Color Invariant Skin Segmentation

Xu, Han

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Abbott, A. Lynn, Chair
Sarkar, Abhijit
Huang, Jia-Bin

Febrary 15, 2022
Blacksburg, Virginia

Keywords: Deep Learning, Image Segmentation, Skin Detection, Image Classification

Copyright 2022, Xu, Han

Color Invariant Skin Segmentation

Xu, Han

(ABSTRACT)

This work addresses the problem of automatically detecting human skin in images without reliance on color information. Unlike previous methods, we present a new approach that performs well in the absence of such information. A key aspect of the work is that color-space augmentation is applied strategically during the training, with the goal of reducing the influence of features that are based entirely on color and increasing more semantic understanding. The resulting system exhibits a dramatic improvement in performance for images in which color details are diminished. We have demonstrated the concept using the U-Net architecture, and experimental results show improvements in evaluations for all Fitzpatrick skin tones in the ECU dataset. We further tested the system with RFW dataset to show that the proposed method is consistent across different ethnicities and reduces bias to any skin tones. Therefore, this work has strong potential to aid in mitigating bias in automated systems that can be applied to many applications including surveillance and biometrics.

Color Invariant Skin Segmentation

Xu, Han

(GENERAL AUDIENCE ABSTRACT)

Skin segmentation deals with the classification of skin and non-skin pixels and regions in a image containing these information. Although most previous skin-detection methods have used color cues almost exclusively, they are vulnerable to external factors (e.g., poor or unnatural illumination and skin tones). In this work, we present a new approach based on U-Net that performs well in the absence of color information. To be specific, we apply a new color space augmentation into the training stage to improve the performance of skin segmentation system over the illumination and skin tone diverse. The system was trained and tested with both original and color changed ECU dataset. We also test our system with RFW dataset, a larger dataset with four human races with different skin tones. The experimental results show improvements in evaluations for skin tones and complex illuminations.

Acknowledgments

First and foremost, I have to thank my research supervisors, Dr. A. Lynn Abbott and Dr. Abhijit Sarkar. Without their assistance and dedicated involvement in every step throughout the process, this thesis would have never been accomplished. I would like to thank you very much for your support and understanding over these past years.

I would also like to show gratitude to the rest of my committee, Dr. Jia-Bin Huang, who was my instructor of classes related to computer vision and deep learning. His teaching style and enthusiasm for the topic made a strong impression on me and I have always carried positive memories of his classes with me.

I am grateful to the Graduate School for their continued assistance at Virginia Tech for concise tutorials and help during office hour sessions.

Most importantly, none of this could have happened without my family and my parents, who offered her encouragement from the other side of Pacific Ocean.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Semantic image segmentation, as an important part of image understanding in computer vision, has become increasingly prominent and popular in the industrial world and in current academic research. Semantic segmentation is classification at the pixel level and has been widely applied in geographic information systems [48], automatic driving, medical image analysis, robots and other various fields. The main task of semantic segmentation is shown in Figure 1.1a.

Before deep learning was applied to the field of computer vision, image segmentation was performed based on low-level visual cues of the pixels [16, 34]. Since early segmentation methods do not have an algorithm training stage, the computational complexity is often not high, but its segmentation effect is not satisfactory on more difficult segmentation tasks (if artificial auxiliary information is not provided). More recently, deep learning methods have achieved great success in semantic segmentation. Initially, patch classification is a commonly used deep learning method, which uses image blocks around each pixel to separate pixels into corresponding categories. The main reason for using image patches is that the classification network usually has a fully connected layer, and its input needs to be a fixed-size image patch. In 2015, the fully convolutional network (FCN) [50] was born. FCN replaced fully connected layers of the network with convolution, so that input of any image size is possible, and the

speed is much faster than the patch classification method. Then, a kind of encoder-decoder architecture (e.g., Segnet [6] and U-Net [62]) and dilated convolutions [85] were proposed to solve the problem of missing dimensions and detailed information of the restoration space in pooling layers. The Conditional Random Field (CRF) [14, 15] method is usually used in post-processing to improve the segmentation effect. The CRF method is a graph model that performs "smooth" segmentation based on the pixel intensity of the underlying image. Points with similar pixel intensity are marked as the same category at runtime. Adding the conditional random field method can increase the final score value by 2%.

Skin segmentation refers to the process of identifying pixels that correspond to human skin within image data or video data as shown in Figure 1.1b. Automated skin detection can play an important role for applications such as face detection and recognition (e.g., [45, 49]), facial expression recognition, gesture recognition [3], content-based image retrieval, filtering of objectionable content [22, 27], skin rendering in computer graphics [7, 21], and virtual reality.

Although the last two decades have seen many efforts related to skin detection and skin modeling [39, 89], it is interesting that almost all techniques for image-based skin detection depend heavily on the use of color information. (Extensive surveys are provided by Mahmoodi and Sayedi [51] and Kakumanu et al. [42].) In order to incorporate other features in skin detection algorithms, researchers have augmented their color-based approaches by incorporating cues related to texture and region shape [17, 28, 38], or by combining two color spaces [80]. In contrast, Sarkar et al. [66] provided an image-specific skin detection algorithm that does not depend on color cues, but instead learns local texture descriptors from skin regions using one-shot methods. However, most of these methods use traditional, hand-crafted feature extractors. Skin modeling, on the other hand, takes into consideration physics-based models of skin reflectance, albedo, and color, but often these models fail to

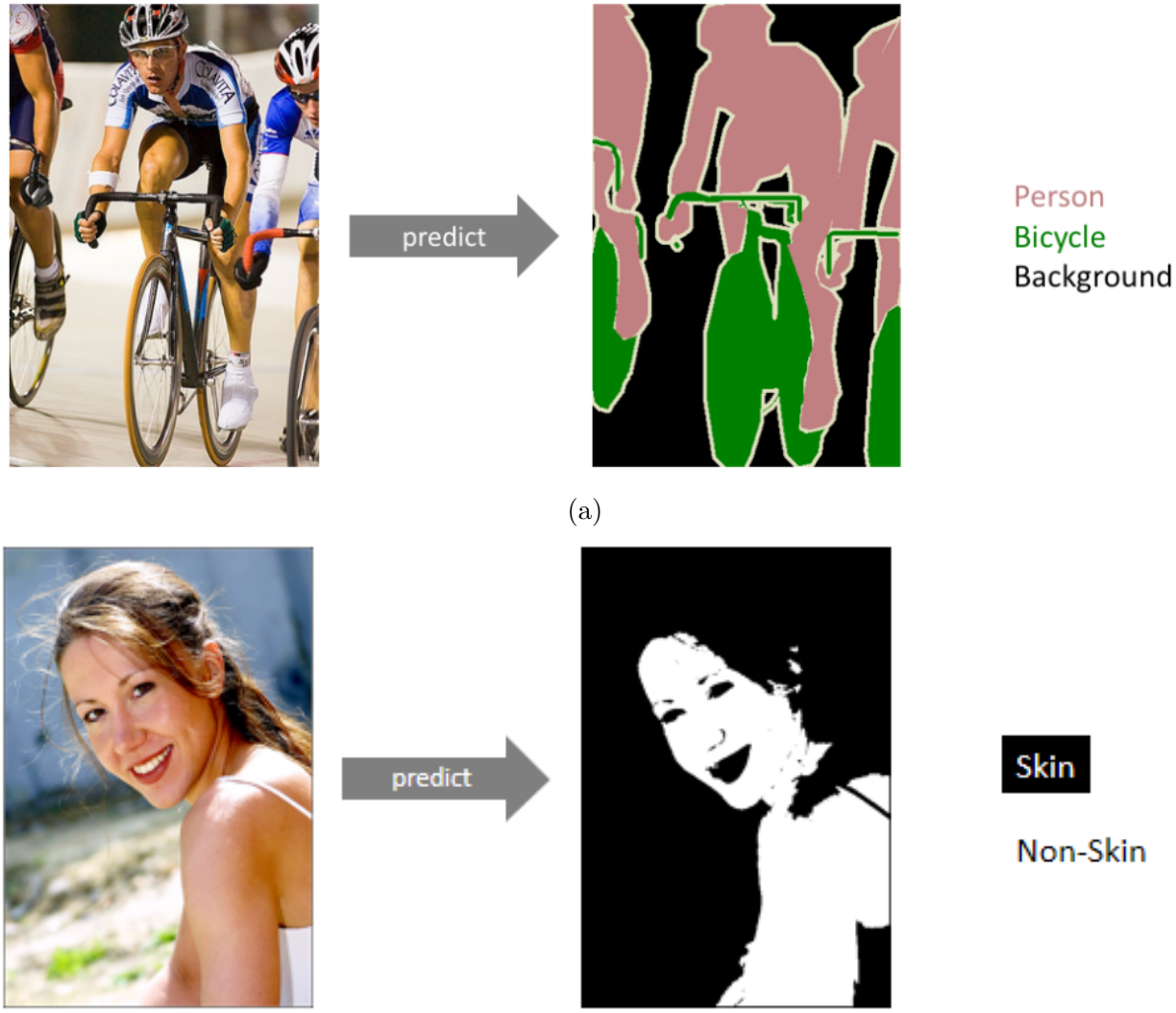


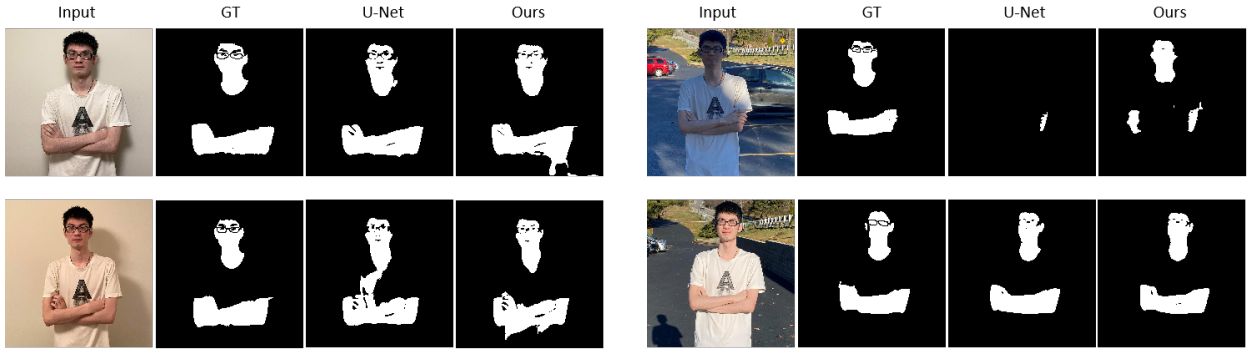
Figure 1.1: The main task of semantic segmentation (a) and skin segmentation (b). In semantic segmentation, pixels of different categories are classified in different colors. In skin segmentation, pixels are only classified into two categories. Image credit: (a): [25].

generalize well [89]. A consequence of over-reliance on color cues is the potential to introduce bias into the system. Bias can result from many perspectives, particularly because the appearance of skin in an image depends on many intrinsic and extrinsic factors. These include the skin tone of the person [64], the environment in which the image was captured (e.g., indoor vs. outdoor), ambient illumination source (sunlight, fluorescent light, neon light), imaging situation (low-light condition, shadows, overexposure to sunlight), scale (distance from the camera), and other factors. Additionally, imaging methods may also introduce variability, including spectral range of sensor arrays (grayscale, near-infrared, RGB) and creative filtering in photography applications (e.g., sepia tones in movies, or Instagram filters).

If all such variations are considered, the appearance of skin covers a much broader operating space that previously considered [24, 41]. Therefore, practically all previous skin-detection systems have been limited to a subspace of possible appearances. Figure 1.2 shows that different skin appearances result in various performances of skin detection systems. For example, an approach based on U-Net [62] predicted redundant false positive results between the neck and arm area in the second row of group (a). In the second row of group (b), the U-Net approach almost failed when the skin areas are covered by shadows.

Apparently, the outdoor environment has a significant impact on human skin, thus changing the color cues in the images. Figure 1.3 shows heatmap distributions of skin pixels in HSV color space. These heatmap distributions show that when human skin is exposed to different light conditions, the appearance of skin pixels in the images varies. Though the changes in group (b) may seem small, significant changes have been shown in group (a). In this case, skin detection systems will obtain different color features and make unexpected predictions. Thus the accuracy will decline.

In addition to the external factors such as sunlight and shadows, internal factors will also change the appearance of skin pixels in the image, which is what we call “skin tone”. The



(a) (b)

Figure 1.2: Example results from skin detection systems working in different environments. Group (a) refers to white and yellow bulb lights in the indoor environment. Group (b) refers to the outdoor environment with intense sunlight and shadow. *Left to right:* Input image; ground truth; U-Net output before training with our augmentation approach; U-Net output after training using our novel augmentation approach.

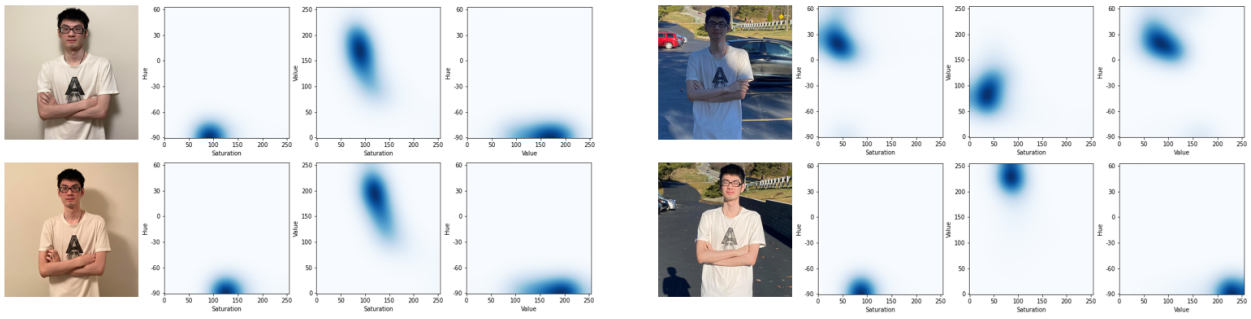


Figure 1.3: Heatmaps of skin pixels in HSV color space. The second to fourth columns in each group refers to the distribution in Saturation-Hue, Saturation-Value, and Value-Hue dimensions respectively.

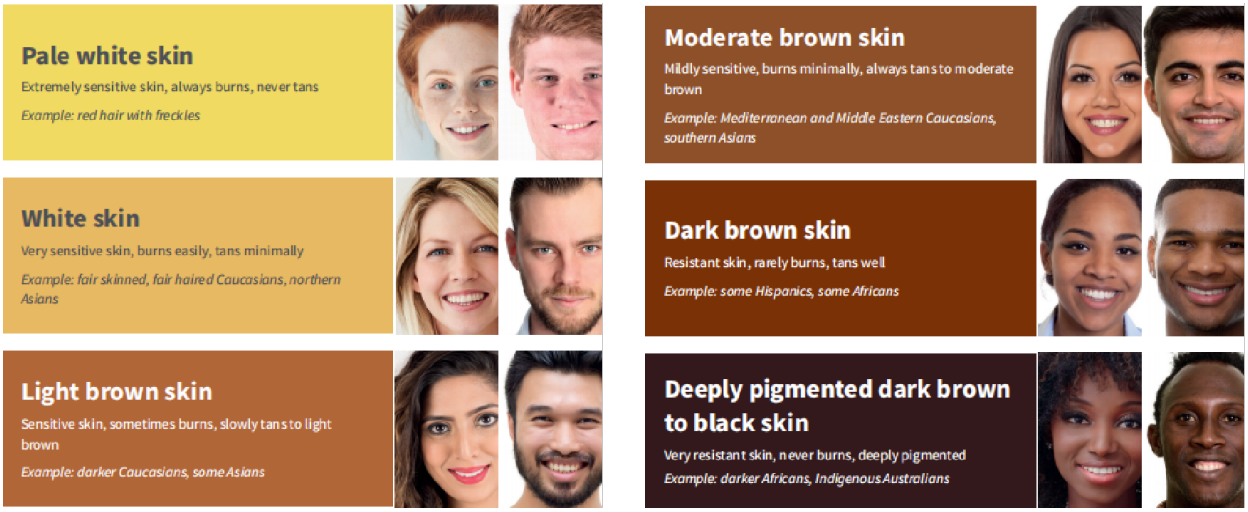


Figure 1.4: Examples of Fitzpatrick [26] skin tones, which is a commonly used system to describe a person's skin type in terms of response to ultraviolet radiation (UVR) exposure. Image credit: [5].

skin is composed of a dermis layer and a thin epidermal layer covering it. Light is absorbed by melanin in the epidermis, while absorption and reflection occur in the dermis at the same time. The difference in skin color of individuals is mainly manifested as brightness changes caused by different concentrations of melanin in the dermis. Fitzpatrick [26] proposed a commonly used system to describe a human's skin type. He divided the human skin type into six classes in terms of response to ultraviolet radiation (UVR) exposure as shown in Figure 1.4. It was widely used to report how human skin responded to sunlight. It also expanded to a broader range of skin types. We will use it to classify human skin types in the dataset. As a result, many skin detection methods fail not only for images obtained "in the wild" and in other unconstrained situations, but may also inadvertently introduce biases due to ethnic and racial differences. Studies [9, 10, 69] have shown that this physiological bias influences computer vision algorithms to some extent. We illustrate the heatmap from the testing set of the ECU dataset in Figure 1.5 to interpret how skin tones [26] affect the appearance of skin pixels. We first classified the testing set in the ECU dataset into six skin tones according to [26]. Every image in the testing set was converted into HSV color space,

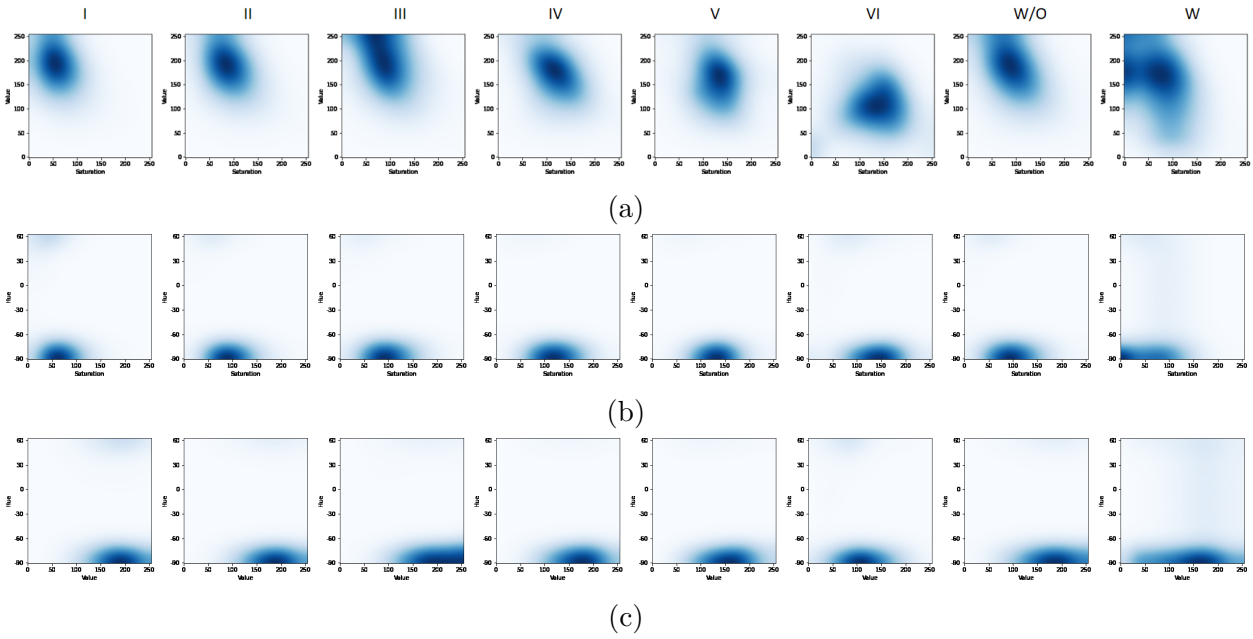


Figure 1.5: Heatmaps from ECU dataset in three dimensions: (a) Saturation-Value, (b) Saturation-Hue, (c) Value-Hue. The first six columns mark the skin pixels distributions of Fitzpatrick [26] skin tones I-VI. The last two columns refer to the skin pixel distribution of the training set before (W/O) and after (W) our color space augmentation.

and its skin pixels were allocated into different bins according to the (S, V), (S, H), and (V, H) value pairs. These heatmaps demonstrate the distribution of skin pixels of different skin tones in HSV color space. The changes in color cues will lead the skin detection system to make unexpected predictions. Take the S-V dimension (Row 1) as an example. From skin tone 1 to skin tone 6, with the skin going darker, the maximum of value channel rises, and the center of the saturation channel moves left. The heatmap in column seven shows that the training set cannot adapt to these changes. So typical skin detection systems will output results with lower accuracy.

To address such problems, our work introduces a new technique for human skin detection that significantly reduces reliance on color information and focuses much more on texture and contextual information to detect the skin pixels in an image. A significant aspect of the system is that color-space augmentation is applied strategically to the training set so that a

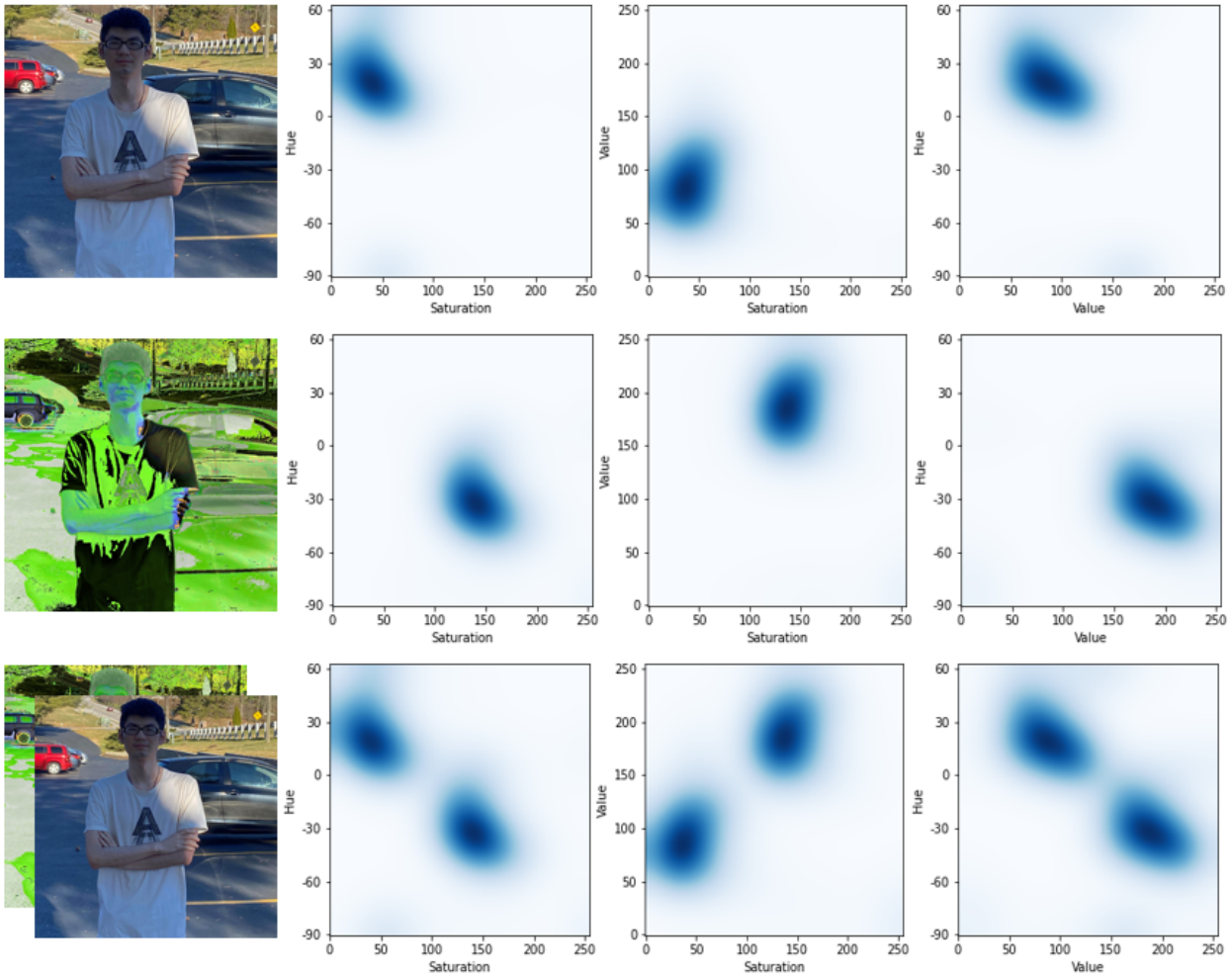


Figure 1.6: Change the HSV values of the image and move the heatmap cloud to desired location.

resulting deep neural network suppresses the system's dependence on color cues. As shown in Figure 1.6, the heatmap clouds can be moved to desired locations by changing the HSV values of the image. In this case, we can enlarge the training set by using multiple (H,S,V) value groups as data augmentation. The augmented training set has a larger heatmap cloud covering more areas. In this way, the influence of various appearances of skin pixels in the images will be eliminated. Hence, our high-level strategy has been to guide the training procedure away from color cues and toward features related to visual texture and context.

We demonstrate our procedure by training the U-Net architecture [6, 62] using ECU [58] datasets, and do testing on both ECU [58] and RFW [79]. The ECU dataset is a common benchmark dataset for skin detection. We have further annotated the ECU dataset by skin tone (Type I through Type VI following Fitzpatrick's method), and we demonstrate that our method performs almost uniformly across all skin types. The RFW dataset contains face images with annotations representing four ethnic groups. We have used the RFW dataset to show that our new algorithm shows virtually no bias to any ethnicity and skin tone.

1.2 Main Contributions

The primary contributions of this work are as follows.

1) *Color invariance.* We describe an approach to automated detection of human skin that does not depend on the color appearance of the skin. The approach does not require additional costly datasets or annotations.

2) *Universality.* The resulting system therefore has potential to operate in environments with relatively unconstrained illumination conditions, including extreme cases of over- and underexposed images, grayscale images, and systems that utilize creative filters (such as Instagram). As such, the system is intended for operation "in the wild," and can relax requirements and reduce costs related to camera selection.

3) *Little or no racial bias.* In our experimental results, we have systematically evaluated the performance of our algorithm for subjects with different skin tones. Using cross-database testing, we have shown that our new algorithm performs virtually uniformly across all of the available annotated skin tones. It is our hope that our color-augmentation strategy for training and testing can be applied widely to other domains, in order to address problems related to racial and social bias.

1.3 Thesis Organization

The whole thesis is organized in the following manner:

1. Chapter 1 presents the motivation of this work. It emphasizes that human skin performance in the image significantly impacts skin detection systems relying on color cues. It also introduced how the proposed method weakens this impact.

2. Chapter 2 sheds more light on the color space and the evolution of skin detection. It first introduces the widely-used color spaces and then covers three main types of methods in skin detection area.

3. Chapter 3 explains what color space augmentation is and how it works with U-Net.

4. Chapter 4 shows the details of our experiments. It first introduces the dataset and evaluation methods used in the experiments and then presents the experimental results.

5. Chapter 5 discuss the future work and usage of our methods.

Chapter 2

Literature Review

2.1 Color Space

A color space is a model to represent properties of visible light. There have been hundreds of color spaces, most of which are dedicated to a certain research field. Here we consider three popular color spaces.

2.1.1 RGB Color Space

The RGB color space (Figure 2.1) is very common in which pixel values of an image can be represented by the combination of three channels, namely red, green and blue. It is the most basic, commonly used, and hardware-oriented color space in image processing and is easier

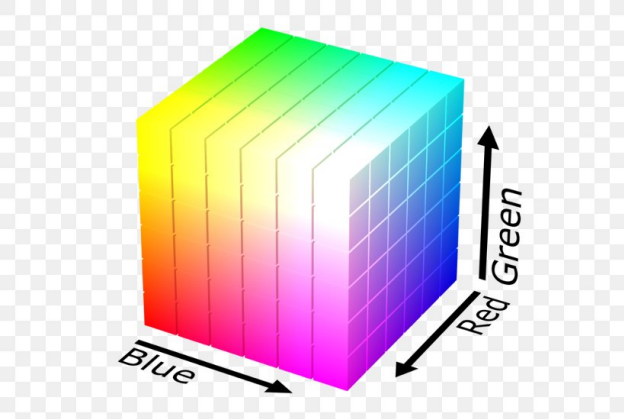


Figure 2.1: RGB color cube. Image credit: [1].

to understand. Any other colors can be described as (R, G, B) triples. For example, (255, 0, 0) refers to red and (255, 255, 255) refers to white.

Unfortunately, the RGB model has some disadvantages when it is used for image processing. One problem in the RGB model is that its three components are susceptible to brightness changes. As long as the brightness changes, the three components will change accordingly, which makes color tracking and analysis an extraordinary task. However, there is no more intuitive way to express this kind of alternative relation. In addition, the sensitivity of the human eyes to these three color components is different. Levine and Shefner [46] and Schwartz [67] have proved that human eyes are more sensitive to green light. If the similarity of colors is directly measured by Euclidean distance, the result will have a significant deviation from human vision. It is difficult for us to infer more accurate values of the three components for a particular color. In summary, the RGB color space may not be the best for the analysis in this thesis.

2.1.2 YCbCr Color Space

The YCbCr space is the color space recommended by the MPEG digital video coding standard. Y refers to the luminance component, Cb refers to the blue chrominance component, and Cr refers to the red chrominance component. The human eyes are more sensitive to the Y component of the video, so after sub-sampling the chrominance component to reduce the chrominance component, the naked eyes will not notice the change in image quality. It has become the most popular color space for analyzing human skin because we can directly analyze its color attributes in the video compression domain without decompression. The

transformation from RGB color space to YCbCr color space is shown below [57].

$$\begin{aligned} Y &= 0.299 \times R + 0.587 \times G + 0.114 \times B \\ Cb &= (R - Y) \times 0.713 + 128 \\ Cr &= (B - Y) \times 0.564 + 128 \end{aligned} \tag{2.1}$$

In the case of the YCbCr color space, compared with HSV, the conversion and effective separation of color and intensity information becomes easier. This color space is effective and efficient for separating image pixels according to colors in a color image. Therefore, the YCbCr color space can be applied to complex color images with uneven illumination [68].

2.1.3 HSV Color Space

HSV color space (Figure 2.2) is a cylindrical coordinate representing the points in Cartesian coordinates based on RGB color space using another set of three components: hue, saturation, and value. Hue is measured by angle, with a value range of 0-360 degrees, and represents color information or the position of the spectrum. Typically, 0 or 360 degrees represent the red color, 120 degrees represents the green color, and 240 degrees represents the blue color. Saturation, valued from 0 to 100, indicates the purity of the color. Generally speaking, it represents the amount of white mixed with pure hue color. The higher the saturation is, the smaller amount of white is mixed. The value determines the lightness and darkness of the color in HSV color space, ranging from 0 to 100. A higher value means a brighter color, and vice versa. The value channel will change when the images taken from the natural environment are affected by the illumination. However, the hue and saturation tend to change slightly [56]. So most image segmentation systems prefer HSV color space. Ganesan and Rajini [29] listed more reasons for using HSV color space in semantic segmen-

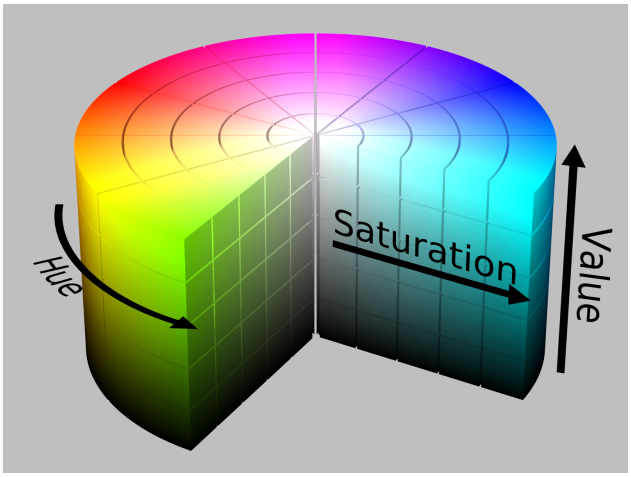


Figure 2.2: HSV color cylinder. Image credit: [83].

tation. The transformation equation from RGB color space to HSV color space are shown below [83].

$$
\begin{aligned}
M &= \min(R, G, B) \\
V &= \max(R, G, B) \\
H &= \begin{cases} 0, & \text{if } R = G = B \\ 60 \times \frac{G-B}{V-M}, & \text{if } V = R \\ 60 \times \frac{B-R}{V-M} + 120, & \text{if } V = G \\ 60 \times \frac{R-G}{V-M} + 240, & \text{if } V = B \end{cases} \\
S &= \begin{cases} 0, & \text{if } V = 0 \\ \frac{V-M}{V}, & \text{otherwise} \end{cases}
\end{aligned}
\tag{2.2}
$$

2.2 Skin Detection

Skin detection refers to the process of selecting pixel regions corresponding to human skin in an image or a video. It has become a popular technique for the wide range of applications in various areas such as face detection and recognition [45, 49], facial expression recognition, gesture recognition [3], Internet pornographic content filtering [27], and content-based image retrieval. Researchers have devoted significant effort to skin detection techniques. We would like to group the skin detection systems into three categories according to the different methods they use.

2.2.1 Threshold Based Methods

Threshold based skin detection methods are fast, straightforward, and not affected by changes in the shape of objects or viewpoints. It is an active research direction in the field of skin detection. Different from most background colors, the skin forms a small and compact cluster in the color space [41]. Angelopoulou [4] and Jablonski and Chaplin [40] have pointed out that individuals from different races, genders, and ages have good correlation and consistency of skin color. The use of color features for skin detection has a robust statistical and physical foundation.

The traditional threshold based skin detection methods always define the exact boundaries of skin pixels in some color space. They delineate skin areas by controlling the range of values for each component. For example, a set of widely-used thresholds for RGB color space was

defined by Kovac et al. [45] as the following:

$$\begin{aligned} &R > 95 \text{ and } G > 40 \text{ and } B > 40 \text{ and} \\ &\max\{R, G, B\} - \min\{R, G, B\} > 15 \text{ and} \\ &|R - G| > 15 \text{ and} \\ &R > G \text{ and } R > B \end{aligned} \tag{2.3}$$

Tsekeridou and Pitas [77] and Chai and Ngan [13] put forward the boundaries in HSV color space and YCbCr color space respectively:

$$\begin{cases} 0^\circ \leq H \leq 25^\circ \text{ and } 335^\circ \leq H \leq 360^\circ \\ 0.2 \leq S \leq 0.6 \text{ and } V \geq 0.4 \end{cases} \tag{2.4}$$

or

$$133 \leq Cr \leq 137 \text{ and } 77 \leq Cb \leq 127 \tag{2.5}$$

Despite these examples, many researchers have devoted themselves to skin detection via color space models and used different thresholds to distinguish skin and non-skin pixels [81, 84]. In addition, various color spaces have been combined for the same goal of skin detection. Garcia and Tziritas [30] proposed a novel scheme for detecting human faces in color images under unconstrained scene conditions (for example, the presence of complex backgrounds and uncontrolled lighting) using YCbCr and HSV color spaces. Considering the illumination, Kolkur et al. [44] proposed novel threshold values under the combination of RGB-HSV-YCbCr color spaces and achieved promising results.

Although these methods are very convenient and have achieved excellent results, their common shortcomings are indeed evident. There may be a clear overlap between skin pixels and

non-skin pixels in the color space. This similarity will lead the system to make misjudgments.

2.2.2 Traditional Machine Learning Methods

Machine learning aims at using certain algorithms to help the computer build an appropriate model with known data and use this model to make judgments about new situations. Traditional techniques refer to constructing the models in a statistical way that summarizes the relationship between data.

Histogram: One of the best ways to do data analysis is to plot the data. The histogram is a valuable way to demonstrate data graphically, counting each value that occurs in the data set and plotting the numbers out. Each vertical bar in the chart represents a unique value in the data set. The frequency distribution of the histogram is regarded as the probability distribution. A higher bin means a larger probability that it is a skin pixel. Pixels with higher probabilities than the designed threshold will be classified as skin pixels, otherwise set as non-skin pixels. Gomez [33] and Jones and Rehg [41] described probability calculation in RGB color space:

$$\begin{aligned} P(rgb|skin) &= \frac{Hist_{skin}[r,g,b]}{Total_{skin}[r,g,b]} \\ P(rgb|nonskin) &= \frac{Hist_{nonskin}[r,g,b]}{Total_{nonskin}[r,g,b]} \end{aligned} \tag{2.6}$$

A pixel which can be labeled as skin should satisfy a given threshold:

$$\frac{P(rgb|skin)}{P(rgb|nonskin)} \geq \theta \tag{2.7}$$

Since images perform differently in different color spaces, the drawing of the histogram and the parameters corresponding to the skin pixels may be different. Though the calculation

process is similar, the results may vary. Zarit et al. [86] has shown that histogram methods perform best in HSV color space since its intensity and chromaticity are separate. Based on this, Soriano et al. [70] solved the problem on normal illumination by adjusting the histogram tracking with skin locus. Liu and Peng [49] apply morphological and blob analysis for further optimization. Histograms are also used in combination with other techniques [11, 73, 78] and get better results.

Gaussian Mixture Model: The Gaussian distribution, sometimes called the normal distribution, is the most common form of distribution that exists in nature. The probability density function (PDF) of Gaussian distribution is as follows:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{2.8}$$

where μ is the mean and σ is the standard deviation.

The Gaussian mixture model (GMM) is a probability model that represent the K sub-distributions in the overall distribution. In other words, the mixed model represents the probability distribution of the observation data in the population. It is a mixture of K sub-distributions. The mixture model does not require the observation data to provide information about the sub-distribution to calculate the probability of the observation data in the overall distribution. The probability distribution of the Gaussian mixture model is:

$$p(x|\theta) = \sum_{k=1}^{K} \alpha_k \phi(x|\theta_k) \tag{2.9}$$

where α_k is the weight of the k^{th} model, and $\phi(x|\theta_k)$ is the Gaussian distribution density function of the k^{th} sub-model. Its expanded form is the same as the single Gaussian model introduced in (2.8). The iterative method of expectation maximization (EM) is usually used to estimate the fitting parameters during the training of GMM. The GMM model has the

characteristics of a simple evaluation process and low memory cost, which has attracted much attention. Zhu et al. [88] proposed a new GMM-based skin color segmentation algorithm. Experiments show that the algorithm can effectively and quickly segment the moving target and has strong robustness. Hossain et al. [37] proposed a method using GMM to extract facial skin color by estimating the illumination changes in images. The experimental results show that this method is suitable for practical applications compared with the traditional GMM-based skin color segmentation method.

However, the calculation process of GMM is long, and the algorithm has limitations. The comparative study of the performance of Gaussian distribution by Caetano et al. [12] shows that the hybrid model only improves the performance in the relevant operating area, and increasing the number of cores does not effectively improve the model efficiency.

Bayesian Classifier: Among various classifiers, the Bayesian classifier has the most negligible probability of classification error and the slightest average risk in the case of a predetermined cost. It uses the Bayesian formula to calculate the posterior probabilities of an object that belongs to different classes. Then, it classify the object to the class with the highest posterior probability. The conditional probability density function of an input to be classified as skin can be expressed as the following Bayes rules:

$$p(skin|x) = \frac{p(x|skin)p(skin)}{p(x|skin)p(skin) + p(x|non\text{-}skin)p(non\text{-}skin)} \tag{2.10}$$

On the basis of this, a pixel will be classified as skin pixel if its conditional probability is larger than the threshold θ.

$$\frac{p(skin|x)}{p(non\text{-}skin|x)} > \theta \tag{2.11}$$

Brand and Mason [8] evaluated the performance of three pixel-level human skin color detec-

tion methods, in which the Bayesian detection rate is as high as 95%. It proves the accuracy of the Bayesian classifier in estimating the skin color probability. Experiments [61] show that the accuracy of the Bayes classifier decreases sharply when the training set becomes smaller.

2.2.3 Neural Network Methods

A neural network is a mathematical or calculation model that imitates the structure and function of a biological neural network (animal's central nervous system, especially the brain). More recently, researchers have applied deep neural network to the problem of skin segmentation.

One outstanding neural network was Fully Convolutional Network (FCN), introduced by Long et al. [50] in 2015, which is a pioneering network model using deep learning for image segmentation and is one of the earliest semantic segmentation models. It replaced the final full-connected layers of the classic CNN classification model with convolutional layers. The training parameters were reduced but the accuracy was not lost. Subsequent CNN-based image segmentation models are mostly improved based on it. Zuo et al. [90] introduced an new end-to-end network for human skin detection by integrating the recurrent neural network (RNN) into FCN and enhanced the skin detection power under complex environments. He et al. [36] proposed a new data-driven skin detection method to achieve robust skin detection combined with human body mask and achieve semi-supervised training address the problem of insufficient samples.

Another well-known network is Regions with Convolutional Neural Network (R-CNN). Girshick et al. [32] designed the R-CNN framework by using region proposal and CNN instead of sliding window and hand-designed features in traditional target detection, which made a

huge breakthrough in target detection. The next year, Girshick [31] improved the R-CNN by adding a ROI pooling layer and applying multi-task to overcome its serious speed bottleneck caused by the repeated calculations when the computer performs feature extraction on all regions. While it still took plenty of time for the network to do selective search. Ren et al. [60] then designed region proposal network for extracting candidate regions to take the place of time-consuming selective search, which greatly improved the detection speed. He et al. [35] extended the Faster-RCNN by adding a new branch for prediction the object mask on the bounding box recognition branch in parallel and designed Mask R-CNN. It runs more efficiently and can also be easily generalized to other tasks. Benefit from the progress in R-CNN family, skin detection systems also develop quickly. Roy et al. [63] used an R-CNN-based approach to reduce the number of false positives by adding a CNN based skin detector and shown a substantial improvement over a baseline of using R-CNN only. Nguyen et al. [54] integrated a mean shift tracker into Mask R-CNN and reported improvements of 5% to 9% in detection accuracy compared by the original Mask R-CNN based method.

Refering to the FCN structure, researchers also built a set of encoder-decoder models. Segnet [6] is a typical encoder-decoder based segmentation network. It is a symmetric network composed of encoder and decoder. The encoder classifies and analyzes the low-level local pixel values of the image to obtain high-level semantic information. The decoder up-samples the reduced feature image and then performs convolution processing on the up-sampled images to make up for the loss of detail caused by the shrinking of the object by the pooling layers in the encoder. The encoder and decoder work together to make the prediction. Nguyen et al. [55] modified the original SegNet architecture by increasing the number of decoders, thereby allowing each encoder to perform multiple tasks at the same time which discriminate skin components in hand area more accurately. U-Net is another typical encoder-decoder based segmentation network and was introduced by Ronneberger et al. [62]. It includes a

contracting path to capture context and a symmetrical expanding path to restore spatial resolution. It use a series of cross-layer connections to fuse low-level detailed information and high-level semantic feature to achieve precise positioning. Topiwala et al. [76] has shown that U-Net stands out among the frequently-used skin detectors on their dataset of human abdomen with different skin colors. Tarasiewicz et al. [74] refined the U-Net architecture [62] by considering large-scale contextual features, using inception blocks and dense blocks to reduce occurrences of false positives significantly while doing skin detection.

2.3 Algorithmic bias

This work has been motivated in part by the need to promote demographic fairness in automated systems, particularly relating to differences in skin tones that result from ethnicity, race, gender, age, and other respects. For tasks such as face recognition, techniques have been developed recently to evaluate bias within algorithms and datasets [9], and to improve fairness with respect to such differences [23]. To eliminate the bias, Steiner et al. [71] proposed a face verification method based on the spectral remission intensities of the different skin tones in the short-wave infrared (SWIR) range. They can reliably distinguish real human skin from other materials, regardless of skin type. More generally, Mehrabi et al. [52] have surveyed the Machine Learning field and have developed a taxonomy of problems that affect bias and fairness within automated systems. Most bias-mitigation systems focus on two types of biases: dataset bias, and task bias. This paper is concerned with the former, which refers to datasets having classes are not represented as well as others within the dataset. Researchers recently have focused on invariant feature learning for protected variables (here, the skin color appearance), and perform database repair to eliminate the representation error [2, 47, 65].

Chapter 3

Methods

In this chapter, we talk about our approach to solving skin detection problems as introduced in Chapter 1.1. We propose a deep learning module that contains a unique data preprocessing called color space augmentation, followed by a traditional neural network to do skin segmentation and output the final results. Our system is based on the U-Net [62] architecture and stands out among the frequently-used neural networks on skin detection tasks [76]. We have used the U-Net architecture to demonstrate the novel skin-detection approach presented here.

3.1 System Architecture

The architecture of our proposed method is shown in Figure 3.1. The design is quite similar to the traditional skin detection system with the training stage and testing stage. Before the training stage, we added a new stage called color space augmentation to enhance the classification ability of the system under different illumination conditions. We use images with different HSV values to simulate colored lighting and occlusions on human skin during the testing stage. In this work, the U-Net structure is identical to the one introduced in [62]

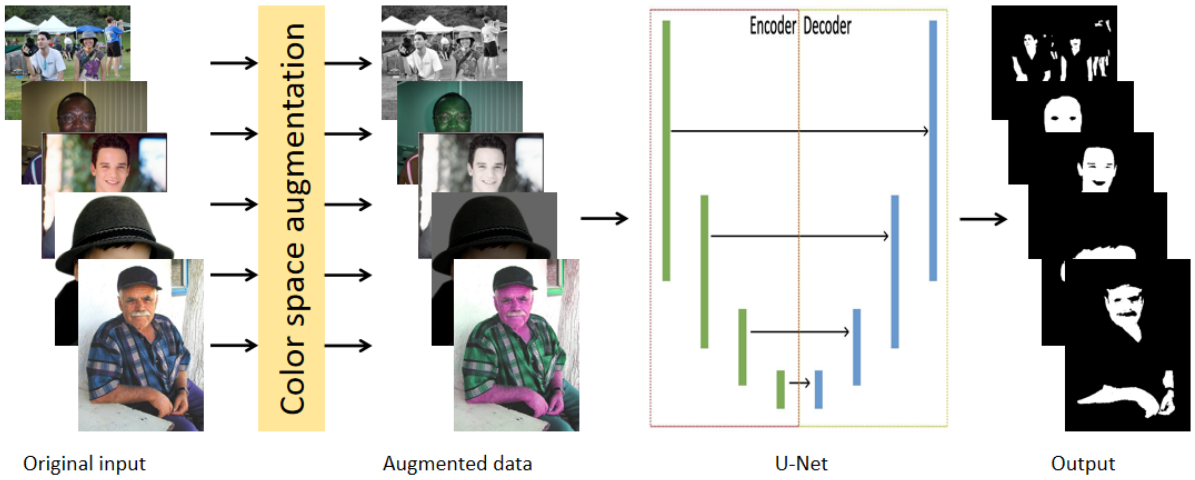


Figure 3.1: The entire process of our method. The training set will be color augmented and then trained by U-Net with its corresponding groundtruth. The output will be the binary images where skin pixels are denoted as 1, and non-skin pixels are denoted as 0.

3.1.1 Color Space Augmentation

Data augmentation has been widely used in computer vision, especially for the purpose of regularizing the network [53, 82]. The augmentation techniques are also useful for reducing generalization error [53, 75], which includes noise augmentation [19], color augmentation [75], pose augmentation, etc. In this work we adopt the color based data augmentations that can create artificial images to mimic alternate representations of the image, in this case, the appearance of the skin. We have implemented the color augmentation in the HSV (Hue, Saturation, Value) space, as illustrated in Figure 3.2.

Suppose the dataset has the training set D and its corresponding ground truth G. For every image $\mathcal{I}$ in the training set D, $\{H, S, V\}$ represents its original hue, saturation and value. We choose a new set of hue ($\mathcal{H} = \{H_1, H_2, ..., H_n\}$), saturation ($\mathcal{S} = \{S_1, S_2, ..., S_m\}$) and value ($\mathcal{V} = \{V_1, V_2, ..., V_l\}$) across its respective range and then transform image $\mathcal{I}$ in the training set and validation set. So the transformed image $\mathcal{I}_{new} = \mathcal{T}(\mathcal{I}, \mathcal{H}, \mathcal{S}, \mathcal{V})$. As an example, if only the hue channel is changed, then the new image is represented as:

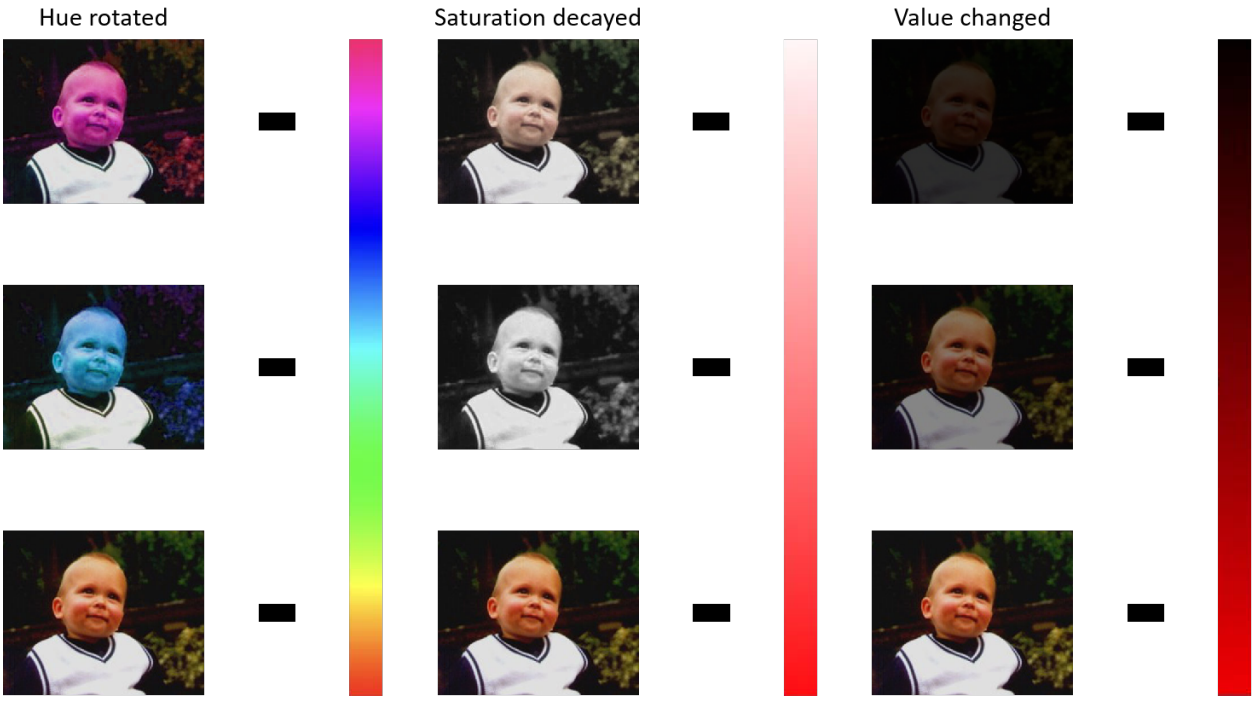


Figure 3.2: Color space augmentation in HSV space containing hue rotation, saturation decay and value change.

$\mathcal{I}^i_{new} = \{H_i, S, V\}$ and formatting a new set of training images $D_h = \{D^1_h, D^2_h, ..., D^n_h\}$. In saturation and value channel the images are changed in the way $\mathcal{I}^j_{new} = \{H, S_j, V\}$, $\mathcal{I}^k_{new} = \{H, S, V_k\}$ and we get another two new sets of training images $D_s = \{D^1_s, D^2_s, ..., D^m_s\}$,$D_v = \{D^1_v, D^2_v, ..., D^l_v\}$. In this way, one training image will become multiple training images which enlarges the training set and provides more features with skin areas covered by different masks and improve the performance of the skin detection model. Then we send the training set $\{D, D^1_h, ..., D^n_h, D^1_s, ..., D^m_s, D^1_v, ..., D^l_v, G\}$ to neural network in the training stage. The images before feeding to the network are again transformed to RGB color space. Note that each image in the dataset generates $m + n + l$ additional images. Figure 3.3 shows example of the augmentation. We have used $m = n = l = 5$ model.

(a) (b) (c)

Figure 3.3: Example of color based augmentation across *hue* (a), *saturation* (b), and *value* (c). The first column of each group shows the changed images $\mathcal{I}_{new}$. The second columns show the skin segmentation results without color space augmentation. The third columns show the results with color space augmentation. The input images are rotated at every 60 degrees in hue channel in group (a). For group (b), the saturation of images are decayed at the ratio of 0.8, 0.6, 0.4, 0.2, and 0.0. For group (c), the value of the images are changed at the ratio of 1.0, 0.8, 0.6, 0.4, and 0.2.. The normal image is on the top of each group.

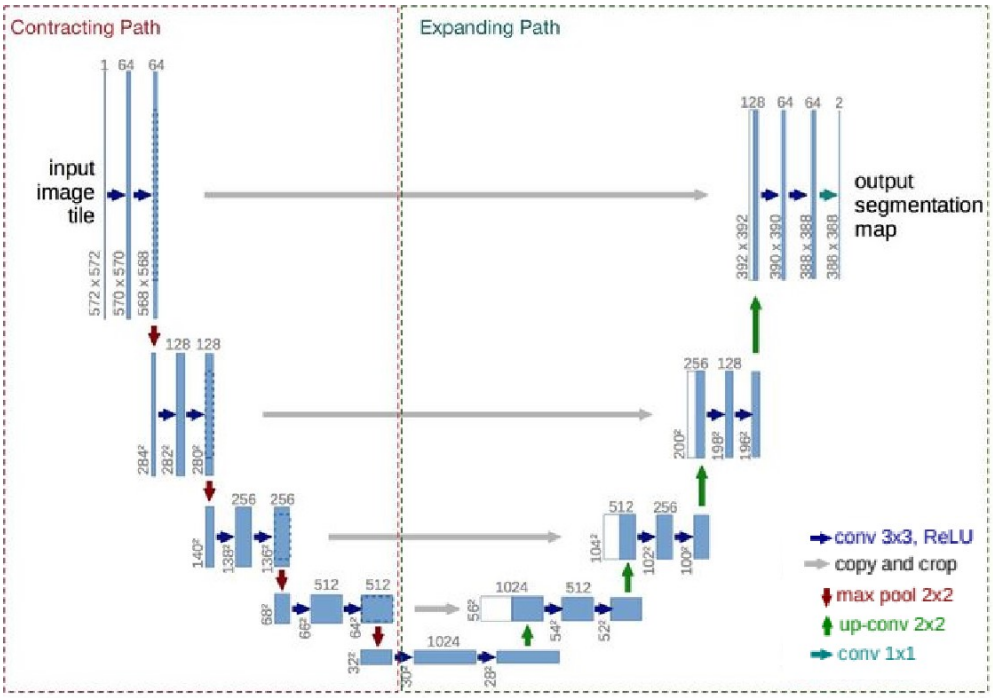


Figure 3.4: An example of the structure of U-Net as shown in [62].

3.1.2 Network Structure

We use a traditional U-Net structure to carry on the training and testing stage in our experiments. The U-shaped structure of the network is shown in Figure 3.4. The network is a classic fully convolutional network (that is, there are no fully connected operations in the network). The left side of the network (red dotted line) is composed of convolution and max-pooling layers. A series of downsampling operations are formed, and this part is called the contracting path in the paper. The contracting path consists of four blocks. Each block uses three effective convolutions and one max-pooling downsampling. After each downsampling, the number of feature maps is multiplied by two, as shown in the figure. Finally, a feature map of size 32× 32 is obtained.

The right part of the network (the green dotted line) is called the expanding path in the paper. It is also composed of four blocks. Before the start of each block, the size of the

feature map is multiplied by two through deconvolution, and the number is halved (the last layer is slightly different). Then the expanding path is merged with the symmetrical contracting path on the left. Since the size of the feature map of the contracting path on the left and the expanding path on the right are different, U-Net normalizes them by clipping the feature maps of the contracting path to the exact size of the feature maps in the expanding path. The convolution operation of the extended path still uses the effective convolution operation, and the size of the finally obtained feature map is 388×388. Since this task is a two-class task, the network has two output feature maps.

Chapter 4

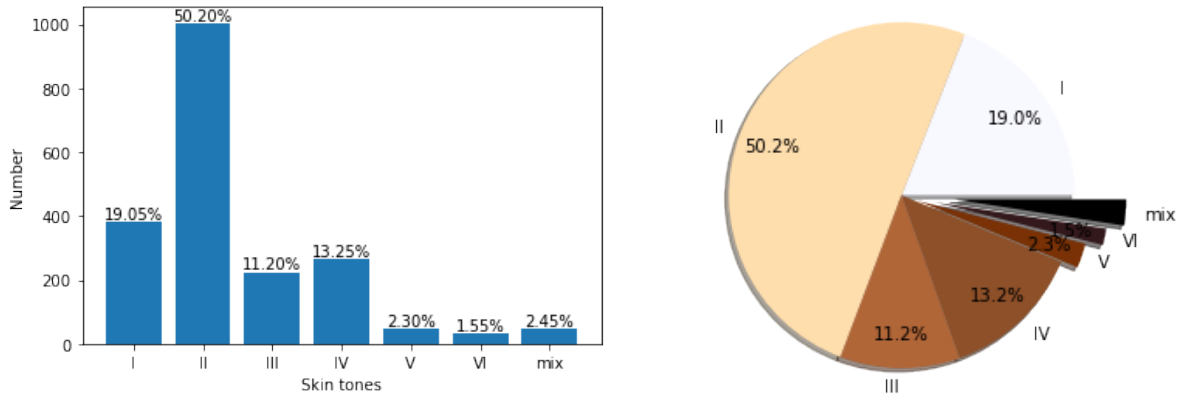
Experiments and Results

4.1 Dataset

In this work, we have used three datasets. The training and initial evaluation were performed using the benchmark ECU [58] dataset. It contains images with diverse attributes including gender, age, skin type, skin-like background, indoor and outdoor images, and images with shadows. The dataset contains 4000 RGB images with manually annotated skin pixels as binary images (see Figure 4.1 as an example). These images are divided into 1600 images for training, 400 for validation, and 2000 for testing. Note that each of these images is used for color space augmentation (18 total). Hence, we have a total of 30400 images for training. To demonstrate the color invariance of the algorithm, we also transformed the test images to augmented

Racial bias is another critical attribute in skin detection systems. To evaluate such bias in our system, we have experimented with six skin types following the Fitzpatrick scale [64] and one 'mix' group. The ECU dataset was divided into 7 groups manually according to the skin tone shown in each image. Images with multiple skin tones will be divided into the 'mix' group. Figure 4.2 shows the distribution of images containing individuals with skin types of Type I (less melanin concentration) to Type VI (high concentration of melanin). The figure clearly shows the representation bias of the ECU dataset that can bias any ML algorithm to Type I and Type II.

Figure 4.1: Results from the ECU dataset. The first row shows the original images. The second row shows the corresponding ground truth containing skin and non-skin pixels. The third row shows output from the proposed algorithm. Results shown for different backgrounds, different numbers of people, various genders, skin tones and gestures.



(a) (b)

Figure 4.2: Distribution of skin types in the ECU dataset. Labels I-VI refer to the six skin tones described by Fitzpatrick [26]. The group "mix" refers to several skin tone categories in a single image.

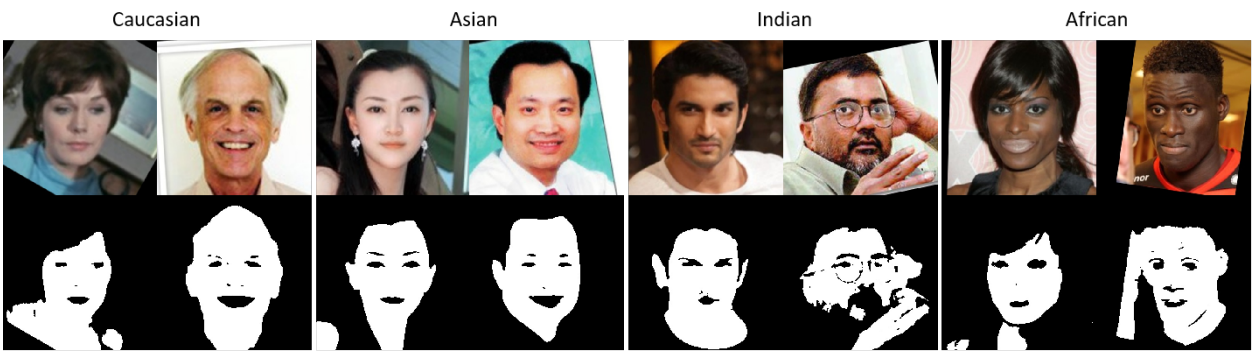


Figure 4.3: Results from the RFW dataset. Faces are divided into four groups: African, Asian, Caucasian and Indian. The first row shows the original images. The second row shows the skin detection output results of our methods.

For further evaluation, we selected the RFW (Racial Faces in the Wild) dataset [79] for *cross dataset validation* of our algorithm in order to test whether the proposed algorithm exhibits bias for ethnicity or skin tones. RFW is a standard test database used to study racial bias in face recognition (see Figure 4.3 as an example). Four test subsets are provided, namely, Caucasians, Asians, Indians, and Africans. Each subset contains about 3000 people and 6000 image pairs for face verification. For our work, RFW dataset provided 10415 African faces, 9688 Asian faces, 10196 Caucasian faces, and 10308 Indian faces, which shows a good balance between skin tones.

Finally, we created a small dataset of 20 pictures with extreme illumination variations. These images were selected to contain either colored neon illumination or artificial filters, as shown in Figure 4.12. Then we performed manual annotation using SuperAnnotate [72] to identify the skin pixels and non-skin pixels. We conducted extensive testing using this dataset and made pixel-wise evaluations using our ground-truth annotations.

Table 4.1: Confusion matrix.

	Predicted as skin	Predicted as non-skin
Ground truth Skin	TP	FN
Ground truth Non-skin	FP	TN

4.2 Evaluation

For the ECU dataset, we used five measures to evaluate the performance: precision, recall, accuracy, F1 score, and IoU. The confusion matrix (Table 4.1) is a situation analysis table that summarizes the prediction results of the classification model in data science, data analysis, and machine learning. In the skin detection problem, there are two types of records in the data set, positive and negative category, so the classification model may make positive judgments (judgment records belong to the positive category) or negative judgments (judgment records belong to the negative category). Therefore, the confusion matrix is a 2×2 situation analysis table that shows the number of the following four sets of records: positive records with correct judgments (true positives), positive records with wrong judgments (false negatives), negative records with correct judgments (true negatives) and negative records with false judgments (false positives). Table 4.1 shows the structure of the confusion matrix. The classification model correctly classified the true-positive and true-negative record groups on the diagonal of the confusion matrix and misjudged the false-positive and false-negative record groups on the anti-diagonal line. A confusion matrix is an essential tool for the performance evaluation of classification models. Various evaluation indicators such as accuracy, precision, and recall can be calculated from the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

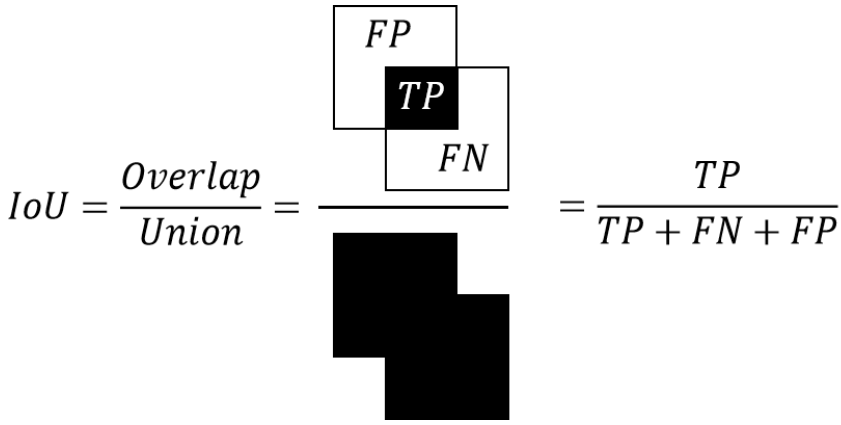


Figure 4.4: The calculation of IoU. It is also can be calculated with the help of confusion matrix.

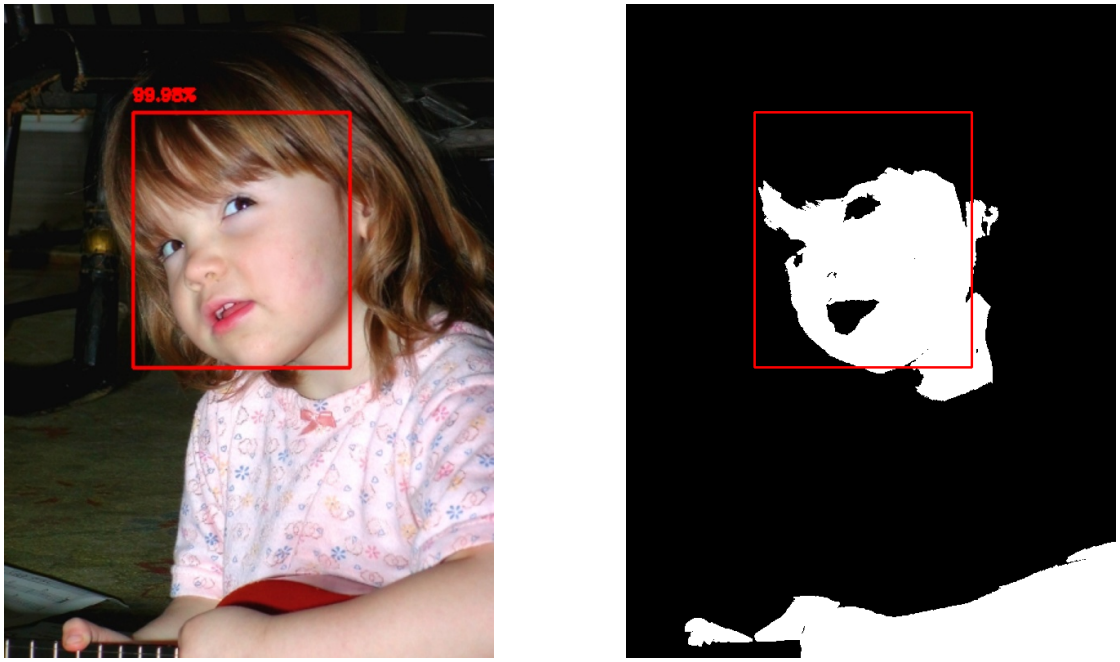
$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 = \frac{Precision \times Recall}{Precision + Recall} \tag{4.4}$$

Intersection over Union is another standard for measuring the accuracy of detecting corresponding objects in a specific data set. It equals dividing the overlapping part of the two regions by the combined part of the two regions. This standard is used to measure the correlation between reality and prediction. The higher the correlation, the higher the value. Figure 4.4 shows how IoU is calculated.

For the RCW dataset, we do not have the ground truth skin annotation. Hence we developed a different method to evaluate the performance. We first used a face detector to get the face area in the image (as shown in Figure 4.5a). Then we run the skin detection algorithm trained on the ECU dataset to get the prediction. Finally, we identify the number of skin pixels in the face boundary (as shown in Figure 4.5b) as an evaluation tool.

$$Skin/Face = \frac{Skin\ pixels\ detected}{Total\ pixels\ in\ face\ rectangle} \tag{4.5}$$



(a) (b)

Figure 4.5: One example from ECU dataset of skin/face evaluation. In first image, we use face detector to detect face boundary. In second image, we move the face box into the results from the skin detector to calculate the skin/face and shows the performance of the skin detection system.

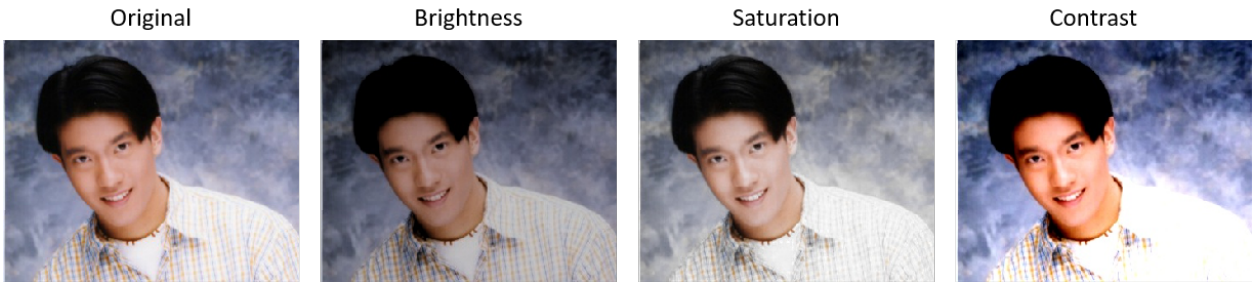


Figure 4.6: An example from ECU dataset of colorjitter. The brightness, saturation, and contrast of the image are changed randomly.

4.3 Training segmentation networks

The experiment uses NVIDIA GeForce RTX 2070 SUPER GPU with 16 GB GPU memory. The algorithm is trained and tested with U-Net [87] and FCN [20]. The U-Net is working under Python 3.8.5 and Tensorflow 2.3.1 environment with no pre-trained models. The FCN is working under Python 3.8.5 and Pytorch 1.70 with a pre-trained VGG-16 network. We train the networks with a fixed learning rate of 10^{-4}, and each epoch takes around 79s when batch size is set to 8. The network will use a module ImageDataGenerator [43] in the Keras to do geometric data augmentation, which includes image rotation, width shift, height shift, shear, zoom, and horizontal flip with nearest fill mode. We used binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(f(y_i)) + (1-y_i)\log(1-f(y_i))\right] \tag{4.6}$$

where N is the number of segmentation classes. The symbol y_i is the label and $f(y_i)$ is the predicted probability of the points belonging to the i^{th} class. The original output of the network will be from 0 to 1. Since the pixels should belong to either skin category or non-skin category, we use the function below to make the output $\mathcal{O}$ binary where 1 refers to skin pixels. The threshold δ we set is 0.5. To make the experiments more convincing, we

also draw precision-recall curve.

$$\mathcal{O} = \begin{cases} 1 & \text{if } f(y_i) \geq \delta \\ 0 & \text{if } f(y_i) < \delta \end{cases} \tag{4.7}$$

Despite normal augmentation methods, we also apply colorjitter as a kind of color augmentation to make a comparison. Figure 4.6 shows an example of colorjitter. Colorjitter augment the training set by randomly change the brightness, saturation, and contrast of training images. For a given image I, its brightness, contrast, and saturation will be jittered uniformly from $[\max(0, 1-\theta), 1+\theta]$. In the following experiments, we choose $\theta = 0.5$ when using colorjitter, and each image will be augmented into 4 images using colorjitter.

4.4 Results

4.4.1 Images in the wild

We compared our method with some state-of-the-art skin segmentation systems, including three traditional methods and one FCN based method. Kolkur et al. [44] and Dahmani et al. [18] are two thresholding methods which establish some rules in several color spaces to classify a pixel is skin or not. Jones and Rehg [41] is a naive Bayes based method, which predicts the probability of a pixel to be skin after training with given skin masks. The problem behind these traditional methods is the lack of high level features during detection tasks, resulting in the weak robustness against light changes, complex backgrounds or skin color diversity. For both FCN based methods and our U-Net based method, we trained three models, one without color augmentation, one with colorjitter and another with color augmentation to confirm the effectiveness of color augmentation.

Table 4.2: Test results for several skin segmentation methods with the ECU dataset. Our results using U-Net are significantly better than previous methods. For both FCN and U-Net, our use of color-based augmentation improved overall performance of the system.

Methods	Accuracy	Precision	Recall	F1	IoU
Kolkur et al. [44]	83.73	57.00	88.38	69.31	53.03
Dahmani et al. [18]	85.95	63.12	77.91	69.74	53.54
Jones and Rehg [41]	89.51	78.23	68.58	73.09	57.59
FCN before aug.	95.78	92.32	86.93	88.66	79.63
FCN colorjitter	**96.40**	91.50	**91.12**	**91.31**	**84.01**
FCN after aug.	95.89	92.14	87.70	89.87	81.60
U-Net before aug.	95.59	89.56	89.15	89.35	80.76
U-Net colorjitter	94.99	82.95	95.56	88.81	79.87
U-Net after aug.	96.33	**92.99**	89.04	90.97	83.44

We first trained the U-Net model with the original RGB images in the ECU dataset (with and without augmentation) and evaluated the performance with the original test set. The precision and recall are shown in Table 4.2. With augmentation, this system yielded a precision of 92.99% and recall of 89.04%, which significantly outperforms the methods of Kolkur et al. [44], Dahmani et al. [18], and Jones and Rehg [41] (Naive Bayes). The FCN model achieves a precision of 92.14% and a recall of 87.70%. For this part of experiment, models with colorjitter reach the highest accuracy.

In Figure 4.7 we show qualitative comparisons, where the examples cover various skin colors, similar colors in the background, and complex illumination. The first row is a girl lying on the grass with her arms open. The second row is a boy holding a cat, and some of his skin areas are covered by shadow and the cat. The third row is an image of three children with dark skin. The fourth row is a man with a large beard area on his face. These challenging conditions make other methods fail or perform poorly. Dahmani et al. [18] and Jones and Rehg [41] fail in the first and third rows. Kolkur et al. [44] classified a large area of background as skin pixels in the fourth row. U-Net (B) works better but still performs poorly in the third row. In contrast, our approach overcame most of the difficulties as stated above, and

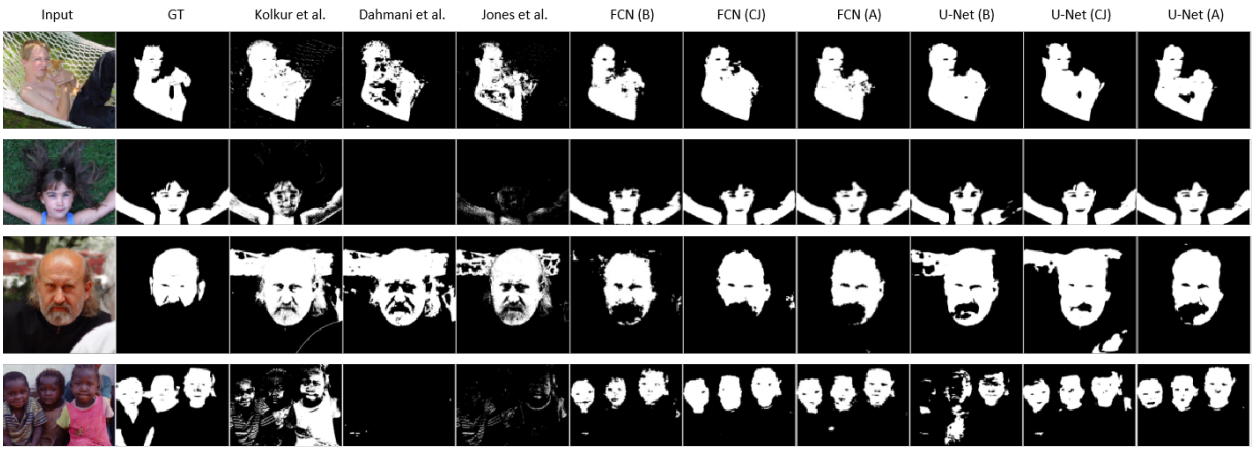


Figure 4.7: Testing results on the ECU dataset, by various skin segmentation methods including Kolkur et al. [44], Dahmani et al. [18], Jones and Rehg [41], FCN before augmentation (B), using colorjitter (CJ), and after (A) augmentation, and U-Net before augmentation (B), using colorjitter (CJ), and after (A) augmentation (Columns 3 to 11). Input and ground truth are shown in columns 1 and 2. Our approaches (marked by "(A)") achieve superior results for different backgrounds, genders, poses, and skin tones.

produced accurate and robust results.

We have set the threshold of predicted probability to be 0.5 to classify the image pixels as skin or non-skin pixels and convert the sigmoid output from the deep learning model to binary. To make the experimental testing results more convincing, we illustrate the precision-recall curve [59] in Figure 4.8.

4.4.2 Across races and skin tones

In order to detect the skin tone bias in the ECU dataset, we further tested the algorithms on different skin tones. Table 4.3 and Table 4.4 show the testing results on different skin tones in the ECU dataset. Among the three baseline methods, Jones [41] shows the best performance for most of the skin types, but all the methods particularly fail for Type VI (dark skin category). Even deep learning skin segmentation methods show an apparent decline in this dark skin category. This also happens to models with colorjitter, which reach

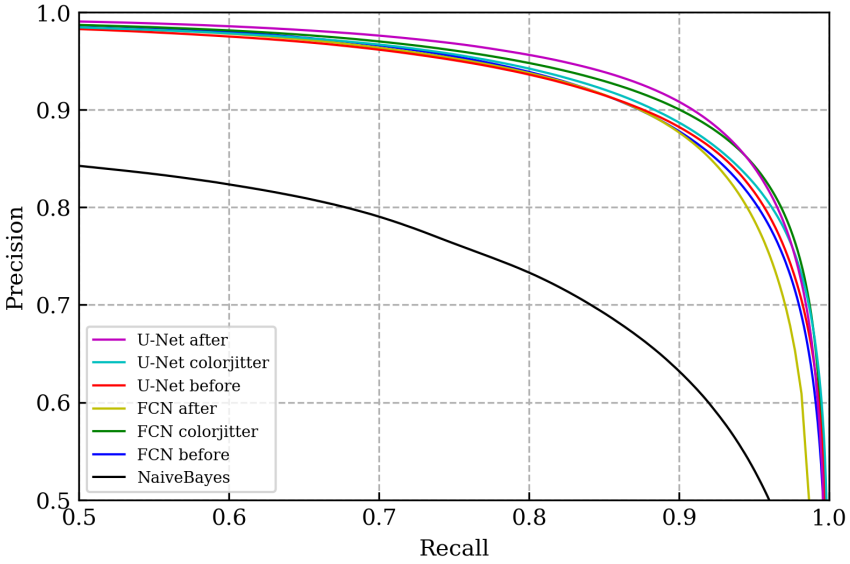


Figure 4.8: Precision-recall curve from testing experiments on the ECU dataset.

the highest F1 and IoU shown in the tables. Our method consistently shows around 90% F1 score and more than 75% IoU for all skin types, which shows its robustness against skin tone changes.

Moreover, the standard derivations in the last column show that deep learning models have more substantial stability over skin tone bias after color augmentation. The standard deviation demonstrates how consistent an algorithm is across the different skin categories. An ideal system should have same IOU and F1 score for every skin category; hence, very little variation (low std). These two tables show that our approach was more effective than all previous systems in reducing bias across skin types. Although models with colorjitter seem to reach the higher F1 and IoU, they appear to have larger standard deviation values, which means models using colorjitter has larger skin tones bias.

As shown in Figure 4.2, one of the significant problems in the ECU dataset is the imbalance in the images for each skin type. To further understand the robustness of our method, we

Table 4.3: F1 scores (%) for the ECU dataset across different skin types. The labels I-VI refer to the six skin tones described by Fitzpatrick [26]. The "mix" column refers to single images containing several individuals with multiple skin categories. The σ column refers to standard derivation of the F1 scores for all columns.

Methods	I	II	III	IV	V	VI	mix	σ
Kolkur et al. [44]	67.61	69.96	70.27	70.44	67.61	46.90	72.42	8.14
Dahmani et al. [18]	66.10	70.52	71.95	71.01	70.46	56.45	70.45	5.07
Jones and Rehg [41]	64.65	75.89	73.99	74.00	73.28	46.82	77.61	9.99
FCN before aug.	89.03	89.90	90.03	89.56	89.59	83.41	87.37	2.20
FCN colorjitter	**91.21**	**91.49**	**91.60**	**91.36**	91.97	85.86	89.32	2.02
FCN after aug.	90.06	90.06	90.34	89.93	90.06	82.98	85.69	2.70
U-Net before aug.	87.16	89.58	90.38	90.99	**91.98**	84.72	88.82	2.29
U-Net colorjitter	88.96	88.86	88.69	88.56	91.42	81.77	89.58	2.80
U-Net after aug.	90.88	91.34	91.21	90.55	89.35	**86.05**	**89.60**	**1.84**

Table 4.4: IoU values (%) for the ECU dataset across different skin types. The column labels are the same as in the previous table.

Methods	I	II	III	IV	V	VI	mix	σ
Kolkur et al. [44]	51.07	53.80	54.16	54.36	51.07	30.64	56.76	8.22
Dahmani et al. [18]	49.37	54.47	56.19	55.05	54.39	39.33	54.38	5.50
Jones and Rehg [41]	47.77	61.15	58.72	58.72	57.83	30.56	63.41	10.60
FCN before aug.	80.24	81.66	81.87	81.09	81.15	71.54	77.57	3.44
FCN colorjitter	**83.84**	**84.32**	**84.50**	**84.09**	85.14	75.22	80.70	3.27
FCN after aug.	81.92	81.91	82.38	81.70	81.91	70.92	74.97	4.22
U-Net before aug.	77.24	81.13	82.44	83.47	**85.16**	73.50	79.89	3.68
U-Net colorjitter	80.12	79.95	79.68	79.47	84.20	69.16	81.13	4.33
U-Net after aug.	83.28	84.06	83.84	82.73	80.75	**75.52**	**81.16**	**2.97**

Table 4.5: Skin/face ratios (%) for the four race groups with different skin tones of the RFW dataset respectively. The last 'Overall' column means data are collected from the whole RFW dataset.

Methods	Caucasian	Asian	Indian	African	Overall
Kolkur et al. [44]	62.34	62.21	64.31	36.78	56.27
Dahmani et al. [18]	60.02	59.19	60.95	49.79	57.44
Jones and Rehg [41]	47.29	45.49	48.45	20.05	40.19
FCN before aug.	67.24	65.78	67.36	64.37	66.20
FCN colorjitter	73.88	70.16	72.80	74.92	73.00
FCN after aug.	73.35	70.59	72.73	73.99	72.71
U-Net before aug.	65.47	65.76	69.82	68.39	67.40
U-Net colorjitter	**82.80**	**78.52**	**81.63**	**84.15**	**81.84**
U-Net after aug.	75.07	71.85	73.49	72.98	73.37

test with the RFW dataset, which has a balanced dataset across ethnicity. Since RFW dataset does contain manual labels, we only compute the skin/face ratio in this part (see Figure 4.5 and Function 4.5). Table 4.5 shows the evaluated skin/face result of RFW dataset. The model is trained on the ECU dataset using data augmentation. The results show that other methods have different degrees of decline in the "African" group, while our method is stable among races with different skin tones. Also, our method is significantly better in all the categories compared to the three baselines. Compared the results shown in the last four rows, color augmentation shows its effectiveness on improving the performance of the models.

We have noticed that U-Net trained with colorjitter has extremely high skin/face values. Although this skin/face ratio can reflect the ability of the detection system, larger values do not mean better prediction entirely. Skin segmentation systems may make over prediction and conduct more false positive pixels, which will increase the skin/face values but have no contribution to accuracy. In Figure 4.9, we extract the skin/face values from various groups of predictions to make it more convincing. In this section, we plot skin/face ratio curve to evaluate the performance of the skin segmentation models. The skin/face ratio curve refers

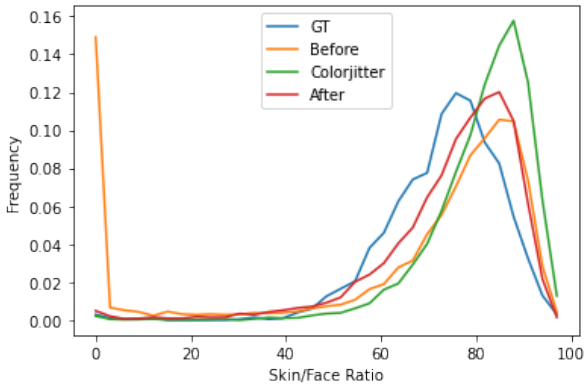
Table 4.6: Kullback–Leibler divergence between the standard probability distribution and that from estimated methods. Results are from the three U-Net models with different groups and the whole RFW dataset.

	Cau	Asian	Ind	Afr	Overall
Before Aug.	2.19	1.84	1.87	1.85	3.27
Colorjitter	1.98	1.67	1.89	2.31	3.20
After Aug.	1.70	1.59	1.68	1.83	3.03

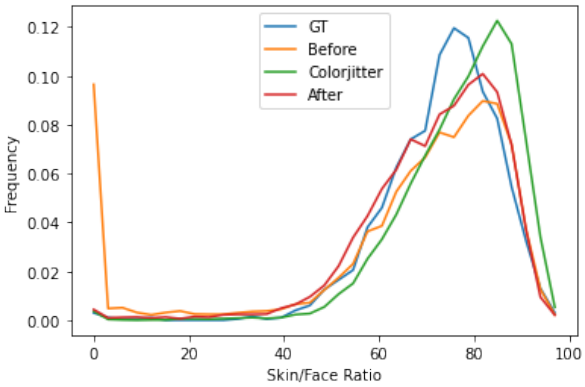
to the probability distribution of the skin/face values from the results.

First, we plot skin/face ratio curve using the annotated ECU dataset and its corresponding ground truth, which will be regarded as a sample or a standard. Then, we plot the exact curve of the results from U-Net before and after color augmentation with RFW dataset. The curves are shown in Figure 4.9. We calculate Kullback–Leibler divergence (D_{KL}) to measure the difference between the standard probability distribution and that from estimated methods. We expect the resulting curve from a better model to be more relevant to the standard curve, that is, has smaller D_{KL} to the standard distribution. The D_{KL} values are listed in Table 4.6. It demonstrates that model after color augmentation is more relevant to the standard distribution in all groups since they have the smaller D_{KL}. This also happens in the whole RFW dataset. Although U-Net has higher skin/face values as listed in Table 4.5, it has lower D_{KL}. Combined the distribution curves in Figure 4.9, we conclude that U-Net trained with colorjitter conducted more over prediction pixels when tested on RFW dataset, which means it has lower accuracy than our proposed method in fact.

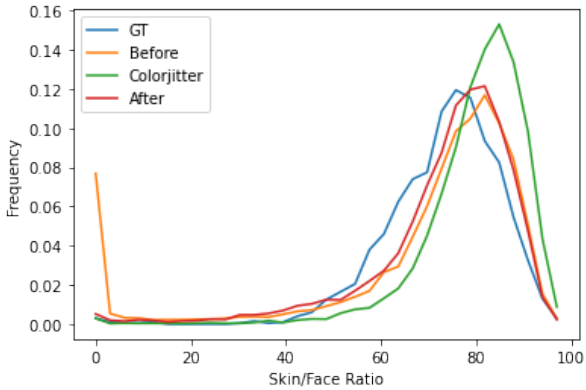
From Figure 4.9, we find that there is a peak at point "0" for the model before augmentation in every group, which does not appear in the standard curve. These peaks indicate that the model does not detect any skin pixels from the face area, which is incorrect. After color augmentation, the model works well, and these peaks disappear. This explains a lot why skin/face ratio increase after color augmentation.



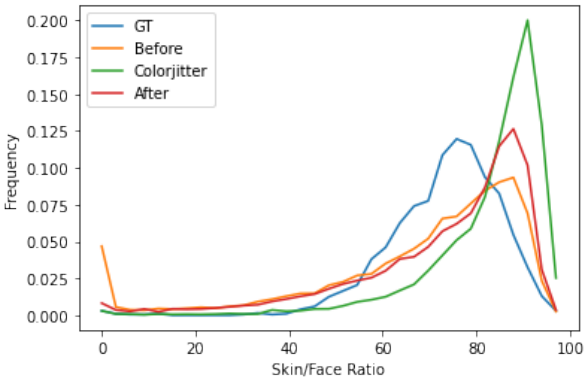
(a) Caucasian



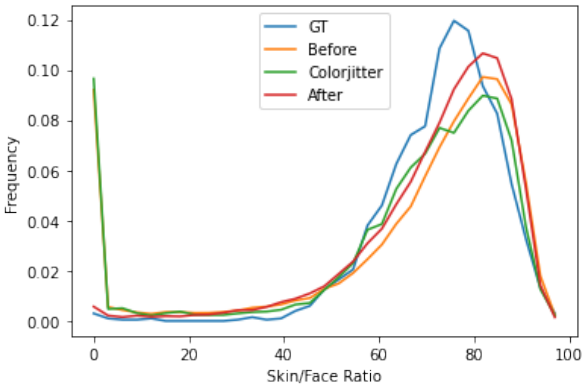
(b) Asian



(c) Indian



(d) African

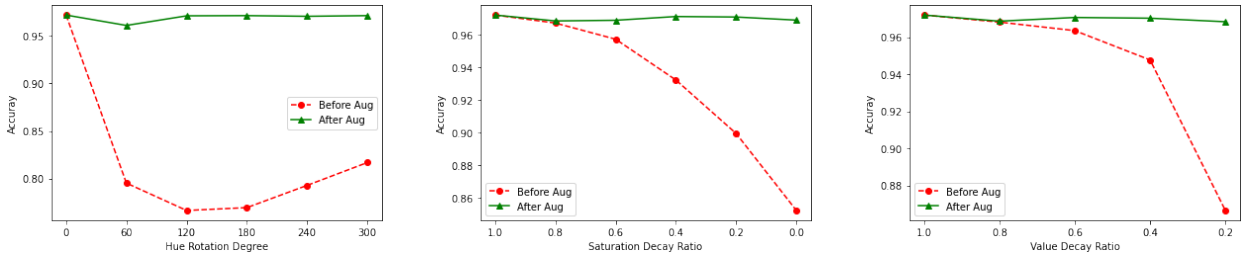


(e) Overall

Figure 4.9: Skin/face ratio distributions curves for overall RFW dataset (e) and the four races with different skin tones in RFW dataset (a to d). Blue curves refer to the sample distribution curve we get from the annotated ECU dataset. Orange and red curves refer to the distribution from testing results before and after color augmentation respectively. Green curves refer to the distribution from the testing results with colorjitter.

Figure 4.10: Experimental results from the RFW dataset using several skin segmentation methods. *Left to right:* Kolkur et al. [44], Dahmani et al. [18], Jones and Rehg [41], FCN before augmentation (B), using colorjitter (CJ), and after (A) augmentation, and U-Net before augmentation (B), using colorjitter (CJ), and after (A) augmentation. Rows 1 to 4 show sample results for the RFW ethnic groups: Caucasian, Asian, Indian, and African.

Figure 4.10 shows some qualitative results containing both various skin colors and complex illuminance. The first row is a Caucasian man with background of color similar to his skin. The second row is an Asian woman with one shoulders in the dark. The third row is an Indian man with strong light on his top head. The fourth row is an African woman with her face in shadow. We find that the three baseline methods in columns 2 to 4 are fully confused by the background in the first row. U-Net (B) fails to detect the skin area with intense light in the fourth row. In contrast, our method works better and outputs accurate and complete results.



(a) (b) (c)

Figure 4.11: Comparison of models that use color augmentation (After Aug) and the one does not use color augmentation (Before Aug) to test robustness for image filtering in the color space. Augmentation shows their effectiveness for all three dimentions (a) Hue, (b) Saturation, and (c) Value.

4.4.3 Unconstrained illuminations and filters

In order to further test the robustness of the algorithm under different illumination and creative filter, we ran experiments by testing images transformed by HSV color space augmentation (similar to the training set). We trained two skin segmentation algorithms: One without color augmentation and one with color-based augmentation. Then we tested them with the color augmented test set. Figure 4.11 shows the comparison between the results. For the model without color augmentation, the accuracy falls immediately as they are tested with image sets that have been modified by hue (Figure 4.11a), saturation (Figure 4.11b), or value (Figure 4.11c). The accuracy remains consistent for the model that was trained with a color augmented set of images. This ablation study shows explicitly the effectiveness of our proposed model. Figure 3.3 shows qualitative examples of how even with drastic changes in the color, the output of our algorithm remains consistent across all the filters. Finally, to test the performance under ambient light across the spectral, we selected random images from the web and tested them.

Figure 4.12 shows the robustness of our methods against the drastic illumination changes. The model without color augmentation fails to detect a single skin pixel in the second and

Table 4.7: Augmentation improves the performance of both U-Net and FCN when tested on images with unconstrained illumination and filters from our self-made dataset.

	IoU before aug.	IoU colorjitter	IoU after aug.
FCN	12.05	31.55	64.92
U-Net	11.76	18.19	35.85

third rows, even colorjitter doesn't conduct any help in some images, while our method (with augmentation) successfully detects skin pixels in most cases. Qualitative evaluations in Table 4.7 shows this improvement. Both colorjitter appplied models and models without color augmentation have poor performance on the extreme illumination images. IoU increase sharply after color augmentation is applied to skin segmentation systems.

4.4.4 Grayscale images

We also conducted tests using grayscale images only. In this part, we used images from the ECU dataset and converted them to grayscale format. Then we performed testing with the U-Net model and the FCN model, with and without color space augmentation. The results are listed in Table 4.8. The table shows that the FCN model without color-space augmentation yields very poor performance, and the resulting IoU is low. For U-Net without augmentation, IoU declines to approximately 0, indicating that the model detected hardly any skin pixels. The performance improves significantly for both of these models, when colorjitter was applied. IoU for the FCN model returned to over 70% and IoU for U-Net returned to over 60%. The saturation part in the colorjitter helps the skin segmentation model perform better on grayscale images. Our proposed methods help the skin segmentation gain more IoU on this basis. The IoU of the FCN model reached 77% and the IoU for U-Net model reached nearly 70%.

Qualitative results are illustrated in Figure 4.13. FCN (B) can detect only a small area of skin

Figure 4.12: Testing results on our self-made dataset by deep learning models FCN and U-Net. The label (B) on the top of the images refers to the results from the model before color augmentation and the label (CJ) refers to the results from mode using color jitter. In comparison, label (A) refers to the model with color augmentation. Input images and ground truth are shown in columns 1 and 2 in each group.

Table 4.8: Augmentation improves the performance of both U-Net and FCN when tested on grayscale images from the ECU dataset.

	IoU before aug.	IoU with colorjitter	IoU after aug.
FCN	47.13	71.76	77.20
U-Net	0.55	60.45	69.42

pixel in grayscale image but U-Net (B) fail to detect any skin pixels. Although improvements are obvious after colorjitter is applied to the network, our methods still output better results with more details. Models with color augmentation work well and correctly detected skin pixels for grayscale images.

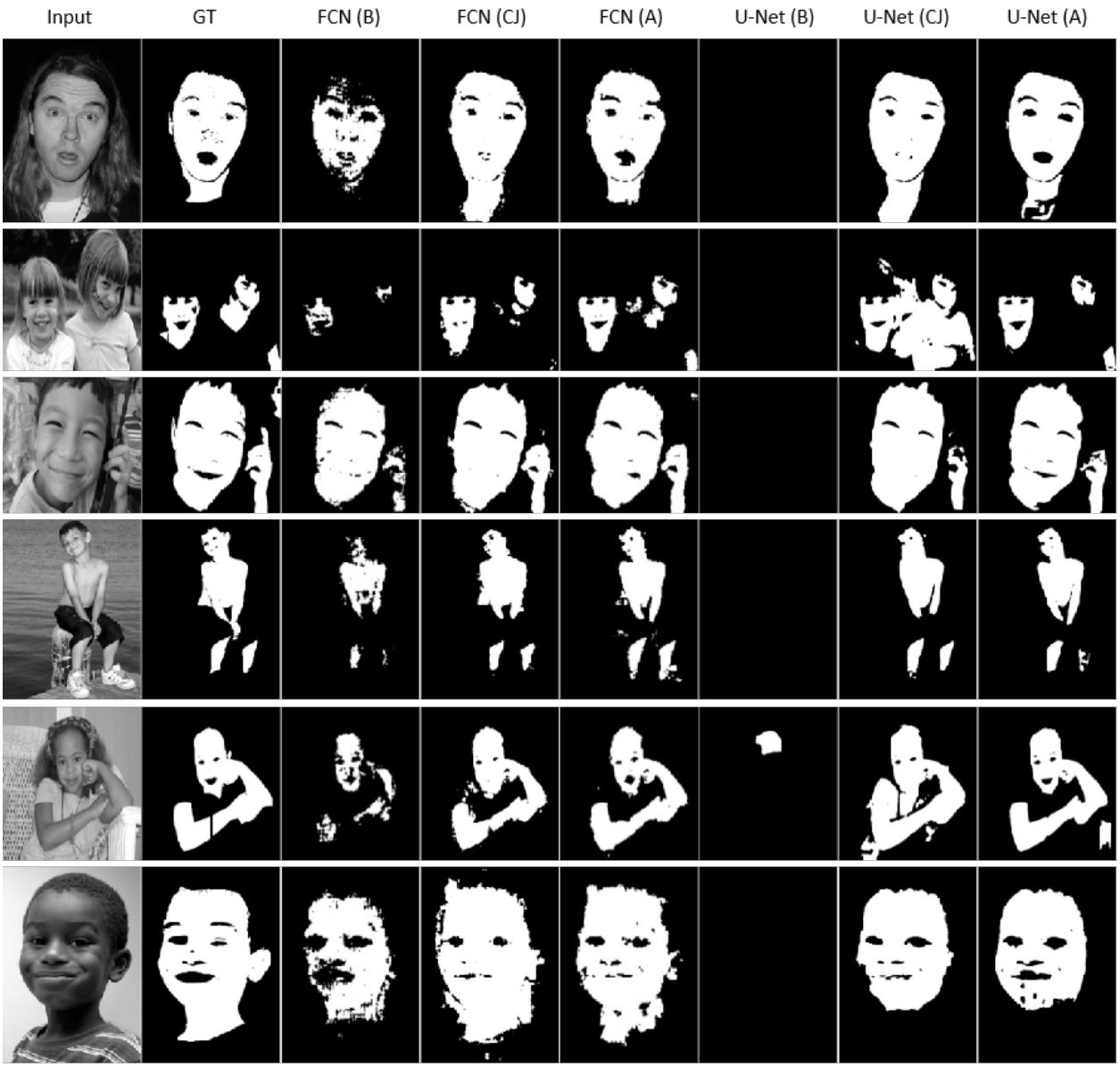


Figure 4.13: Testing results on grayscale images from the ECU dataset by deep learning models FCN and U-Net. The label (B) on the top of the images refers to the results from the model before color augmentation. In comparison, label (A) refers to the model with color augmentation. Input images and ground truth are shown in columns 1 and 2 in each group.

Chapter 5

Conclusions

This work has introduced a new approach for automated detection of skin in images. The system leverages recent innovations involving encoder-decoder networks like U-net. The color-based data-augmentation step strategically reduces dependence by the system on color-based cues and removes racial bias. This in turn addresses problems related to illumination differences (e.g., indoor/outdoor situations, harsh shadows, unnatural lighting), variations in skin tone (especially ethnic/racial variations), and different sensor parameters (e.g., color, monochromatic, varying spectral sensitivities).

To our knowledge, the system presented here is the first to apply deep methods to the problem of skin detection that can be generalized across the color spectrum. Further, this work is the first to directly assess the skin tone bias in skin detection systems. Experimental results are presented using two datasets. Our approach has demonstrated better precision and recall for each of the six annotated skin types than three alternative skin-detection systems for the ECU dataset. Compared with the other color augmentation method, colorjitter, our method still has the smaller skin tone bias and performs better in the grayscale and extreme illumination testing. The experiments on the other network, FCN, also show improvement in skin detection. We anticipate that similar approaches can be applied more broadly to other Computer Vision techniques and tasks.

Bibliography

[1] RGB color cube. https://favpng.com/png_view/space-rgb-color-model-rgb-color-space-png/c2YBj7e1.

[2] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2523, 2021.

[3] Mohamed Alsheakhali, Ahmed Skaik, Mohammed Aldahdouh, and Mahmoud Alhelou. Hand gesture recognition system. *Information & Communication Systems*, 132, 2011.

[4] Elli Angelopoulou. Understanding the color of human skin. In *Human Vision and Electronic Imaging VI*, volume 4299, pages 243–251. International Society for Optics and Photonics, 2001.

[5] Arpansa. Fitzpatrick skin type. https://www.arpansa.gov.au/sites/default/files/legacy/pubs/RadiationProtection/FitzpatrickSkinType.pdf.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[7] George Borshukov and John P. Lewis. Realistic human face rendering for "The Matrix Reloaded". In *ACM Siggraph 2005 Courses*. 2005.

[8] J. Brand and J. S. Mason. A comparative assessment of three approaches to pixel-level

human skin-detection. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 1, pages 1056–1059, vol. 1, 2000.

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

[10] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017.

[11] Emir Buza, Amila Akagic, and Samir Omanovic. Skin detection based on image color segmentation with histogram and k-means clustering. In *Proceedings of the 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 1181–1186. IEEE, 2017.

[12] Tiberio S. Caetano, Sílvia .D Olabarriaga, and Dante A. C. Barone. Do mixture models in chromaticity space improve skin detection? *Pattern Recognition*, 36(12):3019–3021, 2003.

[13] D. Chai and K. N. Ngan. Locating facial region of a head-and-shoulders color image. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 124–129, 1998.

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous

convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[16] Heng-Da Cheng, X. H. Jiang, Ying Sun, and Jingli Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259–2281, 2001.

[17] Oana G. Cula, Kristin J. Dana, Frank P. Murphy, and Babar K. Rao. Skin texture modeling. *International Journal of Computer Vision*, 62(1):97–119, 2005.

[18] Djamila Dahmani, Mehdi Cheref, and Slimane Larabi. Zero-sum game theory model for segmenting skin regions. *Image and Vision Computing*, 99:103925, 2020.

[19] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and Remote Sensing Letters*, 13(3):364–368, 2016.

[20] Yunlong Dong. Trying to be the easiest FCN PyTorch implementation. https://github.com/yunlongdong/FCN-pytorch.

[21] Craig Donner, Tim Weyrich, Eugene d'Eon, Ravi Ramamoorthi, and Szymon Rusinkiewicz. A layered, heterogeneous reflectance model for acquiring and rendering human skin. *ACM Transactions on Graphics (TOG)*, 27(5):1–12, 2008.

[22] Alexandru F. Drimbarean, Peter M. Corcoran, Mihai Cuic, and Vasile Buzuloiu. Image processing techniques to detect and filter objectionable images based on skin tone and shape recognition. In *Proceedings of the International Conference on Consumer Electronics*, pages 278–279. IEEE, 2001.

[23] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[24] Ahmed Elgammal, Crystal Muang, and Dunxu Hu. Skin detection-a short tutorial. *Encyclopedia of Biometrics*, 4:1218–1224, 2009.

[25] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014.

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

[27] David A. Forsyth, Margaret Fleck, and Chris Bregler. Finding naked people. *International Journal of Computer Vision*, 1065(1):593–602, 1996.

[28] Mehran Fotouhi, Mohammad H Rohban, and Shohreh Kasaei. Skin detection using contourlet-based texture analysis. In *Proceedings of the Fourth International Conference on Digital Telecommunications*, pages 59–64. IEEE, 2009.

[29] P. Ganesan and V. Rajini. Assessment of satellite image segmentation in RGB and HSV color space using image quality measures. In *Proceedings of the International Conference on Advances in Electrical Engineering*, pages 1–5. IEEE, 2014.

[30] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999.

[31] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, December 2015.

[32] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[33] G. Gomez. On selecting colour components for skin detection. In *Proceedings of 2002 International Conference on Pattern Recognition*, volume 2, pages 961–964 vol.2, 2002.

[34] Robert M. Haralick and Linda G. Shapiro. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing*, 29(1):100–132, 1985.

[35] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[36] Yi He, Jiayuan Shi, Chuan Wang, Haibin Huang, Jiaming Liu, Guanbin Li, Risheng Liu, and Jue Wang. Semi-supervised skin detection by network with mutual guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2111–2120, 2019.

[37] Md. Foisal Hossain, Mousa Shamsi, Mohammad Reza Alsharif, Reza A. Zoroofi, and Katsumi Yamashita. Automatic facial skin detection using Gaussian mixture model under varying illumination. *International Journal of Innovative Computing, Information and Control*, 8(2):1135–1144, 2012.

[38] Lei Huang, Wen Ji, Zhiqiang Wei, Bo-Wei Chen, Chenggang Clarence Yan, Jie Nie, Jian Yin, and Baochen Jiang. Robust skin detection in real-world images. *Journal of Visual Communication and Image Representation*, 29:147–152, 2015.

[39] Takanori Igarashi, Ko Nishino, and Shree K. Nayar. The appearance of human skin. Department of Computer Science, Columbia University, 2005.

[40] Nina G. Jablonski and George Chaplin. The evolution of human skin coloration. *Journal of Human Evolution*, 39(1):57–106, 2000.

[41] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[42] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.

[43] Keras. Image data preprocessing. https://keras.io/api/preprocessing/image/.

[44] Seema Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and Janvi Jatakia. Human skin detection using RGB, HSV and YCbCr color models. *arXiv preprint arXiv:1708.02694*, 2017.

[45] J. Kovac, P. Peer, and F. Solina. Human skin color clustering for face detection. In *Proceedings of The IEEE Region 8 EUROCON 2003. Computer as a Tool*, volume 2, pages 144–148, 2003.

[46] Michael W. Levine and Jeremy M. Shefner. *Fundamentals of Sensation and Perception*. Brooks/Cole Publishing Company, 1991.

[47] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019.

[48] Bart Liefers, Johanna M. Colijn, Cristina González-Gonzalo, Timo Verzijden, Jie Jin Wang, Nichole Joachim, Paul Mitchell, Carel B. Hoyng, Bram van Ginneken, Caroline C. W. Klaver, and Clara I. Sánchez. A deep learning model for segmentation of geographic atrophy to study its long-term natural history. *Ophthalmology*, 127(8): 1086–1096, 2020.

[49] Qiong Liu and Guang-zheng Peng. A robust skin color based face detection algorithm.

In *Proceedings of the 2nd International Asia Conference on Informatics in Control, Automation and Robotics*, volume 2, pages 525–528. IEEE, 2010.

[50] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[51] Mohammad Reza Mahmoodi and Sayed Masoud Sayedi. A comprehensive survey on human skin detection. *International Journal of Image, Graphics and Signal Processing*, 8(5):1, 2016.

[52] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[53] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *Proceedings of the International Interdisciplinary PhD Workshop*, pages 117–122. IEEE, 2018.

[54] Dinh-Ha Nguyen, Trung-Hieu Le, Thanh-Hai Tran, Hai Vu, Thi-Lan Le, and Huong-Giang Doan. Hand segmentation under different viewpoints by combination of Mask R-CNN with tracking. In *Proceedings of the 5th Asian Conference on Defense Technology*, pages 14–20. IEEE, 2018.

[55] Duong Hai Nguyen, Tai Nhu Do, In-Seop Na, and Soo-Hyung Kim. Hand segmentation and fingertip tracking from depth camera images using deep convolutional neural network and multi-task segnet. *arXiv preprint arXiv:1901.03465*, 2019.

[56] T. Ojala, M. Rautiainen, E. Matinmikko, and M. Aittola. Semantic image retrieval with

HSV correlograms. In *Proceedings of the Scandinavian conference on Image Analysis*, pages 621–627, 2001.

[57] OpenCV. Color space converter. https://docs.opencv.org/3.4/de/d25/imgproc_color_conversions.html.

[58] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005.

[59] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[60] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[61] Siddharth Roheda. A multi-scale approach to skin pixel detection. *Electronic Imaging*, 2017(4):18–23, 2017.

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015.

[63] Kankana Roy, Aparna Mohanty, and Rajiv R. Sahay. Deep learning based hand detection in cluttered environment using skin segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 640–649, 2017.

[64] Silonie Sachdeva. Fitzpatrick skin typing: Applications in dermatology. *Indian Journal of Dermatology, Venereology and Leprology*, 75(1):93, 2009.

[65] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[66] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Universal skin detection without color information. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 20–28, 2017.

[67] Steven H. Schwartz. *Visual perception: A clinical orientation.* McGraw-Hill Medical Pub. Division, 2009.

[68] Khamar Basha Shaik, P. Ganesan, V. Kalist, B. S. Sathish, and J. Merlin Mary Jenitha. Comparative study of skin color detection and segmentation in HSV and YCbCr color space. *Procedia Computer Science*, 57:41–48, 2015.

[69] Tomáš Sixta, Julio C.S. Jacques Junior, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. Fairface challenge at ECCV 2020: Analyzing bias in face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 463–481. Springer, 2020.

[70] Maricor Soriano, Birgitta Martinkauppi, Sami Huovinen, and Mika Laaksonen. Skin detection in video under changing illumination conditions. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 1, pages 839–842. IEEE, 2000.

[71] Holger Steiner, Sebastian Sporrer, Andreas Kolb, and Norbert Jung. Design of an active multispectral SWIR camera system for skin detection and face verification. *Journal of Sensors*, 2016, 2016.

[72] SuperAnnotate. The ultimate training data platform for AI. https://www.superannotate.com/.

[73] Wei Ren Tan, Chee Seng Chan, Pratheepan Yogarajah, and Joan Condell. A fusion approach for efficient human skin detection. *IEEE Transactions on Industrial Informatics*, 8(1):138–147, 2011.

[74] Tomasz Tarasiewicz, Jakub Nalepa, and Michal Kawulok. Skinny: A lightweight U-Net for skin detection and segmentation. In *Proceedings of the 2020 IEEE International Conference on Image Processing*, pages 2386–2390. IEEE, 2020.

[75] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

[76] Anirudh Topiwala, Lidia Al-Zogbi, Thorsten Fleiter, and Axel Krieger. Adaptation and evaluation of deep learning techniques for skin segmentation on novel abdominal dataset. In *Proceedings of the 19th International Conference on Bioinformatics and Bioengineering*, pages 752–759. IEEE, 2019.

[77] Sofia Tsekeridou and Ioannis Pitas. Facial feature extraction in frontal views using biometric analogies. In *Proceedings of the 9th European Signal Processing Conference*, pages 1–4. IEEE, 1998.

[78] Satishkumar L. Varma and Vandana Behera. Human skin detection using histogram processing and gaussian mixture model based on color spaces. In *Proceedings of the International Conference on Intelligent Sustainable Systems*, pages 116–120. IEEE, 2017.

[79] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces

in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702, 2019.

[80] Xiaohua Wang, Xi Zhang, and Jingliang Yao. Skin color detection under complex background. In *Proceedings of the International Conference on Mechatronic Science, Electric Engineering and Computer*, pages 1985–1988. IEEE, 2011.

[81] Yanjiang Wang and Baozong Yuan. A novel approach for human face detection from color images under complex background. *Pattern Recognition*, 34(10):1983–1992, 2001.

[82] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[83] Wikipedia. Hsl and hsv. https://en.wikipedia.org/wiki/HSL_and_HSV.

[84] Kwok-Wai Wong, Kin-Man Lam, and Wan-Chi Siu. A robust scheme for live detection of human faces in color images. *Signal Processing: Image Communication*, 18(2):103–114, 2003.

[85] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *Proceedings of the International Conference on Learning Representations*, 2016.

[86] Benjamin D. Zarit, Boaz J. Super, and Francis K. H. Quek. Comparison of five color models in skin pixel classification. In *Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99*, pages 58–63. IEEE, 1999.

[87] Xuhao Zhi. Implementation of deep learning framework – U-Net, using Keras. https://github.com/zhixuhao/unet.

[88] Qingsong Zhu, Yaoqin Xie, Jia Gu, and Lei Wang. A new video object segmentation algorithm by fusion of spatio-temporal information based on GMM learning. In *Advances in Automation and Robotics, Vol. 2*, pages 641–650. Springer, 2011.

[89] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018.

[90] Haiqiang Zuo, Heng Fan, Erik Blasch, and Haibin Ling. Combining convolutional and recurrent neural networks for human skin detection. *IEEE Signal Processing Letters*, 24(3):289–293, 2017.

Appendices

Appendix A

Additional Results

This section will supply more output results from the skin detection systems mentioned in the main thesis. We illustrate another six examples in Figure A.1. The first row is a girl wearing a skin color-like cloth. The second and third rows contain backgrounds that have similar colors as the people in the image. The fourth row is a girl with brown cloth. The fifth row is a baby with strong lights on his head. The sixth row contains multiple people in various poses and skin colors These challenging conditions make other methods fail or perform poorly. The three baseline methods all fail to classify the skin color-like ground in the first and second rows. U-Net (B) works better, but there is still some false positive noise in rows 1 to 5. Moreover, it fails to detect the people on the right in the last row. In contrast, our approach overcame most of the difficulties mentioned above and produced accurate and robust results. Compared with results before color augmentation, models with color augmentation make less false positive and false negative judgments. For example, the FCN (A) does not detect the baby's hair as skin pixels in the fifth row, and it does not make noise as FCN (B) does in the second row. In this part of view, our method outperforms the traditional skin segmentation methods, and color augmentation helps deep learning methods work better.

We demonstrate more results from the RFW dataset in Figure A.2. The three traditional skin segmentation methods still misclassify the color-like background to be skin pixels. For example, the background of the door is classified as skin areas in the third row by the three

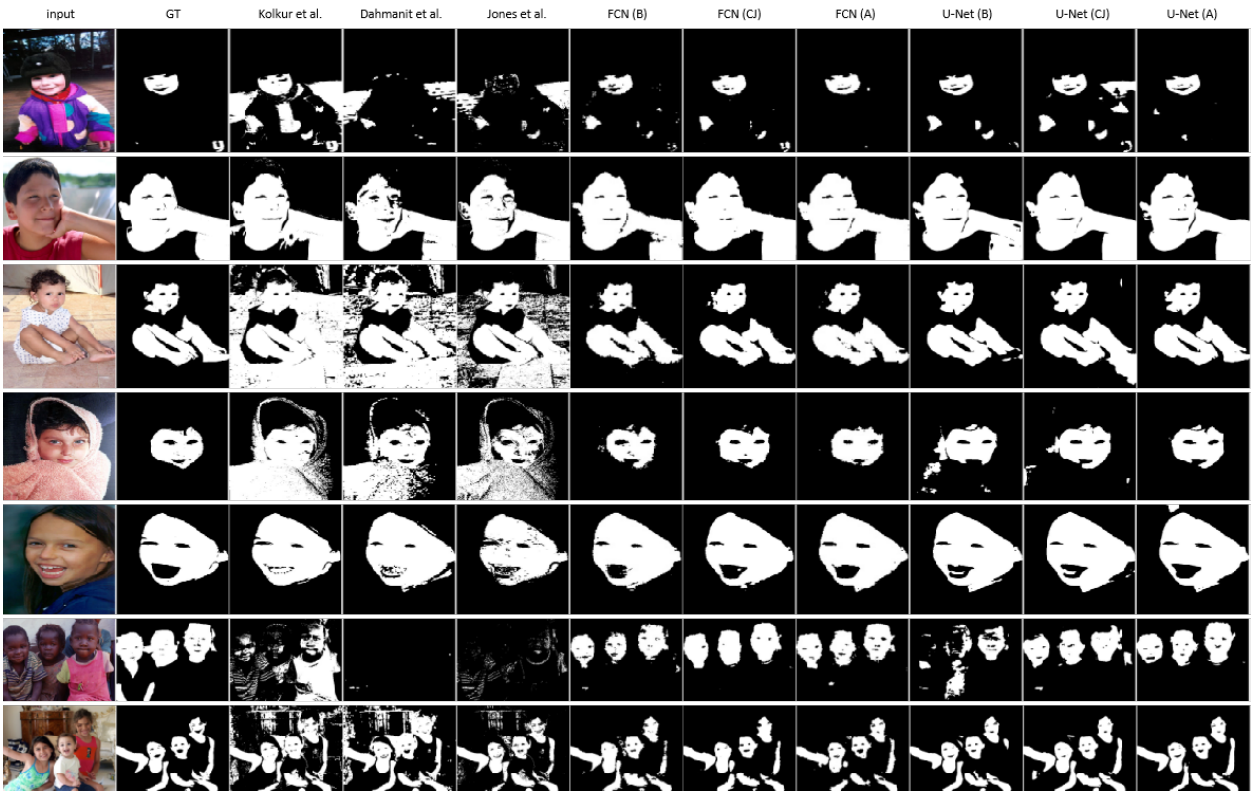


Figure A.1: Additional results on the ECU dataset, by various skin segmentation methods including Kolkur et al. [44], Dahmani et al. [18], Jones and Rehg [41], FCN before augmentation (B), using colorjitter (CJ), and after (A) augmentation, and U-Net before augmentation (B), using colorjitter (CJ), and after (A) augmentation (Columns 3 to 11). Input and ground truth are shown in column 1 and 2.

Figure A.2: Testing results on the RFW dataset, by various skin segmentation methods including Kolkur et al. [44], Dahmani et al. [18], Jones and Rehg [41], FCN before (B) and after (A) augmentation, and U-Net before (B) and after (A) augmentation (Columns 2 to 8). Input are shown in column 1. Result shown for races with different skin tones: Caucasian, Asian, Indian, and African (Row 1 to 4).

baseline methods. On the opposite, in the second row, glasses covered area is not classified as skin areas. What's more, the skin area covered by other items is detected as non-skin pixels. Compared with results after color augmentation, models without color augmentation are more likely to make false-positive judgments. Moreover, in the darker skin group, models after color augmentation can detect more skin pixels. For example, the result from FCN (A) has less false positive noise than that from FCN (B) in the first row. U-Net (A) detects more skin pixels on the man's head in the fourth row.