Addressing Fairness and Data Limitations in Dermatological Diagnosis through Color-Invariant Representation Learning and Synthetic Data Generation

by

Arezou Pakzad

B.Sc., Sharif University of Technology, 2021

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Spring 2024

Declaration of Committee

Name:	**Arezou Pakzad**
Degree:	**Master of Science**
Thesis title:	**Addressing Fairness and Data Limitations in Dermatological Diagnosis through Color-Invariant Representation Learning and Synthetic Data Generation**
Committee:	**Chair:** Maxwell Libbrecht Associate Professor, Computing Science
	Ghassan Hamarneh Supervisor Professor, Computing Science
	Angelica Lim Committee Member Assistant Professor, Computing Science
	Manolis Savva Examiner Assistant Professor, Computing Science

Abstract

While deep learning-based approaches have demonstrated expert-level performance in dermatological diagnosis tasks, they rely on a data-driven learning paradigm that requires large-scale annotated data and mimic the biases therein (e.g., biases towards skin types). Furthermore, existing public dermatological datasets have limitations such as small size, narrow disease coverage, insufficient annotations, and non-standardized image acquisitions. In this thesis, we propose CIRCLe, a skin color-invariant deep representation learning method for improving fairness in skin lesion classification by utilizing a regularization loss to encourage images with the same diagnosis but different skin types to have similar latent representations. Moreover, we introduce DermSynth3D, a novel framework for synthesizing large-scale densely annotated *in-the-wild* dermatological images by blending skin disease patterns onto 3D textured meshes of human subjects using a differentiable renderer and generating 2D images from various camera viewpoints under chosen lighting conditions in diverse background scenes.

Keywords: Skin image analysis; Skin type bias; Dermatology; Classification; Lesion detection; Deep learning

Dedication

To my parents and my sister, without whom none of my success would be possible.

Thank you for all of your support along the way

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Ghassan Hamarneh for his invaluable advice, continuous support, encouragement, and patience during my study. Further, I would also like to thank other members of my examination committee, Prof. Angelica Lim, Prof. Manolis Savva, and Prof. Maxwell Libbrecht, for their time and feedback on this thesis. I want to thank my labmates and collaborators at the Medical Image Analysis Lab for their support and helpful feedback on my project. Lastly, my appreciation also goes out to my family and friends for their encouragement and support throughout my studies.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ACC	Accuracy
AI	Artificial Intelligence
CNN	Convolutional Neural Network
DDI	Diverse Dermatology Images
DL	Deep Learning
EOD	Equal Opportunity Difference
FST	Fitzpatrick Skin Type
GAN	Generative Adversarial Network
ISIC	International Skin Imaging Collaboration
MM	Millions
MulAddOps	Multiply-Add Operations
NAR	Normalized Accuracy Range
SGD	Stochastic Gradient Descent
StarGAN	Star Generative Adversarial Network

std. dev.	Standard Deviation
TPR	True Positive Rate

List of Notations

Chapter 3

$G(.)$	Generator model
$D(.)$	Discriminator model
x	Input image
y	Class label
z	Protected attribute
$\phi_C(.)$	Classifier
$\phi_E(.)$	Feature extractor
r	Feature representation
$\mathcal{L}(.)$	Loss function
λ	Loss weighting hyperparameter
N_c	Number of classes
N_p	Number of protected groups
θ	Learned model parameters
$f(\cdot)$	Classification model
$\mathbb{R}$	The set of real numbers
$\mathcal{S}$	Dataset
X	Set of input images
Y	Set of class labels
Z	Set of protected attributes

Chapter 4

x	2D image
W	Image width

H	Image height
$\tilde{W}$	2D view width
$\tilde{H}$	2D view height
W_T	Texture image width
H_T	Texture image height
V	Set of mesh vertices
F	Set of mesh faces
T	2D texture image
T_b	2D texture image w/ blended skin conditions
T_m	2D texture mask of blended skin conditions
T_{nonskin}	2D texture mask of non-skin regions
U	Set of UV texture coordinates
s	2D binary segmentation mask
$\tilde{a}$	2D view of a 3D mesh
$\tilde{z}$	2D view w/ depth values
$\tilde{a}_{T_b}$	2D view w/ blended skin condition
$\tilde{a}_{T_m}$	2D mask of the skin condition
a_{skin}	2D mask of the skin
a_{nonskin}	2D mask of the non-skin regions
M	3D mesh

Chapter 1

Introduction

1.1 Background and Motivation

Diagnosis and analysis of skin conditions are an enormous burden on the healthcare system, with *at least* 3000 distinct skin diseases identified so far [18]. Both human dermatologists and sophisticated computerized approaches struggle to address this complex task of analyzing skin conditions.

Owing to the advancements in deep learning (DL)-based data-driven learning paradigm, convolutional neural networks (CNNs) can be helpful decision support tools in healthcare. This is particularly true for dermatological applications where recent research has shown that DL-based models can reach the dermatologist-level classification accuracies for skin diseases [22,41,49] while doing so in a clinically interpretable manner [13,77]. Computerized analysis of skin diseases often rely on 2D colored images, with significant research efforts devoted to analysis of conditions within clinical images [72].

However, this data-driven learning paradigm that allows models to automatically learn meaningful representations from data leads DL models to mimic biases found in the data, i.e., biases in the data can propagate through the learning process and result in an inherently biased model, and consequently in a biased output. Although research into algorithmic bias and fairness has been an active area of research, interest in the fairness of machine learning algorithms, in particular, is fairly recent. Multiple studies have shown the inherent racial disparities in machine learning algorithms' decisions for a wide range of areas: pre-trial bail decisions [64], recidivism [9], healthcare [84], facial recognition [23], and college

admissions [65]. Specific to healthcare applications, previous research has shown the effect of dataset biases on DL models' performance across genders and racial groups in cardiac MR imaging [91], chest X-rays [70,100,101], and skin disease imaging [47]. Recently, Groh et al. [47] showed that CNNs are the most accurate when classifying skin diseases manifesting on skin types similar to those they were trained on.

Moreover, this data-driven learning paradigm of DL-based models, requires large-scale annotated training data. Current publicly available datasets of clinical images are used for training DL-based models to perform various tasks, such as classification [37,48,57,108,124], lesion segmentation [51, 80], lesion tracking [44, 107, 129], lesion management [3], and skin tone prediction [62]. While there are numerous publicly available 2D dermatological image datasets [80], existing "in-the-wild" clinical datasets have limitations in creating semantically rich ground truth (GT) labels that can be used for the diverse range of dermatological tasks mentioned.

1.2 Thesis Contributions

In this thesis, we aim to propose methodologies to improve skin type fairness and classification performance in skin condition diagnosis and introduce a novel framework for synthesizing large-scale densely annotated *in-the-wild* dermatological images. The third and fourth chapters of this thesis describe the details of the two contributions. which are briefly described in the following two subsections:

1.2.1 CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions

While deep learning-based approaches have demonstrated expert-level performance in dermatological diagnosis tasks, they have also been shown to exhibit biases toward certain demographic attributes, particularly skin types (e.g., light versus dark), a fairness concern that must be addressed.

In our first contribution, we propose CIRCLe, a skin color invariant deep representation learning method for improving fairness in skin lesion classification. CIRCLe is trained to classify images by utilizing skin type transformations to compute a regularization loss that

encourages images with the same diagnosis but different skin types to have similar latent representations. To the best of our knowledge, this is the first work that uses skin type transformations and skin color-invariant disease classification to tackle the problem of skin type bias present in large-scale clinical image datasets and how these biases permeate through the prediction models. We present a new state-of-the-art classification accuracy over 114 skin conditions and 6 Fitzpatrick skin types (FSTs) from the Fitzpatrick17K dataset. While previous works had either limited their analysis to a subset of diagnoses [16] or less granular FST labels [124], our proposed method achieves superior performance over a much larger set of diagnoses spanning over all the FST labels.

We provide a comprehensive evaluation of our proposed method, CIRCLe, on 6 different CNN architectures, along with ablation studies to demonstrate the efficacy of the proposed domain regularization loss. Furthermore, we also assess the impact of varying the size and the FST distribution of the training dataset partitions on the generalization performance of the classification models. Finally, we propose a new fairness metric called Normalized Accuracy Range that, unlike several existing fairness metrics, works with multiple protected groups (6 different FSTs in our problem).
The code is available at `https://github.com/arezou-pakzad/CIRCLe`.

[87] Arezou Pakzad, Kumar Abhishek, and Ghassan Hamarneh. "**CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions**", In *Proceedings of the 17th European Conference on Computer Vision (ECCV) ISIC Skin Image Analysis Workshop*, 2022. `https://doi.org/10.1007/978-3-031-25069-9_14`

1.2.2 DermSynth3D: Synthesis of in-the-wild Annotated Dermatology Images

Despite the availability of numerous skin image datasets (e.g., [11, 37, 48, 57, 113, 115, 122]), there is a lack of a *large-scale* skin-image dataset that can be applied to a variety of skin analysis tasks, especially in an *in-the-wild* clinical setting. Moreover, existing datasets are limited in their scope and are often task-specific, requiring extensive additional annotation for generalizing them to other dermatological applications.

To address this gap, in our second contribution, we present DermSynth3D, a computational pipeline along with an open-source software library, for generating synthetic 2D skin image datasets using 3D human body meshes blended with skin disorders from clinical images. Our approach uses a differentiable renderer to blend the skin lesions within the texture image of the 3D human body and generates 2D views along with corresponding annotations, including semantic segmentation masks for skin conditions, healthy skin, non-skin regions, and skin condition bounding boxes.

In particular, my contribution to this thesis is demonstrating the effectiveness of the synthesized data by utilizing it in the training process of machine learning models and evaluating them on real-world dermatological images, showcasing that the DermSynth3D-trained model learns to generalize to skin condition detection and segmentation tasks.
The code is available at `https://github.com/sfu-mial/DermSynth3D`.

[105] Ashish Sinha*, Jeremy Kawahara*, Arezou Pakzad*, Kumar Abhishek, Matthieu Ruthven, Enjie Ghorbel, Anis Kacem, Djamila Aouada, and Ghassan Hamarneh *(* joint first authors)*. "**DermSynth3D: Synthesis of in-the-wild Annotated Dermatology Images**", In *Medical Image Analysis*, 2024. `https://arxiv.org/abs/2305.12621`

As indicated in the reference above, I am a joint first author of this work. My contributions to this work that warranted joint first authorship are designing and implementing the experiments, analyzing the results, and writing the manuscript. More specifically, the experiments I designed and conducted were to demonstrate the effectiveness and utilities of the synthesized data in wound bounding box detection with synthetic data augmentation, wound bounding box detection and semantic segmentation using only synthetic data, the utility of synthetic data in pre-training wound detection model, and the ablation studies for parameter choices of wound bounding box detection.

1.3 Thesis Outline

This thesis covers the details of the methods developed to improve skin type fairness in skin condition classification and synthesize in-the-wild annotated dermatology images. In

addition to this introduction chapter, the thesis includes four chapters. The outlines of the following chapters are as follows:

- Chapter 2: Describes the clinical skin condition datasets we used in this thesis, demonstrates the data distribution, and visualizes some image examples.
- Chapter 3: Describes the development of CIRCLe, a method based on domain invariant representation learning for unbiased skin condition classification and the design of fairness metric NAR, and shows that this method improves classification performance and fairness, domain adaptation capability, and generalization ability of the model.
- Chapter 4: Describes the development of DermSynth3D, a framework for synthesis of densely annotated in-the-wild dermatological images, and shows the effectiveness of the generated synthetic data for improving skin condition bounding box detection and segmentation performance.
- Chapter 5: Summarizes the contributions made and presents limitations and potential future research works.

Disclaimer: The author declares that substantial parts of chapters 3, 4, and 5 of this thesis have been borrowed nearly identically from my original first-authored and joint-first-authored publications listed in Section 1.2.

Chapter 2

Datasets Used in The Thesis

There are two primary types of dermatological datasets, clinical and dermoscopic, that offer distinct insights into skin conditions. The datasets used and synthesized in this thesis consist of clinical images of skin conditions. It is important to note the difference between *Dermoscopy* and *Clinical* images when discussing dermatological datasets.

Dermoscopy images generally focus on the analysis of a single lesion, with large scale annotated dermoscopy datasets now available for public use [30, 95, 113]. While dermoscopy has been shown to improve the diagnostic ability of trained specialists, the field-of-view of a dermoscopy image is generally limited to a localized patch of skin on the body (e.g., a mole). In contrast, clinical images vary considerably in their acquisition protocols, ranging from a closeup view focused on a single lesion, to a view that captures a significant portion of the body (Figure 2.1). The contextual information in large-scale clinical images of skin lesions may provide valuable cues regarding the underlying disease that may not be present in dermoscopic images alone [19, 95].

Clinical images exhibit considerable variability across datasets. For example, the public DermoFit Image Library dataset [11, 110] contains 1300 clinical images and manual lesion segmentations from 10 types of skin conditions. These are high-quality images acquired under standardized conditions. In contrast, other clinical datasets, such as SD-198 [108], SD-260 [125], or Fitzpatrick17K [48], contain hundreds of types of skin disorders and are much less standardized, exhibiting a high variability in camera position relative to the lesion, resulting in dramatic changes in the field-of-view. We use the term "in-the-wild

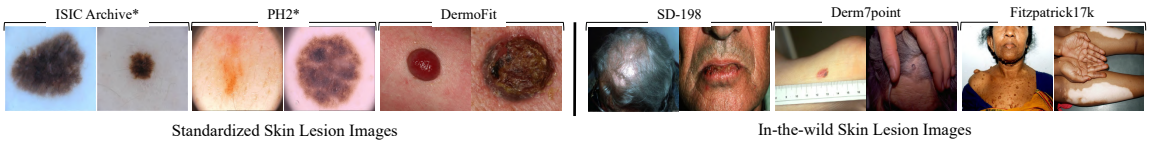


Figure 2.1: Standardized vs in-the-wild skin lesion images ($*$: dermoscopy, all others: clinical).

clinical dataset" to describe these types of image collections, where the camera position, field-of-view, and background, are inconsistent.

The following sections describe the two main clinical datasets of skin conditions used in this thesis.

2.1 Fitzpatrick 17K

The Fitzpatrick17K dataset [47] contains 16,577 clinical images with skin condition labels and skin type labels based on the Fitzpatrick scoring system [43]. The images in this dataset, along with their corresponding skin condition labels, are sourced from two open-source dermatology atlases: 12,672 images from DermaAmin [8] and 3,905 images from Atlas Dermatologico [33].

The images in this dataset are annotated with six Fitzpatrick skin type (FST) labels by a team of non-dermatologist annotators. Figure 2.2 shows some sample images from this dataset along with their skin types. The dataset includes 114 conditions with at least 53 images (and a maximum of 653 images) per skin condition, as shown in Figure 2.3.

The Fitzpatrick labeling system is a six-point scale originally developed for classifying sun reactivity of skin and adjusting clinical medicine according to skin phenotype [43]. In the Fitzpatrick Skin Tone Scale (Table 2.1), different skin types are categorized based on their response to sun exposure. The skin types are categorized into six levels, from 1 to 6, from lightest to darkest skin types. Although Fitzpatrick labels are commonly used for categorizing skin types, we note that not all skin types are represented by the Fitzpatrick scale. [120].

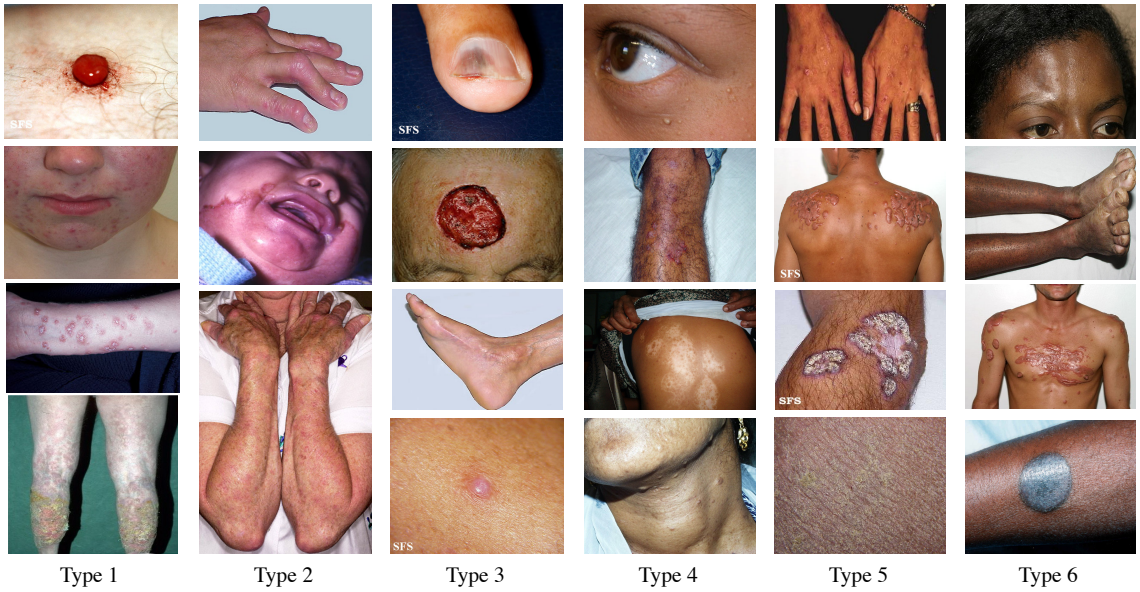


Figure 2.2: Sample images of all six FSTs from the Fitzpatrick17K dataset [47]. Notice the wide variety in disease appearance, field of view, illumination, and presence of imaging artifacts, including non-standard backgrounds consistent with clinical images in the wild and watermarks on some images.

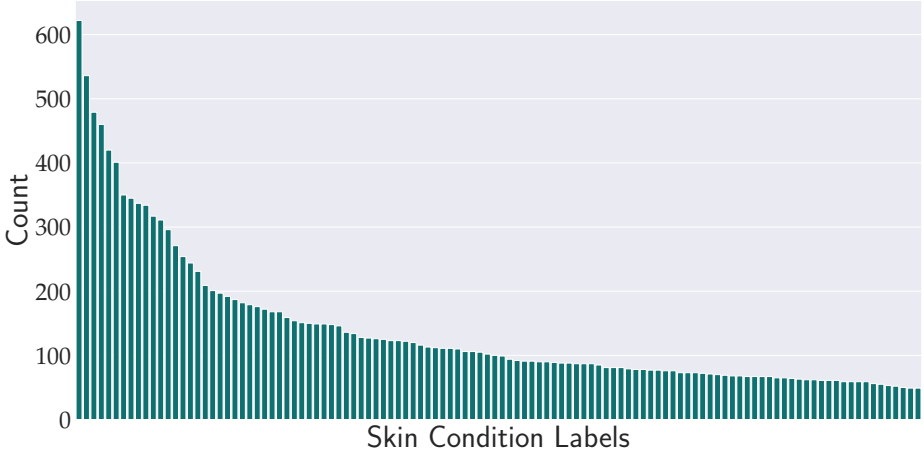


Figure 2.3: Visualizing the distribution of the skin condition labels in the Fitzpatrick17K dataset. Notice that the number of images across different skin conditions is not uniformly distributed.

Table 2.1: Fitzpatrick Skin Tone Scale

Skin Type	**Description**	**Color Sample**
Type 1	Always burns, never tans	
Type 2	Usually burns, tans with difficulty	
Type 3	Burns mildly, tans gradually	
Type 4	Rarely burns, tans with ease	
Type 5	Very Rarely burns, tans very easily	
Type 6	Never burns, tans very easily, deeply pigmented	

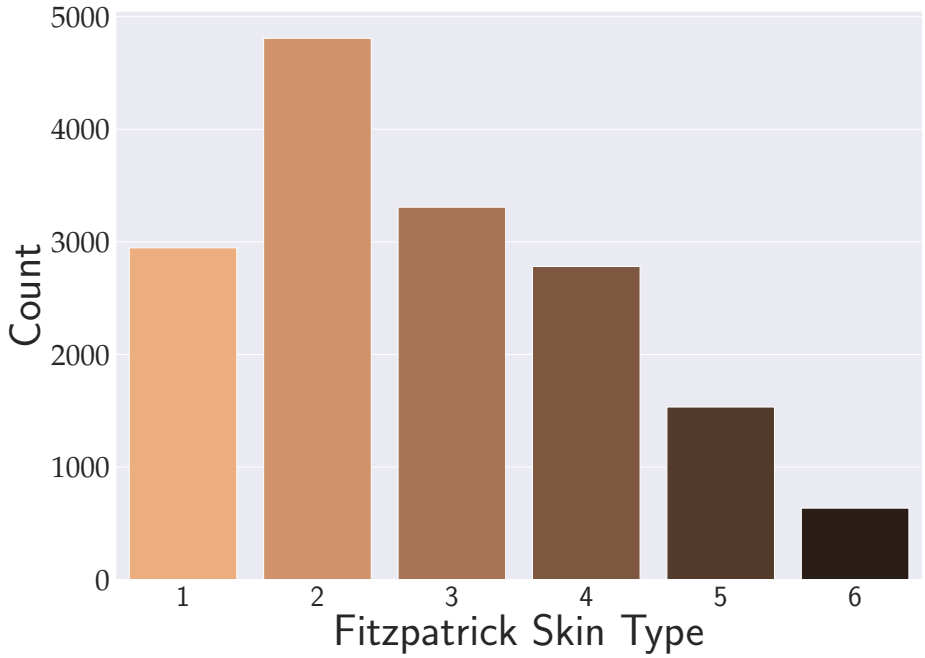


Figure 2.4: Visualizing the distribution of the Fitzpatrick skin type (FST) labels in the Fitzpatrick17K dataset. Notice that the number of images is considerably lower for darker skin types.

In the Fitzpatrick17K dataset, there are significantly more images of light skin types than dark skin. There are 11,060 images of *light* skin types (FSTs 1, 2, and 3), and 4,949 images of *dark* skin types (FSTs 4, 5, and 6), as shown in Figure 2.4.

2.2 Foot Ulcer (FUSeg)

The FUSeg dataset from the *The Foot Ulcer Segmentation Challenge* [115] contains 2D clinical dermatological images of ulcers on the foot and the corresponding wound masks. This dataset includes 1,210 foot ulcer images taken from 889 patients during multiple clinical visits. The raw images in this dataset were taken under uncontrolled illumination conditions with various backgrounds by Canon SX 620 HS digital camera and an iPad Pro camera. The corresponding pixel-wise segmentation mask annotations for each image are acquired manually by wound professionals. Images and their annotations are preprocessed with cropping and zero-padding.

This dataset contains the standard training, validation, and testing partitions of 810, 200, and 200 images, respectively. The annotations for the testing set are kept private and will not be released since the official challenge remains open indefinitely [116].

Figure 2.5 shows some sample images from this dataset along with their segmentation mask annotations.

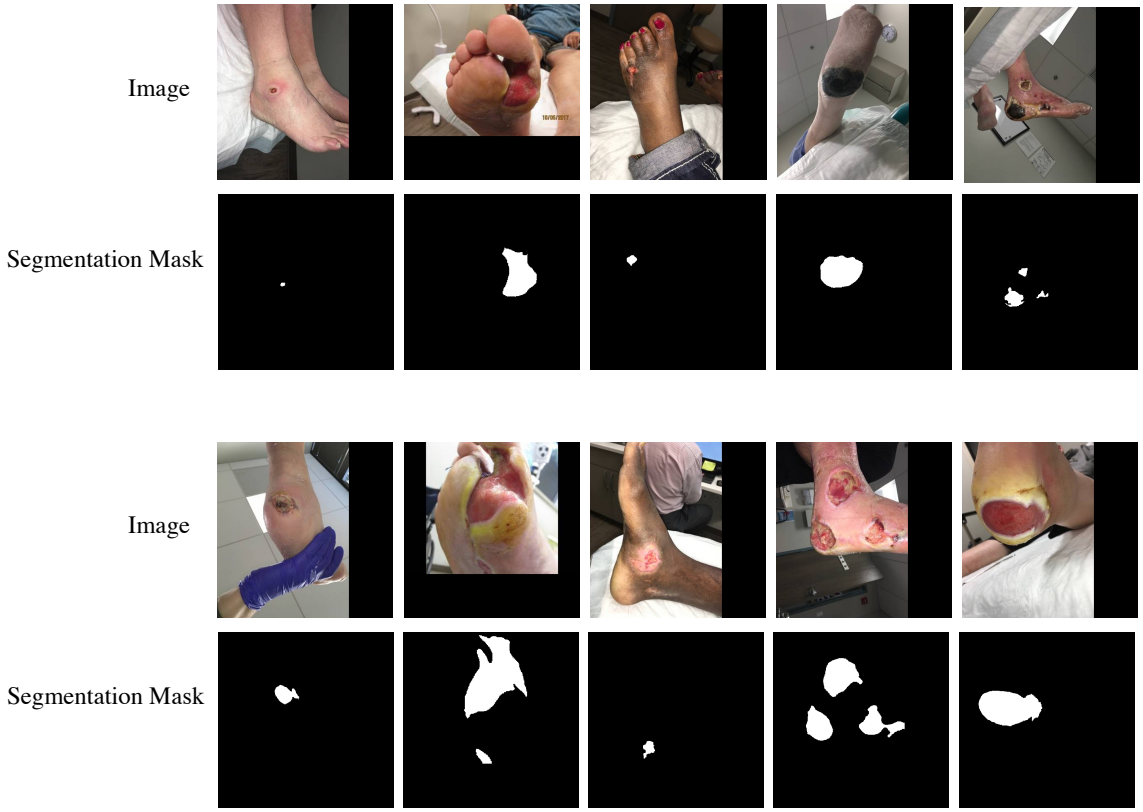


Figure 2.5: Sample images from the FUSeg [115] dataset. The first and third rows contain the preprocessed images in the dataset. The second and fourth rows consist of the corresponding segmentation mask annotations.

Chapter 3

CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions

3.1 Introduction

Most public skin disease image datasets are acquired from demographics consisting primarily of fair-skinned people. However, skin conditions exhibit vast visual differences in manifestations across different skin types [121]. Lighter skinned populations suffer from over-diagnosis of melanoma [5] while darker skinned patients get diagnosed at later stages, leading to increased morbidity and mortality [7]. Despite this, darker skin is under-represented in most publicly available data sets [63, 71], reported studies [35], and in dermatology textbooks [6]. Kinyanjui et al. [63] performed an analysis on two popular benchmark dermatology datasets: ISIC 2018 Challenge dataset [28] and SD-198 dataset [109], to understand the skin type representations. They measured the individual typology angle (ITA), which measures the constitutive pigementation of skin images [85], to estimate the skin tone on these datasets, and found that the majority of the images in the two datasets ITA values between 34.8° and 48°, which are associated with lighter skin. This is consistent with the under-representation of darker skinned populations in these datasets. It has been shown that CNNs perform best at classifying skin conditions for skin types that are similar to those they were trained on [47]. Thus, the data imbalance across different skin types in the majority of the skin

disease image datasets can manifest as racial biases in the DL models' predictions, leading to racial disparities [4]. However, despite these well-documented concerns, very little research has been directed towards evaluating these DL-based skin disease diagnosis models on diverse skin types, and therefore, their utility and reliability as disease screening tools remains untested.

Learning domain invariant representations, a predominant approach in domain generalization [81], attempts to learn data distributions that are independent of the underlying domains, and therefore addresses the issue of training models on data from a set of source domains that can generalize well to previously unseen test domains. Domain invariant representation learning has been used in medical imaging for histopathology image analysis [69] and for learning domain-invariant shape priors in segmentation of prostrate MR and retinal fundus images [76]. On the other hand, previous works on fair classification and diagnosis of skin diseases have relied on skin type detection and debiasing [16] and classification model pruning [124].

One of the common definitions of algorithmic fairness for classification tasks, based on measuring statistical parity, aims to seek independence between the bias attribute (also known as the protected attribute; i.e., the skin type for our task) and the model's prediction (i.e., the skin disease prediction). Our proposed approach, **C**olor **I**nvariant **R**epresentation learning for unbiased **C**lassification of skin **Le**sions (**CIRCLe**), employs a color-invariant model that is trained to classify skin conditions independent of the underlying skin type. In this work, we aim to mitigate the skin type bias learned by the CNNs and reduce the accuracy disparities across skin types. We address this problem by enforcing the feature representation to be invariant across different skin types. We adopt a domain-invariant representation learning method [82] and modify it to transform skin types from clinical skin images and propose a color-invariant skin condition classifier.

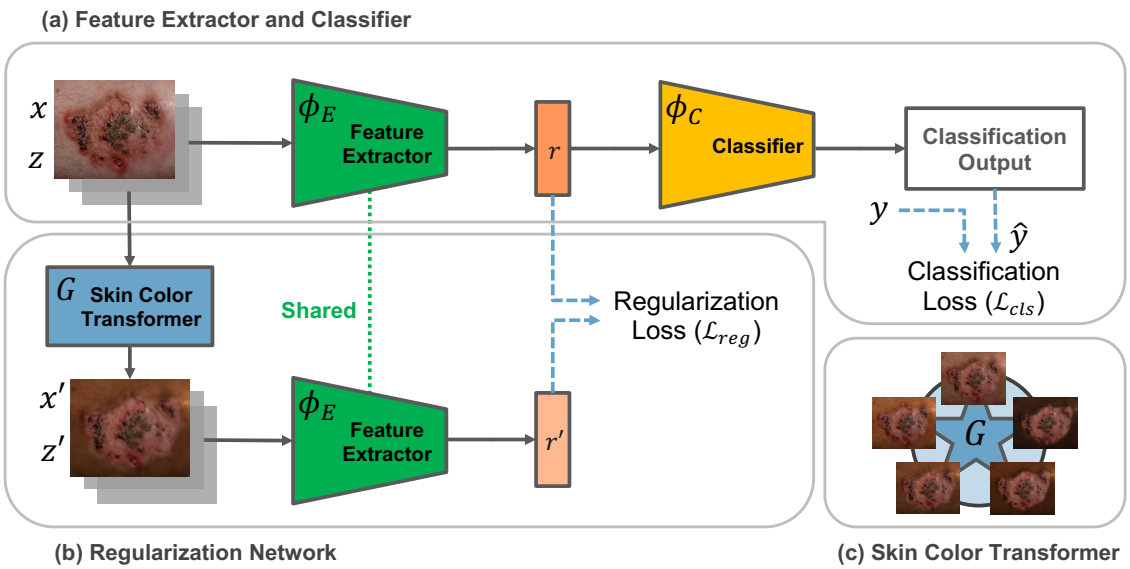


Figure 3.1: Overview of CIRCLe. (a) The skin lesion image x with skin type z and diagnosis label y is passed through the feature extractor ϕ_E. The learned representation r goes through the classifier ϕ_C to obtain the predicted label $\hat{y}$. The classification loss enforces the correct classification objective. (b) The skin color transformer (G), transforms x with skin type z into x' with the new skin type z'. The generated image x' is fed into the feature extractor to get the representation r'. The regularization loss enforces r and r' to be similar. (c) The skin color transformer's schematic view with the possible transformed images, where one of the possible transformations is randomly chosen for generating x'.

3.2 Method

3.2.1 Problem Definition

Given a dataset $\mathcal{S} = \{X, Y, Z\}$, consider x_i, y_i, z_i to be the input, the label, and the protected attribute for the ith sample respectively, where we have N_c classes ($|Y| = N_c$) and N_p protected groups ($|Z| = N_p$). Let $\hat{y}_i$ denote the predicted label of sample i. Our goal is to train a classification model $f_\theta(\cdot)$ parametrized by θ that maps the input x_i to the final prediction $\hat{y}_i = f_\theta(x_i)$, such that (1) the prediction $\hat{y}_i$ is *invariant* to the protected attribute z_i and (2) the model's classification loss is minimized.

3.2.2 Feature Extractor and Classifier

In the representation learning framework, the prediction function $\hat{y}_i = f_\theta(x_i)$ is obtained as a composition $\hat{y}_i = \phi_C \circ \phi_E(x_i)$ of a feature extractor $r_i = \phi_E(x_i)$, where $r_i \in \mathbb{R}^p$ is a learned representation of data x_i, and a classifier $\hat{y}_i = \phi_C(r_i)$, predicting the label $\hat{y}_i$,

given the representation r_i (Figure 3.1(a)). Thus, we aim to learn a feature representation r that is invariant to the protected attributes, and hypothesize that this will lead to better generalization for classification.

3.2.3 Regularization Network

Inspired by the method proposed by Nguyen et al. [82], we use a generative modelling framework to learn a function g that transforms the data distributions between skin types. To this end, we employ a method to synthesize a new image corresponding to a given input image with the subject's skin type in that image changed according to the desired Fitzpatrick skin type (FST) score. We call this model our Skin Color Transformer. After training the Skin Color Transformer model, we introduce an auxiliary loss term to our learning objective, whose aim is to enforce the domain invariance constraint. (Figure 3.1(b))

Skin Color Transformer

We learn the function G that performs image-to-image transformations between skin type domains. To this end, we use a Star Generative Adversarial Network (StarGAN) [26]. The goal of the StarGAN is to learn a unified network G (generator) that transforms the data density among multiple domains. In particular, the network $G(x, z')$ transforms an image x to an output image x' conditioned on the target skin type z'. The generator's goal is to fool the discriminator D into classifying the transformed image as the target skin type z'. StarGAN's model has three main loss functions: (1) Adversarial loss, which is common to all the GAN's. The Discriminator tries to maximize the error while the Generator tries to minimize:

$$L_{adv} = \mathbb{E}_x[logD_{src}] + \mathbb{E}_{x,z'}[log(1 - D_{src}(G(x, z')))], \tag{3.1}$$

where D_{src} is termed as a probability distribution over sources given by D. (2) Domain classification loss, which is associated with classifying and generating images specific to the domains (i.e. skin types in our problem). For a given input image x and a target domain z', the goal is to translate x into an output image x', which is properly classified to the target domain z'. The objective is decomposed into two terms: a domain classification loss of real images used to optimize D, and a domain classification loss of fake images used to optimize

G. In detail, the former is defined as:

$$L_{cls}^{r} = \mathbb{E}_{x,z}[-logD_{cls}(z|x)], \tag{3.2}$$

where the term $D_{cls}(z|x)$ represents a probability distribution over domain labels computed by D. By minimizing this objective, D learns to classify a real image x to its corresponding original domain z. On the other hand, the loss function for the domain classification of fake images is defined as:

$$L_{cls}^{f} = \mathbb{E}_{x,z'}[-logD_{cls}(z'|G(x,z'))], \tag{3.3}$$

where G tries to minimize this objective to generate images that can be classified as the target domain z'. (3) Reconstruction loss to prevent reconstruction errors after changing specified domains:

$$L_{rec} = \mathbb{E}\, x, z', z[||x - G(G(x,z'),z)||_1], \tag{3.4}$$

where G takes in the translated image $G(x,z')$ and the original domain label z as input and tries to reconstruct the original image x. Overall the loss functions combined for the D and G is:

$$L_D = -L_{adv} + \lambda_{cls} L_{cls}^{r}, \tag{3.5}$$

$$L_G = L_{adv} + \lambda_{cls} L_{cls}^{f} + \lambda rec L_{rec}, \tag{3.6}$$

where $\lambda_{cls} = 1$ and $\lambda_{rec} = 10$.

After training, we use G as the Skin Color Transformer. This model takes the image x_i with skin type z_i as the input, along with a target skin type z_j and synthesizes a new image $x_i' = G(x_i, z_j)$ similar to x_i, only with the skin type of the image changed in accordance with z_j.

Domain Regularization Loss

In the training process of the disease classifier, for each input image x_i with skin type z_i, we randomly select another skin type $z_j \neq z_i$, and use the Skin Type Transformer to synthesize

a new image $x'_i = G(x_i, z_i, z_j)$. After that, we obtain the latent representations $r_i = \phi_E(x_i)$, and $r'_i = \phi_E(x'_i)$ for the original image and the synthetic image respectively. Then we enforce the model to learn similar representations for r_i and r'_i by adding a regularization loss term to the overall loss function of the model:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg} \tag{3.7}$$

where $\mathcal{L}_{cls}$ is the prediction loss of the network that predicts $\hat{y}_i$ given $r_i = \phi_E(x_i)$, and $\mathcal{L}_{reg}$ is the regularization loss. In this equation, $\lambda \in [0, 1]$ is a hyper-parameter controlling the trade-off between the classification and regularization losses. We define $\mathcal{L}_{reg}$ as the distance between the two representations r_i and r'_i to enforce the invariant condition. In our implementation, we use cross entropy as the classification loss $\mathcal{L}_{cls}$:

$$\mathcal{L}_{cls} = -\sum_{j=1}^{N_c} y_{ij} \log(\hat{y}_{ij}), \tag{3.8}$$

where y_{ij} is a binary indicator (0 or 1) if class label j is the correct classification for the sample i and $\hat{y}_{ij}$ is the predicted probability the sample i is of class j. The final predicted class $\hat{y}_i$ is calculated as

$$\hat{y}_i = \arg\max_j \ \hat{y}_{ij}. \tag{3.9}$$

We use squared error distance for computing the regularization loss $\mathcal{L}_{reg}$:

$$\mathcal{L}_{reg} = ||r_i - r'_i||_2^2. \tag{3.10}$$

3.3 Experimental Details

3.3.1 Dataset

We evaluate the performance of the proposed method on the Fitzpatrick17K dataset, which we described in Chapter 2. We randomly select 70%, 10%, and 20% of the images for the train, validation, and test splits, where the random selection is stratified on skin conditions. Since the Fitzpatrick17K dataset does not have standard splits, we repeat the experiments with five different random seeds for splitting the data, to ensure the reproducibility and

robustness of our findings. A series of transformations are applied to the training images which include: resize to 128×128 resolution, random rotations in $[-15°, 15°]$, and random horizontal flips. We also use ImageNet [39] training partition's mean and standard deviation values to normalize our images for training and evaluation.

3.3.2 Implementation Details

Feature Extractor and Classifier

We choose VGG-16 [104] pre-trained on ImageNet as our base network. We use the convolutional layers of VGG-16 as the feature extractor ϕ_E. We replace the VGG-16's fully-connected layers with a fully connected 256-to-114 layer as the classifier ϕ_C. We train the network for 100 epochs with plain stochastic gradient descent (SGD) using learning rate 1e-3, momentum 0.9, minibatch size 16, and weight decay 1e-3. We report the results for the epoch with the highest accuracy on the validation set.

Skin Color Transformer

StarGAN [26] implementation is taken from the authors' original source code with no significant modifications. We train StarGAN on the Fitzpatrick17K dataset, using the same train split used for training the classifier. As for the training configurations we use a minibatch size of 16. We train the StarGAN for 200,000 iterations and use the Adam [61] optimizer with a learning rate of 1e-4.

Model Training and Evaluation Setup

We use the PyTorch library [88] to implement our framework and train all our models on a workstation with AMD Ryzen 9 5950X processor, 32 GB of memory, and Nvidia GeForce RTX 3090 GPU with 24 GB of memory.

3.3.3 Evaluation Metrics

We aim for an *accurate* and *fair* skin condition classifier. Therefore, we assess our method's performance using metrics for both accuracy and fairness. We use the well-known Micro-averaged Accuracy, Recall, and F1 metrics for evaluating our model's classification performance. For fairness, we use the Equal Opportunity Difference (EOD) metric [50]. In

addition, since EOD is limited to the assessment of only two protected groups, to measure fairness in the model's accuracy for multiple groups of skin types, we assess the accuracy (ACC) disparities across all six skin types by proposing the Normalized Accuracy Range (NAR).

Equal Opportunity Difference

EOD measures the difference in true positive rates (TPR) for the two protected groups. Let TPR_z denote true positive rate of group z and $z \in \{0, 1\}$. Then EOD can be computed as:

$$EOD = |TPR_{z=0} - TPR_{z=1}|. \tag{3.11}$$

A value of 0 implies both protected groups have equal benefit. Given that the above metric (and other common fairness metrics in the literature [15, 40, 50]) are defined for two groups: privileged and under-privileged, w.r.t the protected attribute, we adopt the light (FSTs 1, 2, and 3) versus dark (FSTs 4, 5, and 6) as the two groups.

Normalized Accuracy Range

In order to measure fairness in the model's accuracy for multiple groups of skin types, we assess the accuracy (ACC) disparities across all the six skin types by proposing the Normalized Accuracy Range (NAR) as follows:

$$NAR = \frac{ACC_{max} - ACC_{min}}{mean(ACC)}, \tag{3.12}$$

where ACC_{max} and ACC_{min} are the maximum and minimum accuracy achieved across skin types and $mean(ACC)$ is the mean accuracy across skin types, i.e.:

$$\begin{aligned} ACC_{max} &= max\{ACC_i : 1 \leq i \leq N_p\}, \\ ACC_{min} &= min\{ACC_i : 1 \leq i \leq N_p\}, \\ mean(ACC) &= \frac{1}{N_p} \sum_{i=1}^{N_p} ACC_i \end{aligned} \tag{3.13}$$

A perfectly fair performance of a model would result in equal accuracy across the different protected groups on a test set, i.e. $ACC_{max} = ACC_{min}$, leading to $NAR = 0$. As the accuracies across protected groups diverge, $ACC_{max} > ACC_{min}$, NAR will change even if the mean accuracy remains the same, thus indicating that the model's fairness is also changed. Moreover, NAR also takes into account the overall mean accuracy: this implies that in cases where the accuracies range ($ACC_{max} - ACC_{min}$) is the same, the model with the overall higher accuracy leads to a lower NAR, which is desirable. In our quantitative results, we report EOD for completeness; however, it is not an ideal measure, given it is restricted to only two protected groups whereas we have six. Therefore, we focus our attention on NAR.

3.3.4 Models

Baseline

For evaluating our method, we compare our results with the method proposed by Groh et al. [47], which has the current state-of-the-art performance on the Fitzpatrick17K dataset. We call their method the *Baseline*. To obtain a fair comparison, we use the same train and test sets they used.

Improved Baseline (Ours)

In order to evaluate the effectiveness of the color-invariant representation learning process, we perform an ablation study, in which we remove the regularization loss $\mathcal{L}_{reg}$ from the learning objective of the model and train the classifier with only the classification objective. We call this model the *Improved Baseline*.

CIRCLe (Ours)

The proposed model for unbiased skin condition classification, CIRCLe, is composed of two main components: the feature extractor and classifier, and the regularization network (Fig. 3.1).

Multiple Backbones

To demonstrate the efficacy of our method, we present evaluation with several other backbone architectures in addition to VGG-16 [104] used by Groh et al. [47]. In particular, we use MobileNetV2 [99], MobileNetV3-Large (referred to as MobileNetV3L hereafter) [54], DenseNet-121 [55], ResNet-18 [52], and ResNet-50 [52], thus covering a wide range of CNN architecture families and a considerable variety in model capacities, i.e. from 2.55 million parameters in MobileNetV2 to 135.31 million parameters in VGG-16 (Table 3.1). For all the models, we perform an ablation study to evaluate if adding the regularization loss $\mathcal{L}_{reg}$ helps improve the performance.

Table 3.1: Comparing the model capacities and computational requirements of different backbones evaluated. For all the six backbones, we report the number of parameters and the number of multiply-add operations (**MulAddOps**). All numbers are in millions (**MM**). Note how the six backbones encompass several architectural families and a large range of model capacities ($\sim$ 2MM to $\sim$ 135MM parameters) and computational requirements ($\sim$ 72MM MulAddOps to $\sim$ 5136MM MulAddOps).

	MobileNetV2	**MobileNetV3L**	**DenseNet-121**	**ResNet-18**	**ResNet-50**	**VGG-16**
Parameters (MM)	2.55	4.53	7.22	11.31	24.03	135.31
MulAddOps (MM)	98.16	72.51	925.45	592.32	1335.15	5136.16

3.4 Results and Analysis

3.4.1 Classification and Fairness Performance

Table 3.2: Classification performance and fairness of CIRCLe for classifying 114 skin conditions across skin types as assessed across five folds (mean ± std. dev.). We compute the overall accuracy based on the micro average accuracy across all skin types. Values in bold indicate the best results. CIRCLe yields the best performance while also improving 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

Table 3.2 shows the accuracy and fairness results for the proposed method in comparison with the baseline. From the table, we can see that our Improved Baseline method recognizably outperforms the baseline method in accuracy and fairness. By using a powerful backbone and a better and longer training process, we more than doubled the classification accuracy on the Fitzpatrick17K dataset for all the skin types. This indicates that the choice of the base classifier and training settings plays a significant role in achieving higher accuracy rates on the Fitzpatrick17K dataset. Moreover, we can see that CIRCLe further improves the performance of our Improved Baseline across all the skin types, as well as the overall accuracy. This significant improvement demonstrates the effectiveness of the color-invariant representation learning method in increasing the model's generalizability. This observation shows that when the model is constrained to learn similar representations from different skin types that the skin condition appears on, it can learn richer features from the disease information in the image, and its overall performance improves. In addition, CIRCLe shows improved fairness scores (lower EOD and lower NAR), which indicates that the model is less biased. To the best of our knowledge, we set a new state-of-the-art performance on the Fitzpatrick17K dataset for the task of classifying the 114 skin conditions.

Different model architectures may show different disparities across protected groups [90]. We can see in Table 3.3 that the color-invariant representation learning (i.e. with the regularization loss $\mathcal{L}_{reg}$ activated) significantly improves the accuracy and fairness results in different model architecture choices across skin types, which indicates the effectiveness of the proposed method independently from the backbone choice and its capacity. We can see that while the regularization loss does not necessarily improve the EOD for all the backbones, EOD is not the ideal measure of fairness for our task since as explained in Section 3.3.3, it can only be applied to a lighter-versus-darker skin tone fairness assessment. However, employing the regularization loss does improve the NAR for all the backbone architectures.

3.4.2 Domain Adaptation Performance

For evaluating the model's performance on adapting to unseen domains, we perform a "two-to-other" experiment, where we train the model on all the images from two FST domains

Table 3.3: Evaluating the classification performance improvement contribution of the regularization loss $\mathcal{L}_{reg}$ with multiple different feature extractor backbones. Best values for each backbone are presented in bold. EOD reported (for two groups of light and dark FSTs) for completeness but evaluation over all the 6 FSTs uses NAR (see text for details). Observe that $\mathcal{L}_{reg}$ improves the classification accuracy and the fairness metric NAR for all backbones.

Model	$\mathcal{L}_{reg}$	**Recall**	**F1-score**	**Accuracy**							**EOD ↓**	**NAR ↓**
				Overall	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**Type 5**	**Type 6**		
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**
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**

Table 3.4: Classification performance measured by micro average accuracy when trained and evaluated on holdout sets composed of different Fitzpatrick skin types (FSTs). For example, "FST3-6" denotes that the model was trained on images only from FSTs 1 and 2 and evaluated on FSTs 3, 4, 5, and 6. CIRCLe achieves higher classification accuracies than Baseline (Groh et al. [47]) and Improved Baseline (also ours) for all holdout partitions and for all skin types.

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**	-	-

and test it on all the other FST domains. Table 3.4 shows the performance of our model for this experiment. CIRCLe recognizably improves the domain adaptation performance in comparison with the Baseline and Improved Baseline, demonstrating the effectiveness of the proposed method in learning a color-invariant representation.

3.4.3 Classification Performance Relation with Training Size

Table 3.5: Total number of training images for each experiment illustrated in Figure 3.2. Note that the test set for all these experiments is the original test split with 3,205 images (20% of the Fitzpatrick17K dataset images), and the number of training images for experiments with 100% of each FST group is the same for all three groups, and is equal to the original train split with 11,934 images (70% of the Fitzpatrick17K dataset images).

	0%	**20%**	**40%**	**60%**	**80%**
FST12	5,964	7,073	8,183	9,293	10,403
FST34	7,088	7,973	8,858	9,743	10,628
FST56	9,974	1,0281	10,589	10,897	11,205

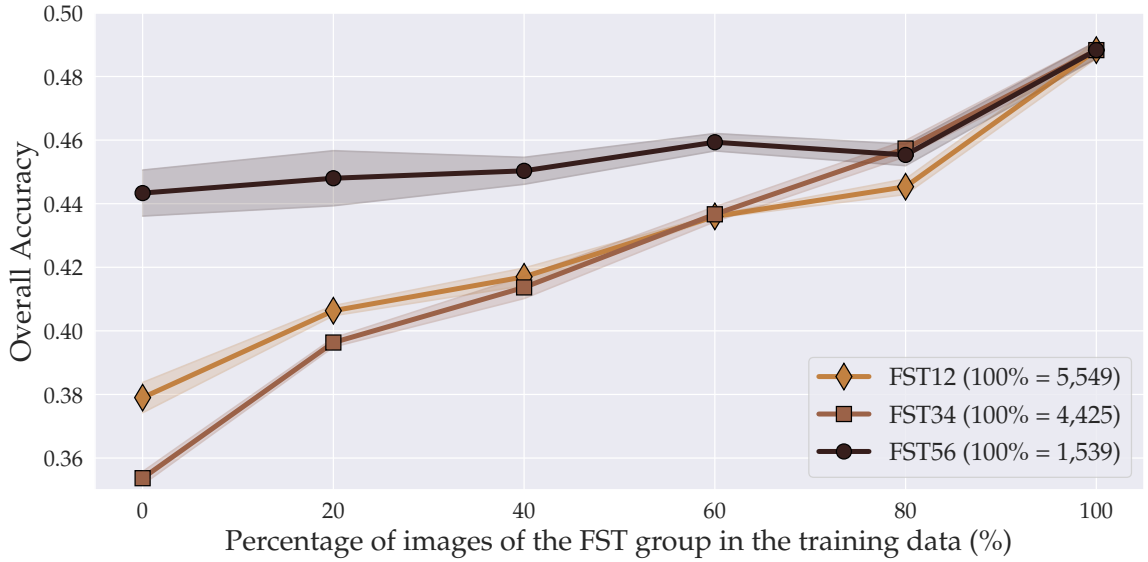


Figure 3.2: Classification performance of CIRCLe on the test set as the number of training images of the FST groups increases. Each FST group line plot indicates the series of experiments in which the percentage of number of training images of that FST group changes as the rest of the training images remain idle. The rightmost point in the plot, with 100%, is identical for all the FST groups, which is the overall accuracy achieved by CIRCLe in Table 3.2. The std. dev. error band, illustrated in the figure, is computed by repetition of experiments with three different random seeds.

As CIRCLe's performance improvement and effectiveness in comparison with the baselines is established in Section 3.4.1, we further analyze the relation of CIRCLe's classification performance with the percentage of images of the FST groups in the training data. To this end, we consider the FST groups of light skin types (FSTs 1 and 2) with 5,549 images, medium skin types (FSTs 3 and 4) with 4,425 images, and dark skin types (FSTs 5 and 6)

with 1,539 images in the training set. For each FST group, we gradually increase the number of images of that group in the training set, while the number of training images in other groups remains unchanged, and report the model's overall accuracy on the test set. The total number of training images for each of these experiments is provided in Table 3.5. As we can see in Figure 3.2, as the number of training images in a certain FST group increases, the overall performance improves, which is expected since DL-based models generalize better with larger training datasets. However, we can see that for the least populated FST group, i.e., dark skin types (FST56) with 13% of the training data, our method demonstrates a more robust performance across experiments, and even with 0% training data of FST56, it achieves a relatively high classification accuracy of 0.443. In addition, note that in these experiments, FST groups with lower number of images in the dataset, would have a larger number of total training images, since removing a percentage of them from the training images will leave a larger portion of images available for training (Table 3.5). This indicates that when the number of training images is large enough, even if images of a certain skin type are not available, or are very limited, our model can perform well overall. This observation signifies our method's ability to effectively utilize the disease-related features in the images from the training set, independently from their skin types, as well as the ability to generalize well to minority groups in the training set.

3.5 Summary

In this chapter, we proposed CIRCLe, a method based on domain invariant representation learning, for mitigating skin type bias in clinical image classification. Using a domain-invariant representation learning approach and training a color-invariant model, CIRCLe improved the accuracy for skin disease classification across different skin types for the Fitzpatrick17K dataset and set a new state-of-the-art performance on the classification of the 114 skin conditions. We also proposed a new fairness metric, Normalized Accuracy Range, for assessing the fairness of classification in the presence of multiple protected groups and showed that CIRCLe improves the fairness of classification. Additionally, we presented an extensive evaluation over multiple CNN backbones as well as experiments to analyze

CIRCLe's domain adaptation performance and the effect of varying the number of training images of different FST groups on its performance.

Having looked at algorithmic approaches to improve the fairness and performance of a skin condition classification model, we next look at leveraging dermatological data synthesis to improve skin condition bounding box detection and semantic segmentation performance.

Chapter 4

DermSynth3D: Synthesis of in-the-wild Annotated Dermatology Images

4.1 Introduction

Clinical images play a vital role in dermatology research, providing crucial insights into various skin conditions. In-the-wild clinical images are commonly utilized to train classification models [37, 48, 57, 108, 124], where the entire image serves as input for predicting skin disorder classes. However, beyond classification, other tasks such as lesion segmentation [51, 80], tracking [44, 107, 129], management [3], and skin tone prediction [62] are important. The dataset for wound segmentation, introduced by Wang et al. [115], presents a valuable resource for automating wound area measurement in therapy monitoring. Additionally, research by Groh et al. [48] highlights the significance of segmenting healthy skin in automated methods aimed at estimating skin tones. This segmentation process has been demonstrated to enhance the accuracy of skin tone prediction in imaged subjects.

Synthesizing images with their corresponding annotations presents a viable strategy for curating the necessary data, proven successful in both medical and non-medical domains. In non-medical contexts, image synthesis with annotations has been applied in face analysis [123] and indoor scene segmentation [78]. For a comprehensive exploration of image synthesis, particularly utilizing generative adversarial network (GAN) models [46, 119], interested readers are directed to the survey conducted by Shamsolmoali et al. [102].

Given the often limited size of medical image datasets [10, 32, 66], the adoption of synthesis techniques for medical image analysis has gained popularity in recent years. This approach facilitates the generation of ground truth-annotated images across various modalities, including MRI [24, 38], CT [27, 83], PET [17, 118], and ultrasound [74, 111]. For a deeper examination of the use of GANs and image synthesis in medical imaging, readers are directed to comprehensive surveys by Yi et al. [128], Kazeminia et al. [59], Wang et al. [117], Rawat et al. [106], and Yang et al. [127].

In the domain of skin image analysis, efforts have been directed toward synthesizing skin lesion images. Initial works utilized noise-based GANs [14] and conditioned output on diagnostic categories [20]. Subsequently, Abhishek et al. [1] proposed a GAN-based framework for generating skin lesion images constrained to binary lesion segmentation masks, while Pollastri et al. [89] employed GANs to generate both skin lesion images and corresponding binary segmentation masks. For a comprehensive review of deep learning-based synthetic data generation techniques for skin lesion images, readers are referred to the survey conducted by Mirikharaji et al. [80].

The current "in-the-wild" clinical datasets are often limited in their scope and tend to be task-specific, thereby limiting their utility in providing semantically rich ground truth (GT) labels suitable for various dermatological tasks. Consequently, there has been relatively less exploration into generating synthetic data for clinical images compared to dermoscopic images. In addressing this gap, Li et al. [73] proposed a method to synthesize 2D data by blending small lesions onto larger torso images, enabling training data creation for lesion mask detection across extensive body regions. Similarly, Dai et al. [34] introduced a technique for generating burn images with automatic annotations, utilizing Style-GAN [56] to synthesize burn wounds integrated with textures from a 3D human avatar. Both approaches emphasized the challenges in obtaining appropriately labeled real training data specific to their dermatological tasks, thus motivating their use of synthetic data.

Our proposed methodology shares similarities with that of Dai et al. [34], where we adopt a comparable pipeline involving the blending of 2D skin disorder images onto 3D textured meshes to create a comprehensive 2D dataset accompanied by corresponding annotations.

However, we expand the scope of our work by incorporating a diverse range of skin tones and background scenes. This expansion enables us to generate semantically rich and meaningful labels for 2D "in-the-wild" clinical images, which are applicable to a variety of dermatological tasks, rather than being limited to a single task. Moreover, the annotated data generated by DermSynth3D in the form of semantic segmentation masks and 3D scene parameters, can be used to train machine learning models for a variety of medical tasks that may benefit clinical practice.

4.2 Method

Our proposed DermSynth3D framework automates the process of blending skin disease regions from 2D images onto 3D texture meshes, while allowing for control over lighting and material parameters, from appropriate camera viewpoints, and renders the resulting 2D image and the corresponding ground truth annotations. Figure 4.1 shows our proposed framework. We describe an overview of our proposed framework in this section. For a more detailed description of the method, we refer interested readers to our published paper [105].

We define a 2D clinical image $x \in \mathbb{R}^{W \times H \times 3}$ as an RGB image with width W and height H that shows a skin condition, and a corresponding binary segmentation mask $s \in \{0, 1\}^{W \times H}$ where pixels with a non-zero value indicate the diseased region (as shown in "2D skin lesions" in Figure 4.1). We define a 3D avatar of a human subject as a mesh M composed of vertices V, faces F, and a UV map U, where the vertices and the faces determine the geometry of the mesh and the UV map determines the mapping between the geometry and a 2D texture image $T \in \mathbb{R}^{W_T \times H_T \times 3}$ that contains pixels representing the surface of the skin. Our goal is to transfer the skin condition within x onto a location on the texture image T of the 3D mesh M. We approach this problem through an image-blending method, where given a 2D binary segmentation mask s indicating the skin condition within x and a target location on the mesh, we blend the diseased region within the mesh's texture image T.

4.2.1 Placing and Blending Skin Conditions on the Mesh

In our framework, we ensure the accurate placement of skin conditions on 3D meshes by enforcing criteria for suitable locations. These criteria dictate that the region for lesion

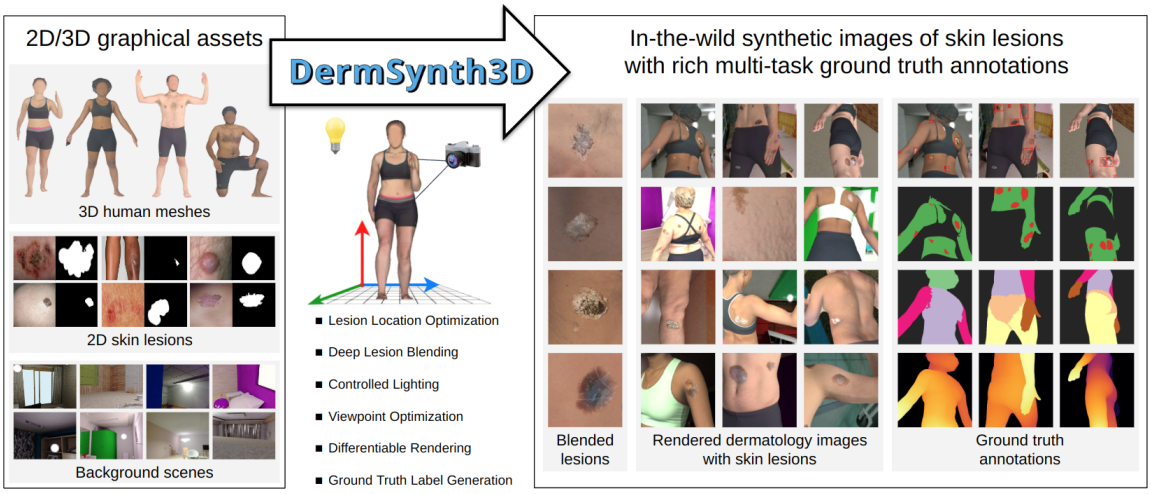


Figure 4.1: Overview of our proposed framework DermSynth3D. The pipeline takes texture images of 3D meshes, 2D segmented skin conditions, and background scenes as input, and blends the skin condition onto texture images to produce lesion-blended texture maps. After blending, 2D views of the meshes are rendered from various camera viewpoints, under different lighting conditions, and combined with background images to create a synthetic dermatology dataset of images with skin lesions and their corresponding ground truth annotations.

placement should; (1) not overlap with clothing, hair, or the background, (2) not overlap with another skin condition (when blending multiple skin conditions), and (3) exhibit minimal depth changes, preventing blending lesions across disjoint anatomy. To validate a location's suitability, we assess if the lesion placement region meets these criteria, leveraging both depth information from the renderer and manual annotations of non-skin regions.

Initially, we assess the feasibility of positioning a scaled clinical image x and its corresponding lesion mask s within the center of the rendered view. Given potential discrepancies in size between x and $\tilde{a}$, we create an image $a_x \in \mathbb{R}^{\tilde{W} \times \tilde{H} \times 3}$, depicting the lesion within the rendered view, along with a corresponding lesion mask $a_s \in \mathbb{R}^{\tilde{W} \times \tilde{H}}$. Subsequently, we evaluate if the region a_s accommodating the skin disorder aligns with the aforementioned suitability criteria.

To mitigate significant depth changes and prevent lesion overlap with the background, we leverage depth information ($\tilde{z}$) provided by the renderer. This depth data aids in identifying local regions with pronounced depth changes or regions exterior to the mesh. Additionally, to avoid lesion overlap with non-skin regions, we rely on manual annotations (referenced in

Section 4.3) delineating non-skin areas within the texture image. These annotations serve to differentiate between skin and non-skin regions, ensuring accurate placement of skin conditions.

Once suitable locations are identified, we blend skin conditions into the texture image of the 3D mesh. Skin conditions manifest across various body locations and can be observed from diverse viewpoints in real-world clinical scenarios. To efficiently synthesize realistic "in-the-wild" clinical images, our framework blends the skin disorder directly into the mesh's texture image. This approach enables the framework to render the blended skin disorder from various viewpoints. The blending process involves updating the texture image to incorporate the skin condition while preserving the original texture in unaffected regions. We follow the deep image blending approach by Zhan et al. [130], where an iterative optimization, minimizes a blending loss function between a foreground object cropped from the source image and the target image which the selected object would be blended onto. Leveraging an iterative optimization technique inspired by deep image blending methods, we achieve a harmonious integration of skin conditions into the texture image.

4.2.2 Synthesizing the 2D Image Dataset

Creating the dataset of 2D rendered images and their corresponding dense annotations involves a methodical two-step process. First, we determine suitable locations for blending skin conditions onto the texture image (T) of the 3D mesh (M). This step, detailed in Section 4.2.1, entails sampling 2D images (x) with skin conditions from real dermatological image collections (Section 4.3) and their respective annotated masks (s). Enhancing color constancy within these images, we employ the Shades of Gray algorithm [42]. Subsequently, iterative blending processes, described in Section 4.2.1, is applied to blend multiple skin conditions at various locations. The first step yields a blended texture image $T_b \in \mathbb{R}^{W_T \times H_T \times 3}$ and a corresponding texture mask $T_m \in \mathbb{Z}^{W_T \times H_T}$, indicating the locations of the skin conditions, where W_T and H_T are the width and the height of the original texture image T respectively.

In the second step, leveraging the blended texture image T_b and texture mask T_m, we generate a dataset of rendered 2D views and target labels. This process involves randomizing

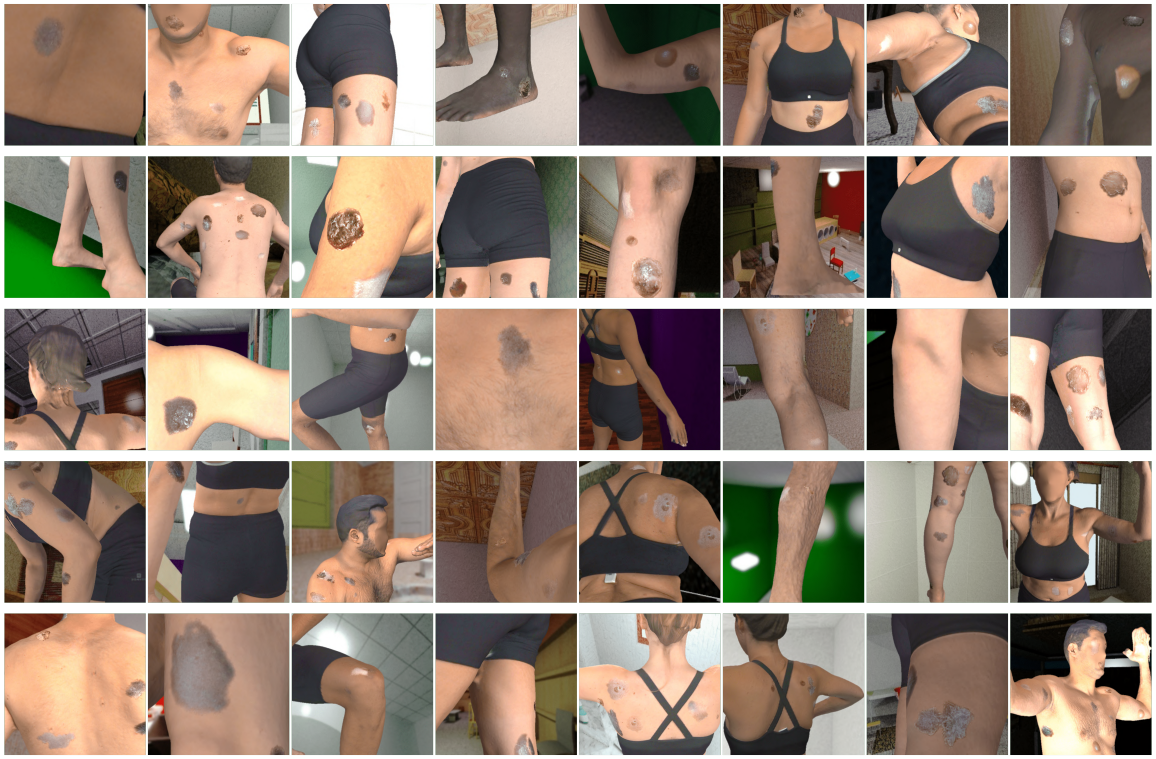


Figure 4.2: Generated synthetic images of multiple subjects across a range of skin tones in various skin conditions, backgrounds, lighting, and viewpoints.

camera positions to render 2D RGB views $\tilde{a}_{T_b}$ from the blended texture image T_b. Variations are introduced through diffuse, ambient, and specular lighting parameters, and for more realistic views and improved illumination, we enforce that the camera is placed outside of the mesh and that the light source reaches the camera without being blocked by the mesh. To create the final image, we combine the foreground with a background image of 2D indoor scene.

Next, we describe each of our different target variables. The skin condition mask $\tilde{a}_{T_m}$ is computed by rendering with the texture mask T_m. The skin mask a_{skin} is computed by excluding both the skin condition regions $\tilde{a}_{T_m}$ and the regions of the body labeled as non-skin. The non-skin mask a_{nonskin} is computed from regions containing neither skin $\tilde{a}_{T_m}$ nor skin conditions a_{skin} (Figure 4.3, third row). Additionally, we obtained bounding boxes around skin condition regions by computing the minimal enclosing box around each skin condition mask (Figure 4.3, second row from the top).

Figure 4.3: A few examples of data synthesized using DermSynth3D. The rows from top to bottom show respectively: the rendered images with blended skin conditions, bounding boxes around the lesions, GT semantic segmentation masks.

Finally, we generate our dataset by rendering a set of 2D images and the corresponding annotations for each mesh, by sampling n times under different camera, lighting, and material parameters, and background scenes. We show some example images from the generated 2D dataset in Figure 4.2.

4.3 Materials for Synthetic Data Generation

We incorporate 3D textured human meshes from the 3DBodyTex dataset [97, 98]. This dataset has 400 high-resolution textured meshes from 200 subjects captured in various poses, wearing sports clothing. These meshes are utilized to introduce realistic human anatomy into synthetic images.

For dermatological image segmentation, we employ the Fitzpatrick17K dataset, a clinical dataset featuring "in-the-wild" images alongside corresponding skin condition labels, which we described in Chapter 2. From this dataset, 75 images are manually segmented into lesion, skin, and background segments. This segmentation aids in accurately representing diverse skin conditions in synthetic images.

To enhance the realism of synthetic images, we integrate backgrounds sourced from 2D indoor scene images available in public datasets [78, 93]. These backgrounds contribute to creating visually convincing synthetic images that closely resemble real-world clinical environments.

4.4 Experimental Details

4.4.1 Synthetic Dataset

We generate a dataset of 10,000 synthetic images using DermSynth3D based on the dataset construction details provided in [105]. The synthetic dataset is generated by capturing randomly rendered views of the 3D meshes with blended skin conditions on them. Images are generated by rendering the blended views with a height $\tilde{H}$ and width $\tilde{W}$ of 512×512, and diversified by placing multiple skin conditions into a single texture image, sampling from a range of camera views and lighting parameters, and various backgrounds. For the experiments, based on the requirements of the experiment, a subset of images was randomly selected from this synthetic dataset of 10,000 images.

4.4.2 Evaluation Dataset

We use the FUSeg dataset, which we described in Chapter 2. As the ground truth annotations for the official test set are not publicly released, we use the official validation set for our evaluation and split the official training set into 610 images for training and 200 images for internal validation. We use common image augmentation and normalization techniques (e.g., rotation, color shifts) on the training images.

4.4.3 Model Training Details

Bounding Box Detection

We convert the masks of the wounds to bounding boxes by labeling the connected regions of the masks and computing the minimal enclosing bounding box. We then train a Faster R-CNN [92] model with pre-trained weights for bounding box detection. We use a mini-batch size of 8 images and train the model for a maximum of 50 epochs using SGD [21, 60, 94] with a learning rate of 0.001. We choose the model weights with the maximum intersection over union (IoU) score over the internal validation set of real images.

Semantic Segmentation

We train a DeepLabV3 [25] network with a ResNet-50 [52] backbone with pre-trained weights as our model. We use a mini-batch size of 8 images and minimize the binary cross

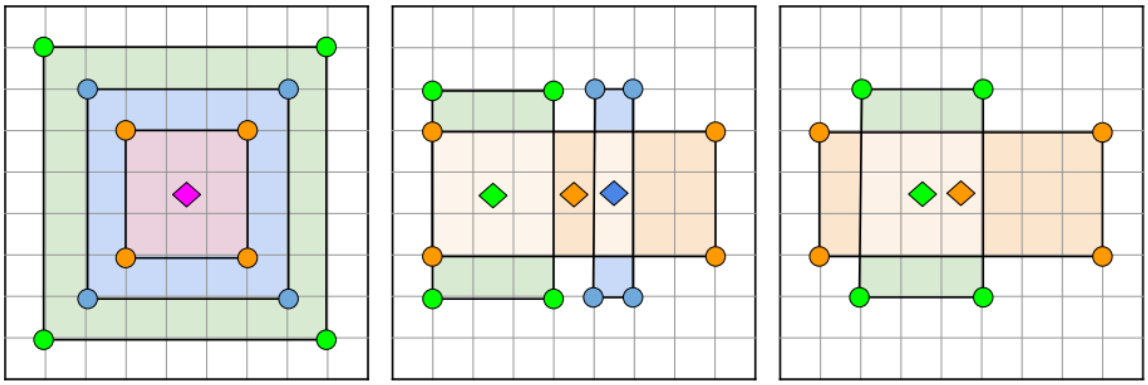


Figure 4.4: An example of the overlapping centroid metric [131]. *Left* shows the difference between IoU and overlapping centroids metric. IoU differs among the green, blue, and orange boxes; however, they have the same centroid (diamond) and are considered as "matching" using the overlapping centroid metric. *Middle* shows an "unmatch" scenario. The orange box contains the centroid for the green and blue boxes; however, the green and blue boxes do not contain the centroid for the orange box, and thus are not considered a match. *Right* shows a "match" scenario. The green and orange boxes match as both contain each other's centroids. Note that the green and orange boxes have the same IoU in the middle and right figures, but only the right figure shows a match using the centroid metric.

entropy loss for a maximum of 250 epochs using the Adam optimizer with a learning rate of 0.00005 and a weight decay of 0.00005. We choose the model weights with the maximum Dice score over the internal validation set of real images.

4.4.4 Evaluation Metrics

For evaluating the bounding box detection performance, we use two metrics: the intersection over union (IoU) score, which measures the exact match between a detected and ground truth bounding box, and the average precision of overlapping centroids ($\mathrm{AP}_{\mathrm{centroid}}$) [131], which determines the bounding box localization performance rather than its precise boundaries and is more suitable for medical applications.

Average precision of overlapping centroids

$\mathrm{AP}_{\mathrm{centroid}}$ metric is based on the overlapping centroids. Specifically, if two centroids of the ground truth bounding box y and the model's predicted box y' are enclosed by both we have,

$$(c(y) \in y') \;\; \& \;\; (c(y') \in y) \tag{4.1}$$

where $c(.)$ computes the centroid of the box, a "correct detection" or "match" occurs. Figure 4.4 shows examples using the overlapping centroid metric on various bounding boxes. After computing the true and false positive detections, we compute *average precision* (AP), which measures the area under the precision-recall curve.

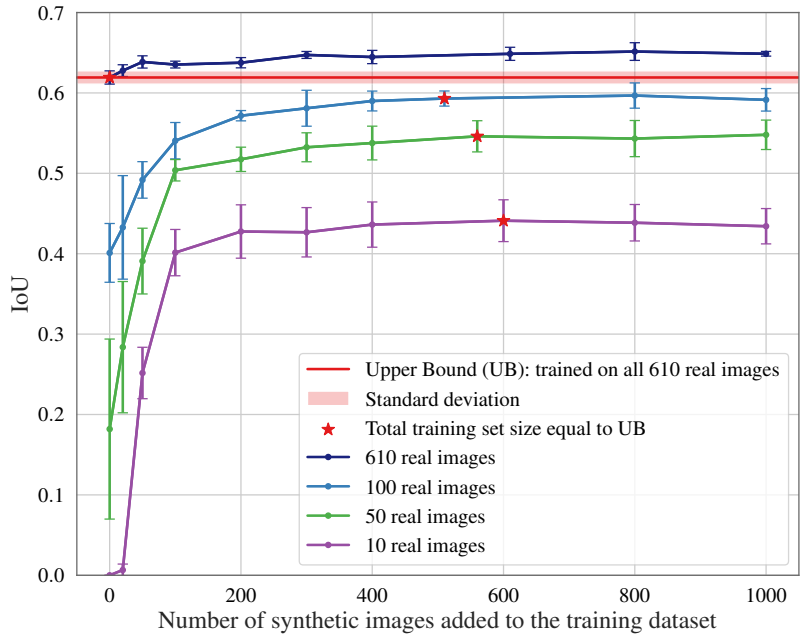
4.5 Experiments and Results

Detecting and segmenting wounds in clinical images is an important step in tracking and extracting morphological features from the wounds, which is crucial for diagnosis and treatment. Bounding box detection and semantic segmentation of wounds can be used to localize the wounds in clinical images and minimize unnecessary information within the scene to improve downstream tasks [115].

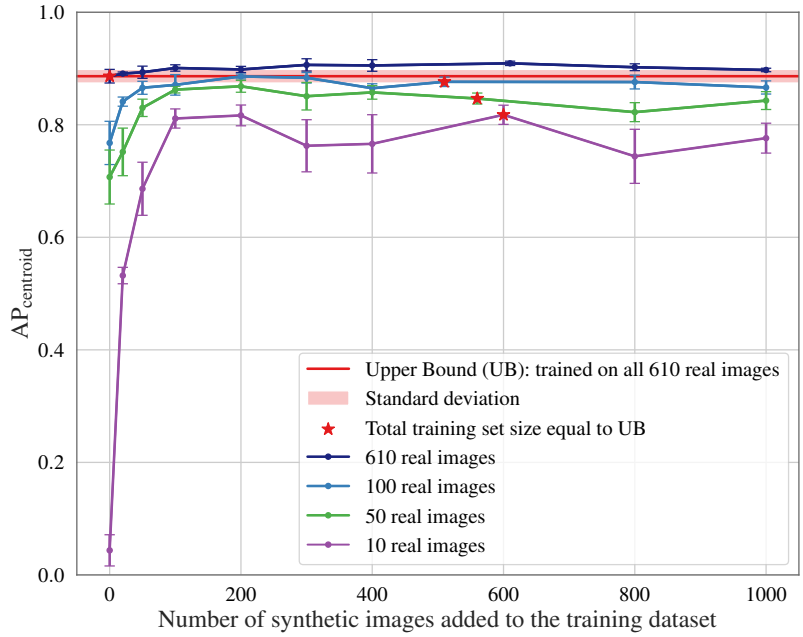
We perform the following experiments in order to evaluate how well a model trained on our generated synthetic data can generalize to unseen real data. We emphasize that our goal in these experiments is not to compete with state-of-the-art performance over these datasets, but rather to show the utility of the generated dataset by assessing the model's ability to generalize to real 2D images when trained on this dataset. Ideally, we would evaluate our approach over an existing "in-the-wild" clinical dermatological dataset with skin conditions, skin, and background segmentation labels. However, to the best of our knowledge, there exists no such dataset, as most skin image datasets contain labels for binary segmentation tasks (e.g., skin vs background or lesion vs background).

4.5.1 Wound Bounding Box Detection with Synthetic Data Augmentation

To assess the performance improvement from using synthetic images in the training process, we gradually increase the number of synthetic images added to the training sets of limited real images. We can see in Figure 4.5 that augmenting the entire real training dataset with synthetic images significantly improves the wound detection performance. This observation highlights the capacity of synthetic images to introduce meaningful information (beyond what is in the real images) during training. Figure 4.5 demonstrates that the addition of



(a)



(b)

Figure 4.5: Wound bounding box detection performance across five folds (mean and standard deviation) on FUSeg dataset, where the number of synthetic images added to a fixed number of real images in the training set gradually increases. Bounding box detection performance is measured by (a) IoU and (b) $AP_{centroid}$ (note that the vertical scales of the two plots are different). The plotted results extend up to the point of convergence. The horizontal red line indicates the results for the model that is trained on 610 real images, which shows the bounded performance using all the real images.

synthetic images consistently improves the detection performance and reduces the standard deviation error in the results, thus leading to more robust and reliable performance.

We note that the performance of the model converges after the addition of 400 synthetic training images and increasing them beyond 1000 did not significantly increase the performance. However, this maybe partly application dependent.

Moreover, using only less than $1/6^{\text{th}}$ of the available real images (100 annotated real images) alongside synthetic ones, we can achieve comparable detection results to the upper bound, which is less than a 2% drop in performance. Note that for generating synthetic training images using DermSynth3D, only 50 lesion annotations were used, which is 8.2% of the cost of dense annotations compared with the real dataset of wounds. Another notable observation in Figure 4.5 is that by adding 100 synthetic images to a very small dataset of 10 real images, we can achieve a similar performance as a dataset of 100 real images. This demonstrates the usefulness of this approach in situations where real data is extremely limited.

4.5.2 Wound Bounding Box Detection and Semantic Segmentation using Only Synthetic Data

To further explore the usefulness of our synthetic images in scenarios where there is no real training data available, we conduct additional experiments. We create a synthetic dataset of 610 images, which is the same size as the "real" wound image training set of the FUSeg dataset. We then evaluate the performance of a model in bounding box detection and segmentation when it is trained on this *synthetic-only* dataset and tested on the real wound image testset. The quantitative results are reported in Table 4.1 alongside the model's performance when trained on the FUSeg training set of real wound images, under the same training settings.

Our experiments show that for wound detection, when only synthetic DermSynth3D data is available, an average precision of 80% in wound localization can still be achieved. We can see in Table 4.1 that the model trained on only synthetic images achieves an $\text{AP}_{\text{centroid}}$ of 0.80 and IoU of 0.42. The significant gap between the IoU and $\text{AP}_{\text{centroid}}$ suggests that the model localizes the wounds, but does not precisely match the bounding boxes encapsulating

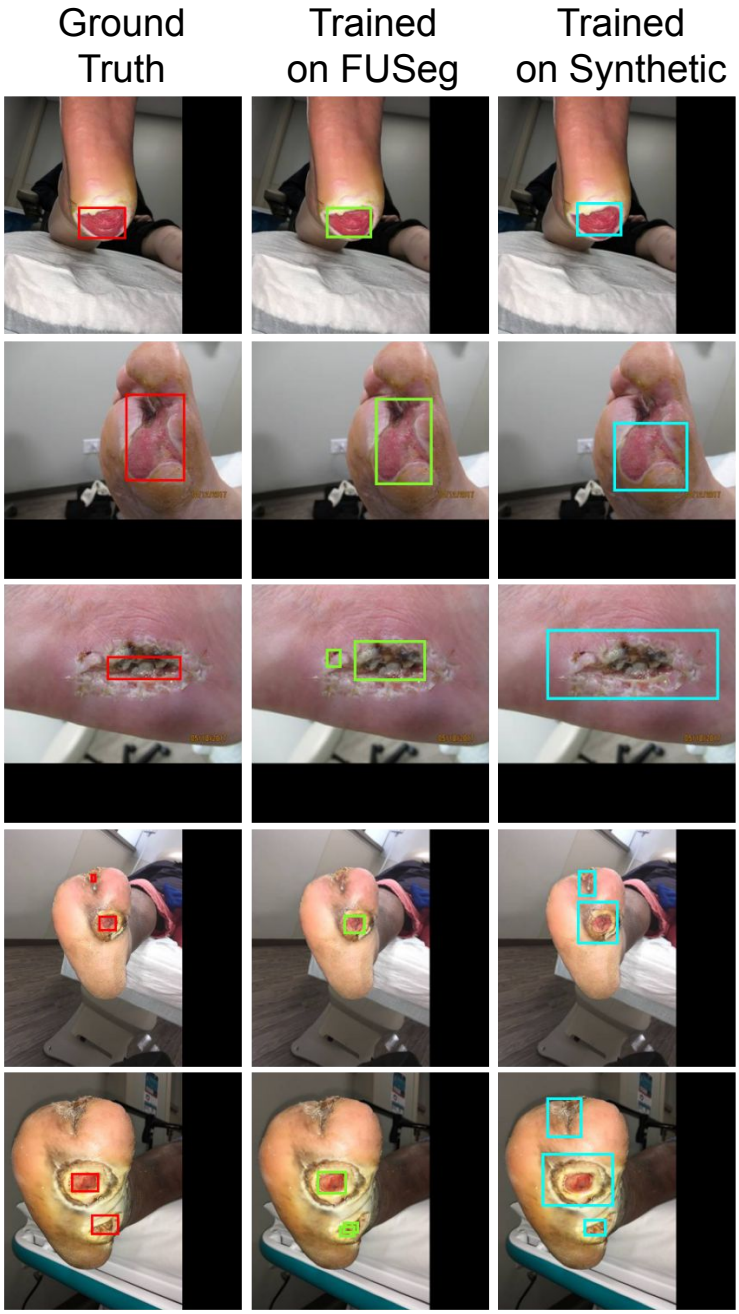


Figure 4.6: Qualitative results for foot ulcer bounding box detection on FUSeg dataset

Table 4.1: Foot ulcer bounding box detection and segmentation performance on the test set of real images of wounds.

	Detection (bounding box overlap)		**Segmentation (pixel-wise comparison)**	
Train dataset	$AP_{centroid}$	IoU	Dice	IoU
Synthetic	0.80 ±0.018	0.42 ±0.011	0.49 ±0.007	0.37 ±0.008
FUSeg	0.88 ±0.012	0.61 ±0.008	0.81 ±0.003	0.71 ±0.004

them. By analyzing the qualitative results of the model's predictions (Figure 4.6), we observe two major trends in the model's failure cases. (1) There seems to be a semantic difference between a skin condition and a wound. In our synthetic dataset, the whole lesion area, including the surrounding affected skin, is annotated as the lesion. However, in the FUSeg dataset, only the open-wound area is covered by the segmentation mask. This mismatch in labeling across these two image domains causes the model to over-segment some images (Figure 4.6 bottom three rows), resulting in a drop in the IoU. (2) As the synthetic data contains a variety of skin conditions across different parts of the body when trained on synthetic images, the model learns to detect other skin conditions within the image that are not of the wound. This can cause the model to over-detect wounds in the images (Figure 4.6 bottom row), resulting in a decrease in both IoU and $AP_{centroid}$.

Additionally, for the segmentation performance, Table 4.1 shows that a model trained on only synthetic images still achieves a Dice score of 0.49, which is more than 60% of the performance on real data (0.81 Dice), despite the differences in semantic content (skin conditions selected from Fitzpatrick17K dataset versus foot ulcers) and source domains (synthetic versus real). This demonstrates that even in the absence of real images, training on synthetic DermSynth3D data can provide more than 60% of the expected performance when trained on real clinical images, despite the significant domain gaps.

4.5.3 Utility of Synthetic Data in Pre-training for Wound Detection

Since the introduction of AlexNet [68], leveraging pre-trained models trained on extensive datasets and fine-tuning them for subsequent tasks has become a widely adopted strategy within the computer vision community [67, 103]. Nevertheless, existing pre-trained models are predominantly trained on natural images, which exhibit a notable domain gap when

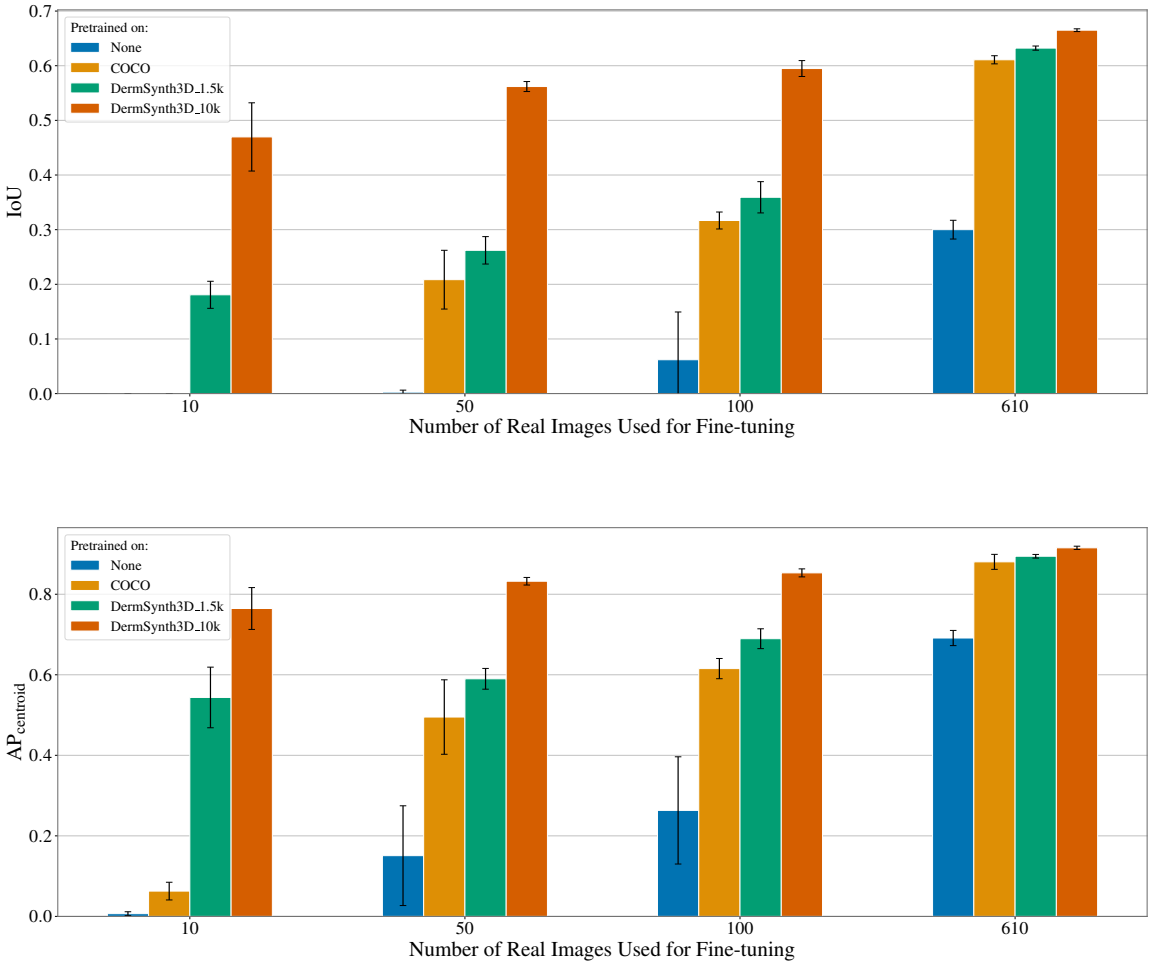


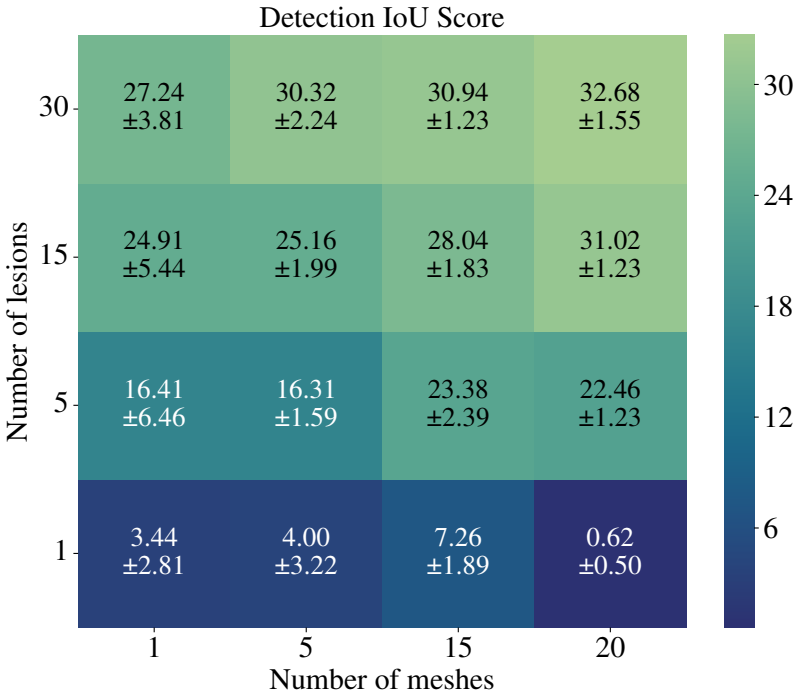
Figure 4.7: Wound bounding box detection performance across three folds (mean and standard deviation) on FUSeg dataset. The pre-training method is changed across experiments with four methods of training from scratch, pretrained backbone on COCO, and two datasets of generated images from DermSynth3D, with sizes of small (1.5k images) and large-scale (10k images).

compared to medical images. The unavailability of pre-trained models tailored to medical data stems mainly from the challenges of annotating such data and the associated costs of constructing large-scale datasets suitable for pretraining models. However, our proposed data synthesis framework, DermSynth3D, can potentially create large-scale data with a relatively much lower cost.

To assess the utility of the synthetic data in pre-training for wound detection, we perform a set of experiments where we use the synthetically generated data from DermSynth3D to pre-train Faster R-CNN [92] model from scratch and fine-tune the model on sets of limited real images. We compare the results obtained on the test set with the other scenarios such that the model is not pre-trained at all and pre-trained on on COCO dataset [75]. We can see in Figure 4.7 that even though the size of our datasets of generated images from DermSynth3D (1.5k and 10k images) is much smaller than the COCO dataset (about 238k images), by pre-training the model on a synthetic dermatological dataset, the model's performance improves noticeably. In addition, in the case of very limited data, with only 10 or 50 real images for fine-tuning, a model pre-trained on DermSynth3D data can achieve comparable performance to that of fine-tuned on the whole dataset of real images (610 images). Therefore, a notable utility of our proposed framework can be in synthesizing large-scale "in-the-wild" clinical datasets for enhancing model performance via pre-training the model on a more specific and similar dataset.

4.6 Ablation Study

To explore the impact of parameter selections in image synthesis on the end results, we conducted an ablation study focusing on a specific application of the proposed framework: foot ulcer bounding box detection. Given the significant cost associated with manual segmentation and acquiring skin lesions and textured meshes, we concentrate on evaluating how varying the number of lesions and meshes influences the performance of foot ulcer bounding box detection. We systematically adjusted the number of lesions and their blending with different numbers of meshes. We generated a training dataset of 1500 synthetic images. We



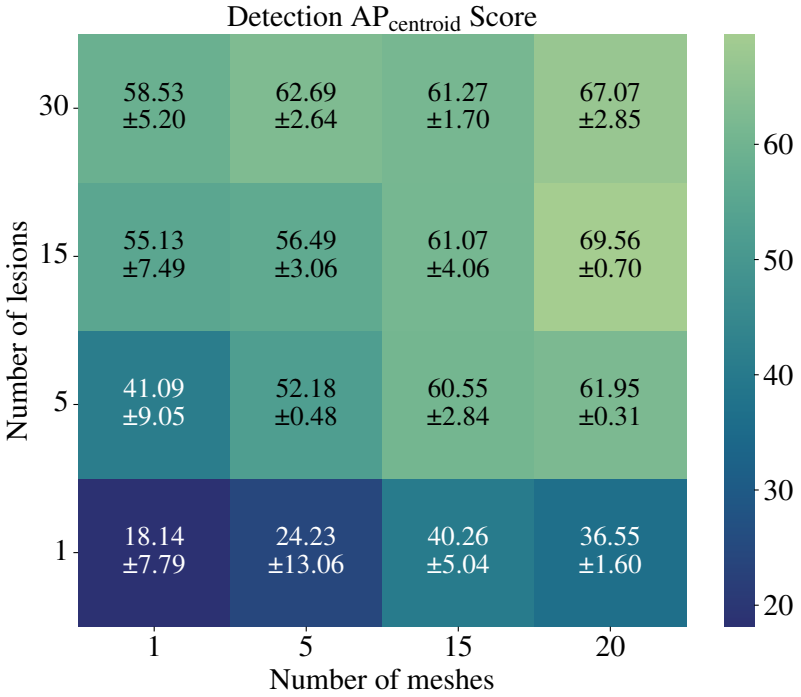


Figure 4.8: An ablation study on the effect of the number of lesions and number of meshes on the downstream task of bounding-box detection is visualized as a heatmap. The darker the shade, the lower the value of the performance metric.

followed the experimental settings outlined in Section 4.4, and used the real test set of the FUSeg dataset for evaluation.

For the wound detection task, we can see in Figure 4.8 that the results improve when generating synthetic images using more lesions and more meshes, which can be attributed to the overall increased diversity of the training images. Furthermore, increasing the lesion count while maintaining a consistent mesh quantity yields a more discernible enhancement in comparison to solely increasing the number of meshes. Moreover, adding more lesions while keeping a consistent number of meshes results in a more noticeable improvement compared to increasing the number of meshes alone. This underscores the significance of lesion variability as a key factor in the efficacy of the data produced by DermSynth3D for the given task. The observed performance gains can be attributed to the diverse image variations achieved by modifying the lighting and viewpoints for each mesh. However, beyond a certain threshold, the benefits diminish when solely augmenting the number of meshes, likely due to differences (e.g., skin tones) between the meshes and the real images.

4.7 Summary

In this chapter we introduced DermSynth3D, a novel framework for synthesizing densely annotated *in-the-wild* dermatological images by blending 2D skin conditions onto textured 3D meshes of human subjects and generating a custom dataset of 2D views with corresponding labels that span across several downstream tasks, such as segmentation and detection. Through extensive evaluation, we show the effectiveness of the generated synthetic data on two main dermatological applications of foot ulcer detection and segmentation, by demonstrating the generalization achieved after training a model on synthetic data and testing on real data. We observed that when the generated synthetic images are added to a small dataset of real images in the training process, they can improve the model's performance. Our results suggest that DermSynth3D has the potential to generate meaningful dermatological data for computerized skin image analysis, especially in resource-constrained or ethically challenging real-world scenarios. We also performed ablation studies to ascertain the contribution of the main components of our image synthesis pipeline.

Chapter 5

Conclusion and Future Work

5.1 Summary of Contributions

In this thesis, we directed our attention to addressing the dermatological data imbalance and dataset availability, and our contribution is two-fold.

In our first contribution, we addressed the problem of mitigating bias in DL-based models for the classification of skin conditions across skin types. We proposed a skin color-invariant model by using a domain-invariant representation-learning method. We proposed a skin color transformer by using a generative model to learn mappings from one skin type to another in a clinical skin condition image, and we enforced the learning objective of the classification model to be invariant across different skin types. We demonstrated that the proposed model enhanced classification performance while improving the results' fairness across skin types, resulting in less biased diagnosis and better model generalization and adaptability.

In our second work, we addressed the problem of a lack of annotated in-the-wild clinical data in the literature. We leveraged textured 3D meshes and blended 2D skin conditions onto them to synthesize densely annotated in-the-wild dermatological images that can be utilized for several downstream tasks. We showed the effectiveness and utility of the synthesized images on two applications of detection and segmentation of skin conditions and demonstrated the model generalization to real data when trained on our synthetic images through extensive evaluation.

5.2 Thesis Limitations and Future Work

Some of the limitations in our work can open up new research directions and potential future works in the field:

5.2.1 Limitations of skin condition image datasets with skin type annotations

In order to develop fair and accurate DL-based data-driven diagnosis methods in dermatology, we need annotated datasets that include a diversity of skin types and a range of skin conditions. However, only a few publicly available datasets satisfy these criteria. Out of all the datasets identified by the Seventh ISIC Skin Image Analysis Workshop at European Conference on Computer Vision 2022 (derm7pt [58], Dermofit Image Library [12], Diverse Dermatology Images (DDI) [36], Fitzpatrick17K [47], ISIC 2018 [28], ISIC 2019 [29,31,112], ISIC 2020 [96], MED-NODE [45], PAD-UFES-20 [86], PH2 [79], SD-128 [109], SD-198 [109], SD-260 [126]), only three datasets contain Fitzpatrick skin type labels: Fitzpatrick17K with 16,577, DDI with 656, and PAD-UFES-20 with 2,298 clinical images. The Fitzpatrick17K dataset is the only dataset out of these three which covers all the 6 different skin types (with over 600 images per skin type) and contains more than 10K images, suitable for training high-capacity DL-based networks and our GAN-based color transformer. It also contains samples from 114 different skin conditions, which is the largest number compared to the other two. For these reasons, we used the Fitzpatrick17K dataset for training and evaluating CIRCLe. However, skin conditions in the Fitzpatrick17K dataset images are not verified by dermatologists, and skin types in this dataset are annotated by non-dermatologists. Also, the patient images captured in the clinical settings exhibit various lighting conditions and perspectives. During our experiments, we found many erroneous and wrongly labeled images in the Fitzpatrick17K dataset, which could affect the training process. Our preliminary investigation into these data discrepancies has been further elaborated upon in the recent work [2]. Fig. 5.1 shows some erroneous images in the Fitzpatrick17K dataset. Therefore, one possible future work can be cleaning the Fitzpatrick17K dataset and verifying its skin conditions and skin types by dermatologists.

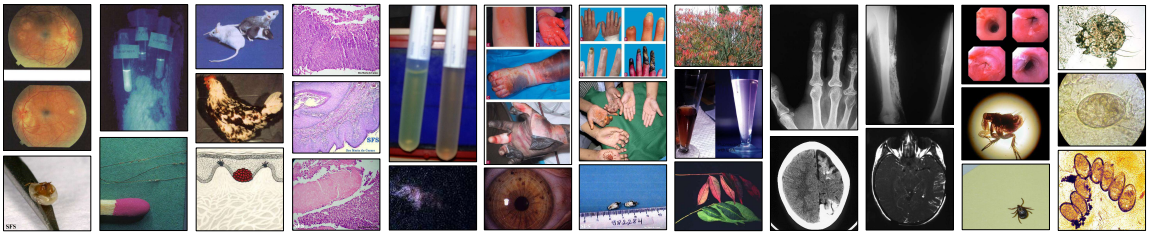


Figure 5.1: Sample erroneous images from the Fitzpatrick17K dataset that are not clinical images of skin conditions, but are included in the dataset and are wrongly labeled with skin conditions.

5.2.2 Extending available annotated data for skin condition classification

As we can see in Section 3.4.3 and Figure 3.2, the number of training images plays a significant role in the model's performance across different skin types. Although we proposed CIRCLe for improving the skin condition classifier's fairness and generalizability, the importance of obtaining large and diverse datasets must not be neglected. Mitigating bias in AI diagnosis tools in the algorithm stage, as we proposed, can be effective and is particularly essential for the currently developed models, however, future research at the intersection of dermatology and computer vision should have specific focus on adding more diverse and annotated images to existing databases.

While our synthetic dataset generated by DermSynth3D included a range of skin tones derived from subjects in the 3DBodyTex dataset, future work may examine extending DermSynth3D to enable selection of specific skin tones, thereby enhancing dataset diversity. Our initial efforts to estimate skin tone based on pixel-level values within the DermSynth3D framework encountered challenges, notably due to the sensitivity of calculation-based methods to image acquisition environments and manual thresholding. These limitations present opportunities for refinement in future research endeavors.

5.2.3 Maintaining skin condition diagnosis in data synthesis

A natural extension to both of our works is to introduce constraints to the data synthesis to maintain the skin condition label. In our first work, while the total loss function of CIRCLe addresses both skin condition classification performance and skin type representation invariance, the skin color transformer is not directly constrained to maintain the skin condition

label during the transformation. Therefore, since skin conditions appear differently across skin types, the images synthesized from the transformation may not be dermatologically correct representations of their original diagnosis. Moreover, in our second work, the design choices of DermSynth3D are as such to randomize the parameter variations (e.g., skin condition type, location on the body, size, etc.) during the dataset creation to diversify the data while synthesizing visually plausible images with high utility in the training process of DL-based models in downstream tasks. However, in reality, different skin conditions might appear in specific parts of the body with certain size limits; therefore, future works can extend our proposed data synthesis framework to address skin condition diagnosis-related constraints to generate more dermatologically correct data.

5.2.4 Domain gap between DermSynth3D data and foot ulcers

In our second work's experiments, while we used the FUSeg dataset as the dataset of real in-the-wild skin condition images, we acknowledge that there is a semantic domain gap between a skin condition and a wound. As we see in Section 4.5.1, the model's performance when trained on synthetic images can partially be attributed to this semantic difference. Moreover, while the FUSeg dataset only contains images of ulcers on the foot (Figure 2.5), the DermSynth3D dataset contains images of different types of skin conditions on various parts of the body (Figure4.2). Future works can explore utilizing domain adaptation methods to improve the segmentation and detection performance on real data (Table 4.1) by leveraging the generated synthetic data.

5.2.5 Other possible future works

In exploring the future directions of this work, it is important to acknowledge the evolving landscape of machine learning architectures. While CNNs have served as the fundamental framework in our study for their efficacy in image classification and object detection tasks, it is essential to recognize the growing prominence of Transformer models [114]. Addressing this concern, future work could explore the integration of Transformer architectures. However, it is worth noting that Transformers typically demand larger amounts

of data and computational resources for effective training, potentially posing challenges in resource-constrained environments.

Moreover, diffusion-based modelling [53], presents a promising avenue for synthesizing dermatological images. Future works can explore more photo-realistic and diverse dermatological image generation using stable diffusion models conditioned on disease class, skin type, location on the body, etc.

Bibliography

[1] Kumar Abhishek and Ghassan Hamarneh. Mask2lesion: Mask-constrained adversarial skin lesion image synthesis. In *Simulation and Synthesis in Medical Imaging: 4th International Workshop, SASHIMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings*, pages 71–80. Springer, 2019.

[2] Kumar Abhishek, Aditi Jain, and Ghassan Hamarneh. Investigating the quality of dermamnist and fitzpatrick17k dermatological image datasets. *arXiv preprint arXiv:2401.14497*, 2024.

[3] Kumar Abhishek, Jeremy Kawahara, and Ghassan Hamarneh. Predicting the clinical management of skin lesions using deep learning. *Scientific Reports*, 11(1):1–14, 2021.

[4] Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA Dermatology*, 154(11):1247–1248, 2018.

[5] Adewole S Adamson, Elizabeth A Suarez, and H Gilbert Welch. Estimating overdiagnosis of melanoma using trends among black and white patients in the US. *JAMA Dermatology*, 158(4):426–431, 2022.

[6] Ademide Adelekun, Ginikanwa Onyekaba, and Jules B Lipoff. Skin color in dermatology textbooks: An updated evaluation and analysis. *Journal of the American Academy of Dermatology*, 84(1):194–196, 2021.

[7] Oma N. Agbai, Kesha Buster, Miguel Sanchez, Claudia Hernandez, Roopal V. Kundu, Melvin Chiu, Wendy E. Roberts, Zoe D. Draelos, Reva Bhushan, Susan C. Taylor, and Henry W. Lim. Skin cancer and photoprotection in people of color: A review and recommendations for physicians and the public. *Journal of the American Academy of Dermatology*, 70(4):748–762, 2014.

[8] Jehad Amin. DermaAmin. https://www.dermaamin.com. Accessed: 2022-04-14.

[9] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. 2016.

[10] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2021.

[11] Lucia Ballerini, Robert B. Fisher, Ben Aldridge, and Jonathan Rees. A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions. In M. Emre Celebi and Gerald Schaefer, editors, *Color Medical Image Analysis*, volume 6, pages 63–86. Springer Netherlands, 2013.

[12] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. 2013.

[13] Catarina Barata, Jorge S Marques, and M Emre Celebi. Deep attention model for the hierarchical diagnosis of skin lesions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2757–2765, 2019.

[14] Christoph Baur, Shadi Albarqouni, and Nassir Navab. Generating highly realistic images of skin lesions with gans. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*, pages 260–267. Springer, 2018.

[15] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[16] Peter J Bevan and Amir Atapour-Abarghouei. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. *arXiv preprint arXiv:2202.02832*, 2022.

[17] Lei Bi, Jinman Kim, Ashnil Kumar, Dagan Feng, and Michael Fulham. Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs). In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment: Fifth International Workshop, CMMI 2017, Second International Workshop, RAMBO 2017, and First International Workshop, SWITCH 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 5*, pages 43–51. Springer, 2017.

[18] David R. Bickers, Henry W. Lim, David Margolis, Martin a. Weinstock, Clifford Goodman, Eric Faulkner, Ciara Gould, Eric Gemmen, and Tim Dall. The burden of skin diseases: 2004. A joint project of the American Academy of Dermatology Association and the Society for Investigative Dermatology. *Journal of the American Academy of Dermatology*, 55(3):490–500, 2006.

[19] Judith S Birkenfeld, Jason M Tucker-Schwartz, Luis R Soenksen, José A Avilés-Izquierdo, and Berta Marti-Fuster. Computer-aided classification of suspicious pigmented lesions using wide-field images. *Computer Methods and Programs in Biomedicine*, 195:105631, 2020.

[20] Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra 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: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*, pages 294–302. Springer, 2018.

[21] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, January 2018.

[22] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Fröhling, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019.

[23] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[24] Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE transactions on medical imaging*, 37(3):803–814, 2017.

[25] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. In *arXiv:1706.05587*, pages 1–14, 2017.

[26] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[27] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 240–244. IEEE, 2018.

[28] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

[29] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 168–172, 2018.

[30] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[31] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig,

et al. BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[32] Clara Curiel-Lewandrowski, Roberto A Novoa, Elizabeth Berry, M Emre Celebi, Noel Codella, Felipe Giuste, David Gutman, Allan Halpern, Sancy Leachman, Yuan Liu, et al. Artificial intelligence approach in melanoma. *Melanoma*, pages 1–31, 2019.

[33] Samuel Freire da Silva. Atlas dermatologico. http://atlasdermatologico.com.br. Accessed: 2022-04-14.

[34] Fei Dai, Dengyi Zhang, Kehua Su, and Ning Xin. Burn Images Segmentation Based on Burn-GAN. *Journal of Burn Care & Research*, 42(4):755–762, 2021.

[35] Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, 157(11):1362–1369, 2021.

[36] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(31):eabq6147, 2022.

[37] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis. *arXiv preprint arXiv:2302.00785*, 2023.

[38] Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10):2375–2388, 2019.

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[40] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[41] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[42] Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color and Imaging Conference*, volume 2004, pages 37–41. Society for Imaging Science and Technology, 2004.

[43] Thomas B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, 124(6):869–871, 1988.

[44] Lauren Fried, Andrea Tan, Shirin Bajaj, Tracey N Liebman, David Polsky, and Jennifer A Stein. Technological advances for the detection of melanoma: Advances in diagnostic techniques. *Journal of the American Academy of Dermatology*, 83(4):983–992, 2020.

[45] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, 42(19):6578–6585, 2015.

[46] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[47] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

[48] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In *ISIC Skin Image Analysis CVPR Workshop*, pages 1–9, 2021.

[49] H.A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

[50] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3323–3331, 2016.

[51] Md. Kamrul Hasan, Md. Asif Ahamad, Choon Hwai Yap, and Guang Yang. A survey, review, and future trends of skin lesion segmentation and classification. *Computers in Biology and Medicine*, 155:106624, 2023.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[53] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[54] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[55] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[56] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[57] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2018.

[58] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[59] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. GANs for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020.

[60] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Estimating skin tone and effects on classification performance in dermatology datasets. *arXiv preprint arXiv:1910.13268*, 2019.

[63] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *Medical Image Computing and Computer-Assisted Intervention*, pages 320–329, 2020.

[64] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

[65] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *American Economic Association Papers and Proceedings*, volume 108, pages 22–27, 2018.

[66] Marc D Kohli, Ronald M Summers, and J Raymond Geis. Medical image data and datasets in the era of machine learning – whitepaper from the 2016 C-MIMI meeting dataset session. *Journal of Digital Imaging*, 30:392–399, 2017.

[67] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[68] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[69] Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, and Mitko Veta. Learning domain-invariant representations of histological images. *Frontiers in Medicine*, 6:162, 2019.

[70] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[71] JC Lester, JL Jia, L Zhang, GA Okoye, and E Linos. Absence of images of skin of colour in publications of COVID-19 skin manifestations. *The British Journal of Dermatology*, 183(3):593–595, 2020.

[72] Hongfeng Li, Yini Pan, Jie Zhao, and Li Zhang. Skin disease diagnosis with deep learning: A review. *Neurocomputing*, 464:364–393, 2021.

[73] Yunzhu Li, Andre Esteva, Brett Kuprel, Rob Novoa, Justin Ko, and Sebastian Thrun. Skin cancer detection and tracking using data synthesis and deep learning. In *AAAI Conference on Artificial Intelligence Joint Workshop on Health Intelligence*, pages 551–554, 2017.

[74] Jiamin Liang, Xin Yang, Yuhao Huang, Haoming Li, Shuangchi He, Xindi Hu, Zejian Chen, Wufeng Xue, Jun Cheng, and Dong Ni. Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. *Medical Image Analysis*, 79:102461, 2022.

[75] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[76] Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1756–1764, 2022.

[77] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Greg S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan J. Huang, Yun Liu, R. Carter Dunn, and David Coz. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020.

[78] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? In *IEEE ICCV*, pages 2697–2706, 2017.

[79] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. PH^2—A Dermoscopic Image Database for Research and Benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5437–5440, 2013.

[80] Zahra Mirikharaji, Catarina Barata, Kumar Abhishek, Alceu Bissoto, Sandra Avila, Eduardo Valle, M Emre Celebi, and Ghassan Hamarneh. A survey on deep learning for skin lesion segmentation. *arXiv preprint arXiv:2206.00356*, 2022.

[81] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on International Conference on Machine Learning*, pages 10–18, 2013.

[82] A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021.

[83] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 417–425. Springer, 2017.

[84] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[85] Muhammad Osto, Iltefat H Hamzavi, Henry W Lim, and Indermeet Kohli. Individual typology angle and Fitzpatrick skin phototypes are not equivalent in photodermatology. *Photochemistry and Photobiology*, 98(1):127–129, 2022.

[86] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomão, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020.

[87] Arezou Pakzad, Kumar Abhishek, and Ghassan Hamarneh. CIRCLe: Color invariant representation learning for unbiased classification of skin lesions. In *Proceedings of the 17th European Conference on Computer Vision (ECCV) - ISIC Skin Image Analysis Workshop*, pages 203–219, 2022.

[88] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035. 2019.

[89] Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Augmenting data with gans to segment melanoma skin lesions. *Multimedia Tools and Applications*, 79:15575–15592, 2020.

[90] Simon Prince. Tutorial #1: Bias and fairness in AI, 2019. https://www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai/. Accessed: 2022-04-14.

[91] Esther Puyol-Antón, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, and Andrew P. King. Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 413–423, 2021.

[92] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[93] Robin Reni. House rooms image dataset. `https://www.kaggle.com/datasets/robinreni/house-rooms-image-dataset`. Accessed: 2022-05-17.

[94] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

[95] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1):34, 2021.

[96] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1):1–8, 2021.

[97] Alexandre Saint, Eman Ahmed, Abd El Rahman Shabayek, Kseniya Cherenkova, Gleb Gusev, Djamila Aouada, and Björn Ottersten. 3DBodyTex: Textured 3D body dataset. In *International Conference on 3D Vision*, pages 495–504, 2018.

[98] Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Gleb Gusev, Djamila Aouada, and Björn Ottersten. BODYFITR: Robust automatic 3D human body fitting. In *IEEE International Conference on Image Processing*, pages 484–488, 2019.

[99] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[100] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *Biocomputing 2021: proceedings of the Pacific symposium*, pages 232–243, 2020.

[101] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

[102] Pourya Shamsolmoali, Masoumeh Zareapoor, Eric Granger, Huiyu Zhou, Ruili Wang, M Emre Celebi, and Jie Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, 72:126–146, 2021.

[103] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[104] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[105] Ashish Sinha, Jeremy Kawahara, Arezou Pakzad, Kumar Abhishek, Matthieu Ruthven, Enjie Ghorbel, Anis Kacem, Djamila Aouada, and Ghassan Hamarneh. Dermsynth3d: Synthesis of in-the-wild annotated dermatology images. *Medical Image Analysis*, 2024.

[106] Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. GANs for medical image synthesis: An empirical study. *Journal of Imaging*, 9(3):69, March 2023.

[107] Wiebke Sondermann, Jochen Sven Utikal, Alexander H. Enk, Dirk Schadendorf, Joachim Klode, Axel Hauschild, Michael Weichenthal, Lars E. French, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Stefan Fröhling, Christof von Kalle, and Titus J. Brinker. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: A call for prospective data. *European Journal of Cancer*, 119:30–34, 2019.

[108] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images. In *European Conference on Computer Vision*, pages 206–222, 2016.

[109] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

[110] The University of Edinburgh. Dermofit Image Library. `https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html`.

[111] Francis Tom and Debdoot Sheet. Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 1174–1177. IEEE, 2018.

[112] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

[113] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:1–9, 2018.

[114] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[115] Chuanbo Wang, D. M. Anisuzzaman, Victor Williamson, Mrinal Kanti Dhar, Behrouz Rostami, Jeffrey Niezgoda, Sandeep Gopalakrishnan, and Zeyun Yu. Fully automatic wound segmentation with deep convolutional neural networks. *Scientific Reports*, 10(21897):1–9, 2020.

[116] Chuanbo Wang, Amirreza Mahbod, Isabella Ellinger, Adrian Galdran, Sandeep Gopalakrishnan, Jeffrey Niezgoda, and Zeyun Yu. Fuseg: The foot ulcer segmentation challenge. *arXiv preprint arXiv:2201.00414*, 2022.

[117] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F Wynne, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of applied clinical medical physics*, 22(1):11–36, 2021.

[118] Yan Wang, Biting Yu, Lei Wang, Chen Zu, David S Lalush, Weili Lin, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *Neuroimage*, 174:550–562, 2018.

[119] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

[120] Olivia R Ware, Jessica E Dawson, Michi M Shinohara, and Susan C Taylor. Racial limitations of Fitzpatrick skin type. *Cutis*, 105(2):77—80, 2020.

[121] Ellen Buchanan Weiss. Brown skin matters. https://brownskinmatters.com/. Accessed: 2022-06-23.

[122] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022.

[123] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake It Till You Make It: Face analysis in the wild using synthetic data alone. In *International Conference on Computer Vision*, pages 3681–3688, 2021.

[124] Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pages 743–753. Springer, 2022.

[125] Jufeng Yang, Xiaoping Wu, Jie Liang, Xiaoxiao Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Self-paced balance learning for clinical skin disease recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2832–2846, 2019.

[126] Jufeng Yang, Xiaoping Wu, Jie Liang, Xiaoxiao Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Self-paced balance learning for clinical skin disease recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2832–2846, 2019.

[127] Xiaofeng Yang. *Medical image synthesis: Methods and clinical applications.* July 2023.

[128] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

[129] Albert T Young, Niki B Vora, Jose Cortez, Andrew Tam, Yildiray Yeniay, Ladi Afifi, Di Yan, Adi Nosrati, Andrew Wong, Arjun Johal, et al. The role of technology in melanoma screening and diagnosis. *Pigment Cell & Melanoma Research*, 34(2):288–300, 2021.

[130] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep Image Blending. In *Winter Conference on Applications of Computer Vision*, pages 231–240, Los Alamitos, CA, USA, 2020. IEEE Computer Society.

[131] Mengliu Zhao, Jeremy Kawahara, Kumar Abhishek, Sajjad Shamanian, and Ghassan Hamarneh. Skin3d: Detection and longitudinal tracking of pigmented skin lesions in 3d total-body textured meshes. *Medical Image Analysis*, 77:102329, 2022.