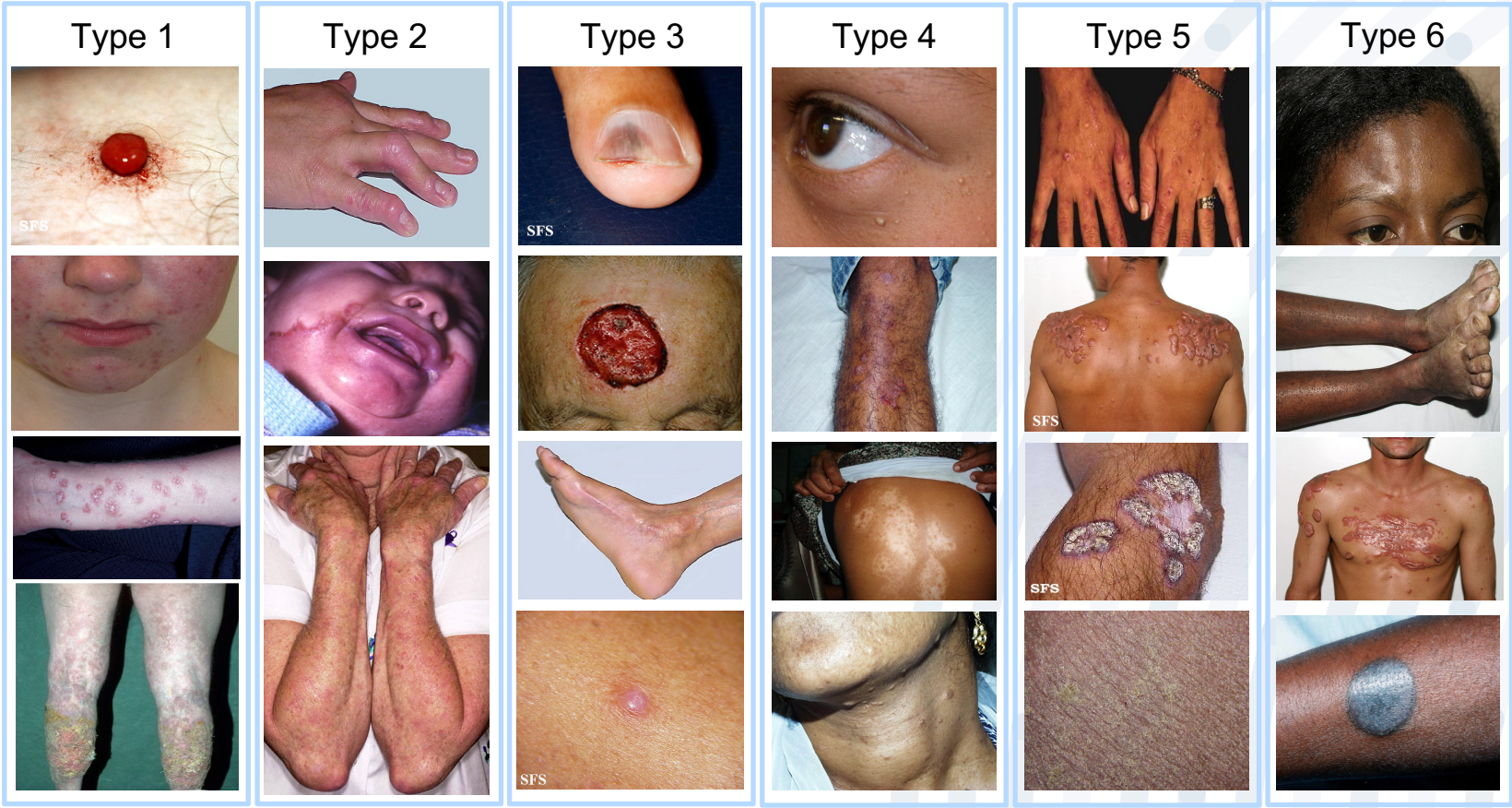
Dataset

- Fitzpatrick17K Dataset [1]
- 16,577 clinical images
- 114 skin conditions
- Each image has Fitzpatrick skin type (**FST**) label



[1] Groh et al., “Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset”, *CVPR* (2021).

Fitzpatrick17K dataset



Metrics

- Accurate and fair skin condition classifier
- **Classification performance**
  - Recall, F1-score, Accuracy

Metrics

- Accurate and fair skin condition classifier
- **Fairness**
  - **Equal Opportunity Difference (EOD)**
    - Difference in TPR rates for the two protected groups
    - Light (FSTs 1, 2, and 3) versus dark (FSTs 4, 5, and 6)

$$\text{EOD} = \left| TPR_{z=dark} - TPR_{z=light} \right|$$

Metrics

- Accurate and fair skin condition classifier
- **Fairness**
  - **Normalized Accuracy Range (NAR)**
    - Assess the accuracy (ACC) disparities across all the six skin types

$$\text{NAR} = \frac{ACC_{max} - ACC_{min}}{mean(ACC)}$$

$$ACC_{max} \approx ACC_{min} \implies NAR \approx 0$$

Models

- **Baseline** [1]
- **Improved Baseline** (Ours)
  - Ablation study → No regularization loss $\mathcal{L}_{reg}$
- **CIRCLe** (Ours)
- **Multiple Backbones**
  - Covering a wide range of CNN architecture families
  - Ablation study for all models

[1] Groh et al., “Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset”, *CVPR* (2021).

Results

- **Classification and Fairness Performance**

- **Improved Baseline** method recognizably outperforms the baseline method in accuracy and fairness.

Model	Recall	F1-score	Accuracy							EOD ↓	NAR ↓
			Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6		
Baseline	0.251	0.193	0.202	0.158	0.169	0.222	0.241	0.289	0.155	0.309	0.652
Improved Baseline (Ours)	0.444 (0.007)	0.441 (0.009)	0.471 (0.004)	0.358 (0.026)	0.408 (0.014)	0.506 (0.023)	0.572 (0.022)	0.604 (0.029)	0.507 (0.027)	0.261 (0.028)	0.512 (0.078)
CIRCLe (Ours)	**0.459** (0.003)	**0.459** (0.003)	**0.488** (0.005)	**0.379** (0.019)	**0.423** (0.011)	**0.528** (0.024)	**0.592** (0.022)	**0.617** (0.021)	**0.512** (0.043)	**0.252** (0.031)	**0.474** (0.047)

Note: values in parenthesis are std. dev. of the results for 5 different random seeds for data splitting

Results

- **Classification and Fairness Performance**

- New state-of-the-art performance on the Fitzpatrick17K dataset for the task of classifying the 114 skin conditions

Model	Recall	F1-score	Accuracy							EOD ↓	NAR ↓
			Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6		
Baseline	0.251	0.193	0.202	0.158	0.169	0.222	0.241	0.289	0.155	0.309	0.652
Improved	0.444	0.441	0.471	0.358	0.408	0.506	0.572	0.604	0.507	0.261	0.512
Baseline (Ours)	(0.007)	(0.009)	(0.004)	(0.026)	(0.014)	(0.023)	(0.022)	(0.029)	(0.027)	(0.028)	(0.078)
CIRCLe	**0.459**	**0.459**	**0.488**	**0.379**	**0.423**	**0.528**	**0.592**	**0.617**	**0.512**	**0.252**	**0.474**
(Ours)	(0.003)	(0.003)	(0.005)	(0.019)	(0.011)	(0.024)	(0.022)	(0.021)	(0.043)	(0.031)	(0.047)

Note: values in parenthesis are std. dev. of the results for 5 different random seeds for data splitting

Results

- **Different Backbones**

Model	$\mathcal{L}_{reg}$	Recall	F1-score	Accuracy							EOD ↓	NAR ↓
				Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6		
2.55M MobileNetV2	✗	0.375	0.365	0.398	0.313	**0.364**	0.409	0.503	0.491	0.333	0.280	0.472
	✓	**0.404**	**0.397**	**0.434**	**0.354**	0.357	**0.471**	**0.559**	**0.544**	**0.421**	0.258	**0.455**
MobileNetV3L	✗	**0.427**	0.403	0.438	0.357	0.388	0.449	0.543	**0.560**	0.413	0.271	0.449
	✓	0.425	**0.412**	**0.451**	**0.369**	**0.400**	**0.464**	**0.565**	0.550	**0.444**	0.275	**0.420**
DenseNet-121	✗	0.425	0.416	0.451	0.393	0.397	0.452	**0.565**	0.522	**0.500**	0.278	0.364
	✓	**0.441**	**0.430**	**0.462**	**0.413**	**0.406**	**0.473**	0.561	**0.550**	0.452	0.294	**0.324**
ResNet-18	✗	0.391	0.381	0.417	0.355	0.353	0.431	0.538	0.516	**0.389**	0.263	0.430
	✓	**0.416**	**0.410**	**0.436**	**0.367**	**0.380**	**0.458**	**0.543**	**0.538**	**0.389**	0.282	**0.395**
24.03M ResNet-50	✗	0.390	0.382	0.416	0.337	0.363	0.422	0.549	0.506	0.389	0.257	0.497
	✓	**0.440**	**0.429**	**0.466**	**0.384**	**0.402**	**0.502**	**0.580**	**0.569**	**0.421**	0.283	**0.411**

Results

- **Different Backbones**

Model	$\mathcal{L}_{reg}$	Recall	F1-score	Accuracy							EOD ↓	NAR ↓
				Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6		
2.55M MobileNetV2	✗	0.375	0.365	0.398	0.313	**0.364**	0.409	0.503	0.491	0.333	0.280	0.472
	✓	**0.404**	**0.397**	**0.434**	**0.354**	0.357	**0.471**	**0.559**	**0.544**	**0.421**	0.258	**0.455**
MobileNetV3L	✗	**0.427**	0.403	0.438	0.357	0.388	0.449	0.543	**0.560**	0.413	0.271	0.449
	✓	0.425	**0.412**	**0.451**	**0.369**	**0.400**	**0.464**	**0.565**	0.550	**0.444**	0.275	**0.420**
DenseNet-121	✗	0.425	0.416	0.451	0.393	0.397	0.452	**0.565**	0.522	**0.500**	0.278	0.364
	✓	**0.441**	**0.430**	**0.462**	**0.413**	**0.406**	**0.473**	0.561	**0.550**	0.452	0.294	**0.324**
ResNet-18	✗	0.391	0.381	0.417	0.355	0.353	0.431	0.538	0.516	**0.389**	0.263	0.430
	✓	**0.416**	**0.410**	**0.436**	**0.367**	**0.380**	**0.458**	**0.543**	**0.538**	**0.389**	0.282	**0.395**
24.03M ResNet-50	✗	0.390	0.382	0.416	0.337	0.363	0.422	0.549	0.506	0.389	0.257	0.497
	✓	**0.440**	**0.429**	**0.466**	**0.384**	**0.402**	**0.502**	**0.580**	**0.569**	**0.421**	0.283	**0.411**

Results

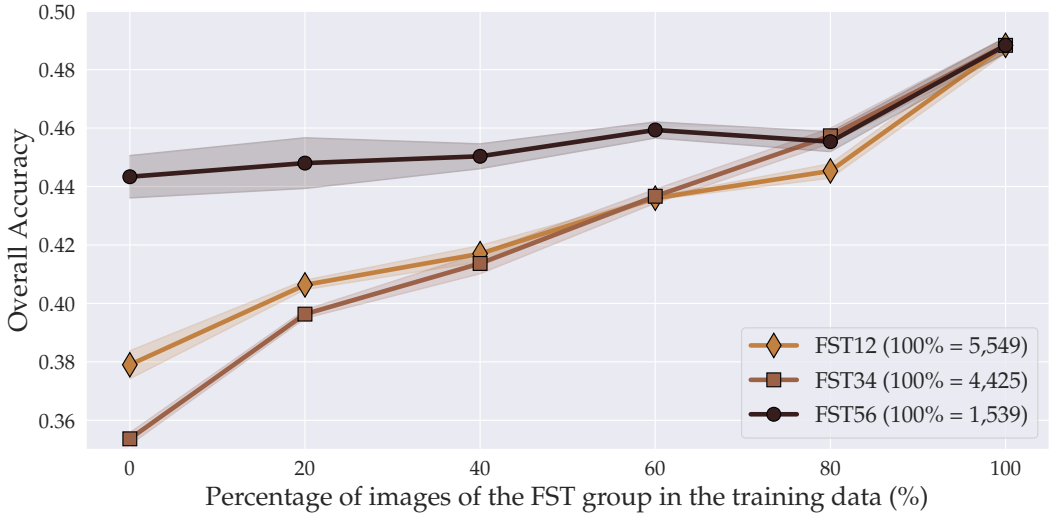
- **Domain Adaptation Performance**
  - "Two-to-other" experiment: train the model on all the images from two FST domains and test it on all the other FST domains.

Holdout Partition	Method	Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
FST3-6	Baseline	0.138	-	-	0.159	0.142	0.101	0.090
	Improved Baseline	0.249	-	-	0.308	0.246	0.185	0.113
	CIRCLe	**0.260**	-	-	**0.327**	**0.250**	**0.193**	**0.115**
FST12 and FST56	Baseline	0.134	0.100	0.130	-	-	0.211	0.121
	Improved Baseline	0.272	0.181	0.274	-	-	0.453	0.227
	CIRCLe	**0.285**	**0.199**	**0.285**	-	-	**0.469**	**0.233**
FST1-4	Baseline	0.077	0.044	0.055	0.091	0.129	-	-
	Improved Baseline	0.152	0.078	0.111	0.167	0.280	-	-
	CIRCLe	**0.163**	**0.095**	**0.121**	**0.177**	**0.293**	-	-

Results

- **Classification Performance Relation with Training Size**

- For each FST group, we gradually increase its number of images in the training set, and report the model's overall accuracy on the test set.
- With very limited or no representation of a skin type, CIRCLe can still perform well overall.



Conclusion

- We proposed **CIRCLe**, a method based on domain invariant representation learning, for mitigating skin type bias in clinical image classification.

- **CIRCLe** sets a new state-of-the-art performance on the classification of the 114 skin conditions in the Fitzpatrick17K dataset.

- We also proposed a new fairness metric **Normalized Accuracy Range** for assessing fairness of classification in the presence of multiple protected groups, and showed that **CIRCLe** improves fairness of classification.

Code:
https://github.com/arezou-pakzad/CIRCLe

Thank You!

Arezou Pakzad. arezou_pakzad@sfu.ca
Kumar Abhishek. kabhishe@sfu.ca
Ghassan Hamarneh. hamarneh@sfu.ca

Website: www.medicalimageanalysis.com

Digital Research Alliance of Canada
NSERC CRSNG
NVIDIA