

IMPERIAL COLLEGE LONDON
DEPARTMENT OF COMPUTING

Neural Modeling and Re-ageing of Bio-physical Facial Appearance

Author

Xiaohui Li

Supervisor

Prof. Abhijeet Ghosh

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing
and the Diploma of Imperial College London, 2024

Declaration of Originality: I, Xiaohui Li, declare that the work in this thesis is my own. The work of others has been appropriately referenced. A full list of references is given in the bibliography.

© Xiaohui Li 2024

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Realistic facial appearance modeling and editing have gained significant attention in computer graphics. Recently, several diffusion-based biophysical skin models have been developed to accurately simulate skin appearance by incorporating multiple spectral parameters representing chromophores in human skin. These chromophores, which are key absorbers in both the visible and invisible spectra, play a crucial role in revealing various facial features under different lighting conditions.

In this thesis, we propose a hyperspectral BSSRDF skin model spanning a wide spectral range from 300nm to 1000nm, as well as a monitor-based practical measurement method. For parameter estimation, we introduce a three-stage LUT search to speed up the conventional LUT search. However, this approach is constrained by the discreteness of Lookup tables. To overcome this limitation, we present an Encoder-Decoder network. This network is further simplified to accept a single image under D65 illumination and output the corresponding skin appearance directly. This simplified network is particularly useful for skin appearance editing, where adjusting melanin and hemoglobin concentrations allows for simulating various skin tones, such as tanned, lightened, and flushed.

Building upon skin editing, we simulate the age-appropriate skin tone changes. However, skin tone variation alone is not the most obvious feature in facial re-ageing. On the other hand, existing re-ageing methods often generate high-quality transformations while neglecting the skin properties changes. To address this issue, we propose a re-ageing

framework that integrates approaches, enabling more realistic age simulations across a wide range from 10 to 80 years old. Additionally, we uplift this method to a 3D by incorporating a geometric network that reconstructs coarse facial shapes with head size control, followed by a lightweight refinement network to generate fine details. This work is the first in computer graphics to combine a biophysical skin model with re-aging, producing high-quality transformations in both facial appearance and geometry.

Acknowledgements

Completing this PhD is an incredible journey, and I would like to thank many people who have supported me along the way. First, my deepest gratitude goes to my supervisor, Prof. Abhijeet Ghosh, for giving me an opportunity to do research in computer graphics. His support and guidance kept me focused and motivated through every stage of this work.

I am heartily thankful to Dr. Giuseppe Claudio Guarnera, who always gave me valuable insights and constructive suggestions on our projects. I would also like to thank my examiners, Prof. William Smith and Prof. Ben Glocker, for their helpful feedback on improving this thesis.

Special thanks to all the past and present members of the Realistic Graphics and Imaging group and Lumirithmic that I have met: Yuliya Gitlina, Arvin Lin, Chongrui Fan, Emilie Nogu  , Ekin Ozturk, Alexandros Lattas, Yiming Luo, Gilles Rainer, Lewis Bridgeman, Luca Filipi, Jayanth Kannan, and Yiming Lin, for their collaboration and many inspiring conversations. I would also like to thank Amani El-Kholy for her support and help over these years.

I am also thankful to all my friends in different stages of my life. Big thanks to Yu Zhao and Lica   Xiao for being my good friends for more than ten years and always supporting me throughout my life no matter what. Thanks to Jiali Zheng and Yicheng Gao who are also PhD students at Imperial and have become my lifelong friends. Additionally, shout

out to my dance course friends: Rita Zhang and Linda Zhao, for bringing many exciting and fun times into my life.

To my parents, words cannot express my gratitude. I would like to give the biggest thanks to my parents for creating a family environment with endless love and providing me with the foundation on which I was able to pursue this dream.

I would like to thank my ever-caring boyfriend Yuheng Wang, who has accompanied me for around half of my life. He always gives me his love, support, and yes, prodding. Without him, I might still be able to finish this PhD, but I cannot imagine what my life would be like.

Finally, I am dedicated to my younger self, who decided to go through this challenging and wonderful life experience.

This thesis is a product of many contributions, and I am forever grateful to those who have helped me along the way.

Publications

Here is a list of publications authored during my PhD studies.

- Xiaohui Li, Giuseppe Claudio Guarnera, Arvin Lin, Abhijeet Ghosh. Realistic Facial Age Transformation with 3D Uplifting. *Computer Graphics Forum (CGF)*, 2024 [\[51\]](#)
- Xiaohui Li, Giuseppe Claudio Guarnera, Arvin Lin, Abhijeet Ghosh. Practical Measurement and Neural Encoding of Hyperspectral Skin Reflectance. *International Conference on 3D Vision (3DV)*, 2024 [\[50\]](#)
- Arvin Lin, Yiming Lin, Xiaohui Li, Abhijeet Ghosh. High Quality Neural Relighting using Practical Zonal Illumination. *Eurographics Symposium on Rendering (EGSR)*, 2024 [\[52\]](#)
- Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, Julie A. McCann. WiMANS: A Benchmark Dataset for WiFi-based Multi-user Activity Sensing. *European Conference on Computer Vision (ECCV) 2024* [\[34\]](#)
- Wang, Yuheng and Ye, Juan and Li, Xiaohui and Borchers, David L. Towards automated animal density estimation with acoustic spatial capture-recapture. *Biometrics 2024* [\[75\]](#)
- Emilie Nogu  , Arvin Lin, Xiaohui Li, Claudio Guarnera, Abhijeet Ghosh. Practical RGB Measurement of Fluorescence and Blood Distributions in Skin. *Color and Imaging Conference*, 2024 [\[57\]](#)

Abbreviations

BRDF Bidirectional Reflectance Distribution Function

BSSRDF Bidirectional Scattering-Surface Reflectance Distribution Function

CIE Commission Internationale de l'Eclairage

SPD Spectral Power Distribution

CSS Camera Sensitivity Curve

CMF Color Matching Functions

LED Light-Emitting Diode

RGBW Red Green Blue White

LUT Lookup Table

UVA Ultra Violet

Vis Visible

NIR Near InfraRed

PCA Principal Component Analysis

FC Layer Fully-Connected Layer

MLP Multiple Layer Perceptron

Conv Layer Convolutional Layer

CNN Convolutional Neural Network

SGD Stochastic Gradient Descent

RMSprop Root Mean Square Propagation

AdaGrad Adaptive Gradient Algorithm

Adam Adaptive Moment Estimation

MSE Mean Squared Error

MAE Mean Absolute Error

AE AutoEncoder

VAE Variational AutoEncoder

VQVAE Vector Quantized Variational AutoEncoder

GAN Generative Adversarial Network

DDPM Denoising Diffusion Probabilistic Model

DDIM Denoising Diffusion Implicit Model

LATS Lifespan Age Transformation Synthesis

SAM Style-based Age Manipulation

FRAN Production Ready Face Re-Aging for Visual Effects

HRFAE High Resolution Face Age Editing

3DMM 3D Morphable Model

BFM Basel Face Model

FLAME Face Learned with Articulated Model and Expression

DECA Detailed Expression Capture and Animation

EMOCA Motion Capture and Animation

MICA Metric Face

COMA Convolutional Mesh Autoencoder

Symbols

$f_r(\vec{\omega}_i, \vec{\omega}_o)$	Bidirectional Reflectance Distribution Function (BRDF)	z_v	Position of the negative point light
$S(x_i, \vec{\omega}_i, x_o, \vec{\omega}_o)$	Bidirectional Scattering Surface Reflectance Distribution Function (BSSRDF)	z_b	Middle distance between positive and negative point lights
x_i	Incident light point on the surface	d_r	Distance to the positive point light from a point on the surface
$\vec{\omega}_i$	Incident light direction	d_v	Distance to the negative point light from a point on the surface
x_o	Outcoming light point on the surface	α'	Reduced albedo
$\vec{\omega}_o$	Outcoming light direction	σ_a	Absorption coefficient
L	Outgoing radiance	σ_t	Extinction coefficient
Φ	Incident flux	σ_s	Scattering coefficient
L_i	Incoming radiance from a light source	σ'_t	Reduced extinction coefficient
L_o	Radiance reflected by a surface	σ'_s	Reduced scattering coefficient
L_{ri}	Reduced incoming radiance	σ_{tr}	Effective transport coefficient
$p(\vec{\omega}_i, \vec{\omega}_i')$	Phase function	$S(\lambda)$	Spectral Power Distribution (SPD)
λ	Wavelength of light	$R(r)$	Diffuse reflectance of the surface
η	Refractive index of the medium	$T(r)$	Diffuse transmittance of the surface
F_{dr}	Diffuse Fresnel reflectance	C_m	Concentration of melanin
F_t	Fresnel transmittance	β_m	Blend fraction of two types melanin
l	Mean free path	C_{he}	Concentration of hemoglobin in epidermis
z_r	Position of the positive point light		

C_{hd} Concentration of hemoglobin in dermis	C_{wd} Concentration of water in dermis
C_{bc} Concentration of β carotene	I_W bluebroadband white illumination
γ^{epi} Ratio of oxygenated and deoxygenated hemoglobin in epidermis	I_{nb} bluenarrowband blue illumination
	I_{ng} bluenarrowband green illumination
γ^{derm} Ratio of oxygenated and deoxygenated hemoglobin in dermis	I_{sb} bluesynthetic blue illumination
	I_{sg} bluesynthetic green illumination
C_{we} Concentration of water in epidermis	

Contents

1 Introduction	1
2 Background and Related Work	6
2.1 Light Diffusion Approximation	7
2.1.1 Light Dipole Diffusion	7
2.1.2 Multipole Diffusion Theory	12
2.1.3 Diffusion in multi-layered materials	14
2.2 Methods of Biophysical Skin Modeling	16
2.2.1 Biophysical Skin Modeling	17
2.2.2 Hyperspectral Skin Modeling	21
2.2.3 Spectral Skin Appearance Measurement	23
2.3 Reconstruction of Hyperspectral Skin Reflectance	24
2.3.1 Visualization of the full spectral	24
2.3.2 Color Correction	28

2.3.3	Dimension reduction of the full reflectance	31
2.4	Neural Network	32
2.4.1	Fundamental Components	32
2.4.2	Introduction to Popular Neural Networks	38
2.5	Face Re-ageing	44
2.6	Facial Geometry Reconstruction	49
3	Practical Measurement and Hyperspectral Skin Model	54
3.1	Bio-physical Skin Model	55
3.2	Acquisition and Measurement	59
3.2.1	Visible domain	59
3.2.2	Invisible domain	63
3.3	Conventional Parameter Estimation and Skin Appearance Reconstruction	63
3.4	Results and Evaluation	65
3.4.1	Results	65
3.4.2	Evaluation	68
3.5	Conclusion	72
4	Neural Parameter Estimation and Reconstruction	74
4.1	Hyperspectral Skin Model	75

4.1.1 Implementation	75
4.1.2 Results and Evaluations	78
4.2 Naive Skin Model	82
4.2.1 Implementation	82
4.2.2 Results and Evaluations	84
4.3 Image-space Scattering Approximation	87
4.4 Application - Skin Appearance Editing	88
4.5 Conclusion	91
5 Realistic 3D Facial Age Transformation	93
5.1 Overview	94
5.2 Appearance Editing	95
5.2.1 Biophysical Skin Editing	96
5.2.2 Ageing Structural editing	99
5.3 Shape Reconstruction	106
5.3.1 Coarse Shape Reconstruction	106
5.3.2 Detailed Shape Reconstruction	110
5.4 Results and Evaluation	113
5.4.1 Results	113
5.4.2 Qualitative Evaluation	117

5.4.3 Quantitative Evaluation	120
5.4.4 User Study	121
5.5 Conclusion	123
6 Conclusion	126
Bibliography	130

List of Tables

3.1	Parameters of skin spectral absorption and scattering.	58
4.1	Accuracy analysis	81
5.1	<i>GeoNet</i> architecture	108
5.2	Quantitative evaluation	121

List of Figures

2.1 Bidirectional Reflectance Distribution Function (BRDF)	7
2.2 Bidirectional Scattering Surface Reflectance Distribution Function (BSS-RDF)	8
2.3 Dipole Diffusion approximation	12
2.4 Comparison of dipole and multiple diffusion theory.	13
2.5 Light transport between the epidermis and dermis computed with Kubelka-Munk theory in Frequency space.	15
2.6 Two-layer skin model proposed by [20]	17
2.7 Spectral absorption coefficients of eumelanin, pheomelanin, oxygenated hemoglobin, deoxygenated hemoglobin, and baseline. This figure is borrowed from [20].	19
2.8 The CIE 1931 XYZ color matching functions.	25
2.9 Comparison of the 2D lookup tables of spectral skin model with C_m (x-axis) and C_h (y-axis) under D65' illumination (a-d) and cold white illumination (e-h) and variation due to β_m ((a,b),(e,f)) and C_{he} ((c,d),(g,h)), respectively.	27

2.10 Visualization of a subject under white (a), red (b), green (c), and blue (d) uniform illuminations in the visible domain.	28
2.11 Camera Sensitivity Curves	29
2.12 Comparison between the uncorrected and corrected color chart.	30
2.13 Comparison between the uncorrected and corrected photo of a subject.	30
2.14 Multiple Layer Perceptrons (MLPs).	33
2.15 The Convolutional Layer	34
2.16 AutoEncoder (AE) Architecture	38
2.17 UNet Architecture	39
2.18 Variational AutoEncoder (VAE) Architecture	40
2.19 Vector Quantized Variational AutoEncoder (VQVAE) Architecture	40
2.20 Generative Adversarial Network (GAN) Architecture	41
2.21 Diffusion Architecture	43
2.22 CLIP Architecture	44
2.23 The overview of LATS	46
2.24 Failure examples of LATS	46
2.25 The overview of SAM	47
2.26 Ageing results generated using SAM.	47
2.27 FRAN.	48

2.28 HRFAE.	48
2.29 MICA	51
2.30 DECA	51
2.31 COMA.	52
3.1 Spectral skin reflectance variations due to C_{bc} . C_m increases along the x-axis, C_{hd} along the y-axis; β_m and C_{he} are kept fixed to 0.5 and 0.25, respectively.	56
3.2 The spectral absorption coefficients curves for the skin chromophores in- cluded in our model from 300nm to 1000nm.	56
3.3 Monitor-based setup and the SPD of white illumination.	60
3.4 Practical spectral measurements of skin reflectance using screen illumina- tion. (a) White (b) Synthetic blue (c) Synthetic green.	62
3.5 Comparison of photograph and reconstructions with different measure- ments. Figure (a) employs broadband white, synthetic blue, and synthetic green illuminations while Figure (c) employs broadband white, direct blue, and direct green illuminations.	62
3.6 Comparison of photos in the Visible domain(a) and reconstructions in the Ultra-Violet (c), Visible (b), and Near-Infrared (d) domains for faces of Asia skin type (first row), Caucasian skin type (second), and South Asia skin type (third row).	66
3.7 The estimated spectral parameters (a - e) for the corresponding faces of Asia skin type (first row), Caucasian skin type (second), and South Asia skin type (third row).	67

3.8	Comparison of a facial photograph captured with Light Stage under white light (b), reconstruction via 4D model [23] (RMSE=0.0311) (a), and via our 5D model (c) (RMSE=0.0164), including absolute error maps in the insets.	69
3.9	Comparison of photograph (a), reconstruction using the naive LUT search method, using white, Synth. blue and Synth. green (b), and reconstruction using the 3-stage search (c).	70
3.10	Comparing UV photo (a) with reconstructions: broadband plus narrow-band inputs (b), and broadband only (c).	70
3.11	Comparisons of Vis, UV, and IR photos (a-c) and reconstructions (d, e) using our proposed method.	71
4.1	The proposed neural architecture consists of two networks. The <i>ChromNN</i> maps multispectral albedo data to the five spectral model parameters. The model parameters are then fed to the <i>SpectraNN</i> to predict a PCA-based, low-dimensional skin reflectance representation, finally reconstructed using the standard PCA reconstruction technique to obtain the full spectral reflectance in the UVA+VIS+NIR range (300nm to 1000nm).	76
4.2	Comparison of the network predictions and references of subject 1 (Mediterranean skin type) and subject 2 (Asian skin type). Figure (a) compares the photograph and visualizations of facial reflectance produced by <i>SpectraNN</i> in the visible domain under white illumination. Given the input image, Figure (b - f) compares the predictions of five spectral parameters by the proposed three-stage LUT searching method and the <i>ChromNN</i> .	79

4.3	Comparison of the network predictions and references of subject 1 (Mediterranean skin type) and subject 2 (Asian skin type). Figure (a - f) compare references and visualizations of facial reflectance produced by <i>SpectraNN</i> in the visible domain under white (a), red (b), green (c), and blue (d) illuminations, and in the UVA (e) and NIR (f) domains. Here, the references for the visible domain are photographs, while the references for the UVA and NIR domains are the reconstructions generated by the proposed three-stage LUT searching method.	80
4.4	Comparison between the generated skin reflectance of our model and Leeds skin dataset of dark and light skin.	82
4.5	The simplified neural architecture consists of an Encoder and a Decoder. Given the images under D65 illumination, <i>ChromNN</i> makes a prediction of the spectral parameters. These estimated parameters are then fed to the <i>ReconNN</i> to reconstruct the input skin appearance.	83
4.6	Comparison of the chromophore prediction between the proposed simplified network <i>ChromNN</i> and LUT search of two subjects with different skin types.	85
4.7	Comparison of the reconstruction between the proposed simplified network <i>ReconNN</i> and LUT search, and the input image.	86
4.8	Approximating profiles (a) with sums of seven Gaussians (b).	88
4.9	Comparison of image-space scattering approximation and photos (or channels) under white (a), red(b), green (c), blue (d), IR (e) illumination.	88
4.10	Editing melanin concentration C_m	89
4.11	Editing melanin blend β_m	90
4.12	Editing β carotene parameters C_{bc}	90

4.13 Editing hemoglobin related parameters C_{hd} and C_{he}	91
--	----

5.1 Overview. Starting with a single image I as input, our method first employs an age predictor [68], to estimate the source age a_s . With a given target age a_t , <i>SkinNet</i> edits the biophysical skin appearance due to ageing based on a skin model while <i>AgeEditNet</i> changes the structural aging effects iteratively from the source age a_s to the target ages a_t . Subsequently, we use the trained <i>Encoder</i> of <i>DECA</i> [22] to predict lighting, camera, and shape information. The next step involves <i>GeoNet</i> , which processes shape (β), expression (ψ), pose (θ), and age (a) parameters to reconstruct a coarse shape. This is followed by <i>RefineNet</i> , a lightweight network designed to predict a displacement map. Finally, we render the resulting image using the re-aged texture and detailed shape, incorporating the camera (c) and lighting (l) parameters. Here, we brighten the displacement map to show the details more clearly.	95
--	----

5.2 Biophysical skin editing. The <i>SkinNet</i> consists of <i>PredNet</i> and <i>ReconNet</i> . <i>PredNet</i> aims to estimate the melanin and hemoglobin concentrations C_m and C_{hd} from the given input image. The estimated chromophores are edited according to the source age a_s and target age a_t . These two parameters are then fed into <i>ReconNet</i> to produce the re-aged skin appearances.	97
--	----

5.3 Comparison between re-ageing results with and without biophysical skin editing.	98
---	----

5.4 An ablation study of the biophysical skin editing with different decreases of chromophore concentrations k_m and k_h . The subject in (a) is 60 years old, and the target age is 30 years old, so we expect an increase in chromophores concentration.	99
--	----

5.5 Ageing Structural editing architecture.	100
5.6 Re-aged result from 20 years old (left) to 40 (middle) and 60 (right) years old, respectively.	102
5.7 Ablation study of age-alpha transformation in <i>AgeEditNet</i> . The image is taken from the CelebA dataset.	103
5.8 Ablation study of the biophysical skin editing before and after the structural changes.	104
5.9 Facail age transformation at high resolution.	105
5.10 Coarse shape reconstruction. Given the shape (β), expression (ψ), pose (θ) parameters predicted by the pre-trained <i>Encoder</i> , along with an age parameter a , <i>GeoNet</i> reconstructs a coarse shape of the input image.	107
5.11 Comparison between FLAME [49] and 3D coarse shapes with different ages from <i>GeoNet</i>	110
5.12 Detailed shape reconstruction. <i>RefineNet</i> extracts the high-frequency information from the input image as a displacement map.	111
5.13 Age transformation results from 14 to 80 years old. The input images are outlined with a thin red border. Our methods generate realistic facial structural changes, such as wrinkles, with good identity preservation, as well as skin color changes due to ageing.	114
5.14 Coarse shape reconstructions with 20 years old and detailed shape reconstructions of 40, 60, and 80 years old.	115
5.15 An example of a round-trip transformation, aging a face from 22 years old (a) to 62 years old (b), then de-aging back to 22 years old (c).	116

5.16 An extreme example of a subject from 80 to 10 years old.	116
5.17 Multi-view rendering results of 20, 40, 60 years old with two environment maps.	117
5.18 Qualitative comparison among our method and three SOTA ageing methods, namely <i>SAM</i> [1], <i>HRFAE</i> [78], and <i>LATS</i> [58]. Notably, <i>SAM</i> shows a large identity loss compared to the other three methods, particularly for the male subject. In contrast, <i>HRFAE</i> demonstrates improved identity preservation, but this method is constrained by the limited age range (20–69 years old). Additionally, challenges in producing significant changes for the oldest age group are observed in both <i>HRFAE</i> and <i>LATS</i> although <i>LATS</i> is the best method among these four to de-age images to the children. Aside from the re-ageing quality, <i>SAM</i> and <i>LATS</i> can NOT preserve the background in the input image, especially for the right subject with a complex background.	118
5.19 Comparison of re-aged results using <i>FRAN</i> [82] and our method. Input image from <i>FRAN</i> [82].	119
5.20 Comparison of 3D detailed reconstruction among <i>DECA</i> [22], <i>EMOCA</i> [16], and ours. <i>DECA</i> and <i>EMOCA</i> can NOT alter ages, therefore we feed them the re-aged images using our methods. The input image is taken from the CelebA dataset.	120
5.21 Bar plot showing the frequency with which each method was chosen as the first choice, along with 95% confidence intervals.	123

Chapter 1

Introduction

Computer graphics is a field of computing that focuses on creating and manipulating images, including image generation, animation, and simulation. These images can either be derived from photographs or completely synthetic, representing two-dimensional or three-dimensional objects and scenes. The applications of computer graphics are wide and extensive, encompassing areas such as virtual avatars, film production, and medical simulations.

Among these areas, the realistic modeling and editing of facial appearances has become a popular but challenging research topic due to the complete structure of human skin. For this task, several biophysically-based spectral skin reflectance models have been proposed in recent years [45, 20, 21, 39, 13, 23, 2]. These approaches can simulate the accurate skin responses varying different lighting conditions, as well as capture temporal changes associated with phenomena such as facial re-ageing.

In this field, current approaches have two directions. One introduced complex biophysical models with numerous spectral parameters to represent various chromophores in the human skin [45, 21, 13]. In contrast, others employed simpler models with fewer parameters that can be measured practically, while reducing model accuracy [39, 23]. These works

mainly focus on the visible domain, however, some facial features that are crucial for skin appearance and protection might be concealed [13]. For instance, ultraviolet light is capable of revealing increased pigmentation, while infrared light can provide insights into the hydration levels of the skin.

Therefore, in Section 3, we employ a diffusion-based hyperspectral BSSRDF skin appearance model spanning a wide spectral range from 300nm to 1000nm, thus encompassing Ultra Violet (UV), Visible (Vis) and Near InfraRed (NIR) domains. Based on this model, we further present a practical measurement approach for acquiring spectral skin reflectance. This setup consists of four desktop LCD monitors for illuminating a face with standard visible spectrum RGB illumination in order to acquire high-quality spectral skin chromophore maps.

For spectral parameter estimation, previous works search the closest RGB values of facial albedo images in the pre-computed LookUp Table (LUT) covering a wide range of skin tones to achieve high-quality appearance reconstructions [39, 23]. However, this approach is time-consuming, particularly for models with higher complexity. To address this issue, we propose a novel three-stage LUT search to efficiently estimate the parameters of our model (see Section 3.3).

An additional issue of the LUT search is due to the discreteness of the results, which is related to the sampling resolutions of parameters used to generate the LUT. To overcome this limitation, we design an Encoder-Decoder network to estimate the five spectral parameters at the encoder stage (*ChromNN*), and a decoder network (*SpectraNN*) for reconstructing the full hyperspectral skin reflectance accurately, which dimension is reduced by Principal component analysis (PCA) to accelerate the training and testing time. We further validate the estimated parameters from RGB measurements by comparing hyperspectral skin appearance reconstructions to reference photographs under various spectral illumination conditions.

This proposed complete network requires strict measurements, but direct reconstruction of skin appearance is often more practical. To address this, we simplified the network to take a single image under D65 illumination as input. This simple version has a variety of applications, with one notable use being skin appearance editing. By adjusting the concentrations of melanin and hemoglobin, the network can modify skin tones, enabling effects such as tanning, lightening, flushing, and paling. Details on both networks and the skin appearance editing are discussed in Section 4.

These biophysical skin properties undergo significant changes with age. The concentrations of principal chromophores (melanin and hemoglobin) typically decrease with age, resulting in lighter and paler skin. A related study by [35] introduced a comprehensive biophysical skin model to simulate the optical effects of ageing by incorporating multiple spectral parameters. However, while this model offers a detailed approach, it is difficult for practical measurement and does not account for the structural changes in the face that occur with age.

In contrast, there exist many works focus on structural changes, such as producing wrinkles [58, 30, 53, 78, 82], while neglecting changes in skin properties. The main challenge for this task is the lack of a large-scale facial image dataset with paired ages spanning a long time. This leads to difficulty in re-ageing faces gradually from childhood to adulthood while preserving the identity at each stage. For example, some studies employ image-to-image translation techniques, such as age-conditioned Generative Adversarial Networks (GANs), which can produce satisfactory re-aging results, they often struggle to preserve identity, especially in transitions from childhood to adulthood [58, 30, 53].

Beyond 2D image editing, we aim to uplift this method into 3d shape reconstruction. While several methods have achieved high-quality 3D coarse shape reconstructions through model-free regression approaches [6], statistics facial models [11, 49, 60], and geometric neural networks [64, 81], they all ignore age as a controllable parameter, particularly to control head sizes, which is an important factor to reconstruct children’s heads. Moreover,

producing realistic 3D shape reconstructions that include age-related facial details, such as wrinkles, remains a challenging task.

To overcome these challenges, we propose a novel method for transforming facial age in both appearance and shape across different ages while preserving individual identity in Section 5. For facial appearance editing, we utilize a skin model to simulate skin tone changes by adjusting the concentrations of the two key chromophores, melanin and hemoglobin. Additionally, we employ an α -(de)blending diffusion network [31] to modify faces within a paired dataset spanning ages 10 to 80, by controlling the α parameter. Given the non-linear relationship between age and α , we introduce an age- α transformation method to produce more realistic re-aging results. This is the first approach to incorporate such a technique into facial re-aging within the field of computer graphics.

For the age-dependent transformation of facial geometry, we build upon part of the trained DECA [22] to capture accurate latent space information, such as camera, lighting, and facial geometry, from a single input image. A geometric neural network is then employed to integrate the age factor into the FLAME model, adjusting the coarse-scale facial geometry in accordance with age progression. This process is further enhanced by a lightweight refinement network that applies age-appropriate skin displacement to the coarse geometry. Both qualitative and quantitative evaluations demonstrate that our method outperforms current state-of-the-art techniques.

The overall structure of this thesis is outlined as follows:

- In Chapter 2, we present background knowledge and related works on diffusion approximation, biophysical skin modeling, facial re-ageing techniques, and facial shape reconstruction methods.
- In Chapter 3, we propose a novel spectral skin model with practical measurement spanning the full spectrum from 300nm to 1000nm. For parameter estimation and skin appearance reconstruction, we demonstrate a three-stage LUT search to speed

up the conventional methods.

- In Chapter 4, we introduce a neural network to estimate spectral parameters and reconstruct the full skin reflectance in both the invisible and visible domains. We also provide a simplified version with a single input image under D65 lighting conditions, as well as a significant application in skin appearance editing.
- In Chapter 5, we propose a re-ageing framework that transforms facial ages in both appearance and structure. This framework combines biophysical skin tone adjustments with the simulation of age-related structural changes, as well as models the coarse and detailed facial features with age-appropriate characteristics.
- In Chapter 6, we summarize the contributions of this thesis, discuss the limitations of our work, and introduce some possible future work.

Chapter 2

Background and Related Work

This chapter makes an introduction to the background and related work of biophysical skin modeling and re-ageing of facial skin appearance. Section 2.1 explains the background theory of the diffusion approximation. Based on Section 2.1, Section 2.2 introduces the biophysical skin modeling methods while Section 2.3 extends the skin model to the hyperspectral domain. Section 2.4 introduces the popular network architectures related to our work. Section 2.4 and Section 2.5 mainly focus on the facial re-ageing approaches in appearance and shape. Section 2.4 introduces the existing 2D age editing methods while Section 2.5 presents 3D facial modeling and reconstruction approaches.

2.1 Light Diffusion Approximation

This section introduces the theory and background of employing diffusion theory to model translucent materials, such as skin, which are most related to our work. We first introduce the dipole diffusion theory and the extended multipole diffusion theory in Section 2.1.1 and Section 2.1.2, respectively. Then we introduce the Kubelka-Munk theory for integrating multi-layered translucent materials in Section 2.1.3.

2.1.1 Light Dipole Diffusion

Before introducing diffusion theory, we first introduce a fundamental concept in computer graphics: the Bidirectional Reflectance Distribution Function (BRDF). It describes how light is reflected at a surface. More specifically, it provides a mathematical model of how light is reflected depending on the incoming light direction and the outgoing view direction. As shown in Figure 2.1, BRDF $f_r(\omega_i, \omega_o)$ is a four-dimensional function where ω_i and ω_o represent the incoming and outgoing light directions, respectively.

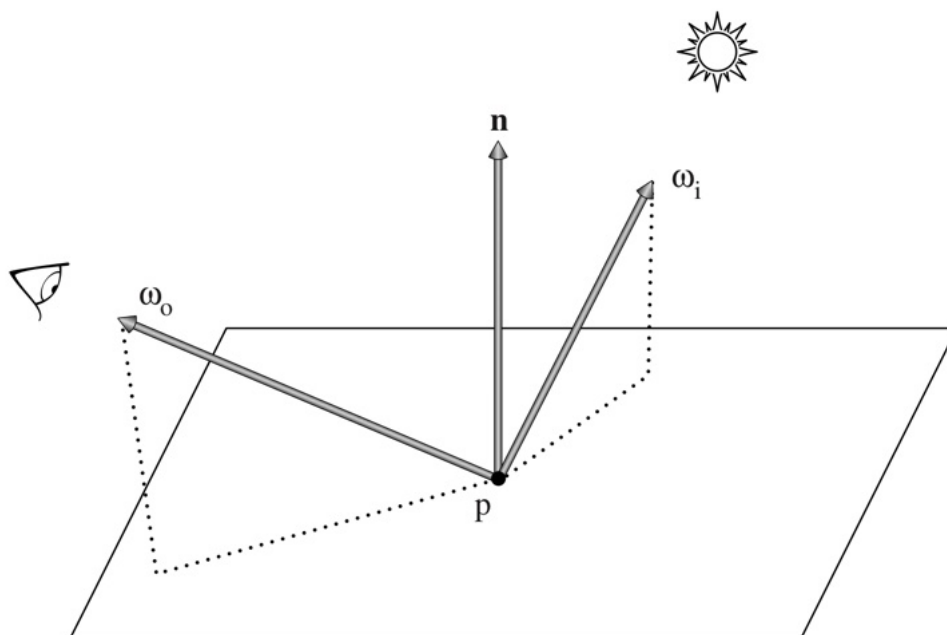


Figure 2.1: Bidirectional Reflectance Distribution Function (BRDF)

The BRDF only focuses on surface reflectance. However, when light is incident on a surface, it not only reflects but also scatters within the material. To account for this, the Bidirectional Scattering Surface Reflectance Distribution Function (BSSRDF) is introduced. It describes light scattering in materials with subsurface light transport, such as skin or leaf, where light penetrates the surface, scatters internally, and exits at a different location, displayed in Figure 2.3.

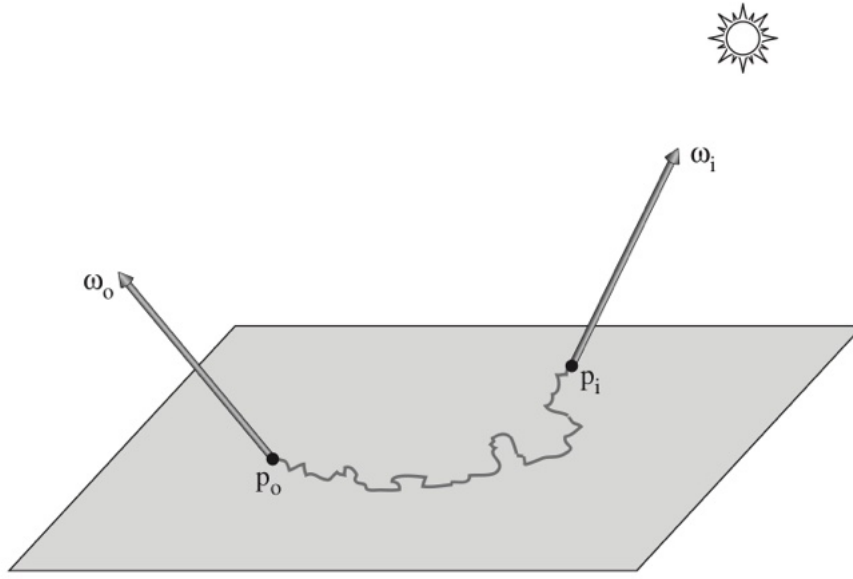


Figure 2.2: Bidirectional Scattering Surface Reflectance Distribution Function (BSSRDF)

The mathematical formulation is defined as

$$S(x_i, \vec{\omega}_i, x_o, \vec{\omega}_o) = \frac{dL(x_o, \vec{\omega}_o)}{d\Phi(x_i, \vec{\omega}_i)}, \quad (2.1)$$

where x_i and x_o are the incident and exitant positions of the incoming and outgoing light, respectively. ω_i and ω_o are the corresponding light directions. L is the outgoing radiance and Φ represents the incident flux.

Given a BSSRDF, the output radiance is computed

$$L_o(x_o, \vec{\omega}_o) = \int_A \int_{2\pi} S(x_i, \vec{\omega}_i, x_o, \vec{\omega}_o) L_i(x_i, \vec{\omega}_i) (\vec{n} \cdot \vec{\omega}_i) d\omega_i dA(x_i) \quad (2.2)$$

Follow the work of [38], we assume the medium is a semi-infinite homogeneous material, the incoming radiance will decrease exponentially with distance, which is referred to the reduced intensity:

$$L_{ri}(x_i + s\vec{\omega}_i, \vec{\omega}_i) = e^{-\sigma_t s} L_i(x_i, \vec{\omega}_i) \quad (2.3)$$

where σ_t is the extinction coefficient, which is the sum of the scattering coefficient σ_s and the absorption coefficient σ_a , respectively.

The first-order scattering of this reduced intensity L_{ri} is

$$Q(x, \vec{\omega}) = \sigma_s \int_{4\pi} p(\vec{\omega}, \vec{\omega}') L_{ri}(x, \vec{\omega}') d\omega' \quad (2.4)$$

where $p(\vec{\omega}, \vec{\omega}')$ is the normalized phase function. Here, we define the 0th-order and 1st-order source term as $Q_0(x) = \int_{4\pi} Q(x, \vec{\omega}) d\omega$ and $Q_1(x) = \int_{4\pi} Q(x, \vec{\omega}) \vec{\omega} d\omega$.

The key idea behind the diffusion approximation is that the light distribution in highly scattering material tends to be isotropic. Depending on this, the radiance is approximated by

$$L(x, \vec{\omega}) = \frac{1}{4\pi} \phi(x) + \frac{3}{4\pi} \vec{\omega} \cdot \vec{E}(x) \quad (2.5)$$

where ϕ is the fluence and $\vec{E}$ is the vector flux [37]. By following the work of [37], we can obtain

$$\vec{\nabla} \phi(x) = -3\sigma'_t \vec{E}(x) + \vec{Q}_1(x) \quad (2.6)$$

where the reduced extinction coefficient $\sigma'_t = \sigma'_s + \sigma_a$. As we assume the sources are isotropic, therefore, we can obtain

$$\vec{E}(x) = -D \vec{\nabla} \phi(x) \quad (2.7)$$

where $D = \frac{1}{3\sigma'_t}$ is the diffuse constant.

The simple solution of a single isotropic point light source in the infinite medium is

$$\phi(x) = \frac{\Phi}{4\pi D} \frac{e^{-\sigma_{tr}r(x)}}{r(x)} \quad (2.8)$$

In this equation, Φ is the power of the light source. $\sigma_{tr} = \sqrt{3\sigma_a\sigma'_t}$ is the effective transport coefficient. r is the distance to the point source.

Then, we focus on the derivation of the dipole model from the boundary conditions to the diffusion equation. Specifically, at the surface of a material, any light that escapes is assumed to never return. Therefore, the total downward diffusion radiance at the surface into the material (+z direction) is equal to the internally reflected upward diffuse radiance

$$\int_{\Omega+} L_d(r, \vec{\omega})(-\vec{n} \cdot \vec{\omega})d\vec{\omega} = F_{dr} \int_{\Omega-} L_d(r, \vec{\omega})(\vec{n} \cdot \vec{\omega})d\vec{\omega} \quad \text{at } z = 0, \quad (2.9)$$

where $\Omega+$ and $\Omega-$ indicate integration over the positive and negative hemispheres, and $\vec{n}$ is the normal at the material surface. F_{dr} is a diffuse Fresnel term that approximates the internal diffuse reflectivity of the slab. Substitution of Equation [2.5](#) into Equation [2.9](#) and simplifying gives [37](#)

$$\phi(r) - 2AD \frac{\partial \phi(r)}{\partial z} = 0 \quad \text{at } z = 0, \quad (2.10)$$

where D is defined as above. Note that this expression is an approximate result, but one widely held as accurate for highly scattering materials. In Equation [2.10](#), A is defined as

$$A = (1 + F_{dr})/(1 - F_{dr}), \quad (2.11)$$

and represents the change in fluence due to internal reflection at the surface. The diffuse

Fresnel reflectance F_{dr} can be approximated by the following polynomial expansions

$$F_{dr} \simeq \begin{cases} -0.4399 + \frac{0.7099}{\eta} - \frac{3319}{\eta^2} + \frac{0636}{\eta^3}, & \eta < 1 \\ -\frac{1.4399}{\eta^2} + \frac{0.7099}{\eta} + 0.6681 + 0.0636\eta, & \eta > 1 \end{cases} \quad (2.12)$$

where η is the ratio of indices of refraction.

We further can obtain the diffuse BSSRDF, R_d , as

$$R_d(r) = -D \frac{(\vec{n} \cdot \vec{\nabla} \phi(x))(x_s)}{d\Phi_i(x_i)} \quad (2.13)$$

where $r = \|x_s - x_i\|$.

Built upon this, Jensen et al. [38] propose that the BSSRDF can be approximated using diffusion theory using two-point sources. In detail, this method consists of two point sources located near the surface, shown in Figure 2.3. The positive real light source is positioned under the surface at the depth z_b while the negative virtual light source is located above the surface at the distance $z_v = z_r + 4AD$. r is the distance between the incident (x_i) and exitant (x_o) points.

The diffuse reflectance profile R is approximated as

$$R(r) = \frac{\alpha' z_r (1 + \sigma_{tr} d_r) e^{-\sigma_{tr} d_r}}{4\pi d_r^3} - \frac{\alpha' z_v (1 + \sigma_{tr} d_v) e^{-\sigma_{tr} d_v}}{4\pi d_v^3}, \quad (2.14)$$

where

- $\alpha' = \sigma'_s / \sigma'_t$ is the reduced albedo.
- $z_r = 1/\sigma'_t$ and $z_v = (1 + 4A/3)/\sigma'_t$ are the z-coordinates of the positive and negative sources relative to the surface at $z = 0$.
- $d_r = \sqrt{r^2 + z_r^2}$ and $d_v = \sqrt{r^2 + z_v^2}$ are the distances to sources from a given point on the surface of the object.

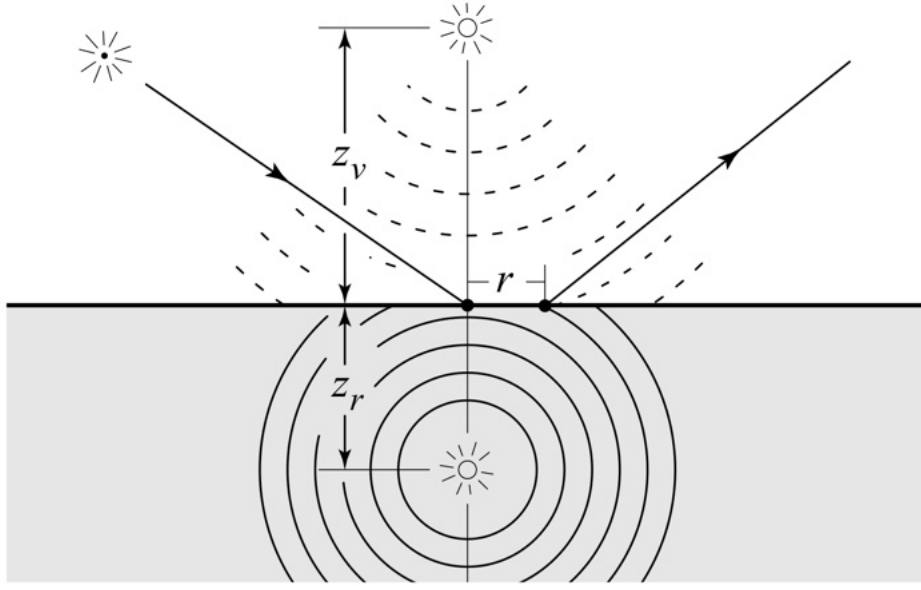


Figure 2.3: Dipole Diffusion approximation

In order to simplify the computation, Donner et al. [20] define the position of the positive point light under the boundary as one mean free path $z_r = l = 1/\sigma'_t$. The middle distance between two light sources refers to $z_b = 2AD$, thus the negative point light is located at $z_v = 2z_b + l = (1 + 4A/3)/\sigma'_t$. The new diagram for dipole diffusion theory is shown in Figure 2.4 (a).

After obtaining the diffuse reflectance R_d , the diffuse scattering can be approximated as

$$S_d(x_i, \vec{\omega}_i; x_o, \vec{\omega}_o) = \frac{1}{\pi} F_t(x_i, \vec{\omega}_i) R(\|x_i - x_o\|_2) F_t(x_o, \vec{\omega}_o). \quad (2.15)$$

where F_t is the Fresnel transmittance.

2.1.2 Multipole Diffusion Theory

This work was extended to multipole diffusion theory to account for multi-layered translucent materials, such as human skin [19]. Unlike the dipole diffusion theory with only two

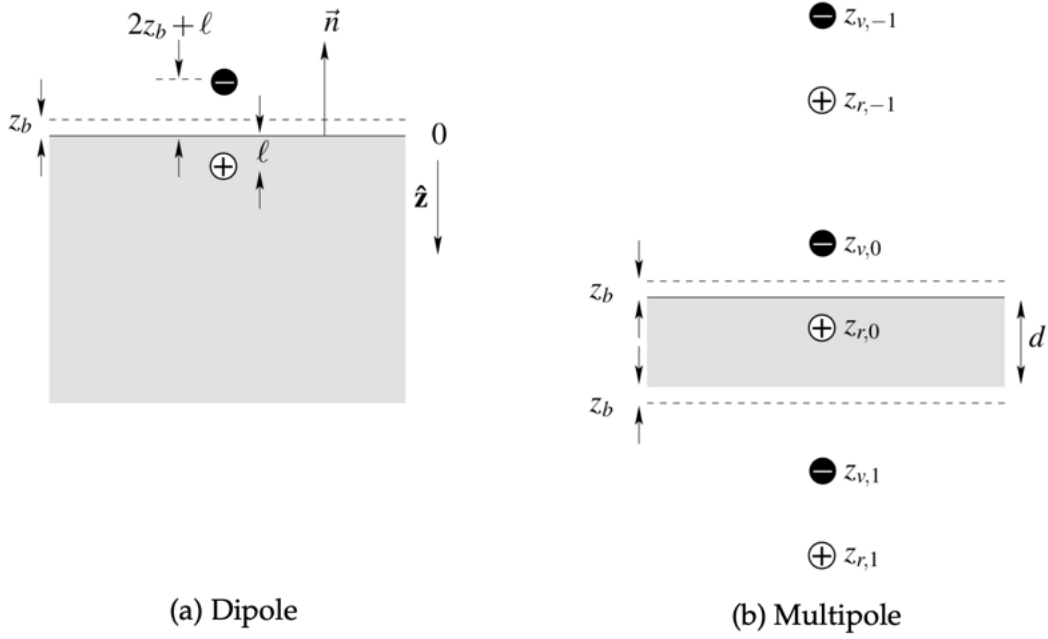


Figure 2.4: Comparison of dipole and multiple diffusion theory.

point lights, we assume there are multiple pairs of point lights in the multiple diffusion theory, shown in Figure 2.4 (b).

Similar to the boundary condition of the dipole diffusion, we define a boundary condition of each thin layer with thickness d as

$$\int_{\Omega^-} L_d(r, \vec{\omega})(\vec{n} \cdot \vec{\omega}) d\vec{\omega} = F_{dr} \int_{\Omega^+} L_d(r, \vec{\omega})(-\vec{n} \cdot \vec{\omega}) d\vec{\omega} \text{ at } z = d, \quad (2.16)$$

This equation can be simplified as

$$\phi(r) + 2AD \frac{\partial \phi(r)}{\partial z} = 0 \text{ at } z = d, \quad (2.17)$$

where we make the assumption that the non-scattering mediums above and below the slab have the same index of refraction.

Each pair of point lights are located at

$$\begin{aligned} z_{r,i} &= 2i(d + 2z_b) + l \\ z_{v,i} &= 2i(d + 2z_b) - l - 2z_b, \quad i = -n, \dots, n, \end{aligned} \quad (2.18)$$

where $2n + 1$ is the number of dipoles, d is the slab thickness, and $z_b = 2AD$ is the extrapolation distance.

Here, we simply sum the contributions of each pair to obtain the final reflectance

$$R(r) = \sum_{i=-n}^n \frac{\alpha' z_{r,i} (1 + \sigma_{tr} d_{r,i}) e^{-\sigma_{tr} d_{r,i}}}{4\pi d_{r,i}^3} - \frac{\alpha' z_{v,i} (1 + \sigma_{tr} d_{v,i}) e^{-\sigma_{tr} d_{v,i}}}{4\pi d_{v,i}^3}, \quad (2.19)$$

where $d_{r,i} = \sqrt{r^2 + z_{r,i}^2}$ and $d_{v,i} = \sqrt{r^2 + z_{v,i}^2}$ are the distances to dipole sources from a given point on the surface of the object.

Similarly, the diffuse transmittance can be computed as

$$T(r) = \sum_{i=-n}^n \frac{\alpha' (d - z_{r,i}) (1 + \sigma_{tr} d_{r,i}) e^{-\sigma_{tr} d_{r,i}}}{4\pi d_{r,i}^3} - \frac{\alpha' (d - z_{v,i}) (1 + \sigma_{tr} d_{v,i}) e^{-\sigma_{tr} d_{v,i}}}{4\pi d_{v,i}^3}, \quad (2.20)$$

2.1.3 Diffusion in multi-layered materials

After obtaining the reflectance and transmittance of each layer, we can compute the reflectance and transmittance profiles of the multi-layered materials by Kubelka-Munk theory in the Frequency space [6].

Assume that all interactions between two layers are due to multiple scattering, we can compute the profile of T_{12} of the light transmitted through two slabs with transmittance profiles T_1 and T_2 by convolving the profiles

$$T_{12}(r) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T_1(r') T_2(r'') dx' dy' = T_1(r) * T_2(r), \quad (2.21)$$

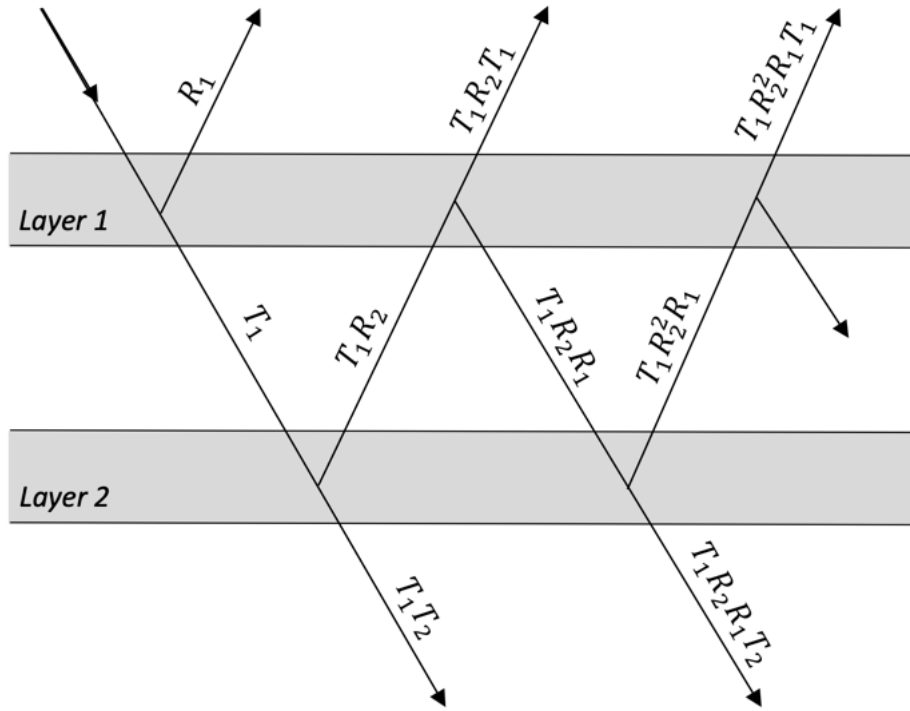


Figure 2.5: Light transport between the epidermis and dermis computed with Kulbeka-Munk theory in Frequency space.

where $r'' = \sqrt{(x - x')^2 + (y - y')^2}$. However, this equation cannot represent all the transmittance in light transport. When a light beam enters the first layer, some light will be absorbed by the first layer, some will transmit to the second layer, and others will reflect back to the air, which will be continued permanently by the unabsorbed light (see Figure 2.5).

As can be seen, the transmittance is

$$T_{12} = T_1 * T_2 + T_1 * R_2 * R_1 * T_2 + T_1 * R_2 * R_1 * R_2 * R_1 * T_2 + \dots \quad (2.22)$$

Since the convolution in real space can be converted into the multiplication in frequency space by employing Fourier transform, the Equation [2.22](#) can be converted to

$$\begin{aligned} T_{12} &= T_1 T_2 + T_1 R_2 R_1 T_2 + T_1 R_2 R_1 R_2 R_1 T_2 + \dots \\ &= T_1 T_2 (1 + R_2 R_1 + (R_2 R_1)^2 + (R_2 R_1)^3 + \dots) \end{aligned} \quad (2.23)$$

If $\mathcal{R}_2\mathcal{R}_1 < 1$, the this equation can be simplified as

$$\mathcal{T}_{12} = \frac{\mathcal{T}_1\mathcal{T}_2}{1 - \mathcal{R}_2\mathcal{R}_1} \quad (2.24)$$

Similarly, the reflectance is

$$R_{12} = R_1 + T_1 * R_2 * T_1 + T_1 * R_2 * R_1 * R_2 * T_1 + \dots \quad (2.25)$$

This equation can be converted it into frequency space, shown as follows:

$$\begin{aligned} \mathcal{R}_{12} &= \mathcal{R}_1 + \mathcal{T}_1\mathcal{R}_2\mathcal{T}_1 + \mathcal{T}_1\mathcal{R}_2\mathcal{R}_1\mathcal{R}_2\mathcal{T}_1 + \dots \\ &= \mathcal{R}_1 + \mathcal{T}_1\mathcal{R}_2\mathcal{T}_1(1 + \mathcal{R}_1\mathcal{R}_2 + (\mathcal{R}_1\mathcal{R}_2)^2 + \dots) \end{aligned} \quad (2.26)$$

Then, we can obtain the final reflectance

$$\mathcal{R}_{12} = \mathcal{R}_1 + \frac{\mathcal{T}_1\mathcal{R}_2\mathcal{T}_1}{1 - \mathcal{R}_2\mathcal{R}_1} \quad (2.27)$$

Finally, we convert these two equations from the frequency space to the real space by applying the inverse Fourier transform.

This method can be extended to multiple layers by recursively substituting Equations [2.24](#) and [2.27](#) for R_1 or T_1 , and then recalculating the formulas. The real-space reflectance and transmittance profiles for a multi-layered material are obtained by taking the inverse Fourier transform of the total frequency response.

2.2 Methods of Biophysical Skin Modeling

Following the diffusion theory, we introduce a series of the current biophysical skin appearance modeling methods and the corresponding mathematical formulations. This section

also explores extension to the hyperspectral appearance modeling, covering a wide spectral range including UV and IR, along with practical measurement techniques.

2.2.1 Biophysical Skin Modeling

Built upon the dipole and multipole diffusion theory, Donner et al. proposed a biophysical skin model to simulate realistic skin appearance [20]. This skin model consists of two layers: epidermis and dermis, shown in Figure 2.6.

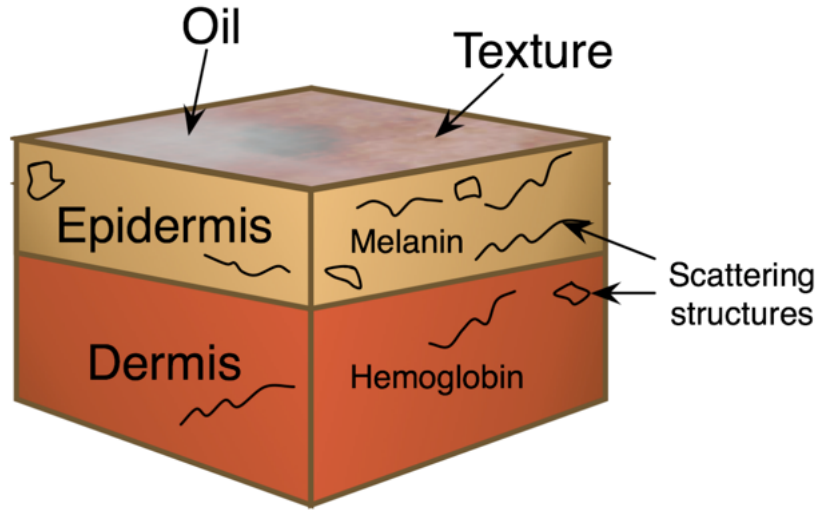


Figure 2.6: Two-layer skin model proposed by [20]

For the epidermis, they assumed it is a multi-layered medium that can be approximated by the multipole diffusion theory. In this layer, the significant chromophore that affects the skin tone is two types of melanin, including eumelanin and pheomelanin. The eumelanin is for brown or black color while the pheomelanin is for lighter skin tone. The absorption of the epidermal layer is computed as

$$\sigma_a^{epi} = C_m(\beta_m\sigma_a^{em}(\lambda) + (1 - \beta_m)\sigma_a^{pm}(\lambda)) + (1 - C_m)\sigma_a^{baseline} \quad (2.28)$$

where

- C_m is the melanin concentration,
- β_m is the blend fraction between eumelanin and pheomelanin,
- σ_a^{em} is the absorption coefficient of eumelanin,
- σ_a^{pm} is the absorption coefficient of pheomelanin,
- $\sigma_a^{baseline}$ is the absorption coefficient of other skin tissue.
- λ refers to the light wavelength in nanometers.

The absorption coefficients of eumelanin and pheomelanin are defined as

$$\sigma_a^{em}(\lambda) = 6.6 \times 10^{10} \times \lambda^{-3.33} \text{ mm}^{-1}, \quad (2.29)$$

$$\sigma_a^{pm}(\lambda) = 2.9 \times 10^{14} \times \lambda^{-4.75} \text{ mm}^{-1} \quad (2.30)$$

The absorption of the baseline skin tissue is

$$\sigma_a^{base}(\lambda) = 0.0244 + 8.53 \times e^{-(\lambda-154)/66.2} \text{ mm}^{-1} \quad (2.31)$$

For the dermis, they assumed it is a semi-infinite transparent medium that can be simulated using dipole diffusion theory. In this layer, the significant chromophore is the hemoglobin. This also consists of two types: oxygenated and deoxygenated hemoglobin. These two chromophores mainly affect the redness of the skin tones. The absorption of the dermal layer is computed as

$$\sigma_a^{derm} = C_h(\gamma\sigma_a^{oxy}(\lambda) + (1 - \gamma)\sigma_a^{deoxy}(\lambda)) + (1 - C_h)\sigma_a^{baseline} \quad (2.32)$$

where

- C_h is the hemoglobin concentration,
- γ is the blend fraction between oxygenated and deoxygenated hemoglobin fixed at 0.7,
- σ_a^{oxy} is the absorption coefficient of oxygenated hemoglobin,
- σ_a^{deoxy} is the absorption coefficient of deoxygenated hemoglobin.

The absorption coefficients of these chromophores are shown in the following figure.

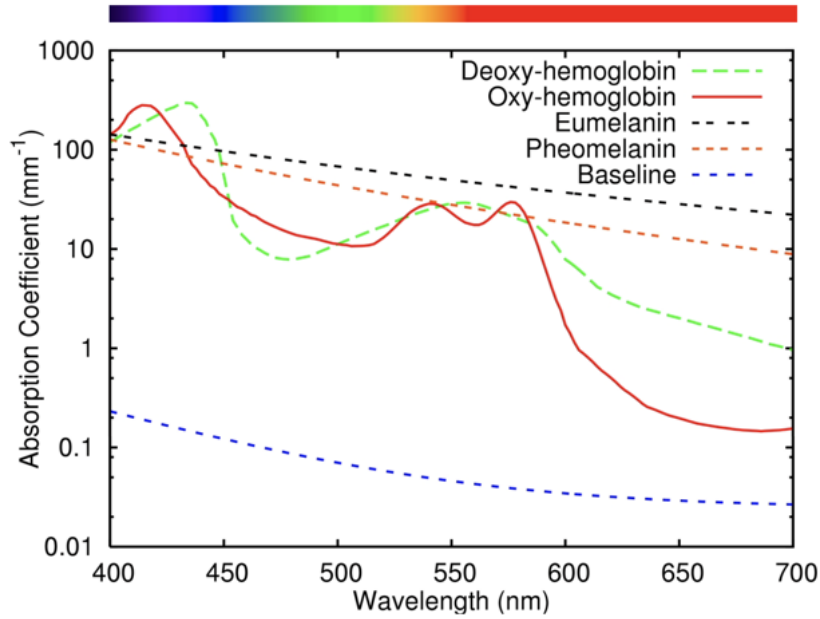


Figure 2.7: Spectral absorption coefficients of eumelanin, pheomelanin, oxygenated hemoglobin, deoxygenated hemoglobin, and baseline. This figure is borrowed from [20].

The reduced scattering coefficient of the epidermis is

$$\sigma_s'^{epi}(\lambda) = 14.74 \times \lambda^{-0.22} + 2.2 \times 10^{11} \times \lambda^{-4} \text{ mm}^{-1} \quad (2.33)$$

The scattering coefficient of the dermis is approximated as half of epidermal scatter-

ing [20] [5], shown below:

$$\begin{aligned}\sigma_s'^{derm}(\lambda) &= 0.5 \times \sigma_s'^{epi}(\lambda) \\ &= 7.37 \times \lambda^{-0.22} + 1.1 \times 10^{11} \times \lambda^{-4} \text{ mm}^{-1}\end{aligned}\tag{2.34}$$

Once computed, the spectral absorption and reduced scattering coefficients σ_a^{derm} , σ_a^{epi} , $\sigma_s'^{epi}$ and $\sigma_s'^{derm}$ can be used as inputs for the dipole and multipole models to generate layer-specific diffusion profiles. These profiles are then convolved using the Kubelka-Munk equation (Equation 2.27) to compute the total diffuse reflectance of human skin, then generated LookUp Tables (LUTs) with SPDs of different illuminations.

It is important to note that the reflectance and transmittance derived from diffusion theory, as well as the convolved expressions for multi-layered materials, are functions of the radial distance r . To convert these diffuse distributions, represented as $R(r)$, into the more conventional form of spectral material reflectance R_d , we need to integrate the functions as follows:

$$R_d = \int_0^\infty rR(r)dr.\tag{2.35}$$

The following integration is performed for each sampled wavelength of the visible light. The final spectral reflectance can then be used to generate diffuse albedo color of skin for a given set of biophysical parameters.

This model was further extended and simplified by Jimenez et al. [39] to make it suitable for practical measurement and prediction of skin color change during facial skin appearance simulation. This model only employs two free parameters to generate a LookUp Table, and starts to estimate the parameters and reconstruct the skin albedo by LUT search.

Recently, Gitlina et al. [23] employed a more complete spectral model of [39], with four free parameters to more accurately reconstruct skin appearance using measurements in a multispectral LED sphere or a hand-held on-the-shelf device Antera3D. Besides employing LUT searching, they also proposed a cascaded neural network for model fitting and RGB albedo reconstruction. Each sub-network predicts one chromophore parameter in their model. However, this network architecture focuses on reconstructing RGB facial appearance from the estimated chromophores, under a specific illumination spectrum (e.g., D65’).

The spectral skin chromophores modeled in these works primarily focus on melanin and hemoglobin concentrations. However, other blood-borne pigments, such as beta-carotene and bilirubin, also contribute significantly to skin color [45, 21, 13]. Krishnaswamy et al. [45] introduced a parametric five-layer skin model, BioSpec, to approximate light interaction within human skin. Later, Donner et al. [21] proposed a more detailed layered heterogeneous reflectance model that incorporated lateral inter-scattering of light between skin layers, using six parameters and additional inter-layer absorption to account for features such as veins and tattoos [21].

While these models can simulate highly realistic skin appearance, their complexity makes practical measurements challenging. The high number of parameters and intricate light interaction modeling require significant computational resources and detailed inputs, which can limit their usability in real-time applications or large-scale studies.

2.2.2 Hyperspectral Skin Modeling

Prolonged exposure to UV light can lead to deep skin damage, resulting in gradual changes to facial appearance and overall skin health. One of the early indicators of such damage is the appearance of pigmentation irregularities, such as freckles, which are particularly useful for detecting signs of skin aging [7, 36]. Identifying these irregularities not only helps

with monitoring the aging process but also plays a crucial role in protecting the skin from further photo-damage, which can elevate the risk of conditions like skin cancer [79]. Early prediction and detection of these changes are essential for taking preventive measures to maintain skin health and reduce UV-related damage.

Beyond pigmentation, skin hydration is another key factor that reflects skin health and vitality [13]. Appropriate hydration levels are important for maintaining skin elasticity and appearance, and deviations in hydration can indicate various dermatological issues. Since water exhibits strong absorption in the infrared (IR) spectrum, IR photography has proven to be a valuable tool for assessing skin hydration levels [4]. By capturing IR images, it is possible to gain insights into the water content within the skin layers, making it an effective non-invasive method for evaluating skin health and diagnosing dehydration or other hydration-related conditions. At the same time, monitoring both pigmentation and hydration offers a comprehensive approach to understanding and maintaining skin health.

The spectral parameters, such as melanin and hemoglobin, are significant absorbers in both the visible and invisible domains varying from the UVA to the NIR domain. Chen et al. [13] are the first to introduce a hyperspectral skin appearance model in computer graphics. This model incorporates a large set of skin chromophores to comprehensively forward simulate the light interacting with human skin from the UV to IR domain. Such a complex model is challenging for practical measurement, therefore they employed some synthetic images to evaluate this method qualitatively. Additionally, this work provides a dataset for spectral parameters in the hyperspectral domain, which are later borrowed in our model.

Recently, Aliaga et al. [2] proposed an accurate skin model in the hyperspectral domain based on the Monte-Carlo light simulation. They also presented a network that directly predicts spectral skin reflectance in the Vis and NIR range. However, they ignored the practical measurement and also lacked the validations for reconstructions in the full spec-

trum.

To address these issues, we determinate to follow the line of previous diffusion-based biophysical skin appearance models. We employ a practical hyperspectral skin model with five free parameters to balance model complexity and practicality for measurements. We also propose a novel Encoder-Decoder network to predict chromophore maps and reconstruct full spectral skin reflectance given practical RGB measurements. These predicted skin reflectances can be visualized by various uniform illuminations in the visible domain and also the invisible (UVA and NIR) domains. Please refer to Section 3 and Section 4 for more details.

2.2.3 Spectral Skin Appearance Measurement

In this part, we make a brief introduction to the recent practical measurement approaches for spectral skin models. These setups capture photographs of skin appearance provided as measurements and qualitative comparisons.

Jensen et al. [38] introduced a simple method for measuring the RGB optical properties of materials, including two skin samples. In their approach, they illuminated the material’s surface with a tightly focused beam of white light and used a color camera to capture photographs, allowing them to analyze the radiant exitance across the surface. Tsumura et al. [74] applied independent component analysis (ICA) to photographic data to estimate the distributions of melanin and hemoglobin, assuming that these chromophores independently influence skin color, while neglecting the impact of light scattering and depth variation within the skin layers.

Donner et al. [21] started to utilize multispectral imaging of skin patches illuminated by a broadband flash illumination along with nine narrow spectral bands to measure spectral skin reflectance and derive parameters for their skin model. Jimenez et al. [39]

employed a non-contact SIAscope™ system [15], which used cross-polarized flashes to measure hemoglobin and melanin concentrations in facial skin. In contrast, Alotaibi and Smith [3] used a single LED illumination spectrum from photographs for similar measurements. Gotardo et al. [26] introduced a multi-view setup with static illumination to estimate dynamic skin albedo, while Park et al. [59] used multiplexed spectral illumination techniques to estimate skin reflectance more accurately.

More recently, Gitlina et al. [23] introduced a measurement technique utilizing two complementary broad and narrow-band spectral illumination conditions with a multispectral LED sphere. Additionally, they proposed a practical method for capturing skin patch measurements using the commercially available Miravex Antera 3D camera, a dermatological imaging device. Building on these advancements, Aliaga et al. [2] employed RGB albedo measurements of facial skin to estimate chromophore parameters for their skin model, which is based on Monte Carlo simulations.

Our method builds on these approaches by utilizing photographs taken under complementary broad and narrow-band illuminations, produced by a setup with four LCD screens and a centrally positioned camera. This configuration enables us to efficiently estimate parameters for a diffusion-based hyperspectral skin reflectance model, providing a practical and accurate solution for analyzing skin reflectance.

2.3 Reconstruction of Hyperspectral Skin Reflectance

2.3.1 Visualization of the full spectral

After acquiring the spectral skin profiles, the next step is to visualize these profiles by applying specific illumination and color matching functions. This process involves converting the spectral data from the CIE XYZ color space to the sRGB color space [65].

The equations are shown as below:

$$\begin{aligned}
 X &= \frac{\sum_{\lambda} S(\lambda)R(\lambda)\bar{x}(\lambda)\Delta\lambda}{N} \\
 Y &= \frac{\sum_{\lambda} S(\lambda)R(\lambda)\bar{y}(\lambda)\Delta\lambda}{N} \\
 Z &= \frac{\sum_{\lambda} S(\lambda)R(\lambda)\bar{z}(\lambda)\Delta\lambda}{N} \\
 N &= \sum_{\lambda} S(\lambda)\bar{y}(\lambda)\Delta\lambda
 \end{aligned} \tag{2.36}$$

where:

- X , Y , and Z represent tristimulus values,
- λ represents wavelength,
- $S(\lambda)$ represents spectral power distribution,
- $R(\lambda)$ represents skin reflectance,
- $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ represent color matching functions (see Figure 2.8).

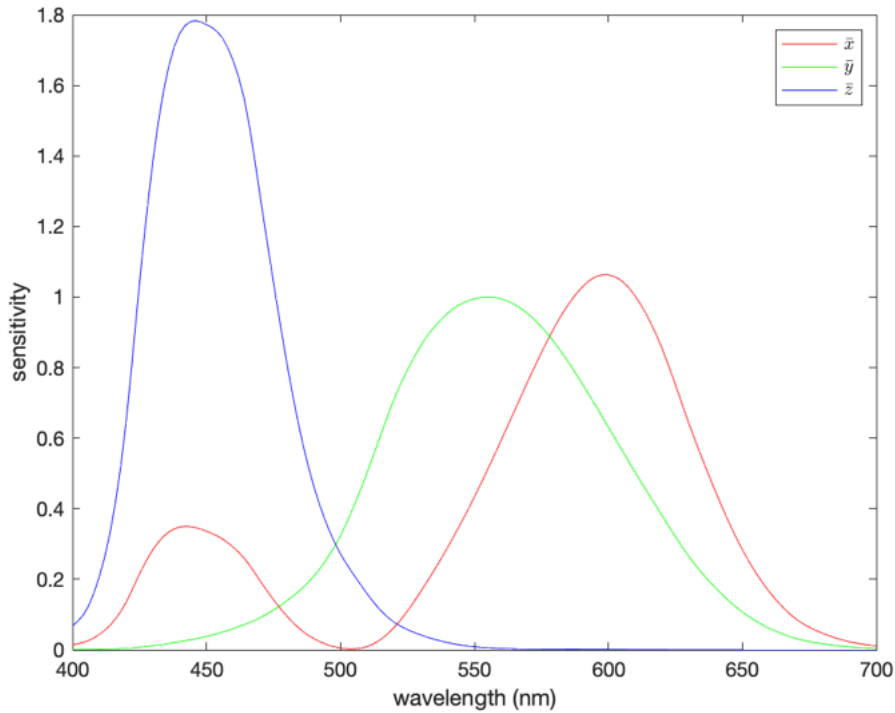


Figure 2.8: The CIE 1931 XYZ color matching functions.

The process of converting to sRGB color space contains two steps. First, we need to transform the XYZ color space into the linear RGB color space, described in Equation [2.37](#).

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.2404542 & -1.5371385 & -0.4985314 \\ -0.9692660 & 1.8760108 & 0.0415560 \\ 0.0556434 & -0.2040259 & 1.0572252 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (2.37)$$

The next step is to apply sRGB companding to these values, converting them into a non-linear sRGB color system using Equation [2.38](#).

$$V = \begin{cases} 12.92v & \text{if } v \leq 0.0031308 \\ 1.055v^{1/2.4} - 0.055 & \text{otherwise} \end{cases} \quad (2.38)$$

where V is an individual red, green or blue color channel. Gamma correction is applied similarly to each component, and the result in sRGB color space is represented as $[R_{sRGB}, G_{sRGB}, B_{sRGB}]$.

Figure 2.9 illustrates the visualized lookup tables under two illuminations: D65' illumination in the top row and cold white illumination from the light stage in the bottom row. By comparing these two rows, we can clearly observe how different lighting affects skin tones.

In each patch, the color variations are primarily influenced by changes in the parameters C_m (x-axis) and C_{hd} (y-axis), while keeping the values of β_m and C_{he} constant. Notably, the concentration of melanin is directly related to the pigmentation level of the skin, whereas the concentration of hemoglobin plays a significant role in determining the skin's redness.

The parameter β_m represents the blend fraction between eumelanin and pheomelanin. As mentioned earlier, eumelanin contributes to darker skin tones, while pheomelanin is

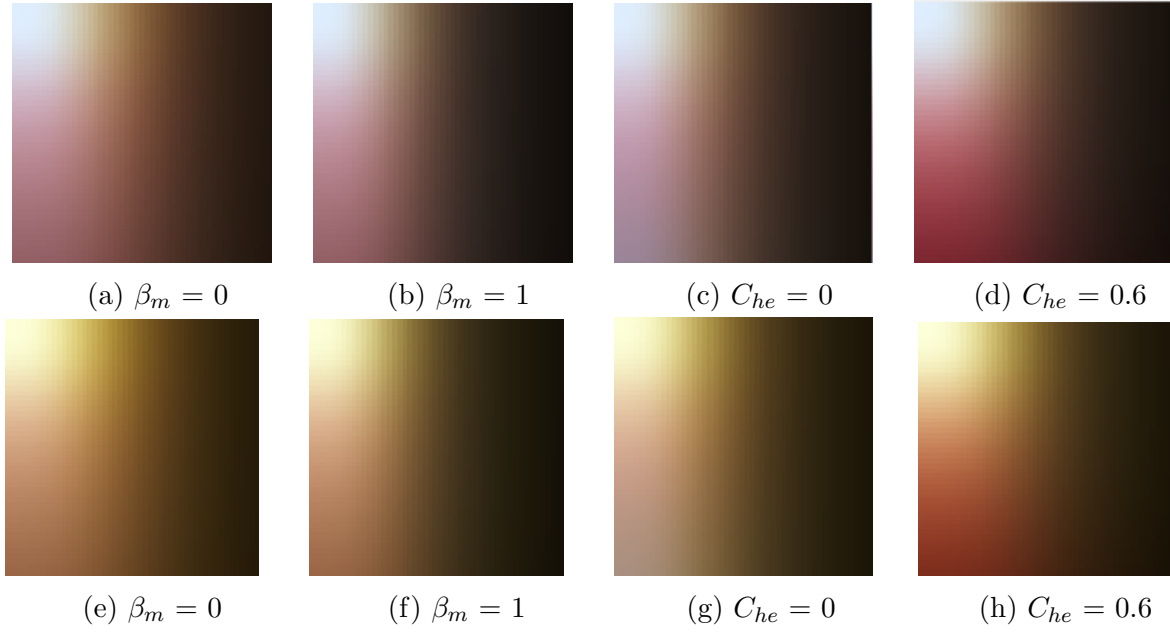


Figure 2.9: Comparison of the 2D lookup tables of spectral skin model with C_m (x-axis) and C_h (y-axis) under D65' illumination (a-d) and cold white illumination (e-h) and variation due to β_m ((a,b),(e,f)) and C_{he} ((c,d),(g,h)), respectively.

associated with lighter shades, such as blonde and red. Therefore, when β_m increases, the skin color tends to appear darker, as illustrated by the comparisons between patches (a) and (b) or patches (e) and (f). On the other hand, C_{he} denotes the concentration of hemoglobin in the epidermis. The differences between patches (c) and (d) or patches (g) and (h) highlight that a higher concentration of C_{he} results in a more obvious reddish in the skin. These patches demonstrate how variations in these four spectral parameters affect the overall appearance of skin tones.

Additionally, we can not only visualize the lookup tables but also reconstruct the full skin reflectance of a subject from the UVA domain to the NIR domain. Figure 2.10 shows the reconstruction of a subject under white (a), red (b), green (c), and blue (d) illuminations in the visible domain. We kindly remind the readers that the visualization in the visible spectrum can be implemented with any uniform illumination.

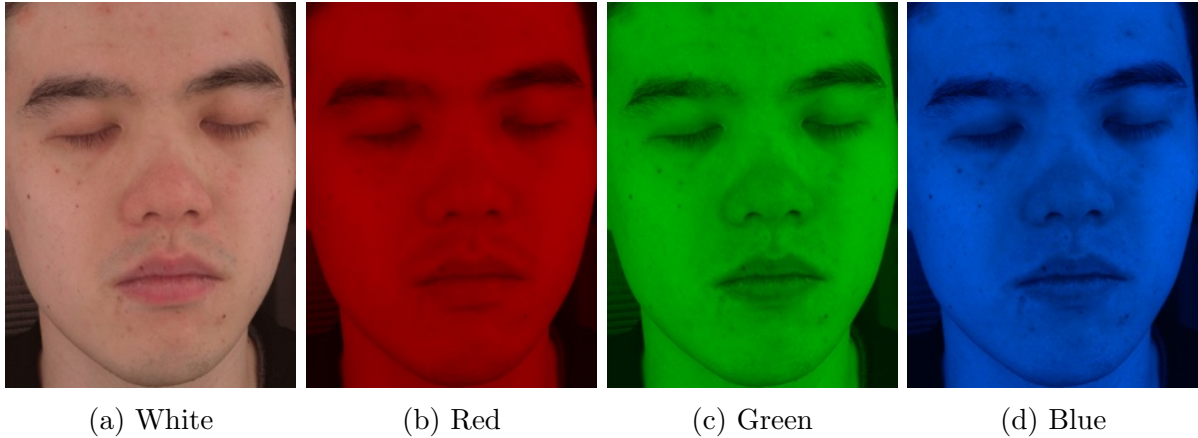


Figure 2.10: Visualization of a subject under white (a), red (b), green (c), and blue (d) uniform illuminations in the visible domain.

2.3.2 Color Correction

In practical measurement, various setups are employed to capture photos. While the standard color matching functions (CMFs) are typically used to represent human color vision, they are not suitable for visualization in this measurement because cameras perceive color differently than the human eye.

Instead of using CMFs, we rely on the camera sensitivity curves, which describe how the camera’s sensors respond to different wavelengths of light across the visible spectrum. These sensitivity curves are essential for accurately simulating how the camera captures color information. Figure 2.11 illustrates the sensitivity curves of the camera employed in our monitor setup. As shown in this figure, the curves differ significantly from the standard CMFs (see Figure 2.8).

Therefore, directly using the camera sensitivity curves to replace the color matching function would lead to significant color inaccuracies. To address this issue, we need to apply color correction on both the captured photos and lookup tables. The procedure of the color correction is as follows:

- we first capture a photo of a color chart, read the raw file of the photo, and convert

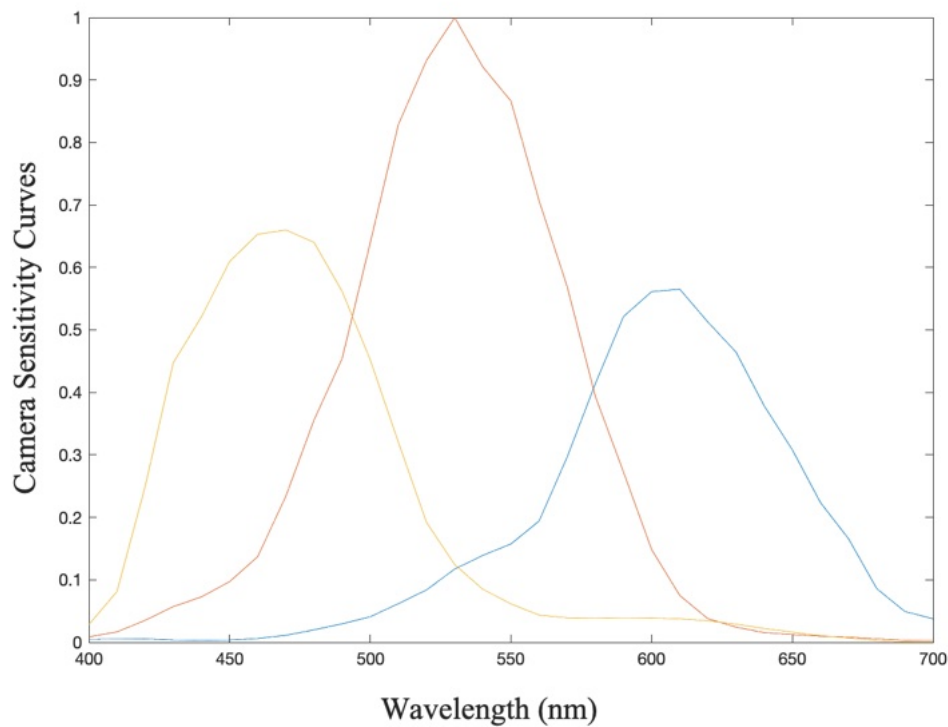


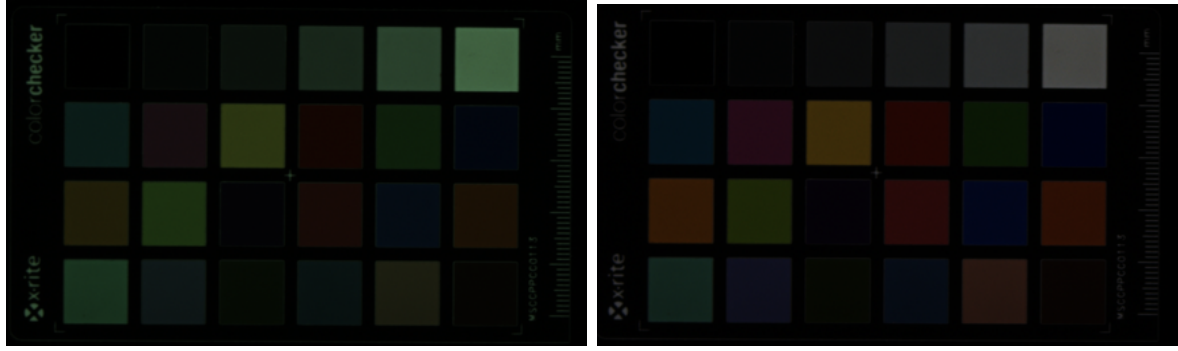
Figure 2.11: Camera Sensitivity Curves

it to a tiff file without white balance.

- Then, we compute the LUT with the standard color matching functions and the spectral power distribution of the corresponding illumination.
- We further calculate the color correct matrix by comparing the same patch of the photo of the color chart and the lookup table.
- At the end, we use this matrix factor to color correct the photo of subjects and lookup tables.

The photo of the color chart without color correction and the corrected result are shown in Figure 2.12. As shown, the uncorrected color chart under white illumination is greenish while the corrected one displays the normal color.

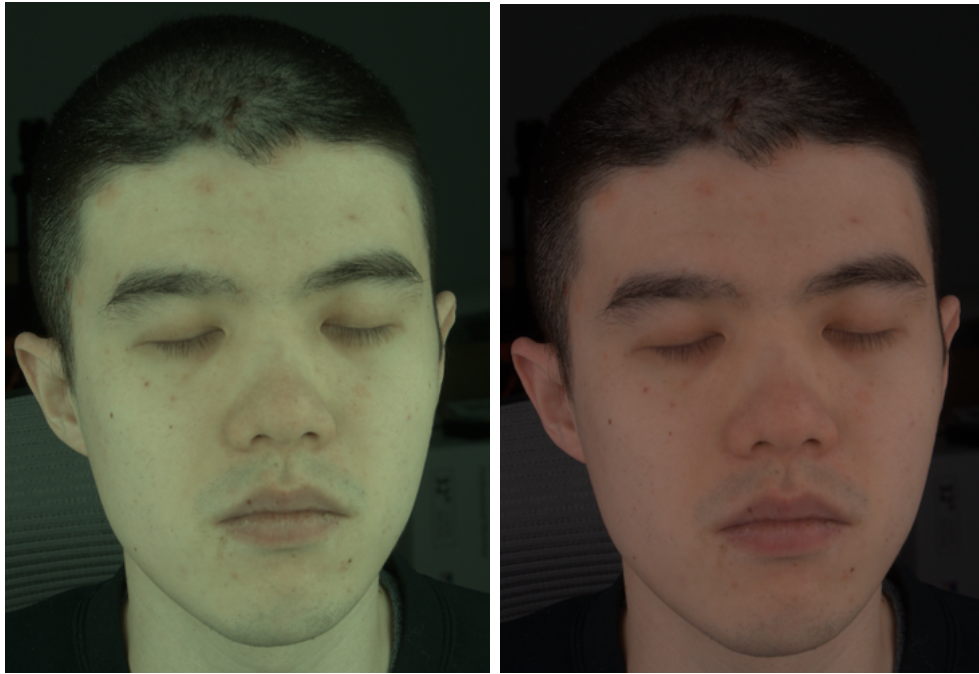
Except for the color chart, we can also apply this matrix to the photos of subjects shown in Figure 2.13. Except for converting photos from the camera space to the color space, we can



(a) Uncorrected color chart (b) Corrected color chart

Figure 2.12: Comparison between the uncorrected and corrected color chart.

also apply a similar approach to transit among multiple illuminations, such as converting from a white illumination of the monitor setup to the standard D65 illuminations. In this case, we simply need to replace the corresponding light source with D65 to maintain consistency in color representation under varying lighting environments.



(a) Uncorrected photo of a subject (b) Corrected photo of a subject

Figure 2.13: Comparison between the uncorrected and corrected photo of a subject.

2.3.3 Dimension reduction of the full reflectance

Currently, the full skin reflectance spectrum spans a dimension of 151D, spanning wavelengths from 300 nm to 1000 nm, with a 2 nm interval. This high dimension is significantly larger than that of chromophores, leading to a challenge in reconstructing the full skin reflectance from chromophore predictions. To address this issue, reducing the dimensionality of the reflectance is crucial.

One straightforward approach is to increase the sampling interval across the hyperspectral range, for example, reducing the dimension to 36D by using a 20 nm step size. However, this down-sampling may miss finer spectral details, potentially overlooking important local variations that are critical for accurately capturing the nuances of the reconstructed reflectance.

Alternatively, we can also apply a more advanced dimension reduction approach, such as Principal Component Analysis (PCA), for feature extraction and to compress the spectral data effectively. Here, we make a brief introduction to PCA. The original data is n dimension and the expected dimension is q . The goal of PCA is to find the direction with the largest variance in the data and project the data along these directions. The main step is as follows:

- Step 1: Data centering.
- Step 2: Compute the covariance matrix.
- Step 3: Calculate the eigenvalues and eigenvectors of the covariance matrix.
- Step 4: Sort the eigenvectors according to the corresponding eigenvalues and choose the first q features.

With PCA, we can extremely reduce the dimension of the full skin reflectance from 351-dimensional to 10-dimensional, and reconstruct the accurate full reflectance with the

reduced data. Please find the details in Section 4.2.

2.4 Neural Network

In recent years, deep learning has become a popular technique in computer graphics and vision. Researchers employ various neural networks to solve complex problems that traditional methods struggle to address. In this thesis, we built upon some networks for skin appearance and shape reconstructions, therefore we decided to introduce the popular network architecture with the fundamental components related to our thesis.

2.4.1 Fundamental Components

Here, we first introduce some basic components: the fully connected layers, the convolutional layer, several activation functions, optimizers, and also loss functions.

Fully Connected (FC) Layer

The Fully Connected (FC) Layer is a fundamental component of the feedforward neural networks. Each neuron in this layer is connected to all neurons in the previous layer. This layer enables the dimension change from the input to the output. We usually employ this layer in the final stages of a network to implement classification or regression tasks.

Figure 2.14 introduces Multiple Layer Perceptrons (MLPs), where the fully connected layers are the main components. In this figure, the output y is computed as a linear combination of the inputs x with a weight matrix W and bias b :

$$y = Wx + b \tag{2.39}$$

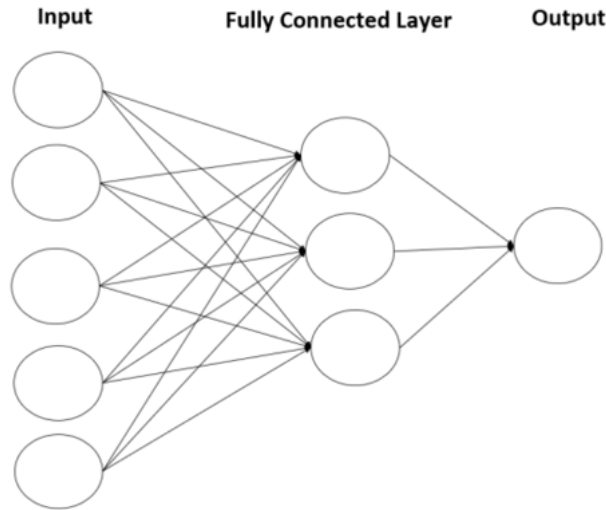


Figure 2.14: Multiple Layer Perceptrons (MLPs).

Convolutional Layer

The Convolutional Layer is a fundamental block of the Convolutional Neural Networks (CNNs), which are effective for a wide range of tasks, such as image and video processing. A convolutional layer applies a series of kernels to the input data and preserves the spatial relationships between pixels by learning image features in small, local patches. Unlike the fully connected layers, the convolutional layer employs weight sharing and sparse connectivity, which makes it more computationally efficient and suitable for high-dimensional data like images.

For an input I and a filter K , the convolution operation is defined as:

$$S(i, j) = (I * K)(i, j) \quad (2.40)$$

$$= \sum_m \sum_n I(i + m, j + n) K(m, n) \quad (2.41)$$

where $S(i, j)$ is the result of the convolution at position (i, j) , and K is the filter applied to the input. The convolutional layer is shown in Figure 2.15.

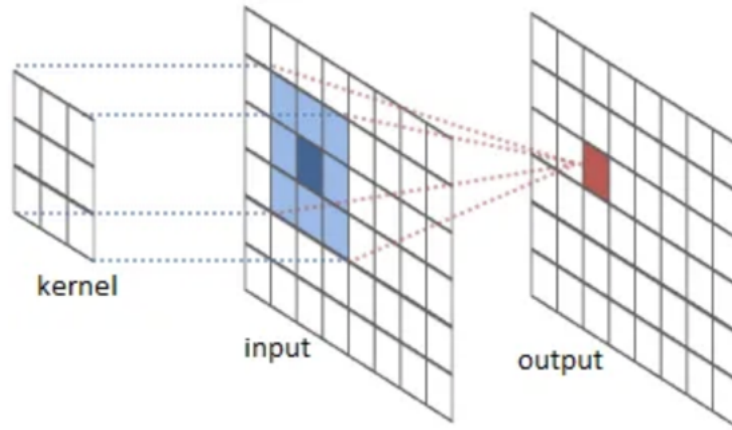


Figure 2.15: The Convolutional Layer

Activation Function

Activation functions involve non-linearity in the neural network to learn more complex patterns and relationships. In the following, we introduce some common-used activation functions:

- The **Sigmoid** function maps input values to a range between 0 and 1, which is popular in binary classification problems. It is defined as:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.42)$$

However, it is possible to cause a gradient vanishing problem, especially when inputs fall into the saturation region (close to 0 or 1).

- The **Tanh** function maps inputs to a range between -1 and 1. It is often preferred over sigmoid due to its zero-centered output:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2.43)$$

Tanh is effective when the data is approximately centered around zero, but it also

suffers gradient vanishing issues.

- The **ReLU** function is one of the most widely used activation functions. It is defined as:

$$f(x) = \max(0, x) \quad (2.44)$$

ReLU is a simple and effective activation function. However, it activates neurons only when their input is positive, leading to sparsity in activations. Hence, it helps with mitigating the vanishing gradient problem and speeding up convergence. To address this issue, there is an advanced function, such as LeakyReLU defined as:

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases} \quad (2.45)$$

where α is a small constant value, typically set to 0.01.

Optimizers

Optimizers are the algorithms that adjust the weights and biases of the network in the training process. The choice of optimizer makes a significant impact on the convergence speed and the final performance of the model. Here, we include some of the most common optimizers as follows:

- The **Stochastic Gradient Descent (SGD)** is a basic optimizer where the network's parameters are updated using the gradient of the loss function with respect to the weights. The update is performed for each mini-batch of data, hence the term "stochastic." While simple, SGD can converge slowly and is sensitive to the choice of learning rate.
- The **Root Mean Square Propagation (RMSprop)** is an adaptive learning rate optimization method that adjusts the learning rate dynamically for each param-

ter, allowing for more efficient learning. This optimizer can address some of the shortcomings of the traditional Stochastic Gradient Descent (SGD).

At each iteration, the squared gradient for each parameter is accumulated using an exponentially decaying average:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1 - \gamma)g_t^2 \quad (2.46)$$

where g_t is the gradient at time step t , and γ is a decay factor typically set to 0.9. The parameter update is then performed as

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} \quad (2.47)$$

where η is the learning rate, ϵ is a small constant to prevent division by zero, and $E[g^2]_t$ is the moving average of the squared gradients.

- The **AdaGrad (Adaptive Gradient Algorithm)** is another optimization algorithm that adapts the learning rate for each parameter based on the previous gradients. Different from RMSprop, AdaGrad accumulates the squared gradients over the entire training process and adjusts the learning rate for each parameter accordingly. The squared gradients for each parameter are accumulated over time:

$$G_t = G_{t-1} + g_t^2 \quad (2.48)$$

where G_t is the sum of the squared gradients up to time step t and g_t is the gradient at time step t . The parameter update rule in AdaGrad is:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} \quad (2.49)$$

where η is the learning rate and ϵ is a small constant to avoid division by zero.

- The **Adam (Adaptive Moment Estimation)** is a widely-used optimizer that

combines the advantages of both the Adaptive Gradient Algorithm (AdaGrad) and RMSProp. It contains two moving averages, one for the gradients and another for the squared gradients, which helps it adaptively adjust the learning rate for each parameter. Adam typically converges faster than SGD and is less sensitive to the choice of learning rate.

Loss functions

Loss functions are used to compare the difference between prediction and the ground truth. Minimizing the loss function is the goal of training a network. Here, we introduce three popular loss functions.

- **Cross Entropy Loss** is the common loss function in the classification task, defined as

$$L = - \sum_i^n \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}) \quad (2.50)$$

where $y_{i,c}$ is the reference label and $\hat{y}_{i,c}$ is the predicted probability for class c , and C is the number of classes.

- **Mean Squared Error (MSE)** is a popular loss function in the regression task. It computes the average squared difference between the prediction and the ground truth. The MSE is defined as:

$$L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.51)$$

- **Mean Absolute Error (MAE)** is another popular loss function for regression, which computes the absolute values between the prediction and ground truth.

$$L = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.52)$$

2.4.2 Introduction to Popular Neural Networks

AutoEncoder (AE)

Hinton et al. first proposed the concept of AutoEncoder[32]. This network is composed of an Encoder and a Decoder. The Encoder aims to predict the features of the input image with dimension reduction, while the Decoder aims to reconstruct the input image. The structure of AutoEncoder is shown in Figure 2.16.

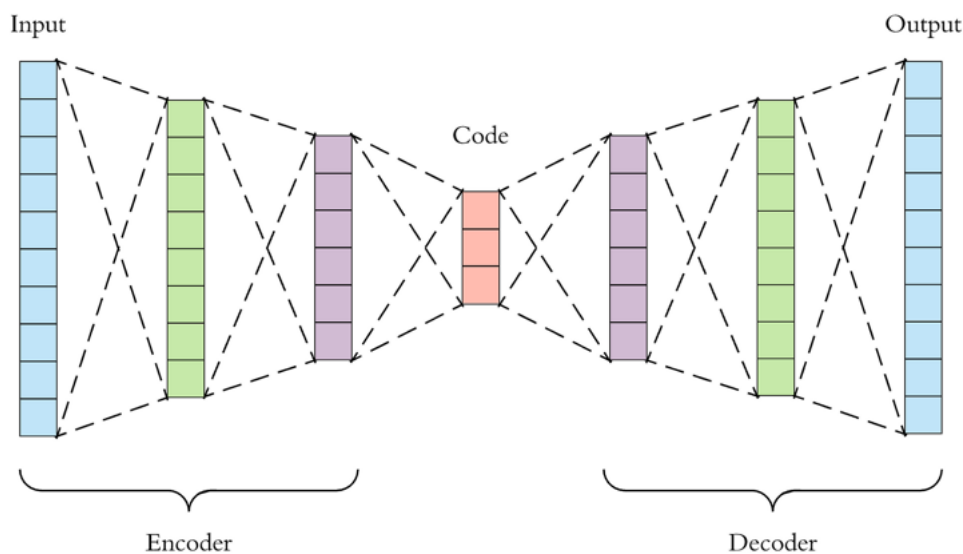


Figure 2.16: AutoEncoder (AE) Architecture

The AutoEncoder has several applications. For example, we can add noise to the input images and employ the AutoEncoder to reconstruct the images without noise, thus denoise the input images.

Another popular application is image semantic segmentation. U-Net which is a popular neural network architecture designed for these tasks[67]. The U-Net architecture is characterized by its symmetric U-shaped structure shown in Figure 2.17. It consists of an encoder and a decoder with skip connections between corresponding layers.

Recently, the U-Net has been employed in diffusion models for iterative image denoising, such as Diffusion, and modern image generation models, such as Stable Diffusion.

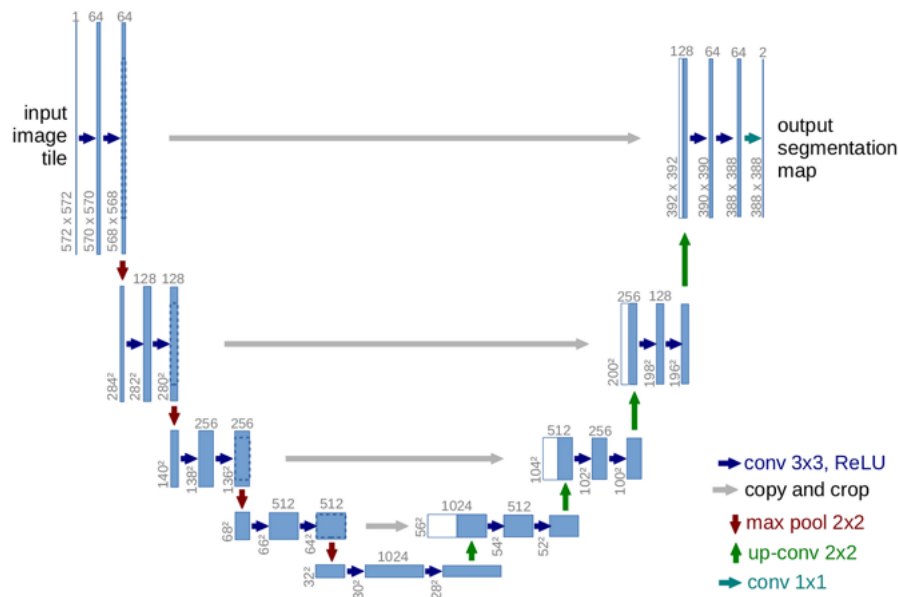


Figure 2.17: UNet Architecture

Variational AutoEncoder (VAE)

Similar to the traditional AutoEncoders, the Variational Autoencoder (VAE) also consists of an Encoder-Decoder architecture (see Figure 2.18). However, it is a probabilistic generative model that learns a continuous latent space representation of input data. The latent space of VAEs is a known distribution, usually a standard Gaussian distribution with a mean μ and a variance σ^2 , that allows for efficient sampling and generation of new data.

The loss function involves two parts. One is the reconstruction loss while another is KL divergence Loss for minimizing the difference between the predicted latent distribution and the reference distribution.

VAEs are employed for learning compact and meaningful representations of data, as well as sampling diverse and realistic outputs. However, compared to the AEs, VAEs sometimes lead to blurred images. Furthermore, the training process of VAEs is sometimes unstable due to the imbalance between two parts of the loss functions.

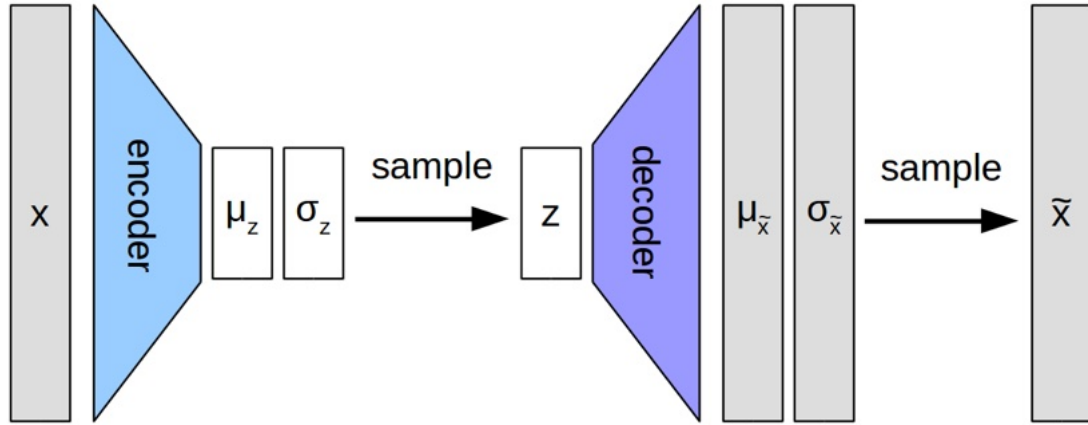


Figure 2.18: Variational AutoEncoder (VAE) Architecture

One possible solution is to disentangle the encoded features in the latent space into different "meanings". One popular example is the Vector Quantized Variational AutoEncoder (VQVAE). This network first encodes the input images into a vector, then search the codebook to find the most similar one and put it back to the Decoder. Different from the continuous values in the latent space in the AutoEncoder, the vectors in the codebook are discrete values and represent multiple features. The network structure is shown below:

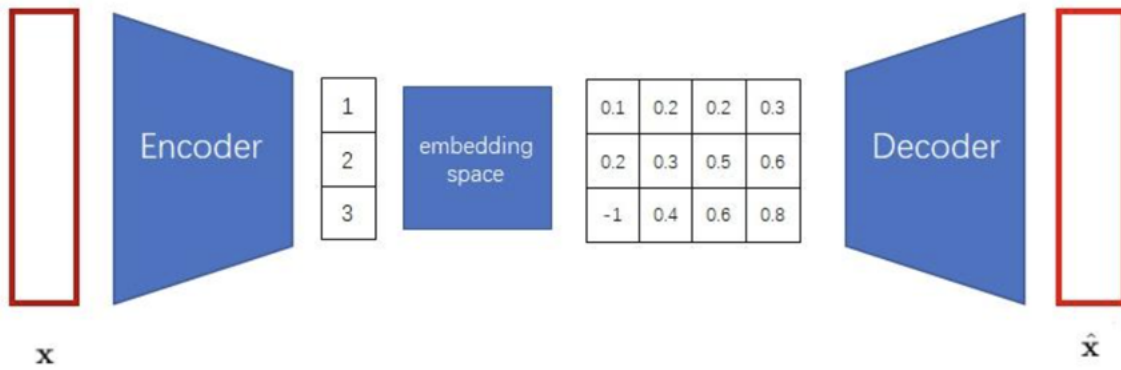


Figure 2.19: Vector Quantized Variational AutoEncoder (VQVAE) Architecture

Generative Adversarial Network (GAN)

Goodfellow et al. proposed Generative Adversarial Networks (GANs) to produce highly realistic data, particularly images [25]. This network consists of a generator and a discriminator. Given noise maps z sampled from a simple distribution, such as Gaussian, the

Generators G mapping from X domain to Y domain and F mapping from Y domain to X domain, and two discriminators D_x and D_y . The key idea of CycleGAN is cycle consistency $F(G(x)) \approx x$.

- **StyleGAN** borrows from style transfer literature to allow better control over the generated image styles. This network is widely-used in generating high-quality and realistic faces with fine details [41].

Diffusion

Diffusion networks become increasingly popular in recent years. The key idea behind diffusion models is to add noise to the input data iteratively in the forward process and learn to reverse this denoising process to recover the original data in the backward process. Both the forward and backward processes of the traditional diffusion networks (Denoising Diffusion Probabilistic Model, DDPM) are modeled as a Markov chain in that the state of each time step relies on the previous time step [33]. Once trained, the network can start from random noise and generate new, realistic samples by following the learned reverse diffusion steps. The training and sampling processes are shown in the following tables:

Training

repeat

$$x_0 \sim q(x_0)$$

$$t \sim \text{Uniform}(1, \dots, T)$$

$$\epsilon \sim \mathcal{N}(0, I)$$

Take gradient descent step on

$$\nabla_{\theta} \left\| \epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t) \right\|^2$$

until converged

The DDPM requires thousands of time steps. To overcome this limitation, the Denoising Diffusion Implicit Model (DDIM) was proposed to speed up the diffusion process, making the models more efficient [69]. A more efficient approach is Stable Diffusion or Latent Diffusion Models [66], which performs the diffusion process in the low-dimensional

Sampling

```

 $x_T \sim \mathcal{N}(0, I)$ 
for  $t = T, \dots, 1$  do
   $z \sim \mathcal{N}(0, I)$  if  $t > 1$ , else  $z = 0$ 
   $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$ 
end for
return  $x_0$ 

```

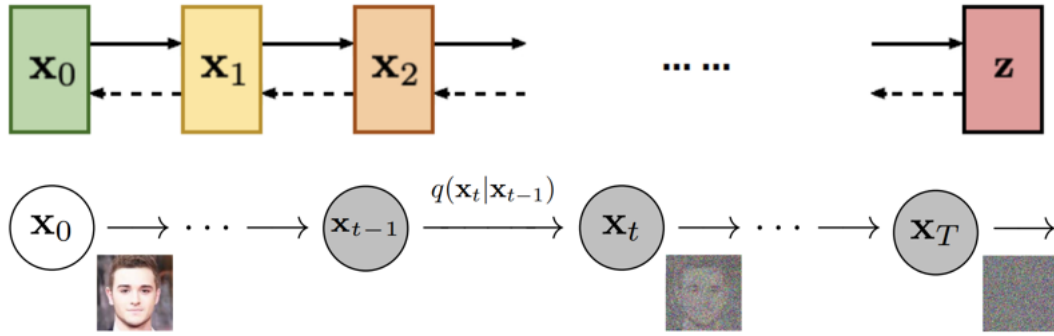


Figure 2.21: Diffusion Architecture

latent space rather than the high-dimensional pixel space. This significantly reduces the computational cost while maintaining high-quality image generation.

In addition, α -(de)blending diffusion is a minimalist deterministic diffusion model is a model introduced by Heitz et al. [31]. This model builds upon the classic α -blending technique from computer graphics, where two images are linearly combined based on a transparency factor α . When $\alpha = 1$, the result fully reflects the foreground; when $\alpha = 0$, it represents the background. By extending this concept, α -(de)blending diffusion iteratively blends and deblends samples from two different distributions, establishing a deterministic path between them. Unlike traditional stochastic diffusion models, this method offers a simple and intuitive framework derived from basic calculus and probability, yet achieves high-quality results with improved numerical stability.

We adopt this technique in Section 5 for the task of re-ageing facial images. By treating facial appearance at different ages as two separate distributions, the α -(de)blending framework allows us to generate realistic transformations between young and old appearances.

This approach enables precise and controllable age progression and regression, making it a powerful tool for facial editing applications.

With the rapid development of large language models, many multimodal learning methods are proposed to connect different information. For example, Contrastive Language–Image Pre-training (CLIP) is a powerful technique in this field. It consists of an image encoder and a text encoder for extracting visual and textual embeddings separately, displayed in Figure 2.22. This network is a bridge that aligns with the text and images by optimizing the cosine similarity between these two embeddings.

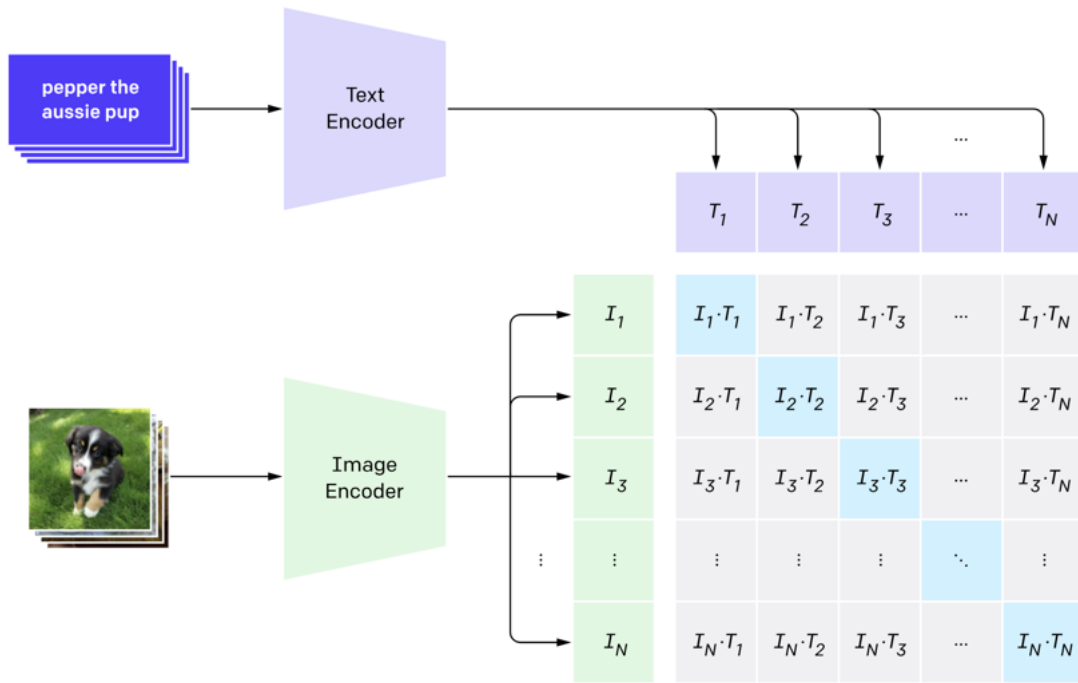


Figure 2.22: CLIP Architecture

2.5 Face Re-ageing

Face re-ageing is one of the most challenging tasks in facial image manipulation, primarily due to the lack of paired datasets recording individuals at different stages throughout their entire lifespan. Furthermore, several factors affect the apparent age of a face, which

may significantly differ from its chronological age. The possible reasons include genetics and individual lifestyle habits, such as diet and exercise. Additionally, various external environmental influences, such as exposure to UV light, and the applications of makeup or cosmetic procedures can also hide the actual ages, making the realistic age transformation task even more challenging.

Early works in this field focus on developing either physics-based models or prototype-based methods to simulate the facial age transformation. Physics-based models [46, 70] aim to capture the underlying biological mechanisms of ageing, using mathematical and physical principles to emulate changes in skin texture, facial structure, and the appearance of wrinkles. These models typically rely on predefined rules and algorithms to predict how facial features evolve over time.

In contrast, prototype-based methods [43, 71, 72] utilize a set of representative facial templates or prototypes corresponding to different age groups. These approaches simulate aging by morphing between prototypes, adjusting facial details such as wrinkle patterns, skin sagging, and other age-related features. While physics-based models offer a more detailed and theoretically grounded approach to aging, prototype-based methods benefit from simplicity and computational efficiency, allowing for quicker yet less anatomically precise simulations of the aging process.

In recent years, several researchers have employed data-driven methods with deep learning techniques. Some use age-conditioned Generative Adversarial Networks (GAN)-based [25] methods to generate realistic re-age images by editing latent representation [58, 30, 53]. For example, Or-El et al. propose a method called Lifespan Age Transformation Synthesis (LATS) [58]. This method involves age as a condition to guide a GAN to extract facial identity information in the latent space, and generates the re-ageing results in the target age group. An overview of this method is shown in Figure 2.23.

LATS can generate realistic re-ageing results, however, there are also some limitations.

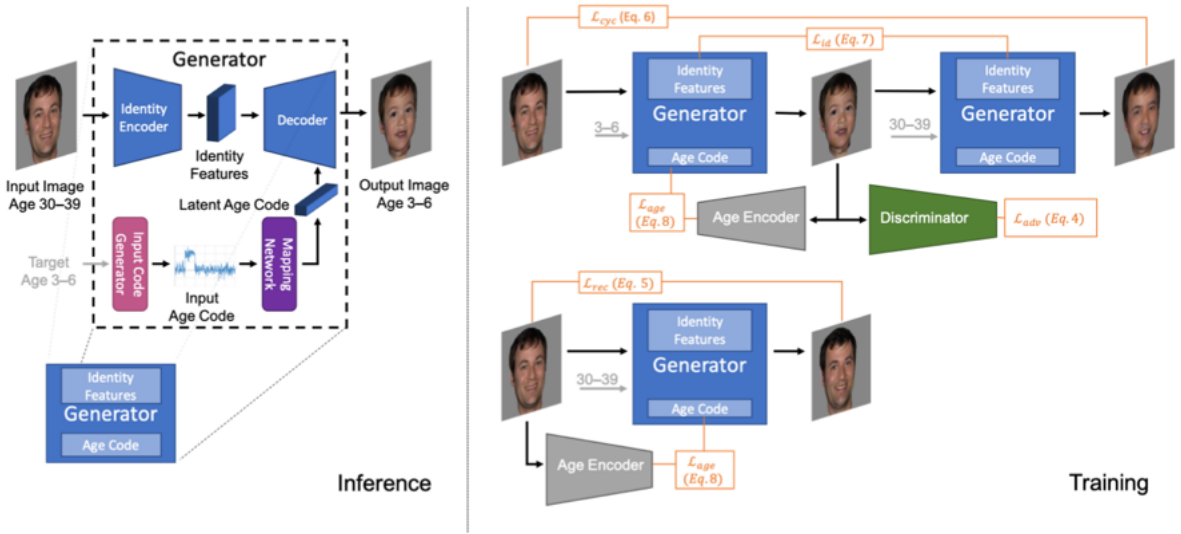


Figure 2.23: The overview of LATS

First, the prior age of LATS is split into different groups, therefore it is difficult to transform faces to an accurate target age. Furthermore, it is challenging for LATS to preserve the features with details. As shown in Figure 2.24, the re-aged faces wrongly wear glasses that are not included in the original faces. However, this limitation is one of the common issues of GANs.



Figure 2.24: Failure examples of LATS

To address this issue, various approaches have adopted an Encoder-Decoder architecture, specifically employing an Encoder paired with a generative Decoder [1, 29, 24]. For example, SAM (Style-based Age Manipulation) framework integrates a pre-trained pSp Encoder, which is responsible for extracting facial identity features, and a StyleGAN as a decoder to produce age-appropriate facial transformations. The use of these two complementary networks enables SAM to effectively generate high-fidelity, realistic age-transformed images while preserving individual identity features. Furthermore, SAM

addresses the challenges of unsupervised datasets by incorporating a cycle consistency loss of ages. This ensures that the age transformation is reversible and maintains consistency across the generated images.

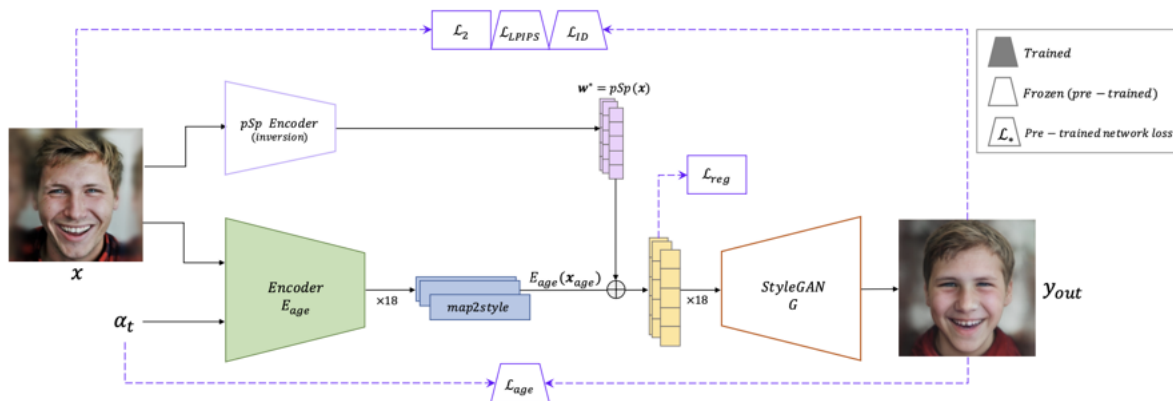


Figure 2.25: The overview of SAM

Figure 2.26 displays some ageing results of SAM. As shown in this figure, this approach can generate high-quality re-ageing results with good identity preservation across various age ranges. Therefore, this method is efficient for generating large facial datasets with paired ages [82]. These datasets are extremely beneficial for training and evaluating models in this field.



Figure 2.26: Ageing results generated using SAM.

Benefits from these datasets, ZOSS et al. recently proposed a production-ready face re-ageing network FRAN to implement facial age transformation [82]. As shown in Fig-

ure 2.27, they employ a UNet to learn the facial differences between the source age and target age, which is robust enough to produce high-quality results with fine details. This method is efficient to be applied on videos depicting faces under free viewpoint, depth, motion, and illumination conditions. Except for FRAN, we also designed a re-ageing method based on this dataset. Please see more details in Section 5.

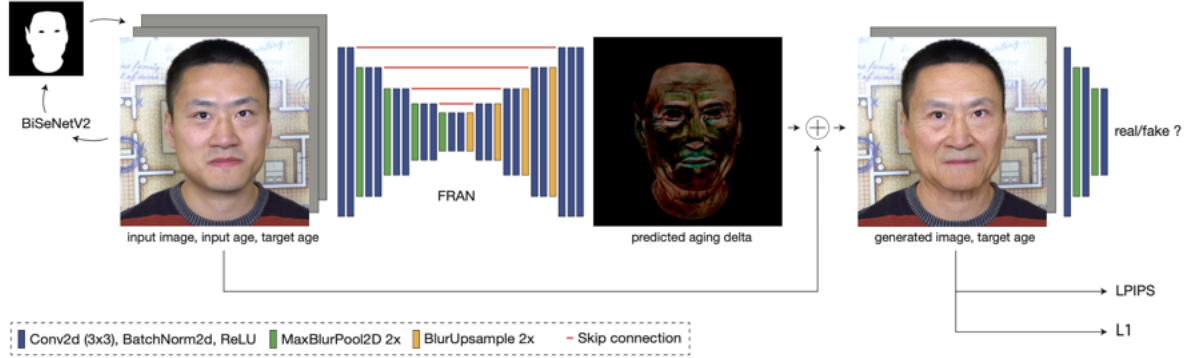


Figure 2.27: FRAN.

Similar to SAM, HRFAE also uses an Encoder-Decoder architecture for face age editing [78]. This method focuses on the quality of the re-ageing results. Instead of low-resolution results with visual artifacts, FRAN trains on the high-resolution image dataset by modifying the latent space which saves the facial identity features, and allows for continuous age editing with high frequency details.

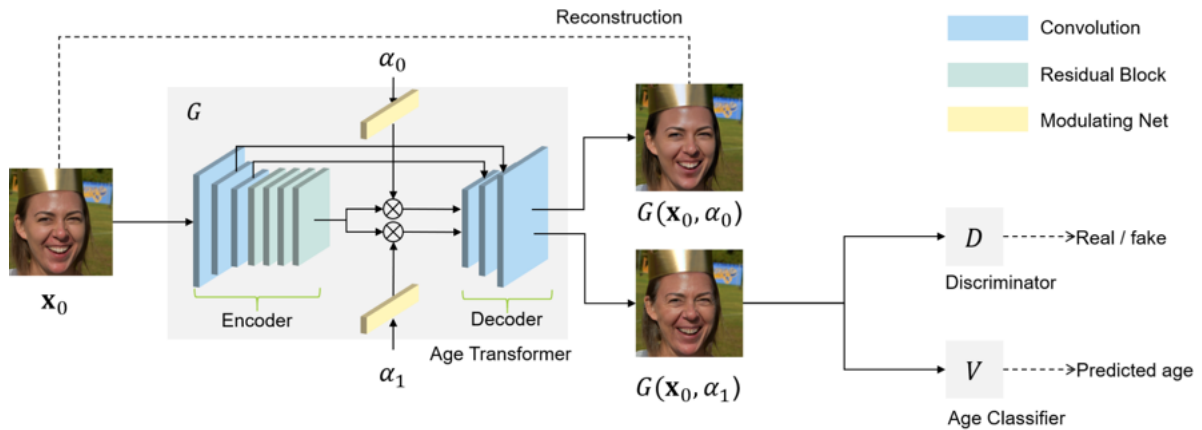


Figure 2.28: HRFAE.

These methods can produce high-quality re-ageing results. However, they overlook the

optical changes in the skin due to ageing. As described in Section 2.1, several biophysical skin appearance models have been proposed in recent years [20, 39, 23, 2, 50]. These methods primarily focus on simulating accurate skin appearance by leveraging practical measurements and models. However, these approaches ignore the effects of ageing. In contrast, [35] proposes a complex biophysical skin model to simulate skin appearance changes due to ageing. However, this method does not take any structural changes, such as wrinkles, into consideration.

In addition to changes in skin appearance, human faces also experience significant structural transformations with age, particularly in terms of shape. Therefore, some research, such as [29], also considers changes in shape features. However, these studies primarily concentrate on 2D image editing.

To summarize, the majority of the current approaches proposed for age editing tasks focus on 2D image manipulation while ignoring the 3D geometry changes. Moreover, several conventional methods categorize individuals into discrete age groups rather than a continuous age-editing framework. In addition, generative-network-based methods are difficult to preserve facial identity while encoder-decoder-based methods are mostly limited by the narrow age range from 18-60 years old.

2.6 Facial Geometry Reconstruction

There are many excellent approaches have been proposed for 3D face reconstruction. Here, we make a brief introduction to some famous methods in this field.

Several works employ model-free approaches to regress 3D shapes. The 3D Morphable Model (3DMM) is such a famous model, built upon a dataset of 3D face scans, that represents a 3D face as a linear combination of a set of basis shapes, derived from a statistical analysis of facial geometry and textures. The key idea of this method is to

decompose the 3D face space into principal components through Principal Component Analysis (PCA), allowing new faces to be generated as weighted sums of these basis components within the parameter space.

Additionally, this model defines many facial attributes in this space, such as gender, to control face reconstruction. These techniques are highly effective in representing facial geometry and variations, compensating for variations in pose, lighting, and expression, without relying on predefined parametric models. Although this method is widely used, it is still limited by the variations of the captured dataset.

Extending on the 3DMM, The Basel Face Model (BFM) was proposed to improve the representation of facial identity and expression variations [60]. It is a high-quality statistical model of human faces based on a comprehensive dataset of 3D face scans. Similar to 3DMM, BFM also employs PCA to decompose the principal components of facial shape and texture. However, BFM incorporates more representations of facial geometry, especially in terms of facial identity and expression separation. Therefore, this method allows for more accurate and flexible facial reconstructions with better control of facial features.

The Face Learned with Articulated Model and Expressions (FLAME) [49] is another statistical model. Unlike the traditional 3DMM and BFM, FLAME focuses not only on facial shape identity and expressions but also on the full head pose, such as the jaw direction. This makes FLAME capable of controlling the shape, expression, and pose reconstructions separately, resulting in higher quality and realistic facial reconstructions. Therefore, FLAME is one of the most popular methods for 3D reconstruction and has various applications, such as facial animation and virtual reality.

Moreover, some approaches employ deep learning networks to extract facial shape, expression, and pose information, and feed them to these models to reconstruct the 3D geometry. MICA is one of the methods for metrical human face shape estimation from a single image, leveraging the ArcFace architecture as an encoder to extract facial iden-

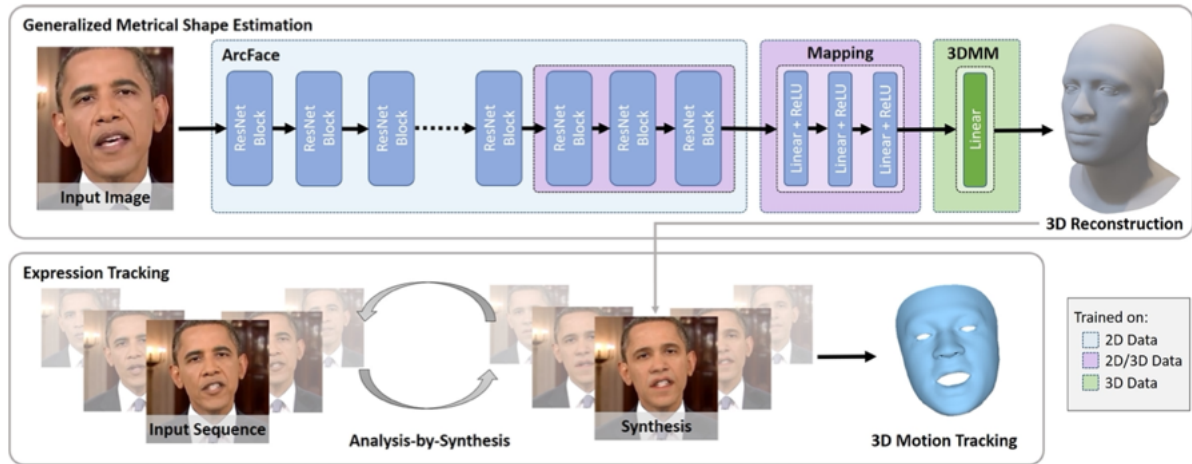


Figure 2.29: MICA

tity, which is then mapped into the parametric space of 3DMM [81]. Rather than using a dataset of 3D facial scans, this method trains on a mixture dataset of 2D and 3D datasets. This makes MICA an efficient way to track facial expressions, as well as the per-frame illumination, facial expression and pose.

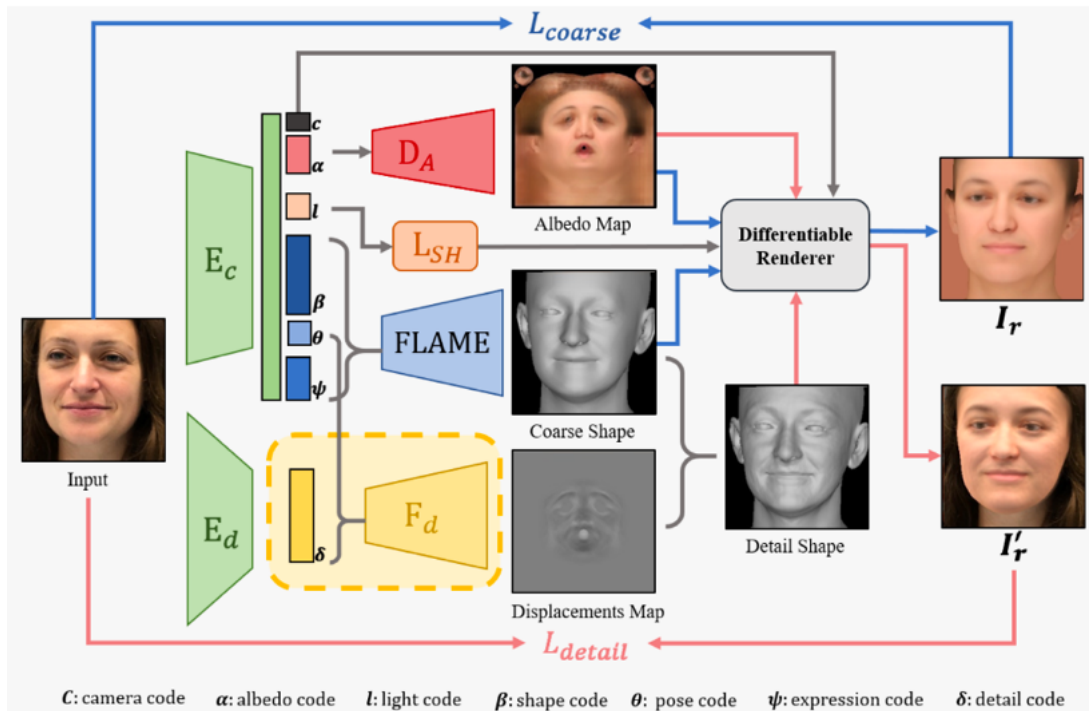


Figure 2.30: DECA

Similar to MICA, DECA is another well-known approach with the architecture in Fig-

ure 2.30 [22]. With a single 2D image as input, DECA trains an Encoder to learn the representations of camera, albedo, lighting, shape, expression, and pose vectors. The last three vectors align with the parametric space of FLAME, thus this method can generate accurate 3D facial geometry. Once trained, they feed the reconstructed shapes and albedos into a differentiable renderer to have 2D rendering results. This makes DECA only require 2D image datasets. Based on DECA, Danecek et al. later propose EMOCA to have better reconstructions of expressions with a novel deep perceptual emotion consistency loss [16].

Recently, several works employ deep learning methods to reconstruct 3D mesh directly, such as COMA. This method uses a convolutional mesh autoencoder consisting of mesh downsampling and mesh upsampling layers with fast localized convolutional filters defined on the mesh surface [64], shown in Figure 2.31. Instead of applying PCA, COMA directly learns a latent representation from the input 3D facial shapes with a mesh Encoder, and reconstructs the input mesh with a mesh Decoder. The key ideas behind COMA is the fast spectral convolutions using a recursive Chebyshev polynomial in Fourier space and hierarchical multi-scale mesh sampling operations. Benefits from these ideas, COMA is an efficient and accurate method for 3D facial reconstruction.

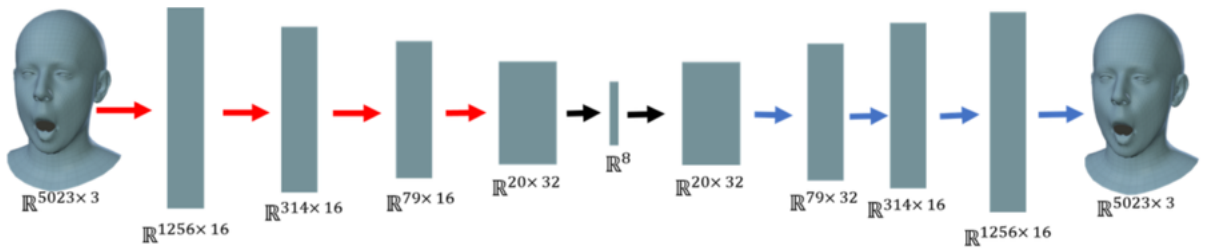


Figure 2.31: COMA.

After obtaining the coarse shape, some works further add refinement methods to obtain more detailed shape features. Some employ or capture a 3D high-quality face dataset, and then compute facial details map from images [12, 27, 9, 48, 73]. For example, Extreme3D first fits 3DMM parameters to generate a coarse shape, then models local deformations

of the face surface as depth displacements to reconstruct detailed shape [73].

On the other hand, some methods design neural networks to predict facial detail maps, such as displacement maps. As shown in Figure 2.30, DECA [22] also includes an Encoder to reconstruct displacement maps from the input images, then combines with the coarse shape to generate the detailed shapes. This makes DECA a state-of-the-art method to reconstruct an animatable detailed 3D Face Model from a single in-the-wild image.

Although these methods can reconstruct high-quality coarse or detailed shapes, they have not taken age into account. However, the head sizes of children and adults are different, therefore, directly using these methods is not beneficial for re-ageing tasks. We require to present a novel approach to incorporate an age factor to control the head shape.

Chapter 3

Practical Measurement and Hyperspectral Skin Model

In this chapter, we propose a practical skin reflectance model based on diffusion theory, which covers a wide hyperspectral range from 300nm to 1000nm in Section 3.1. The proposed method employs a desktop monitor-based capture setup to illuminate human faces with multiple uniform light conditions within the visible spectrum. The acquisition and measurement are described in Section 3.2. Using the captured photographs, we design a novel three-stage LookUp Table (LUT) search approach for parameter estimation and the reconstruction of spectral skin reflectance spanning from ultraviolet, visible, and near-infrared domains in Section 3.3. We make a qualitative evaluation of our method in Section 3.4. The results demonstrate that our method can generate accurate skin reflectance and, at the same time, speed up the conventional LUT search. The advantages and disadvantages of the proposed model and three-stage LUT search are discussed and concluded in Section 3.5.

3.1 Bio-physical Skin Model

We propose a hyperspectral skin appearance model spanning a wide range from 300nm to 1000nm. Building upon the work of Donner et al. [20, 21], our model consists of two skin layers: the dermis and epidermis. We employ dipole diffusion theory on the dermis assumed as a semi-infinite material while applying multipole diffusion theory on the epidermis to simulate the subsurface scattering in these translucent layers. We would like to refer the reader to Section 2.1 and Section 2.2 for the theory behind the biophysical skin model and a detailed description of generating LookUp Tables (LUTs) and visualizations in the hyperspectral domains.

Our model contains five key biophysical parameters that are essential in determining skin appearance. These include the melanin concentration (C_m), the melanin type blend (β_m), which represents the ratio between eumelanin and pheomelanin, the dermal hemoglobin concentration (C_{hd}), the epidermal hemoglobin concentration (C_{he}), and the β -carotene fraction (C_{bc}). In contrast to the recent model proposed by Gitlina et al. [23], our model integrates β -carotene for a more accurate simulation of the appearance in yellowish and olive skin tones, thereby expanding the skin model color gamut.

We employ this skin model to generate LUTs with multiple uniform lighting conditions. Figure 3.1 shows four LUT patches under white illumination of our setup with varying values for β carotene. As shown in this figure, even minor adjustments to the β -carotene fraction can lead to significant changes in skin color, highlighting its significant role in skin appearance.

In addition to the visible domain, these parameters are also important absorbers that affect skin color in the invisible spectral domains. To capture the full spectral range of skin appearance, we extend our model to cover the ultraviolet (UVA) domain from 300nm to 400nm and the near-infrared (NIR) domain from 700nm to 1000nm. In the NIR domain, water concentration is a key factor in skin appearance due to the reduced absorption of

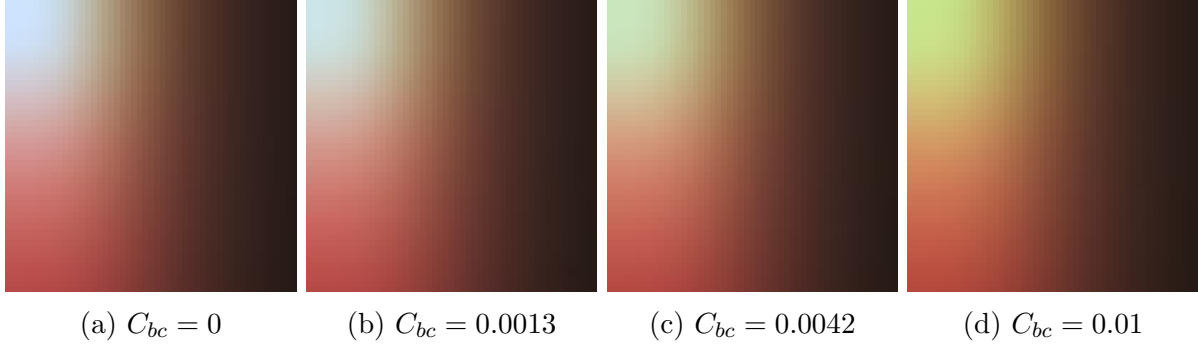


Figure 3.1: Spectral skin reflectance variations due to C_{bc} . C_m increases along the x-axis, C_{hd} along the y-axis; β_m and C_{he} are kept fixed to 0.5 and 0.25, respectively.

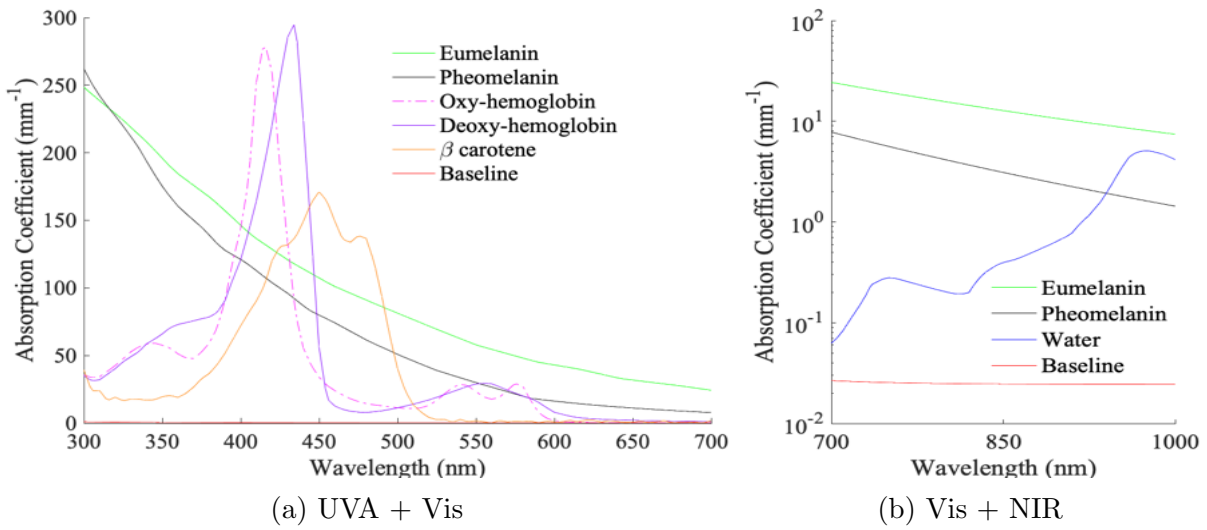


Figure 3.2: The spectral absorption coefficients curves for the skin chromophores included in our model from 300nm to 1000nm.

hemoglobin and β -carotene. Therefore, we incorporated two additional parameters, C_{we} and C_{wd} , representing the water concentrations in the epidermis and dermis, respectively.

The spectral absorptions of these chromophores are borrowed from the work of Chen et al. [13], as shown in Figure 3.2. This extension allows us to simulate skin appearance in both the visible domain and invisible domain, providing a more accurate and realistic model of skin reflectance.

In detail, we define the absorption of the epidermis and dermis in the hyperspectral range as follows:

$$\begin{aligned}
\sigma_a^{epi}(\lambda) = & (1 - C_{we})(C_m(\beta_m\sigma_a^{em}(\lambda) + (1 - \beta_m)\sigma_a^{pm}(\lambda))) \\
& + C_{he}(\gamma^{epi}\sigma_a^{oxy}(\lambda) + (1 - \gamma^{epi})\sigma_a^{deoxy}(\lambda)) + C_{bc}\sigma_a^{bc}(\lambda) \\
& + (1 - C_m - C_{he} - C_{bc})\sigma_a^{base} + C_{we}\sigma_a^w(\lambda) \text{ mm}^{-1}
\end{aligned} \tag{3.1}$$

$$\begin{aligned}
\sigma_a^{derm}(\lambda) = & (1 - C_{wd})(C_{hd}(\gamma^{derm}\sigma_a^{oxy}(\lambda) + (1 - \gamma^{derm})\sigma_a^{deoxy}(\lambda))) \\
& + (1 - C_{hd})\sigma_a^{base} + C_{wd}\sigma_a^w(\lambda) \text{ mm}^{-1}
\end{aligned} \tag{3.2}$$

where

- λ is the light wavelength in nanometers,
- C_m represents melanin fraction,
- β_m represents melanin type blend between eumelanin and pheomelanin,
- C_{hd} and C_{he} refer to hemoglobin fraction in the dermis and epidermis,
- γ^{epi} and γ^{derm} is the blood oxygenation ratio between deoxy- and oxy-hemoglobin of epidermis and dermis, fixed as 0.7 and 0.75 respectively,
- C_{bc} represent the fraction of β carotene,
- C_{we} and C_{wd} are water fractions in the epidermis and dermis fixed as 0.2 and 0.6, respectively.

In these equations, σ_a^x represents the corresponding absorption coefficients. The absorption coefficients of two types of melanin (eumelanin and pheomelanin) are defined as

$$\sigma_a^{em}(\lambda) = 6.6 \times 10^{10} \times \lambda^{-3.33} \text{ mm}^{-1}, \tag{3.3}$$

$$\sigma_a^{pm}(\lambda) = 2.9 \times 10^{14} \times \lambda^{-4.75} \text{ mm}^{-1} \tag{3.4}$$

σ_a^{base} is the baseline skin tissue absorption, defined as follows:

$$\sigma_a^{base}(\lambda) = 0.0244 + 8.53 \times e^{-(\lambda-154)/66.2} \text{ mm}^{-1} \quad (3.5)$$

After the absorption coefficients, the reduced scattering coefficient of the epidermis is

$$\sigma_s^{epi}(\lambda) = 14.74 \times \lambda^{-0.22} + 2.2 \times 10^{11} \times \lambda^{-4} \text{ mm}^{-1} \quad (3.6)$$

The scattering coefficient of the dermis is approximated as half of epidermal scattering [20] [5].

These spectral parameters have different ranges, sampling methods, and spectral domains, shown in Table 3.1. For C_m , C_{hd} and C_{bc} , the parameters are sampled cubically in their respective ranges, according to the following equation:

$$c_i = C_{max} \times \left(\frac{i}{N}\right)^3 \quad (3.7)$$

where c_i is the value of the i^{th} sample for chromophore C , $i \in \{0, 1, \dots, N\}$, $N + 1$ is the number of samples and C_{max} is the maximum value C can assume in our model. C_m , C_{hd} and C_{bc} have 51, 51, 5 samples, respectively.

Unlike the C_m , C_{hd} and C_{bc} , β_m and C_{he} are uniformly sampled in their ranges with 11 and 4 samples, respectively. In addition, C_{wd} and C_{we} are fixed as 0.2 and 0.6.

Parameters	Sampling	Spectral Domains
C_m	0 - 0.5 (cubically)	UV-Vis-IR
β_m	0 - 1 (uniformly)	UV-Vis-IR
C_{hd}	0 - 0.5 (cubically)	UV-Vis
C_{he}	0 - 0.3 (uniformly)	UV-Vis
C_{bc}	0 - 0.01 (cubically)	UV-Vis
C_{we}, C_{wd}	0.2, 0.6	Vis-IR

Table 3.1: Parameters of skin spectral absorption and scattering.

The spectral ranges for these chromophores are also different. Two melanin-related parameters C_m and β_m are sampled in the full spectra covering UV, Visible, and IR domains. Two hemoglobin-related parameters C_{hd} and C_{he} and β carotene parameter C_{bc} are employed in the UV and Visible domains. And, two water-related parameters are applied in the Visible and IR domains.

This makes a 5D spectral skin appearance model. After computing the reflectance and transmittance from this model, we convolve the spectral profiles with the Kubelka-Munk theory [28] to obtain the total diffuse skin reflectance of the full spectral domains. Then, we generate LUTs with SPDs of different uniform illuminations. The details of this part are described in Section 2.2.

3.2 Acquisition and Measurement

In this section, we introduce a practical and efficient measurement approach for capturing human facial images with a monitor-based setup in the visible domain, as well as employ a full spectral camera to acquire photos for qualitative evaluation.

3.2.1 Visible domain

To illuminate the human faces, we employ a practical desktop capture setup consisting of four 4K LCD monitors (Asus ProArt PA279CV) arranged around the subject, with a camera (Canon EOS M6 Mark II) positioned in the center of the monitors for facial measurements, as illustrated in Figure 3.3 (a). This setup is similar to the work of Lattas et al. [47], however, unlike the previous work, the camera in our setup is cross-polarized, while in their setup, the illumination emitted by the monitors is linearly polarized. Cross-polarization in our setup helps minimize surface reflections, enhancing the accuracy of the captured skin reflectance data by isolating subsurface scattering effects, which are crucial

for detailed skin analysis.

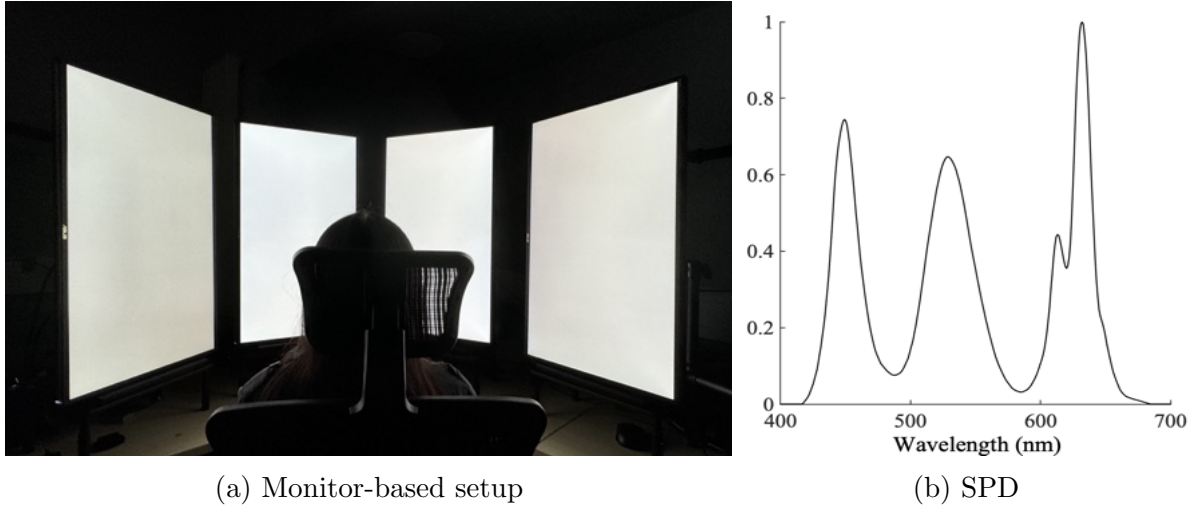


Figure 3.3: Monitor-based setup and the SPD of white illumination.

As presented in previous work by Preece and Claridge [62], narrowband illuminations with peaks around 450nm and 560nm are beneficial for measuring melanin and hemoglobin concentrations, respectively. Narrowband lighting in these wavelengths captures sharper facial details, such as skin blemishes and fine textures, compared to broadband lighting. This increased sharpness is essential for accurately rendering skin appearance, as the blurring introduced by subsurface scattering simulation aligns closely with the ground truth. In contrast, measurements taken under broadband illumination usually appear overly blurred when rendered, leading to missing some skin details.

Therefore, we employ multiple illuminations, including white, blue, and green illuminations, in our measurement strategy. The broadband white illumination serves as a baseline for capturing a general skin reflectance profile. The Spectral Power Distribution (SPD) of this white illumination (6500K CCT) is displayed in Figure 3.3 (b), and the peaks of blue and green illuminations are around 450nm and 530nm, respectively. The blue and green illuminations are supposed to excite the melanin and hemoglobin, respectively. However, according to the work [62], the narrowband green and blue illuminations of this setup are slightly biased to the best illumination conditions. This means that our setup is not optimal for melanin and hemoglobin, but it is beneficial to measure the concentration of

β carotene.

Directly employing narrowband illuminations results in differing color gamuts across RGB cameras. To mitigate this issue, we adopt the measurement strategy proposed by Gitlina et al. [23], which synthesizes blue and green illuminations instead of utilizing them directly. Accordingly, the final pipeline of the proposed practical measurement setup includes the following illumination conditions:

- broadband white illumination (I_W),
- a mixed illumination (I_{W+nb}) of direct narrowband blue (I_{nb}) and broadband white illumination (I_W),
- a mixed illumination (I_{W+ng}) of direct narrowband green (I_{ng}) and broadband white illumination (I_W).

To illustrate the synthesis process, we use the synthetic blue illumination as an example. After image acquisition, we apply a Chromatic Adaptation Transform (CAT) independently to the images under the broadband white light (I_W) and the mixed light (I_{W+nb}) to predict their appearance under the narrowband illumination and map them within the sRGB gamut [23]. The desired narrowband response I_{nb} is then computationally recovered as follows:

$$I_{nb} = \delta \times (CAT(I_{W+nb}) - CAT(I_W)/\delta) \quad (3.8)$$

where the factor δ compensates for the difference in LED intensities between the two illumination conditions. The synthetic green image is obtained similarly by combining broadband white and narrowband green illuminations. The resulting white, synthetic blue, and synthetic green images are presented in Figure 3.4.

Figure 3.5 displays the photograph (b) and reconstructions of the cheek area using broadband white and synthetic illuminations (a), and broadband white and direct illuminations

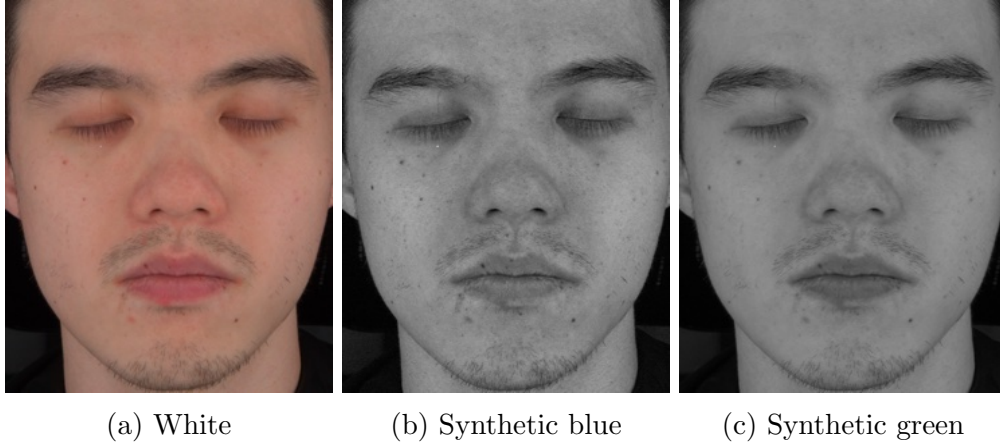


Figure 3.4: Practical spectral measurements of skin reflectance using screen illumination. (a) White (b) Synthetic blue (c) Synthetic green.

(c), respectively. As can be seen, the reconstruction with direct illuminations is closer to the ground truth, however, this reconstruction is blurred slightly. In contrast, the reconstruction with synthetic illuminations shows sharper skin details, especially around skin moles and pores.

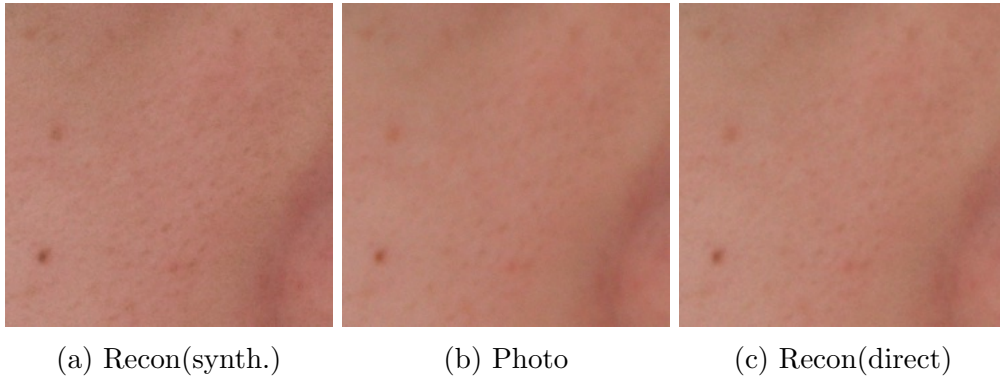


Figure 3.5: Comparison of photograph and reconstructions with different measurements. Figure (a) employs broadband white, synthetic blue, and synthetic green illuminations while Figure (c) employs broadband white, direct blue, and direct green illuminations.

To summarize, this multiple illumination strategy allows for more realistic and detailed skin measurements, which are significant for producing high-fidelity skin appearance reconstructions. By employing this approach, we can later achieve a more accurate estimation for the chromophore parameters and also improve the quality of both skin appearance reconstructions and rendering.

3.2.2 Invisible domain

Our skin model extends beyond the visible domain but also the UVA and NIR domains. Therefore, we can obtain the reconstructions in both visible and invisible domains after estimating the chromophores maps. However, the proposed measurement strategy does not include the invisible range as both the color gamut of the common cameras and the illuminations from the monitor-based setup are in the visible spectrum.

For qualitative evaluation in the invisible domains, we employ a UV camera (fxo487MXGE) without an IR-cutoff filter to acquire facial images in the full spectral domains. For the UV domain, we illuminate the subjects by using five UV LED tubes with emission in the $[300 - 400]nm$ range. For the IR domain, we illuminate the faces with an IR LED in $[700 - 1000]nm$ range. This setup allows us to capture high-quality photos in both the UV and IR domains.

In this case, the measurement of the visible domain and invisible domain applies different strategies, leading to the mismatch of the captured photographs in the various spectrums. Therefore, it is impossible to evaluate the results within the full spectrum quantitatively, but we would like to note that the qualitative evaluation between the reconstructions in both the visible and invisible domains and captured photographs can be found in Section 3.4.

3.3 Conventional Parameter Estimation and Skin Appearance Reconstruction

In this part, we present an advanced method for parameter estimation and skin appearance reconstruction. The conventional parameter estimation relies on a precomputed Lookup Table (LUT) search.

A significant limitation of LUT search is the discretization, which can adversely affect the accuracy of parameter estimation and the subsequent reconstruction of skin appearance. The simplest solution to this issue is to expand the LUTs by increasing the sampling densities or enlarging the sampling ranges of various chromophores. This approach enhances the resolution of the LUT, allowing for more accurate searches and leading to better skin appearance reconstruction. However, the scales of LUTs directly affects the efficiency of this LUT search directly. Moreover, our proposed measurement method utilizes multiple illuminations while our model involves five spectral parameters. This increases the model complexity, thus leading to a more time-consuming process of the LUT search.

To balance the accuracy and efficiency of the LUT search, we develop a novel three-stage LUT search method that significantly speeds up the parameter estimation process while maintaining accuracy.

Here, we first introduce the conventional LUT search. As described in Section 3.2, the given measurement is 5D data including the images under broadband white illumination, grayscale synthetic blue illumination, and grayscale synthetic green illumination. With these illuminations, we generate the corresponding lookup tables. The conventional LUT search is a brute-force search that search the closest values of each pixel in the skin albedo from LUTs by minimizing the L2 distance in CIE LAB color space (i.e. ΔE_{94} color difference), thereby obtaining the corresponding spectral parameters, as well as the reconstructions under white illumination.

In order to speed up this process, we propose a novel three-stage LUT search method that reduces the computational cost without affecting the quality of the estimated parameter maps. As described before, blue illumination helps excite the response of melanin and β carotene while green illumination enhances the response of hemoglobin. Therefore, the key idea is to estimate various spectral parameters with the corresponding beneficial lighting conditions, employing blue illumination for C_m , β_m and C_{bc} , and using green illumination for C_{hd} and C_{he} in multiple stages.

According to this idea, we make the detailed strategy of this method as follows:

- Stage 1: fix β -carotene as a specific value, then search broadband white and synthetic blue LUTs to predict two melanin parameters C_m and β_m ;
- Stage 2: fix two melanin parameters predicted by Stage 1, further search broadband white and synthetic blue LUTs to predict β carotene C_{bc} ;
- Stage 3: fix C_m , β_m , and C_{bc} predicted by Stage 1-2, then search white and synthetic green LUTs to estimate two hemoglobin parameters C_{hd} and C_{he} .

About the specific value for β -carotene in Stage 1, we randomly choose a data set including subjects with different skin types. Then, we employ the conventional full LUT search on this dataset and estimate the β -carotene maps for each subject. We later vote the optimal value for, and finally set C_{bc} as $2e^{-4}$ in Stage 1.

3.4 Results and Evaluation

In this part, we show the results of estimated parameters and reconstructions in both the invisible and visible domains in Section 3.4.1 and quantitative and qualitative evaluations in Section 3.4.2.

3.4.1 Results

With the monitor-based setup, we captured multiple subjects with different genders and various skin types, including Caucasian skin type, Asia skin type, and South Asia skin type. These three skin types exhibit significant variations in the chromophore concentrations and thus the overall skin tones. In the following, we display the results of the

proposed three-stage LUT search for parameter estimation and skin appearance reconstruction.

Figure 3.6 (a) shows the photograph captured with our setup in the visible domain. Following the proposed three-stage LUT search method, we match each pixel of the input photograph to the generated LUTs to obtain the reconstructions displayed in Figure 3.6 (b). As shown in this comparison, all reconstructions are close to the photographs, which proves that our proposed method is suitable for different skin types.

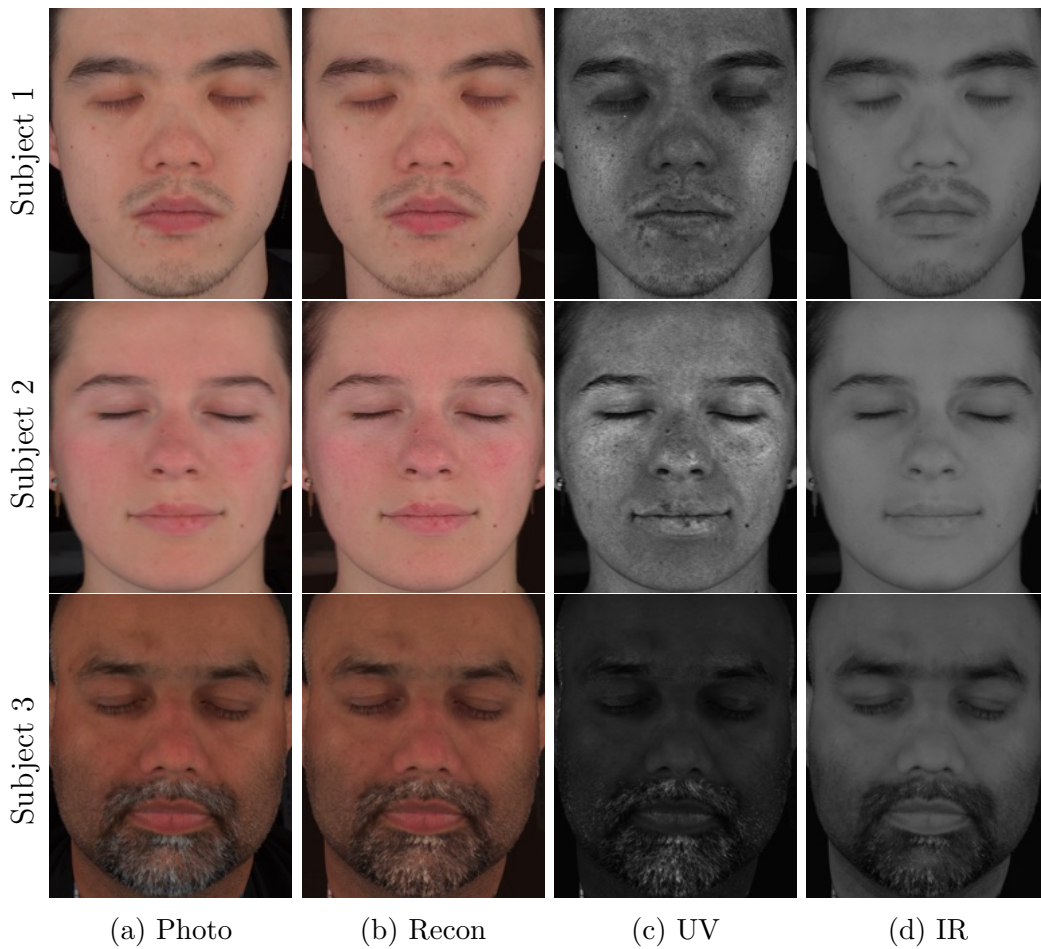


Figure 3.6: Comparison of photos in the Visible domain(a) and reconstructions in the Ultra-Violet (c), Visible (b), and Near-Infrared (d) domains for faces of Asia skin type (first row), Caucassian skin type (second), and South Asia skin type (third row).

After showing the reconstruction, we display the estimated chromophores in Figure 3.7 (a - e). In the maps of the two melanin-related parameters, C_m and β_m (Figures 3.7 a and b), distinct pigmentations can be observed that correspond closely to the input image. For

example, in the case of the first subject, visible moles around the left cheek are accurately captured and shown clearly in the melanin parameter maps.

Similarly, in the maps representing the two hemoglobin-related parameters, C_{hd} and C_{he} (Figures 3.7 c and d), the reddish areas are clearly modeled, aligning with the input image. An illustrative example is the second subject, where a visible rash on both sides of the cheeks is accurately reflected in the hemoglobin parameter maps.

In Figure 3.7 (e), the third subject, who has darker skin, exhibits a prominent β carotene map. This highlights the benefits of incorporating β carotene into our model, which is useful for accurately simulating olive and dark skin tones. These results demonstrate the capacity of our model to capture fine details and chromophore variations across different skin types, further validating its applicability for diverse skin appearance modeling.

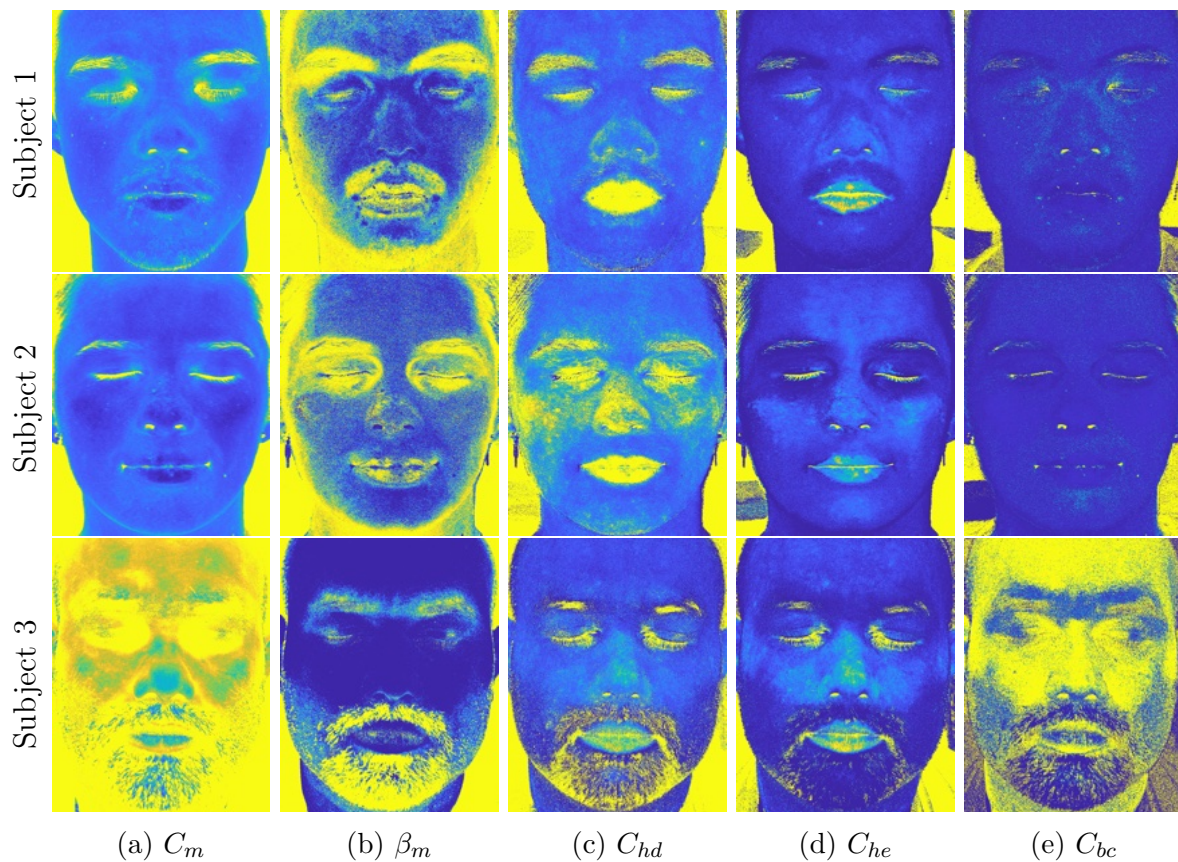


Figure 3.7: The estimated spectral parameters (a - e) for the corresponding faces of Asia skin type (first row), Caucassian skin type (second), and South Asia skin type (third row).

Furthermore, we also reconstruct the faces in the UV and IR domains shown in Figure 3.6 (c, d), respectively. Compared to visible light, UV light has shorter wavelengths and penetrates less deeply into the skin, therefore, UV reconstructions show more detailed and sharper features, such as pigmentations, that may not be as obvious in the visible domain. In contrast, IR light has longer wavelengths and penetrates deeper into the skin layers, reaching the dermis where blood vessels reside, hence we can find that IR reconstructions are softer than the visible photos and show some veins around the temple area.

The skin tones of Subject 1 and 2 are lighter than that of Subject 3, which leads to a large difference in both chromophore maps and reconstructions in the invisible range. Dark skin has higher melanin concentration than light skin, therefore C_m of Subject 3 is much brighter than the other two. At the same time, the UV response of Subject 3 is lower as high melanin concentration absorbs less UV radiation, which is consistent with [13].

3.4.2 Evaluation

Here, we present four qualitative evaluations of the proposed three-stage LUT search in the following content.

We start by comparing the reconstructions generated by our 5D skin model with those of the 4D skin model proposed by Gitlina et al. [23]. This work is the previous state-of-the-art approach that achieves a good balance between the model complexity and reconstruction accuracy. However, their model lacks the inclusion of β carotene, which limits the range of skin tones. In contrast, our model covers a wider range of skin colors, making it more versatile for diverse skin types.

In Figure 3.8, we present the comparison between a photograph (b) and the reconstructions from both models: our 5D model (c) and the 4D model (a). To clearly illustrate the differences in accuracy, we also include absolute error maps in the bottom right corner of each reconstruction. These maps highlight that our 5D skin model consistently produces lower error values compared to the 4D model, demonstrating that our skin model is more accurate than Gitlina et al.’s.

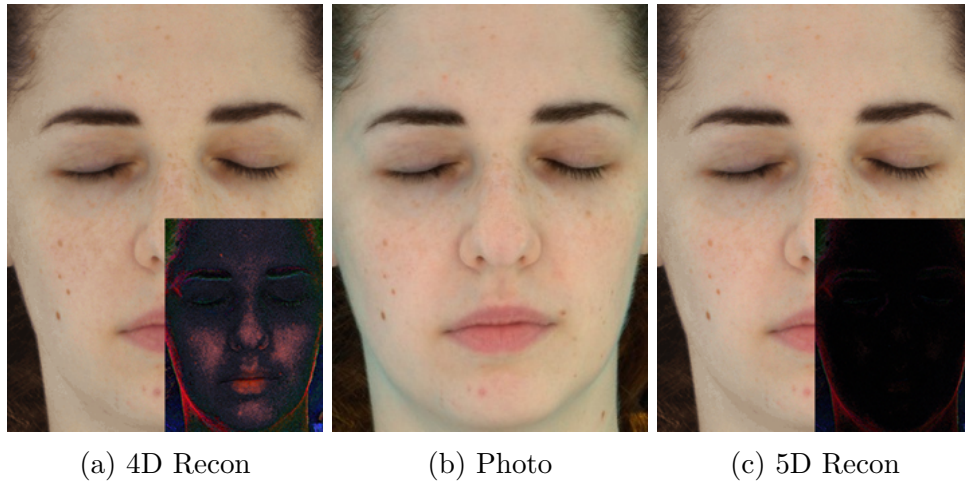


Figure 3.8: Comparison of a facial photograph captured with Light Stage under white light (b), reconstruction via 4D model [23] (RMSE=0.0311) (a), and via our 5D model (c) (RMSE=0.0164), including absolute error maps in the insets.

Secondly, we compare the results between the proposed three-stage LUT search and those of the conventional full LUT search, as shown in Figure 3.9. Both methods have qualitatively similar reconstructions to the photograph, which validates that our novel approach maintains high quality in skin appearance reconstruction. The main advantage of the proposed method is its efficiency. By strategically narrowing the search space in multiple stages, this approach significantly reduces computation time without sacrificing the quality of the final reconstruction.

Furthermore, we discuss the benefit of employing multiple illuminations as inputs in the LUT search. While reconstructions using multiple inputs and a single input in the visible domain are quite similar, the difference becomes more apparent in the invisible domain. As shown in Figure 3.10, the reconstruction generated from narrowband synthetic data



Figure 3.9: Comparison of photograph (a), reconstruction using the naive LUT search method, using white, Synth. blue and Synth. green (b), and reconstruction using the 3-stage search (c).

(incorporating multiple illumination types) shows more and sharper facial details, such as pores, and demonstrates a closer method to the reference photograph compared to the reconstruction derived from a single white illumination input. The observation highlights the benefit of multiple inputs, particularly in enhancing the reconstruction accuracy in the UV domain. This aligns with the previous conclusion that blue and green illuminations help with exciting melanin and hemoglobin.

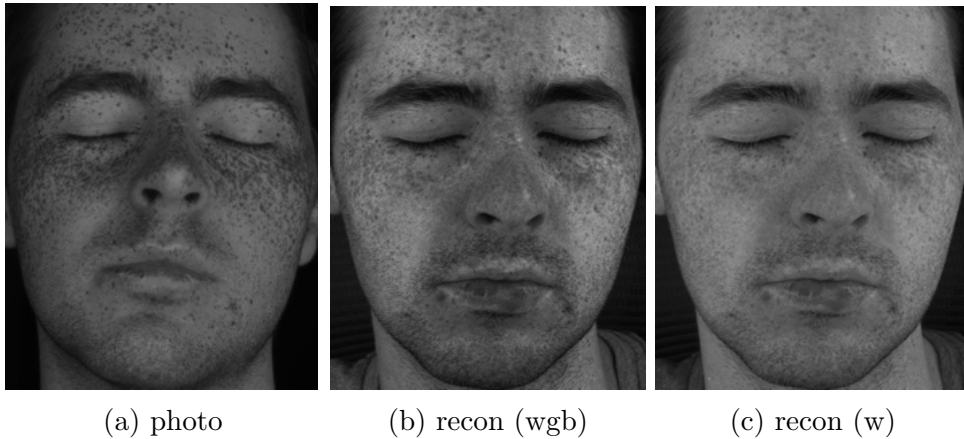


Figure 3.10: Comparing UV photo (a) with reconstructions: broadband plus narrowband inputs (b), and broadband only (c).

Finally, we make a qualitative comparison between the facial reconstructions in the UVA and NIR domains and the photographs captured using the measurement setup described in Section 3.2.2. The cameras used for the invisible domains (UVA and NIR) are not

cross-polarized, which results in specular reflections, unlike the cameras for the visible domain, which are cross-polarized to remove such effects. To ensure a fair comparison, we simulate shading in the reconstructions by adding a small frontal area light and rendering specular reflections to match the UV and IR photographs. Despite differences in camera setups and lighting directions, the reconstructions are close to the photographs in both UVA and NIR domains.

As previously noted, the UVA photograph (b) and its corresponding reconstruction (c) show more skin details, such as blemishes and spots around the cheeks and forehead, compared to the visible photograph (a). In contrast, the NIR photograph (d) and its reconstruction (e) present a softer appearance, with fewer visible pigmentations but more obvious blood vessels. This figure validates that our method can generate high-quality reconstructions in the invisible domains, which are comparable to the captured photographs.

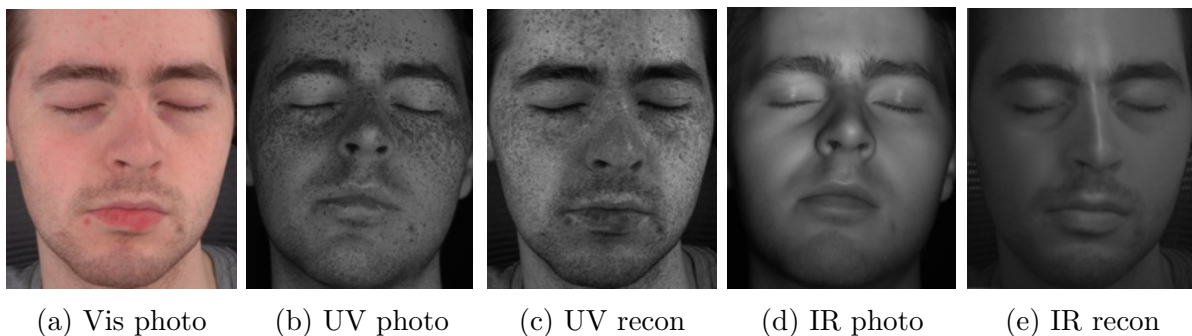


Figure 3.11: Comparisons of Vis, UV, and IR photos (a-c) and reconstructions (d, e) using our proposed method.

3.5 Conclusion

To conclude this chapter, we proposed a practical diffusion-based hyperspectral skin reflectance model driven by five biophysical chromophore parameters. With these parameters, this model enables accurate modeling of skin appearance across a broad spectral range, spanning from ultraviolet (300 nm) to near-infrared (1000 nm).

We also designed a practical measurement setup that utilizes four LCD monitors with a camera positioned at the center to provide spectral illumination. The monitors are calibrated to emit broadband white light, and narrowband blue and green light at selected wavelengths. These two narrowband lights are further synthesized into sharper lights, adapting to the sRGB color gamut of the camera while simultaneously exciting the chromophore responses.

For parameter estimation, we initially precomputed Lookup Tables (LUTs) based on the proposed biophysical skin model. Then, we employed the conventional LUT search method, which involves a brute-force search for each pixel in the captured images to find the closest color match in the LUT, thus obtaining the corresponding chromophore parameters. This method can produce plausible chromophore predictions and skin appearance reconstructions. However, due to the model complexity, the search process is computationally intensive, meanwhile, the discreteness of LUTs affects the accuracy of the reconstruction.

To improve the efficiency and accuracy of the process, we explored several approaches. The most straightforward method is to interpolate more pixels from the Lookup Tables (LUTs). However, each point in the LUTs represents more than just an RGB color. It corresponds to specific chromophore parameters. The relationship between the RGB colors and chromophore parameters is non-linear. Therefore, although there exist many methods to sample colors, it is challenging to map these sampled colors to their corresponding chromophore values. We also investigated more efficient data structures, such

as KD-trees, to optimize the search process. However, these structures require the sorting of 3D RGB color data or splitting them into different subspaces. As some colors in a LUT are extremely similar, even when an optimal match is found, it often requires backtracking through multiple branches to confirm the closest color, which is also time-consuming.

Given these challenges, we shifted our focus from optimizing the search process for color-to-parameter mapping (the inverse of LUT precomputation) to improving the process from the perspective of the chromophore parameters themselves. To this end, we propose a novel three-stage LUT search algorithm for efficient parameter estimation. Rather than relying on exhaustive brute-force searches, which become computationally expensive with high-resolution LUTs, our method progressively refines the search space over multiple stages. As shown in Section 3.4, this approach significantly accelerates the parameter estimation process while maintaining high accuracy for a range of skin types.

Despite these improvements, several limitations remain. First, while the biophysical skin model provides an accurate representation of skin appearance, it is highly sensitive to lighting conditions. Consequently, variations in lighting or camera settings need to color correct the color chart and facial images again to compute a new scaling factor, followed by the re-computation of LUTs. Second, the validation process needs further refinement. We did not include a subject with dark skin tones in our evaluation, and since our model spans a wide hyperspectral range, a hyperspectral camera would be more appropriate for validation, rather than a visible-spectrum camera. As a result, our evaluation was limited to a qualitative evaluation using a visible-domain camera. Third, while the proposed three-stage LUT search method accelerates the search process compared to the conventional brute-force approach, it still faces the inherent limitation of LUT discreteness, which is a common issue in LUT-based methods. Further acceleration is needed to handle larger datasets more efficiently.

Chapter 4

Neural Parameter Estimation and Reconstruction

This chapter addresses two main limitations of Lookup Table (LUT) search methods: 1) the LUT search process is time-consuming, with search time scaling according to the scale of the LUT; and 2) the discrete values within the LUT lead to reduced accuracy in both parameter estimation and skin appearance reconstruction. To overcome these challenges, we propose an Encoder-Decoder network that given multiple measurements to estimate spectral parameters and reconstruct skin reflectance in both visible and invisible domains, which can be further visualized under various uniform illuminations (Section 4.1). Moreover, we introduce a simplified version of the method with a single input, allowing for direct skin appearance reconstruction (Section 4.2), which is followed by image-space scattering approximation in Section 4.3. We further demonstrate the practical application of the biophysical skin model in skin appearance editing. This simulates different skin appearances, such as tanned, lightening, flushed, and pale, by modifying different chromophore parameters (Section 4.4). Finally, we conclude the findings and limitations in Section 4.5.

4.1 Hyperspectral Skin Model

4.1.1 Implementation

In this part, we introduce an Encoder-Decoder network to increase the efficiency and accuracy of the LUT search. Same with the LUT search, the inputs are 5D measurements including broadband white, grayscale synthetic blue, and grayscale synthetic green images. Given these measurements, the encoder *ChromNN* aims to predict five biophysical parameters C_m , β_m , C_{hd} , C_{he} , C_{bc} .

With the predicted chromophore parameters, we can reconstruct the complete hyperspectral skin reflectance profiles spanning both visible and non-visible spectral regions. One straightforward approach to achieve this is to follow the same procedure for LUT pre-computation. However, this method remains computationally intensive, especially when applied to high-resolution images or large datasets. To address this inefficiency, we designed a neural network, *SpectraNN*, with a similar structure to *ChromNN*. Once trained, *SpectraNN* enables fast and accurate generation of full spectral reflectances, which can then be visualized under various uniform spectral illuminations across the full wavelength range, including ultraviolet (UVA), visible, and near-infrared (NIR) domains.

The main challenge for *SpectraNN* is the large disparity between the input and output dimensions. Specifically, the input consists of five biophysical parameters, while the output comprises densely sampled spectral data from 300nm to 1000nm, with a resolution of 2nm. This results in a significantly higher output dimension compared to the input, making accurate estimation a complex task.

To overcome this dimensionality challenge, we apply Principal Component Analysis (PCA) to the output spectral data, reducing its complexity while preserving the essential information. We can obtain nearly 100% of the total variance by retaining the first 10 principal components. This transformation allows *SpectraNN* to not only simplify the training and

testing processes but also enhance the efficiency and accuracy in predicting the spectral skin reflectance. The proposed network structure is shown in Figure 4.1. Further details of the PCA application and its implementation are provided in Section 2.3.3.

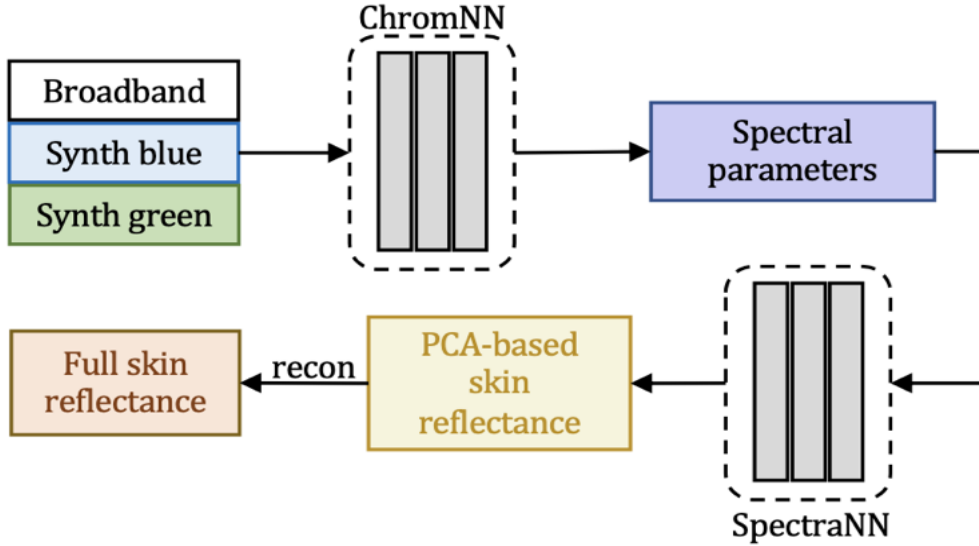


Figure 4.1: The proposed neural architecture consists of two networks. The *ChromNN* maps multispectral albedo data to the five spectral model parameters. The model parameters are then fed to the *SpectraNN* to predict a PCA-based, low-dimensional skin reflectance representation, finally reconstructed using the standard PCA reconstruction technique to obtain the full spectral reflectance in the UVA+VIS+NIR range (300nm to 1000nm).

Training and Testing Datasets

For these two sub-networks, the training and testing datasets are different. The training input dataset for *ChromNN* is 572,220 groups of 5-dimensional Lookup Tables (LUTs), including red, green, and blue channels of broadband white illumination, grayscale synthetic blue illumination (SB), and grayscale synthetic green illumination (SG). The training reference dataset for the outputs of *ChromNN* is the corresponding five spectral parameters. This dataset is also the training input dataset for *SpectraNN* while the reference dataset for outputs is 10D PCA-based full spectral skin reflectance profiles.

Different from the training dataset, the testing datasets for *ChromNN* are the image

datasets. For *ChromNN*, the input testing dataset of *ChromNN* is 5D data concatenated with the broadband white images, the grayscale synthetic blue image, and the grayscale synthetic green image from the acquisition and measurement method. The reference of the output testing dataset is the parameter maps from the LUT search. The outputs of the *ChromNN* are the inputs for *SpectraNN*. The outputs of the *SpectraNN* are 10D PCA-based data, which are then converted to the full hyperspectral skin reflectances.

These predicted skin reflectances can be visualized in the invisible domains, including UVA and NIR domains, as well as the visible domain with various uniform illuminations. The references for these reconstructions are also from the LUT search, but they are not the ground truth, therefore we will compare the outputs qualitatively to the measurements in the invisible domain.

Implementation Details

Both *ChromNN* and *SpectraNN* are designed as three-layered Multilayer Perceptrons (MLPs). Each fully-connected layer in the middle of networks is followed by a Rectified Linear Unit (ReLU) activation function. We initially trained this Encoder-Decoder network as an end-to-end network, however, the results are inferior. As a result, we chose to train them independently as two sub-networks. This approach allows for better control over the training of each individual component, enabling more precise optimization and improving the overall performance of the system. The loss function used to optimize both networks is Mean Squared Error (MSE), while the optimizer is Adam solver with weight decay = $1e-6$ and learning rate $3e-4$ for stable convergence. All experiments and model training were conducted on an NVIDIA Titan X GPU.

4.1.2 Results and Evaluations

Figure 4.2 (a) presents a comparison between the photograph and the reconstructed image under white illumination in the visible spectrum. As can be seen, the reconstructions are close to the photograph. This validates that our network can generate high-quality predictions of the full skin reflectance.

Moreover, Figure 4.2 (b-f) displays the comparison of estimated chromophore maps for five spectral parameters, obtained via the traditional Lookup Table (LUT) search method and the proposed network *ChromNN*. The small difference demonstrates the ability of *SpectraNN* to closely approximate the LUT search results while significantly improving computational efficiency. Please note that we scaled all chromophore maps to display clearly, especially C_{bc} maps which should have little difference shown in the range from 0 to 0.01.

Figure 4.3 shows the comparison of photographs and reconstructions under white (a), red (b), green (c), blue (d) illumination in the visible domain, as well as the UVA (e) and NIR (f) domains. We can reconstruct results with any uniform illuminations using the predicted skin reflectance. Here, red, green, and blue reconstructions are examples. Additionally, we do not have ground truth for reconstructions in the invisible domains, therefore, we employ the reconstructions estimated from the proposed three-stage LUT search as references. As displayed in these figure, the reconstructions in both invisible and visible domains are close to the photographs or references. These figures demonstrate that the proposed network is capable of accurately estimating the spectral parameters and reconstructing the skin appearance under any uniform illuminations.

In addition to the qualitative evaluation, we also make three quantitative evaluations of efficiency and accuracy among the conventional LUT search, the proposed three-stage LUT search, and the neural network.

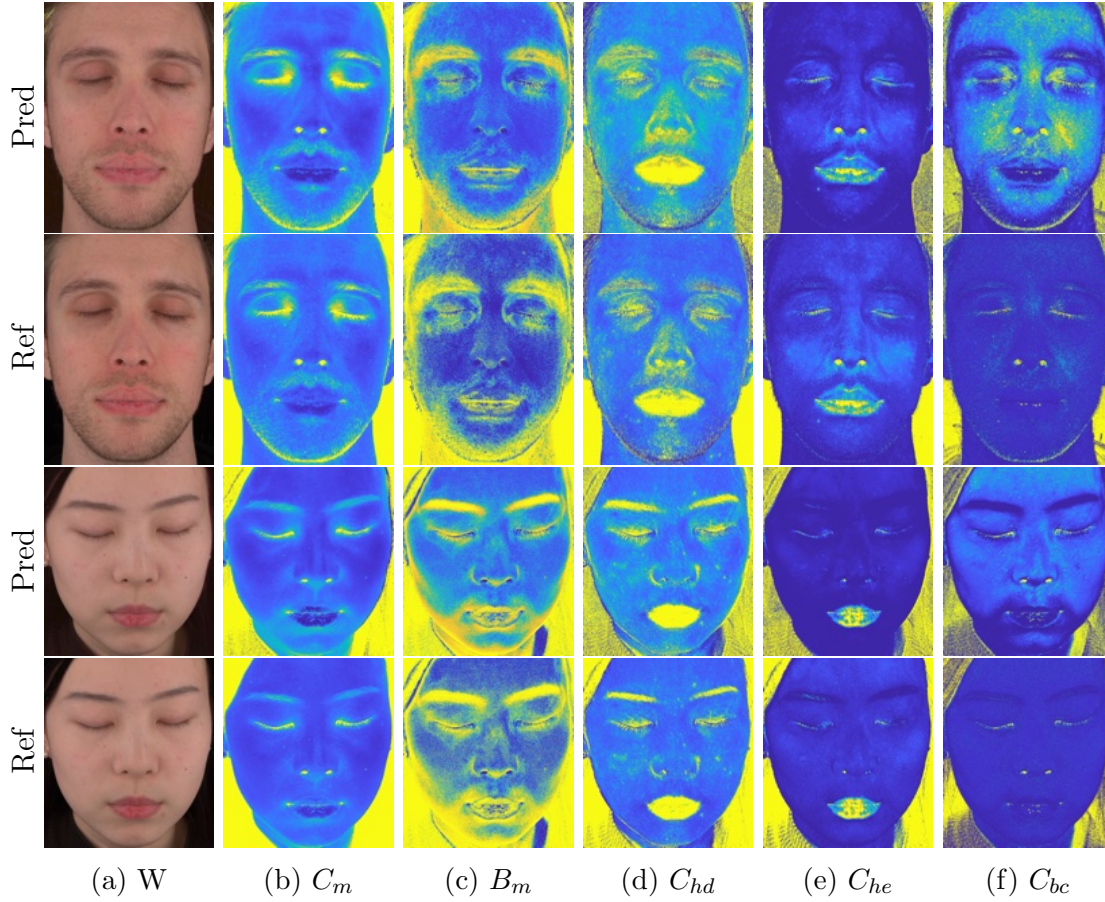


Figure 4.2: Comparison of the network predictions and references of subject 1 (Mediterranean skin type) and subject 2 (Asian skin type). Figure (a) compares the photograph and visualizations of facial reflectance produced by *SpectraNN* in the visible domain under white illumination. Given the input image, Figure (b - f) compares the predictions of five spectral parameters by the proposed three-stage LUT searching method and the *ChromNN*.

We first compare the computational efficiency of three different methods on a diverse dataset comprising various skin types and genders. For high-resolution images with $2K$ resolution, the conventional LUT search method is the most time-consuming, requiring an average of 2.88 hours per image. In contrast, the proposed three-stage LUT search significantly reduces the processing time to 13.91 minutes, demonstrating a notable improvement in computational efficiency while maintaining accuracy. The PCA-based neural network further accelerates the process, completing the parameter estimation and skin appearance reconstruction in a mere 2.27 seconds on average. The comparison clearly illustrates that the proposed Encoder-Decoder network provides a more efficient solution

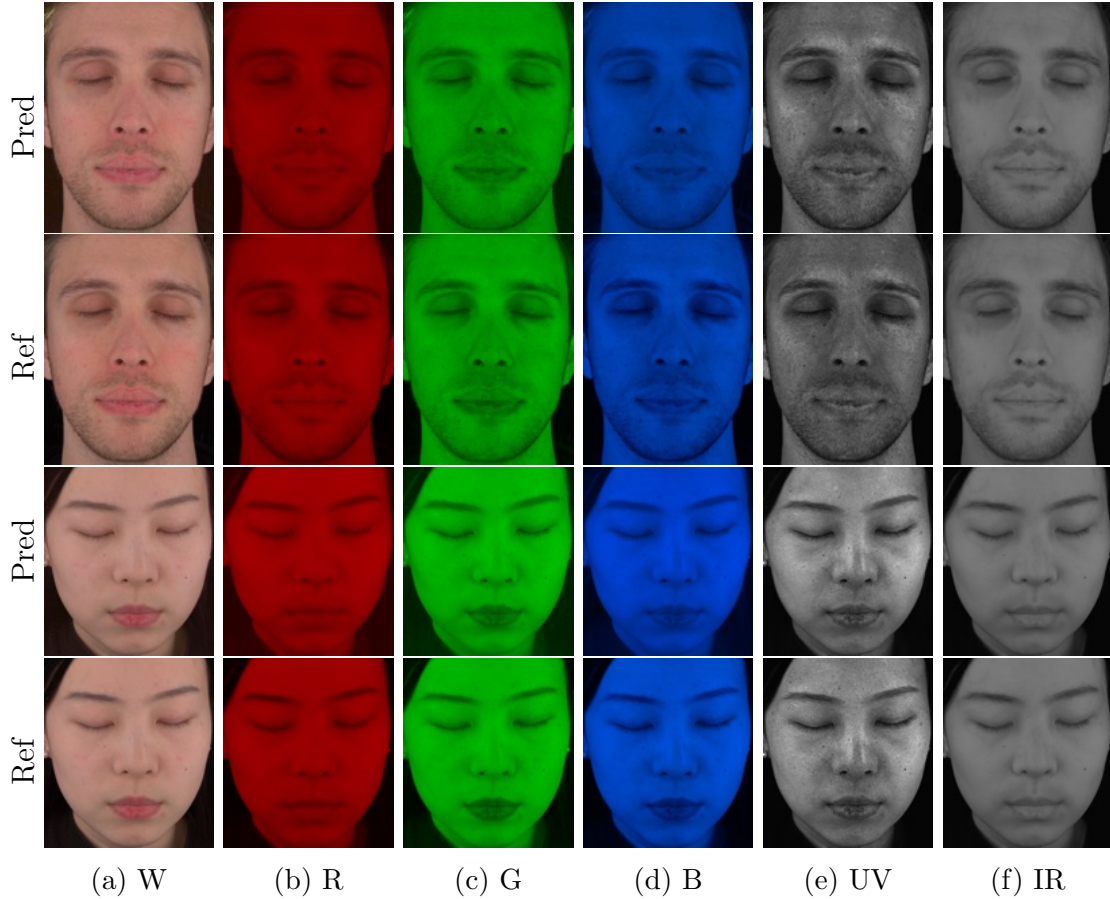


Figure 4.3: Comparison of the network predictions and references of subject 1 (Mediterranean skin type) and subject 2 (Asian skin type). Figure (a - f) compare references and visualizations of facial reflectance produced by *SpectraNN* in the visible domain under white (a), red (b), green (c), and blue (d) illuminations, and in the UVA (e) and NIR (f) domains. Here, the references for the visible domain are photographs, while the references for the UVA and NIR domains are the reconstructions generated by the proposed three-stage LUT searching method.

for hyperspectral skin reflectance modeling .

Secondly, we analyze the accuracy of the reconstructions of the spectral skin reflectance in the visible domain generated by these three methods. We chose a data set that included multiple subjects with different skin types. Table 4.1 shows the MSE loss for three methods on average. The results demonstrate that the proposed three-stage LUT search method can achieve good results comparable to those from the conventional LUT search. Despite the significant reduction in computational time, the three-stage LUT search maintains high accuracy, proving its effectiveness for hyperspectral skin reflectance

Table 4.1: Accuracy analysis

	conventional	three-stage	our network
Caucassian	1.8345e-04	1.8150e-04	3.11e-08
Mediterranean	1.8916e-04	2.0859e-04	3.02e-08
Asian	2.1900e-04	2.2337e-04	3.33e-08
South Asian	2.4326e-04	2.4697e-04	3.96e-08

estimation.

Furthermore, the neural network-based approach provides a distinct advantage over both LUT search methods by addressing the inherent limitation of discrete values associated with lookup tables. By leveraging continuous parameter estimation, the neural network can produce more accurate reconstructions. Moreover, the neural network’s capacity to generalize across different skin tones ensures that the reconstruction quality is consistent and unbiased, regardless of the subject’s skin type. Additionally, the network’s significantly reduced runtime, makes it an optimal choice for large-scale datasets where both speed and accuracy are critical.

The third quantitative evaluation compares our network and the most recent work [2]. In the work of Aliaga et al., they make a comparison with the data measured in the visible domain in the Leeds skin dataset [76], therefore we also display two examples with light and dark skin types of our reconstructions of skin reflectance and Leeds dataset in Figure 4.4. The maximum and average RMSE between our model and all the data in the Leeds dataset is 0.036 and 0.008, respectively, which is lower than that of Aliaga et al. [2] (mostly lower than 0.05). In addition to the accuracy, our PCA-based network has $\sim 2\%$ of the parameters compared to the network used by [2], indicating a more efficient approach.

Besides these quantitative evaluations, Aliaga et al. [2] propose a Monte Carlo simulation-based model with a single white image as input while our method employs a diffusion-based model with multiple inputs. In addition, our model spans a broader spectral domain from UVA to NIR, as well as practical measurement and validations in both visible and

invisible domains.

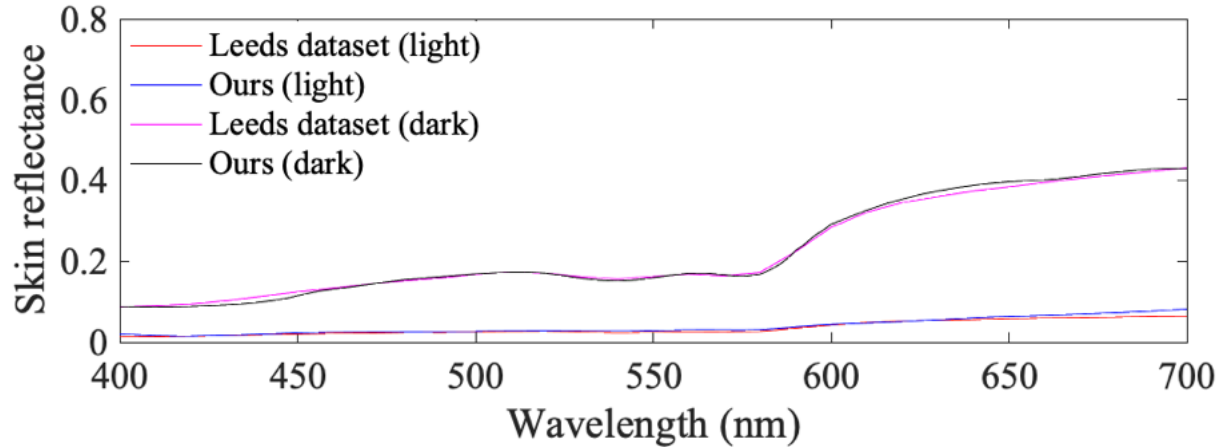


Figure 4.4: Comparison between the generated skin reflectance of our model and Leeds skin dataset of dark and light skin.

4.2 Naive Skin Model

4.2.1 Implementation

The previous section introduced a complete network architecture designed for parameter estimation and full skin reflectance reconstruction with multiple inputs. The requirements of this architecture are strict, primarily due to the difficulty in obtaining synthetic data and capturing photographs under controlled illumination and specific setups. This complexity leads to significant challenges in practical applications, particularly when reproducing realistic skin reflectance. Furthermore, reconstructing the original skin appearance directly is a more practical application. In response to these practical considerations, we propose a simplification of the initial neural network to focus on reconstructing the skin’s appearance using a single white light input, thereby reducing the complexity while still achieving accurate skin reflectance reproduction.

This simplified network architecture has a similar structure to the complete one in Figure 4.1, consisting of both an Encoder and a Decoder. In this design, the input to the

Encoder is a 3D measurement under D65 illumination. We assume that the Lookup Table (LUT) is constructed using the same illumination and the standard color matching functions. Given this measurement, the Encoder outputs a set of spectral parameters, which are then passed to the Decoder to reconstruct the original skin appearance.

The skin model behind this network is also simplified. Although the incorporation of β carotene in our model helps with high-quality reconstructions, it is still less important than the melanin and hemoglobin. Therefore, we fix the β carotene as the same value as the first stage of the proposed three-stage LUT search, thus obtaining a 4D skin model (C_m , β_m , C_{hd} , and C_{he}). This simplified network provides an efficient and direct approach to reconstructing the skin appearance, illustrated in Figure 4.5.

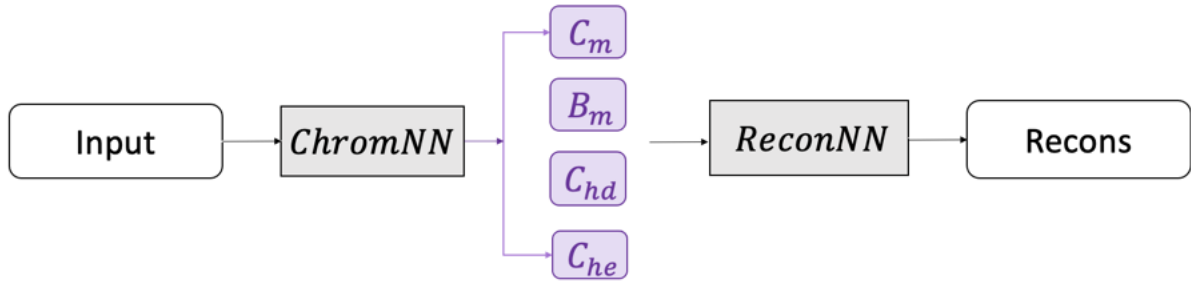


Figure 4.5: The simplified neural architecture consists of an Encoder and a Decoder. Given the images under D65 illumination, *ChromNN* makes a prediction of the spectral parameters. These estimated parameters are then fed to the *ReconNN* to reconstruct the input skin appearance.

Training and Testing Datasets

The training procedure for this simplified network follows that of the complete network architecture. The training input dataset for *ChromNN* is three channels R, G, B of LUTs under D65 illumination. The training reference dataset for the outputs of *ChromNN* is the corresponding four spectral parameters.

In contrast, the testing procedure has a significant difference. The D65 illumination used for the training dataset is a standardized illuminant, commonly representing average

daylight conditions. It is widely utilized in color science for its consistency and reliability in replicating natural lighting. However, in practical scenarios, this specific illuminant is not easily replicated or captured by various imaging setups due to differences in lighting equipment and camera sensitivities. Therefore, the testing dataset can be images captured by any setup with the known illumination and camera sensitivity curves. However, these images require to be color-corrected to the same color gamut as the LUTs before passing to this network. Please see Section 3.2 for a detailed explanation of the color correction procedure.

Implementation Details

Similar to the complete network, both *ChromNN* and *ReconNN* are designed as three-layered MLPs with ReLU activation functions. These two sub-networks are also trained separately using Adam optimizer with 3e-4 learning rate, and apply MAE as the loss function. All experiments and model training were conducted on NVIDIA Titan X GPUs.

4.2.2 Results and Evaluations

As mentioned before, the testing image dataset can be captured by any setup with the known lighting conditions and camera sensitivity curves. Here, we use the images from the monitor-based setup employed in the proposed LUT search as examples.

Figure 4.6 displays a comparison between the predictions of four spectral parameters generated by the encoder, *ChromNN*, and those derived from the conventional LUT search method. The figure includes two subjects with different skin types, including Asian skin type and Mediterranean skin type, for a comprehensive evaluation of the network’s performance across diverse skin tones.

As shown in this figure, the spectral parameter predictions from *ChromNN* are close to those obtained through the LUT search method. However, the estimated chromophores

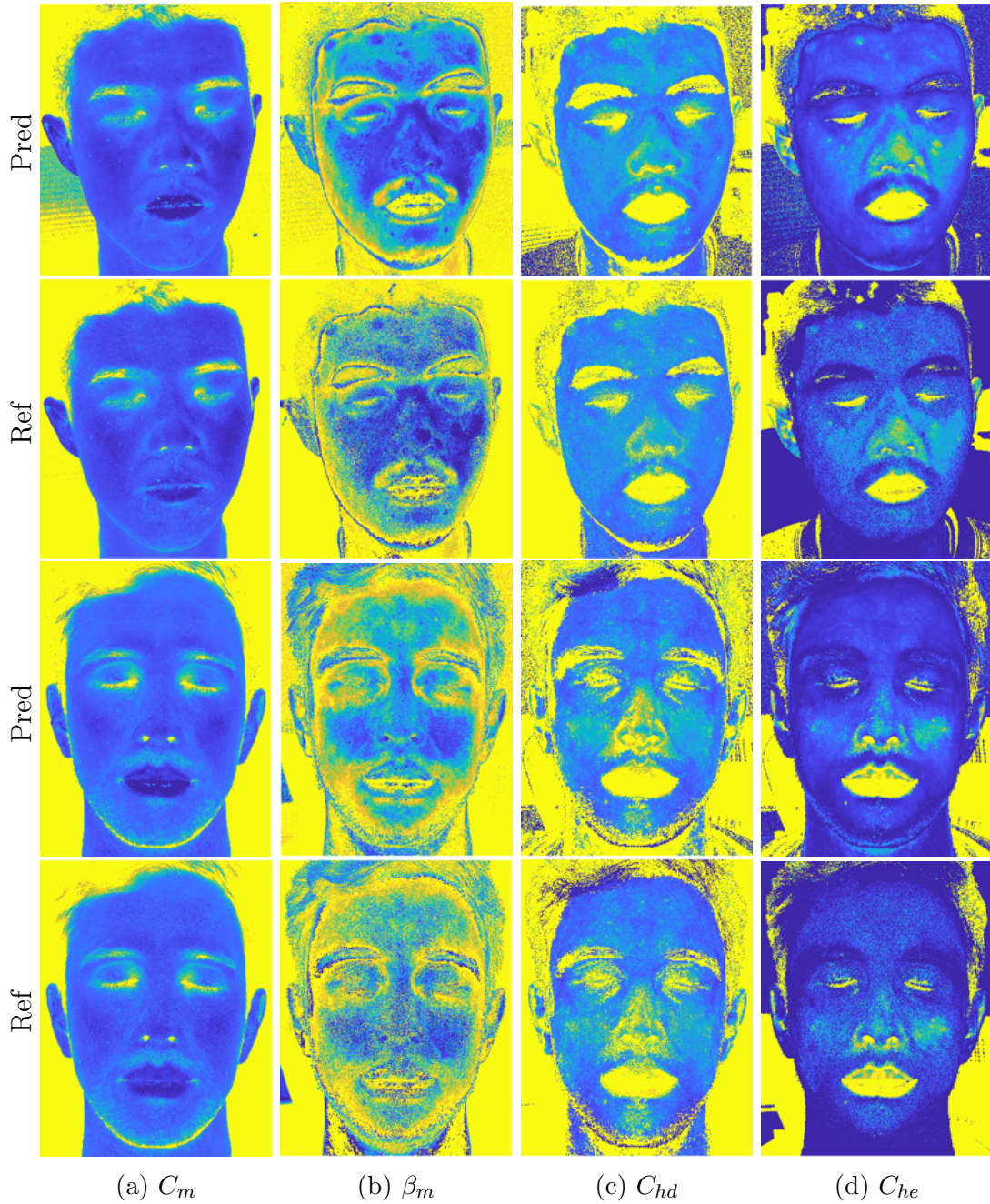


Figure 4.6: Comparison of the chromophore prediction between the proposed simplified network *ChromNN* and LUT search of two subjects with different skin types.

from the LUT search have several noises, especially in β_m and C_{he} maps. This aligns with the limitation of the LUT search that all the LUTs are discrete samplings. In contrast, *ChromNN* overcomes this limitation by generating continuous values, thereby the predictions are less noisy. This comparison underscores the advantages of using a neural network for spectral parameter estimation, particularly in terms of accuracy and

quality.

Figure 4.7 compares between the original input image and the reconstructions generated by these two methods. As illustrated in the figure, the reconstructions produced by both approaches are comparable to the original input image.

These two figures highlight the advantages of using a neural network for parameter estimation and skin appearance reconstructions, particularly in terms of accuracy and quality. Additionally, the robustness of these two sub-networks across varying skin types, as demonstrated in this evaluation, further validates its potential for generalization in practical applications, such as skin appearance editing in the following content.

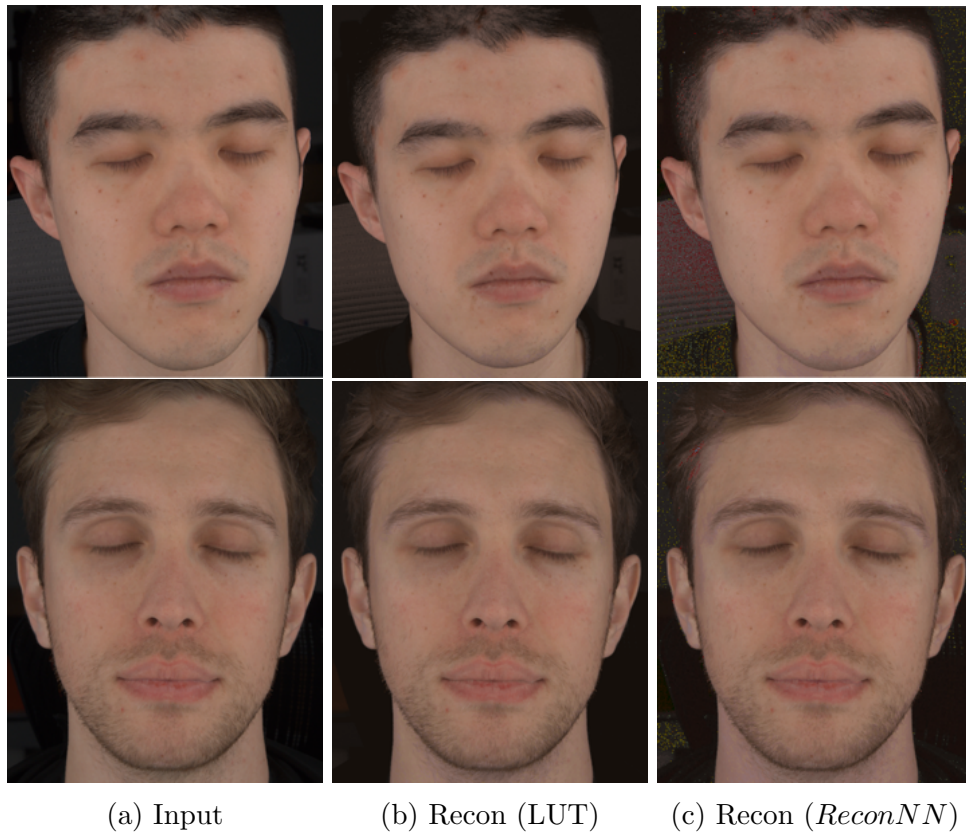


Figure 4.7: Comparison of the reconstruction between the proposed simplified network *ReconNN* and LUT search, and the input image.

4.3 Image-space Scattering Approximation

In this work, both the proposed LUT-based method and the neural network are designed to produce sharp reconstructions, effectively reducing the impact of lateral light scattering. However, as noted in the previous section, this sharpness can diminish the perceived realism when the reconstructed albedos are directly compared to ground-truth photographs, which inherently include scattering effects. To address this, we apply an image-space scattering approximation to the reconstructed albedos, yielding renderings that more closely match the appearance of real skin in the validation images.

We begin by identifying the skin reflectance curves that most closely correspond to the light and dark skin patches in a standard color chart. For each patch, we extract reflectance values at 450 nm, 530 nm, 630 nm, and 835 nm—corresponding to the dominant wavelengths of blue, green, red, and near-infrared (NIR) illumination in our setup. Following the approach of [56], we fit each curve using a linear combination of seven Gaussian functions. These Gaussians share the same variance, while each reflectance profile assigns different weights, as shown in Figure 4.8. The reconstructed albedos are then blurred with each Gaussian kernel, and the final result is obtained by linearly combining the blurred outputs. We omit this blurring step for the UVA reconstruction, as it is derived from measurements under longer wavelengths (i.e., visible light), which already encode a degree of spatial smoothing.

Figure 4.9 presents a qualitative comparison between the reconstructed results and the reference photographs under five illumination conditions: white (a), red (b), green (c), blue (d), and infrared (e). For the broadband white image, the RGB channels are processed independently, since the monitor’s white illumination is composed of three separate narrowband RGB peaks. For the IR LED panel, we additionally simulate shading effects due to its limited frontal area and incorporate specular highlights. The images are brightened slightly to enhance the visibility of fine details. As illustrated in the figure, the

application of image-space scattering improves the visual fidelity of the reconstructions, bringing them closer to the appearance of real skin in the photographs.

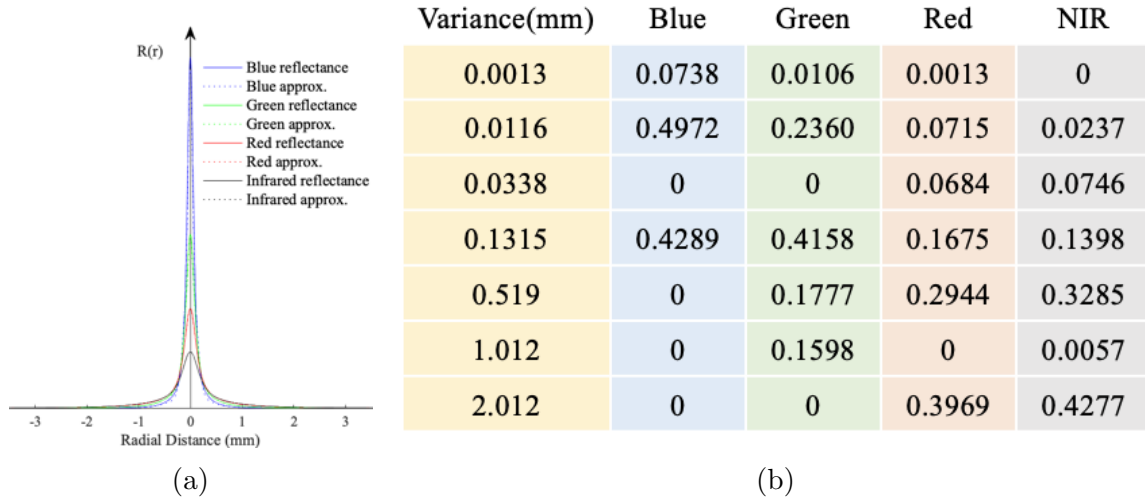


Figure 4.8: Approximating profiles (a) with sums of seven Gaussians (b).

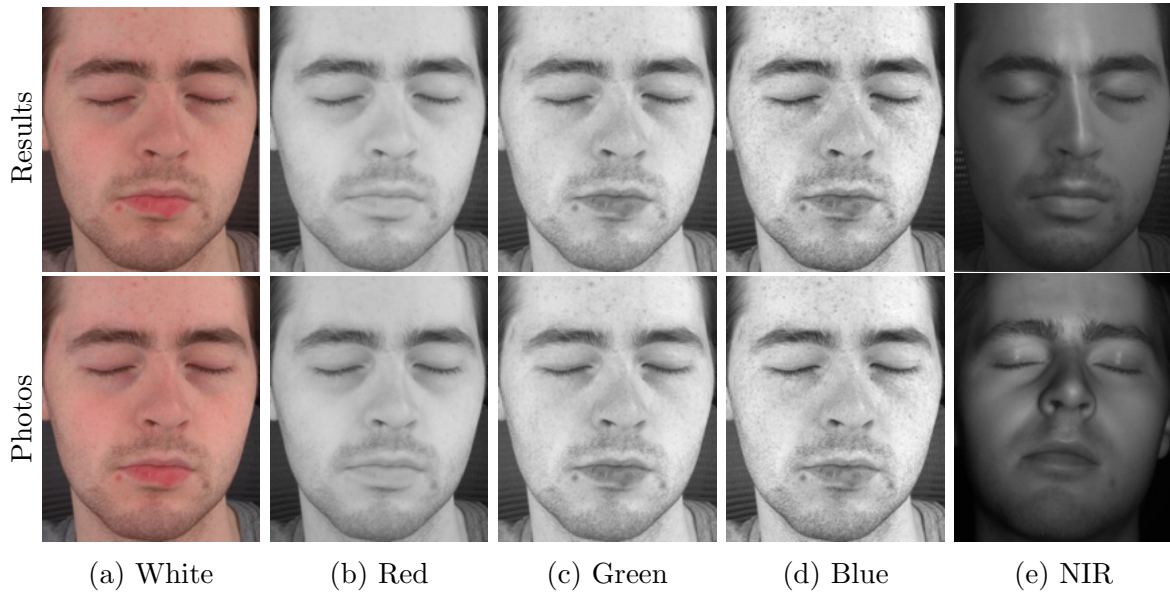


Figure 4.9: Comparison of image-space scattering approximation and photos (or channels) under white (a), red(b), green (c), blue (d), IR (e) illumination.

4.4 Application - Skin Appearance Editing

The biophysical skin model has a wide range of practical applications. One significant application is in the monitoring of skin health. For instance, individuals can track changes in

their skin's status, such as hydration levels, by observing reflectance in the NIR spectrum. This is useful when assessing the efficacy of skincare treatments, cosmetic products, or medical beauty procedures.

Moreover, the model has potential applications in the early detection of skin conditions. For example, by analyzing the skin's reflectance in the UVA spectrum, it is possible to identify early signs of pigmentation abnormalities. Such abnormalities can serve as indicators of potential skin damage, including the precursors to skin cancer. Early detection of these changes allows individuals to take preventative measures to reduce the risk of developing more serious conditions.

Here, we present an attractive application of this model: skin appearance editing. In this application, we can simulate biophysical changes in skin colors by modifying these parameters.

By adjusting two melanin-related parameters, we can simulate changes in skin color by modifying the melanin concentration, C_m , while also altering facial hair, such as stubble, through the parameter β_m . Figure 4.10 illustrates various skin appearances, ranging from lightening to tanned skin due to changes in melanin concentration. Figure 4.11 demonstrates the impact of β_m , showing how it can increase or decrease facial hair density, particularly in the stubble area for male subjects. These visualizations highlight the effectiveness of using melanin-related parameters for realistic skin and facial hair editing.



Figure 4.10: Editing melanin concentration C_m

Figure 4.11: Editing melanin blend β_m

Similar to editing melanin-related parameters, we can also edit the concentration of β carotene to change the skin appearances displayed in Figure 4.12.

Figure 4.12: Editing β carotene parameters C_{bc}

The two hemoglobin-related parameters, C_{hd} and C_{he} , are crucial in simulating variations in facial redness, as illustrated in Figure 4.13. This figure displays a set of skin appearances from pale to flushed tones. Unlike the uniform changes observed with melanin, alterations in hemoglobin typically manifest as localized variations, which can significantly influence skin health and emotional expression. To effectively capture these changes, we employ a simple but efficient method to adjust the hemoglobin parameters across the entire facial surface non-linearly. The equation for editing C_{hd} and C_{he} is

$$C_{mod} = C \times \left[1 + \frac{(s - 1) \times i}{N} \right] \quad (4.1)$$

where i is the index of the value C in the array of C_{hd} or C_{he} with the size of N while s is the scaling factor. With this scaling method, the local variation of hemoglobin distributions is accentuated in the edits.



Figure 4.13: Editing hemoglobin related parameters C_{hd} and C_{he}

4.5 Conclusion

In this Chapter, we first introduced an Encoder-Decoder neural network architecture designed to estimate spectral parameters and reconstruct full hyperspectral skin reflectance from multiple measurements. The predicted skin reflectances can be further visualized to the invisible domain (UV and NIR domains), as well as the visible domain under various uniform illuminations. This approach addresses two main limitations of the LUT search described in Chapter 3: it significantly improves the accuracy of spectral reconstruction by overcoming the discretization inherent in LUTs, and it reduces the parameter estimation time from hours to seconds, enabling practical use in real-time or large-scale applications.

Later, we proposed a simplified version of the method that takes a single input to directly reconstruct skin appearance, which enhances the efficiency of the process for practical use cases. To further improve the realism of the reconstructed skin appearance, we incorporated an image-space scattering approximation, providing a more accurate representation of light interactions within the skin tissue.

Finally, we demonstrated the practical utility of our biophysical skin model in the context of skin appearance editing. By adjusting chromophore parameters, we simulate various skin changes, including tanning, lightening, flushing, and paleness. This highlights the potential of our model for biophysical skin editing, paving the way for a wide array of applications in fields such as virtual reality, the cosmetic industry, and personalized

healthcare.

Except for its accuracy and effectiveness, the proposed method has several limitations. First, the training data for both networks are generated from precomputed Lookup Tables (LUTs) based on our biophysical skin model. As a result, the input conditions are strictly constrained to match the assumptions of the model. Even in the case of the simplified network, input still requires a facial image captured under white illumination, which limits generalization to uncontrolled lighting conditions. Second, unlike recent works incorporating generative models to handle entire facial images, our approach focuses exclusively on the skin region. This deliberate exclusion of non-skin areas, such as hair and clothing, can lead to less accurate reconstructions outside the targeted skin regions, reducing overall image realism. Furthermore, although the current framework adds image-space scattering approximation to improve visual fidelity, it does not leverage local spatial context, which could potentially enhance robustness and visual consistency, especially in high-frequency regions.

Chapter 5

Realistic 3D Facial Age Transformation

While existing facial re-aging methods can produce realistic results, they purely focus on the 2D age transformation. In this work, we introduce a novel approach that transforms both the facial appearance and geometry across different ages while preserving the identity of each subject (Section 5.1). Our method employs an α -(de)blending diffusion network *AgeEditNet* with an age- α transformation to simulate structural changes, such as wrinkles, while also editing biophysical skin properties, specifically melanin and hemoglobin concentrations, by *SkinNet* to simulate age-related skin color changes (Section 5.2). This enables realistic re-ageing from 10 to 80 years. Additionally, we propose a geometric neural network *GeoNet* to adjust coarse facial geometry according to age, followed by an efficient network *RefineNet* that adds high-frequency skin displacement (Section 5.3). Through both qualitative and quantitative evaluations, we demonstrate that our method outperforms current state-of-the-art methods (Section 5.4). In the end, we make a conclusion and discussion in Section 5.5.

5.1 Overview

In overview, we present a novel 3D re-aging method editing in both appearance and shape from a single in-the-wild image. Our approach begins by estimating the subject’s age using a pre-trained age predictor, namely DEX [68]. The predicted source age and the specific target age are then input into *SkinNet*, which simulates biophysical changes in skin appearance related to aging, including variations in melanin and hemoglobin concentrations. We further apply α -(de)blending through our diffusion-based network, *AgeEditNet*, with age- α modification to simulate the facial structure changes due to age.

Simultaneously, we utilize the pre-trained encoder from DECA [22] to extract parameters, including lighting, camera settings, shape, expression, and pose. Of these, shape, expression, and pose parameters are used in our geometric network, *GeoNet*, which reconstructs the subject’s coarse 3D shape. GeoNet is particularly useful for controlling head size when dealing with younger age groups, ensuring appropriate face reconstructions for children.

To add the coarse shape with appropriate age-related details, we extract displacement maps from re-aged results using our *RefineNet*. This enables us to capture high-frequency geometric features such as wrinkles. The resulting detailed 3D shape is then combined with re-aged textures, lighting, and camera parameters to render a realistic final image. This pipeline allows for accurate and realistic age transformations in both appearance and geometry, covering a wide range of ages from children to the elderly.

The entire pipeline is illustrated in Figure 5.1, which provides a clear overview of the key stages: from age prediction, texture and shape editing, to the final rendering of the re-aged 3D model. This pipeline allows for accurate and realistic age transformations in both appearance and geometry, covering a wide range of ages from children to the elderly.

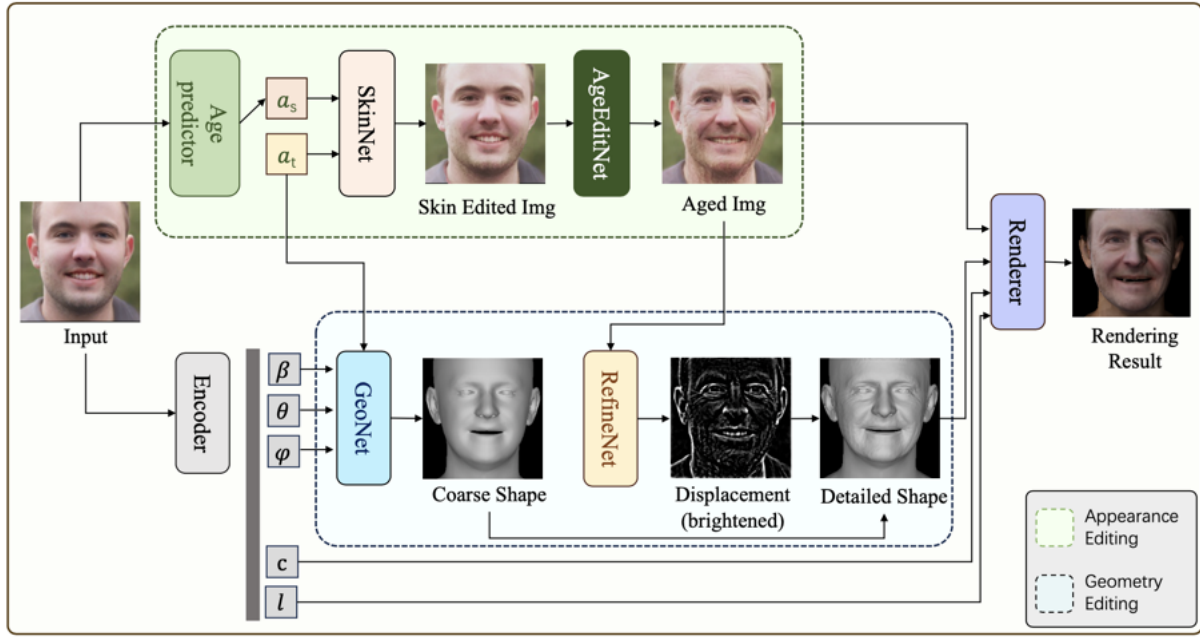


Figure 5.1: **Overview.** Starting with a single image I as input, our method first employs an age predictor [68], to estimate the source age a_s . With a given target age a_t , *SkinNet* edits the biophysical skin appearance due to ageing based on a skin model while *AgeEditNet* changes the structural aging effects iteratively from the source age a_s to the target ages a_t . Subsequently, we use the trained *Encoder* of *DECA* [22] to predict lighting, camera, and shape information. The next step involves *GeoNet*, which processes shape (β), expression (ψ), pose (θ), and age (a) parameters to reconstruct a coarse shape. This is followed by *RefineNet*, a lightweight network designed to predict a displacement map. Finally, we render the resulting image using the re-aged texture and detailed shape, incorporating the camera (c) and lighting (l) parameters. Here, we brighten the displacement map to show the details more clearly.

5.2 Appearance Editing

The facial age transformation contains both coarse structural changes, such as wrinkles, and skin tone changes resulting from the optical properties varying from different ages. In this part, we simulate the effects of ageing on facial appearance through two primary directions: biophysical skin editing and structural aging modifications.

5.2.1 Biophysical Skin Editing

Chapter 3 and Chapter 4 make a detailed introduction to the biophysical skin model and several methods for parameter estimation and reconstruction of full spectral skin reflectance and appearances. Built upon this model, we present an application focusing on skin appearance editing, where we demonstrate the effects of editing various chromophore parameters, such as melanin and hemoglobin, to simulate targeted skin tone changes.

As the optical properties of skin tissue change due to age, we follow this application to present an approach to simulating the effects of skin ageing by adjusting the concentrations of main chromophores, particularly melanin, and hemoglobin.

Similar to the network structure in Section 4.2, the *SkinNet* also consists of two sub-networks *PredNet* and *ReconNet*. Given a single input image, *PredNet* estimates the chromophore parameters, which are further fed into *ReconNet* to reconstruct the skin appearances. Please see Section 4.2 to find the detailed description of the training and testing datasets and the corresponding implementation.

So far, the network is limited to reconstructing reconstruct the original skin appearance. To extend its functionality and simulate skin tone variations due to aging, we introduce an additional editing stage between the two sub-networks. This intermediate process allows us to model the aging effects on facial appearance, capturing the changes in skin tone related to age. The structure of this part is shown in Figure 5.2.

According to the work of [35], melanin and hemoglobin are the two most important chromophores affecting skin colors. Therefore, rather than modifying several parameters, such as the thickness of each skin layer, we mainly focus on adjusting the concentrations of these two chromophores to simulate changes in skin tones effectively. Please find the details of this excellent work [35] in Section 2.4. The equations for computing the edited melanin and hemoglobin parameters are shown below:

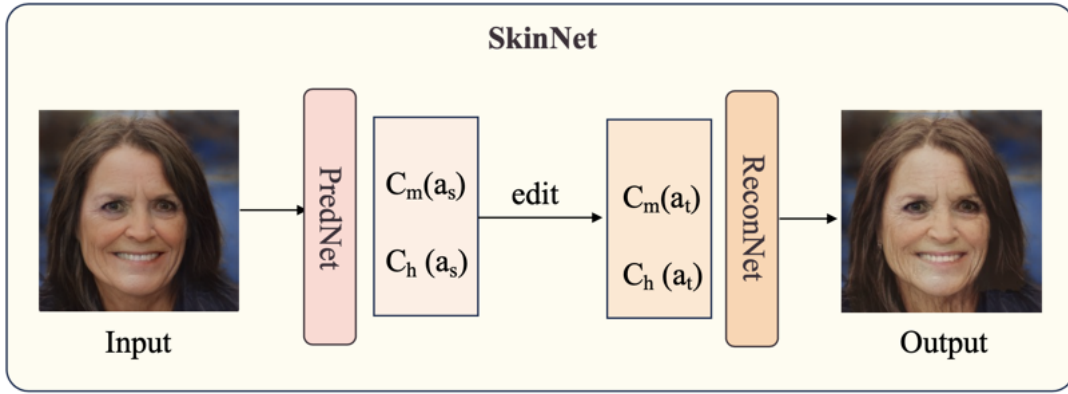


Figure 5.2: **Biophysical skin editing.** The *SkinNet* consists of *PredNet* and *ReconNet*. *PredNet* aims to estimate the melanin and hemoglobin concentrations C_m and C_h from the given input image. The estimated chromophores are edited according to the source age a_s and target age a_t . These two parameters are then fed into *ReconNet* to produce the re-aged skin appearances.

$$\begin{aligned} C_m(a_t) &= \left(1 + \frac{k_m(a_s - a_t)}{10}\right) \times C_m(a_s), \\ C_h(a_t) &= \left(1 + \frac{k_h(a_s - a_t)}{10}\right) \times C_h(a_s) \end{aligned} \quad (5.1)$$

where

- C_m represent the concentration of melanin, and C_h represent the concentration of hemoglobin;
- $k_m = -0.08$ and $k_h = -0.06$ refer to the decrease in concentration of melanin and hemoglobin per decade as a function of age [35];
- a_s and a_t are the source and target ages, respectively.

Ablation Study

Here, we implement two ablation studies. The first study evaluates the impact of biophysical skin editing on the realism of simulated aging effects. To illustrate this, we compare results with and without the biophysical skin editing between the two sub-networks, as shown in Figure 5.3. Given an input image of a 20-year-old subject (Figure 5.3(a)), we

re-age the subject to 60 years old. The result obtained with biophysical skin editing (Figure 5.3(c)) is a significantly more realistic representation of ageing, displaying a lighter skin tone. In contrast, the output without biophysical skin editing (Figure 5.3(b)) lacks the skin tone changes, resulting in a less convincing aged appearance. This comparison highlights the important role that biophysical skin editing plays in simulating the realistic effects of ageing.

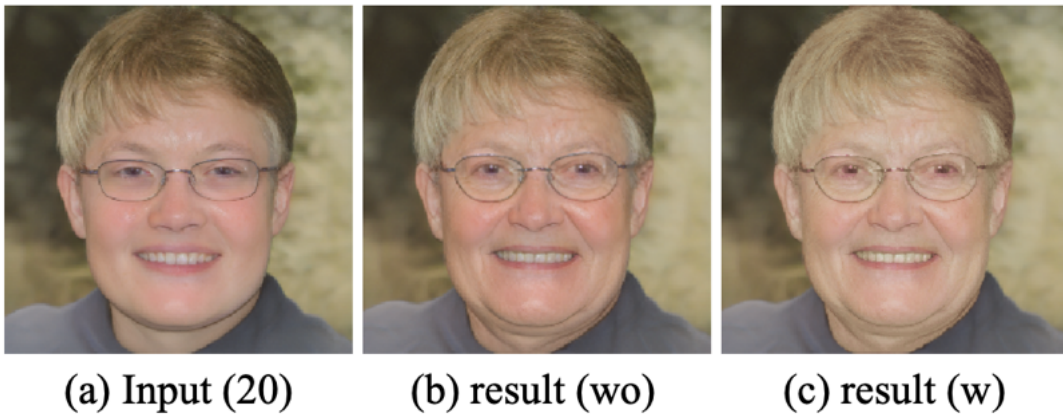


Figure 5.3: Comparison between re-ageing results with and without biophysical skin editing.

In the second study, we discuss the impact of the decrease in melanin and hemoglobin concentrations k_m and k_h , respectively. Figure 5.4 illustrates an example where the source age is 60 years old, and the target age is 30 years old. In the first row of the figure, we fix the hemoglobin concentration k_h as -0.06 while decreasing the melanin concentration k_m from 0 to -0.16. This modification results in progressively darker skin tones, as seen in images (b) through (d). In the second row, we fix the melanin concentration k_m as -0.08 and decrease k_h , which leads to increasingly flushed facial appearances from images (e) to (h). These observations align with the previous conclusion in Section 4.3. According to [35], we select $k_m = -0.08$ and $k_h = -0.06$ as optimal values for simulating realistic age-related changes in skin tone.

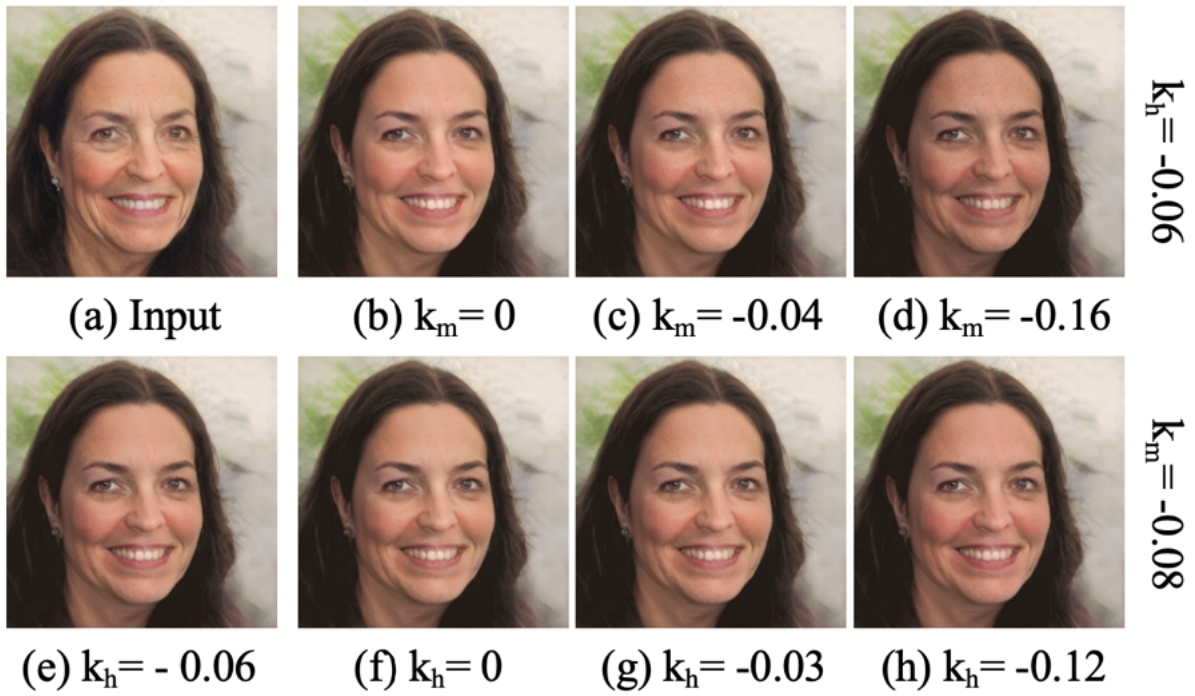


Figure 5.4: An ablation study of the biophysical skin editing with different decreases of chromophore concentrations k_m and k_h . The subject in (a) is 60 years old, and the target age is 30 years old, so we expect an increase in chromophores concentration.

5.2.2 Ageing Structural editing

After addressing skin tone alterations through biophysical editing, we next introduce the method for simulating the structural ageing changes in the following content. The main challenge for this part is the absence of large-scale facial image datasets with paired ages over a long period. The lack of such datasets makes it difficult to simulate gradual age progression. Additionally, it is also a complex task to maintain individual identity consistency.

To address these two challenges, we employ *AgeEditNet* to simulate facial structural changes due to age. *AgeEditNet* is an α -(de)blending diffusion network [31], a robust and efficient framework designed to iteratively map data between arbitrary densities. This network provides significant advantages in facial re-aging tasks, as it allows for smooth and gradual transformations while maintaining the features of the input data. By leveraging

the strengths of this network, we can effectively preserve the identity of the input faces throughout the re-aging process.

One of the main advantages of *AgeEditNet* is its ability to precisely edit facial features from a source age to a target age by adjusting the blending parameter α . However, directly adjusting the parameter from α_s to α_t does not produce realistic results when transitioning from the source age a_s to the target age a_t . This is because the relationship between α and age is non-linear. To address this, we introduce an age- α modification step before feeding the data into the network. This pre-processing step ensures that the transformation between ages is more accurately mapped, resulting in smoother and more realistic age progression or regression.

Figure 5.5 illustrates the structure of this part. Given an input image, we first utilize a trained age predictor [68] to estimate the source age a_s . Once the source age is determined, we select a target age a_t , and through the age- α modification process, derive the corresponding controllable parameters α_s and α_t . These parameters, along with the input image, are then fed into the *AgeEditNet* to generate the re-aged result. This framework allows for precise control over the age transformation, ensuring a smooth and realistic progression from the source age to the target age.

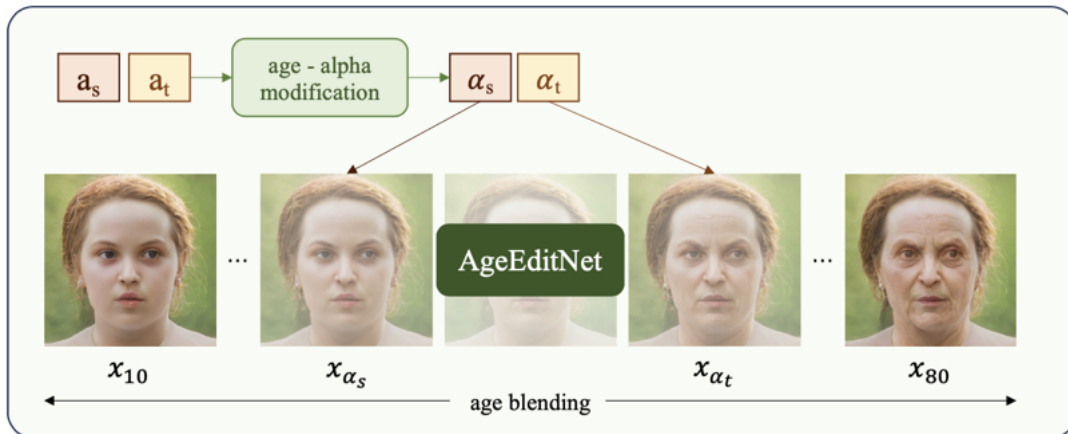


Figure 5.5: Ageing Structural editing architecture.

Training and Testing Datasets

As previously described, the main challenge is the lack of a large image dataset with paired ages. To address this issue, we follow a methodology of [82] by employing SAM [1] to create a paired dataset, which we denote it as *Lifespan Dataset*, based on the FFHQ-Aging dataset [58]. This dataset is an extension of the Nvidia FFHQ dataset [42] and contains 70,000 images with age and gender information. The generated *Lifespan Dataset* is a high-quality facial dataset that provides paired images ranging from ages 10 to 80. This synthetic dataset facilitates the training and evaluation of this task, allowing for a large dataset with paired ages and preserving the facial identity of each subject.

Implementation Details

The detailed description of α -(de)blending diffusion network is shown in Section 2.3.2. For this task, the training process is

$$x_\alpha = (1 - \alpha)x_{10} + \alpha x_{80} \quad (5.2)$$

where x_{10} and x_{80} are the facial images with 10 and 80 years old, respectively, and α is a random value from 0 to 1.

To train this network, we employ Adam optimizer with a learning rate of $4e - 3$ to minimize the differences between the predictions and the references. We utilize the MSE loss function defined as

$$\mathcal{L}_{age} = ||O_{80} - I_{80}||_2 + ||O_{\alpha_r} - I_{\alpha_r}||_2 \quad (5.3)$$

where O_{80} and I_{80} represent the predicted and reference images at 80 years old, respectively. Similarly, O_{α_r} and I_{α_r} denote the predicted and reference images corresponding to the actual age, respectively.

We train this network for 200 epochs. Once trained the network, we leverage *AgeEditNet* to sample the source image I_s from α_s to α_t iteratively. As the training process converts faces from 10 years old to 80 years old, the sampling processes have different orders depending on the source age and target age. If the target age is older than the source age, the sampling is performed in ascending order. Conversely, if the target age is younger than the source age, the sampling is in descending order. This adaptive sampling process of the age transformation in both directions, is defined as follows:

Sampling process

Require: $I_s, \alpha_s := \frac{s}{T}, \alpha_t := \frac{t}{T}$
if $\alpha_t > \alpha_s$ **do**
 for $i = s, \dots, t$ **do**
 $I_{\alpha_{i+1}} = I_{\alpha_i} + D_\theta(x_{\alpha_i}, \alpha_i)$
else do
 for $i = t, \dots, s$ **do**
 $I_{\alpha_{i+1}} = I_{\alpha_i} - D_\theta(x_{\alpha_i}, \alpha_i)$

Results

Figure 5.6 presents a comparison between the source image of a 20-year-old (left) and the re-aged results for 40 and 60 years old (middle and right, respectively). As this subject ages, her face appears several obvious structural changes, including wrinkles around her chin and eyes. The figure clearly demonstrates that this network is capable of producing high-quality re-ageing results across multiple ages while preserving the facial identity.



Figure 5.6: Re-aged result from 20 years old (left) to 40 (middle) and 60 (right) years old, respectively.

Ablation Studies

Here, we implement two ablation studies. In the first study, we discuss the significance of the age- α modification in the re-ageing task. It is well-recognized that facial changes vary across different age intervals. For instance, individuals typically have more obvious facial structural changes between the ages of 10 and 20 compared to those between 20 and 30 years old.

Figure 5.7 illustrates the re-aging results from 20 years old to both 10 and 30 years old, comparing results obtained with and without the age- α modification. Notably, the results in 30 years old for both with and without the modification are quite similar. However, the de-ageing result with the age- α modification (Figure 5.7 (b)) is more realistic than the result generated without the modification (Figure 5.7 (d)). This demonstrates the necessity of this age- α adjustment, thus enabling to simulate more believable age transformation.

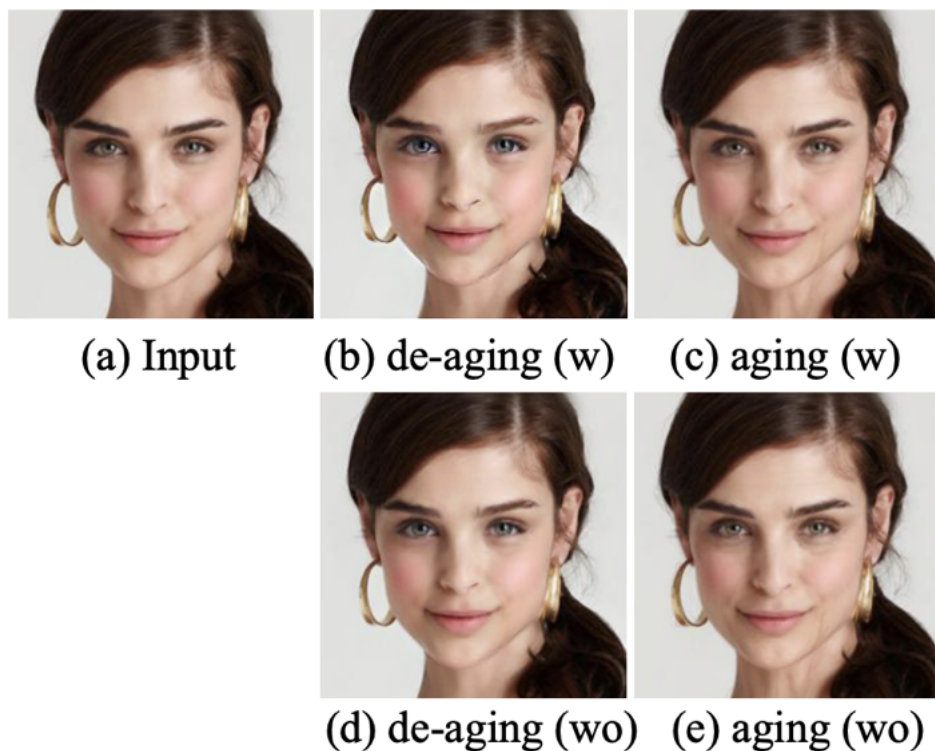


Figure 5.7: Ablation study of age-alpha transformation in *AgeEditNet*. The image is taken from the CelebA dataset.

In the second ablation study, we discuss the order of the *SkinNet* and *AgeEditNet* within our pipeline. As illustrated in Figure 5.8, the pre-processing result (c) shows more obvious ageing effects compared to the post-processing result (b). To display the details clearly, we crop the areas of the left eye and chin of the results where the differences are particularly evident, shown at the bottom of Figure 5.8 (d, e).

The main reason is that *SkinNet* is designed to adjust skin tones based on the specific source and target ages. For instance, when re-aging an input image from 20 years old to 60 years old, *SkinNet* effectively produces a lighter and paler skin tone, including in the regions where wrinkles are present.

Therefore, we determine that placing *SkinNet* before *AgeEditNet* in our pipeline yields superior results. This order allows for more effective editing of skin tones prior to the application of age modifications, ensuring to generate realistic re-ageing results with both structural changes and skin tone editing.

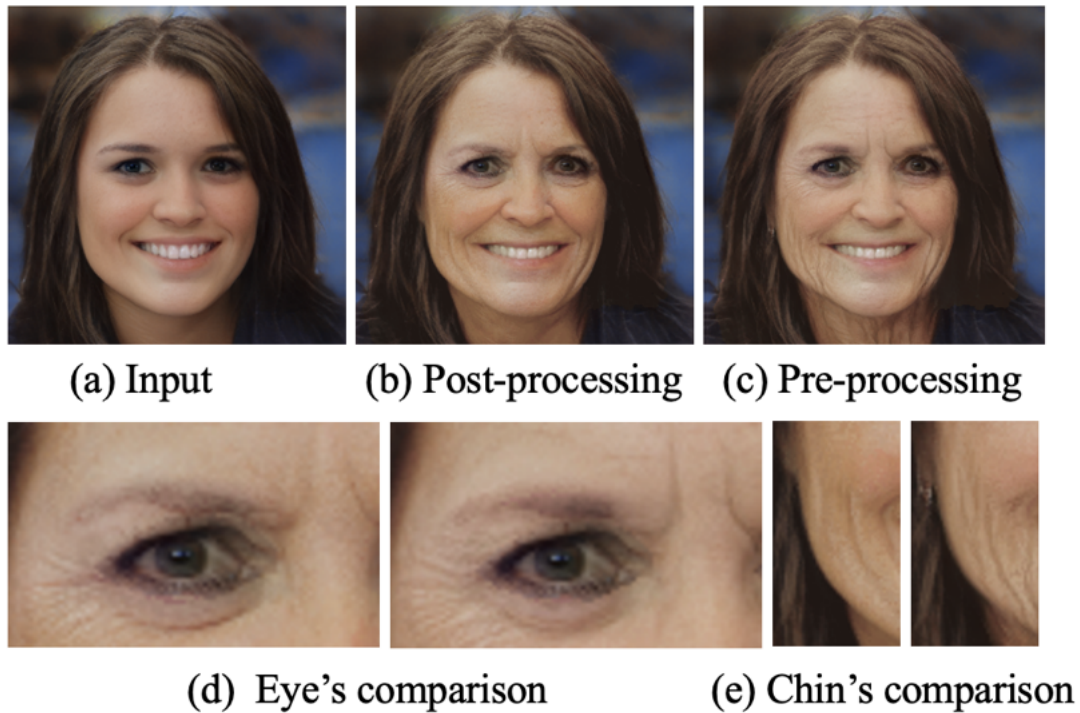


Figure 5.8: Ablation study of the biophysical skin editing before and after the structural changes.

Final Results

Figure 5.9 show the results of two subjects with different genders. The male subject is aged from 20 years old to 80 years old, while the female subject is transformed reversely from 60 years old to 30 years old. These results are realistic resulting from the combination of the facial structural changes and the biophysical skin tone modification due to age.

We would like to remind the readers that these two examples are generated at a higher resolution of 512×512 . Other results are produced at low resolution mainly due to the hardware limitations. In contrast, these higher-resolution examples demonstrate that our method can be adapted to high resolution. Moreover, our method with high resolution has a better performance in showing more and clearer ageing effects.

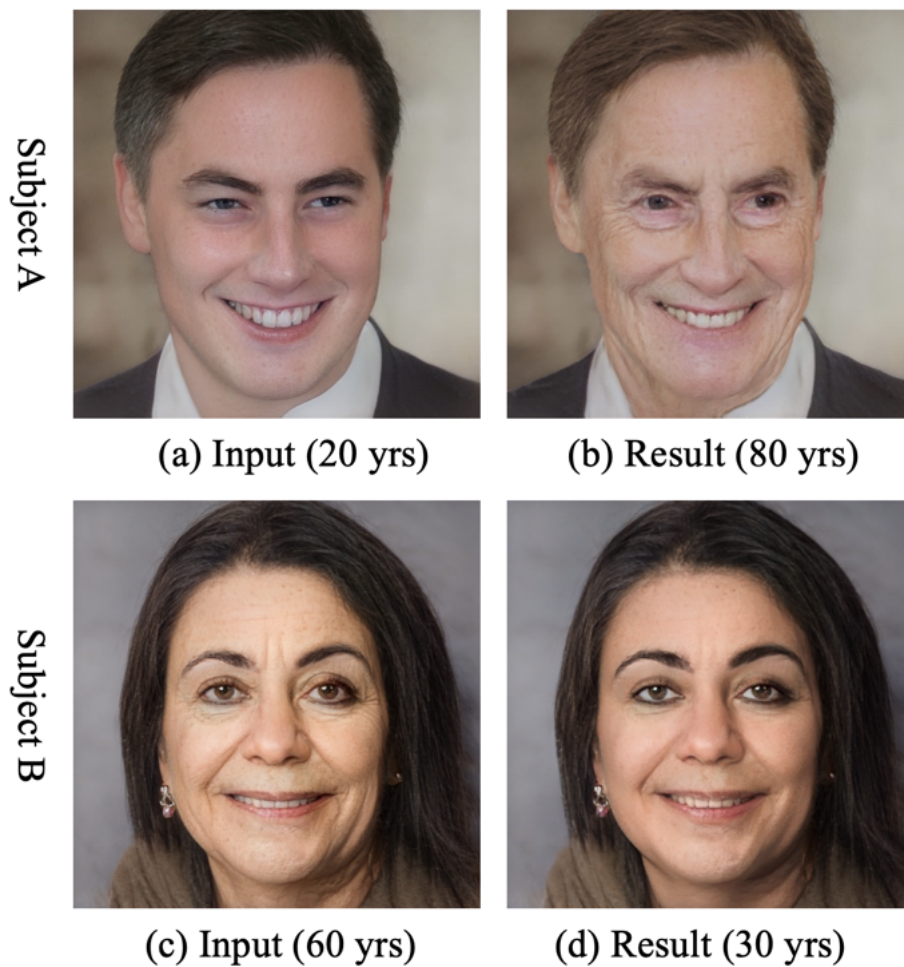


Figure 5.9: Facial age transformation at high resolution.

5.3 Shape Reconstruction

In addition to editing textures, we uplift 2D facial images to 3D models. This process requires age-appropriate details in both the coarse and detailed shape reconstructions. To achieve this, we first focus on reconstructing the coarse shape from a single input image, as discussed in Section 5.3.1. This step establishes a foundational 3D model with the general feature of the face. Then, we generate a displacement map from the 2D re-aged results, which allows us to refine the model by adding more ageing details. This ensures that the final 3D shape incorporates age-related features, such as wrinkles, described in Section 5.3.2.

5.3.1 Coarse Shape Reconstruction

To reconstruct the coarse shape from a single input image, we leverage the pre-trained Encoder from [22], which predicts the camera (c), lighting (l), shape (β), expression (ψ), and pose (θ) parameters. The last three parameters are further passed through *GeoNet*, which is capable of generating the coarse 3D facial shape. At the same time, the camera and lighting parameters are employed for rendering the image, ensuring the reconstructed model aligns properly with the visual perspective and lighting conditions of the original input.

Current methods for coarse shape reconstruction typically do not account for age variations, especially when considering head size, which is important for accurately modeling children’s faces. Therefore, except for the three shape-related parameters, we propose to incorporate age as an additional factor into *GeoNet*.

To ensure that age-specific adjustments are appropriately applied, we introduce a gate that allows for selective control of the head size based on the age parameter. This gate ensures that head size remains constant for adult subjects while allowing for proportional

changes in younger subjects, where head size differences are more obvious. This modification introduces a more precise representation of facial geometry across a wide age range, enabling age-consistent facial reconstructions and enhancing the realism of our 3D models.

GeoNet

GeoNet is a convolutional mesh network similar to the architecture of the Decoder proposed in [64], with its architecture shown in Figure 5.10. This network starts with a fully connected layer that maps the input facial images to a latent space representation, followed by five blocks and a Chebyshev Convolution layer [17]. Each block in *GeoNet* includes an Up-Sampling layer, which aims to increase the vertices numbers of the mesh. This is followed by a Chebyshev Convolution layer [17] with $K = 3$ Chebyshev polynomials, and also an Exponential Linear Unit (ELU) activation function [14]. This network enables to reconstruction of a 3D coarse shape from a single 2D facial image with controlling head size for children. Table 5.1 shows the detailed architecture of *GeoNet*.

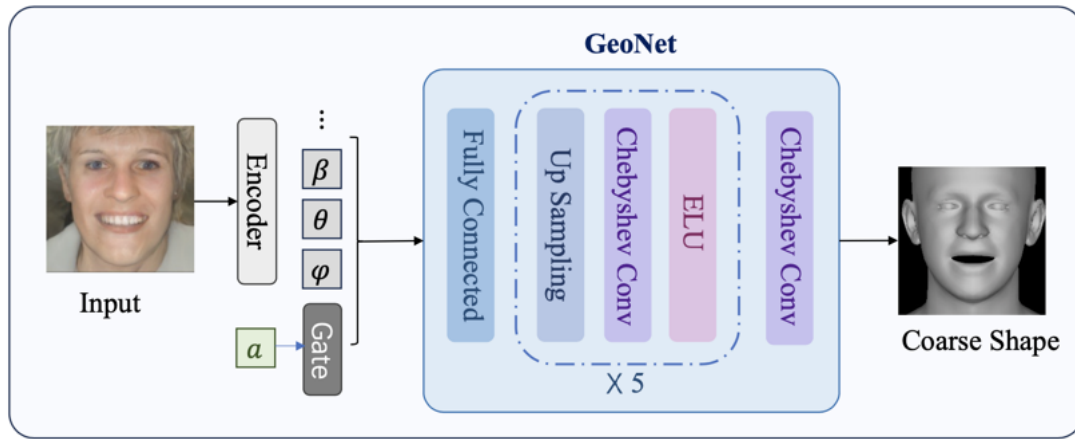


Figure 5.10: **Coarse shape reconstruction.** Given the shape (β), expression (ψ), pose (θ) parameters predicted by the pre-trained *Encoder*, along with an age parameter a , *GeoNet* reconstructs a coarse shape of the input image.

After obtaining the coarse shape, we integrate the camera parameter (c), lighting parameter (l), and the coarse geometry (G) to render the final result. The rendering process is defined by the equation $\mathcal{R}(G, S, c)$, where S represents the shaded face, which is computed

Table 5.1: *GeoNet* architecture

Layer	Input Size	Output Size
Fully Connected	160	5×128
Up-Sampling	5×128	20×128
Chebyshev Convolution	20×128	20×128
Up-Sampling	20×128	79×128
Chebyshev Convolution	79×128	79×64
Up-Sampling	79×64	314×64
Chebyshev Convolution	314×64	314×16
Up-Sampling	314×16	1256×16
Chebyshev Convolution	1256×16	1256×16
Up-Sampling	1256×16	5023×16
Chebyshev Convolution	5023×16	5023×3

by simulating the interaction of light with the facial surface geometry, defined as

$$S(T, l, N)_{i,j} = T_{i,j} \odot \sum_{k=1}^9 l_k H_k(N_{i,j}) \quad (5.4)$$

T is diffuse texture, l and H are the SH coefficients and basis [63], and N is the normal map of pixel (i,j) in the UV coordinate space.

Training and Testing Datasets

Reconstructing a high-quality 3D shape from a single image is a significant challenge. To address this limitation, we leverage the multi-view facial dataset, Facescape [77], which contains 359 subjects captured with 20 distinct facial expressions across three different views (front, left, and right). This multi-view dataset provides an opportunity for accurate estimation of facial geometry. Additionally, we also employ the generated *Lifespan Dataset* with three different ages when training this network. The combination of multi-view and age-varied data allows for improved accuracy of shape reconstruction while ensuring age consistency.

Implementation Details

When training this network, we evaluate the quality of the coarse shape reconstruc-

tions and the rendering results. The rendering quality ensures that the geometry is well-reconstructed. Additionally, we also take the accuracy of the age prediction into account. This ensures the re-ageing results align with the target age. The final loss function is

$$\mathcal{L}_{shape_c} = \lambda_v \mathcal{L}_v + \lambda_{pho} \mathcal{L}_{pho} + \lambda_{age} \mathcal{L}_{age} + \lambda_{reg} \mathcal{L}_{reg} \quad (5.5)$$

where $\mathcal{L}_v$ is the vertices loss, $\mathcal{L}_{pho}$ is the photometric loss, and $\mathcal{L}_{age}$ is the age prediction loss; each loss component is weighted by a factor λ_x , including $\lambda_v = 1.0$, $\lambda_{pho} = 0.2$, $\lambda_{age} = 0.05$, and $\lambda_{reg} = 0.1$.

- **Vertices loss** compares the predicted vertices $\tilde{V}$ from *GeoNet* to those V from FLAME computed as $\mathcal{L}_v = \|\tilde{V} - V\|_1$.
- **Photometric loss** $\mathcal{L}_{pho} = \|M \odot (\tilde{R} - R)\|_2$, where M is the face mask. R are the reference images and $\tilde{R}$ are the rendering results. These images vary from different datasets to ensure shape and age consistency.
- **Age prediction loss** $\mathcal{L}_{age} = \|\tilde{a} - a\|_2$ computes the difference between the predicted ages a and the reference ages $\tilde{a}$ predicted by the same age predictor as in the age editing part.
- **Regularization** regularizes the predicted mesh we use $\mathcal{L}_{reg} = \mathcal{L}_{normal} + \mathcal{L}_{laplacian}$, where $\mathcal{L}_{normal}$ is the mesh normal consistency, computed for each pair of neighboring faces and $\mathcal{L}_{laplacian}$ for mesh smoothing [55].

We train this network using the Adam optimizer with a learning rate of $1e - 4$ and a learning rate decay of 0.99 every epoch.

Ablation study

This ablation study evaluates the necessity of this part, particularly the age factor incorporated in this network. Figure 5.11 presents the 3D coarse shape reconstructions

generated by *GeoNet* with inputs of multiple ages, compared to reconstructions from FLAME [49]. Like FLAME, *GeoNet* generates high-quality shape reconstructions, especially for adult subjects. However, when given a younger age, the head size becomes smaller. This reflects our network is capable of reconstructing high-quality coarse shapes with age-specific differences.

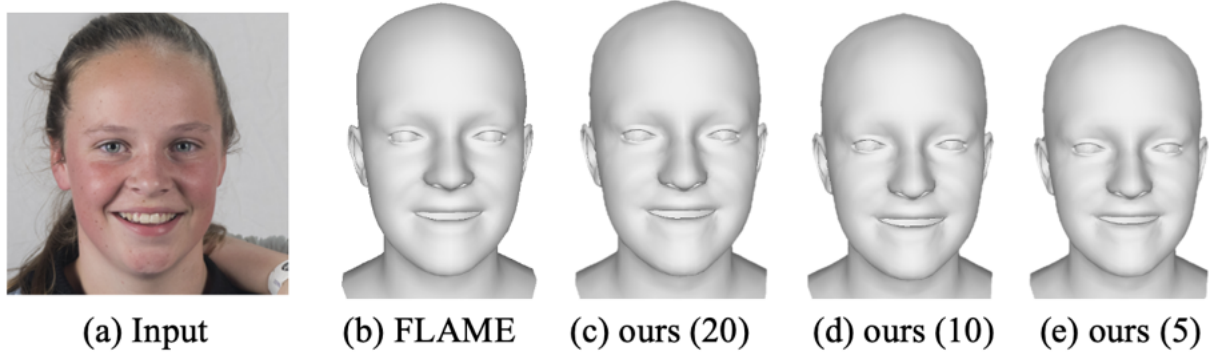


Figure 5.11: Comparison between FLAME [49] and 3D coarse shapes with different ages from *GeoNet*.

5.3.2 Detailed Shape Reconstruction

The coarse shapes generated by *GeoNet* represent high-quality reconstructions from a single image. However, they primarily encompass low- and mid-frequency information, such as pose. To enhance these reconstructions by incorporating age-appropriate high-frequency details, we propose the *RefineNet* architecture, which is designed to extract and synthesize detailed shape information.

The idea of this network *RefineNet* is inspired by some previous works that utilized low-pass filters, such as the Difference of Gaussians (DoG), to blur images and subsequently subtract these blurred versions to isolate high-frequency information [10]. While effective, they require users to manually select parameters like μ and σ depending on their experience. In contrast, more recent approaches have employed deep neural networks to extract displacement maps, such as DECA [22]. Although these networks can produce detailed results, they require large image datasets and long training and testing times.

To address these challenges and combine the benefits of both traditional filtering techniques and deep neural networks, we designed a lightweight and efficient network *RefineNet*. It consists of two convolutional layers that facilitate the extraction of high-frequency information while minimizing computational complexity. We initialize the weights of *RefineNet* using a Gaussian filter with parameters ($\mu = 0, \sigma = 3$), which allows for a systematic approach to high-frequency extraction without the need for extensive parameter tuning.

The architecture of *RefineNet* is displayed in Figure 5.12. In this figure, the input image is the re-aged result from the texture pipeline, while the corresponding displacement map is adjusted brightness to ensure clarity. By integrating this lightweight network into our pipeline, we can obtain realistic 3D facial reconstruction with age-appropriate details.

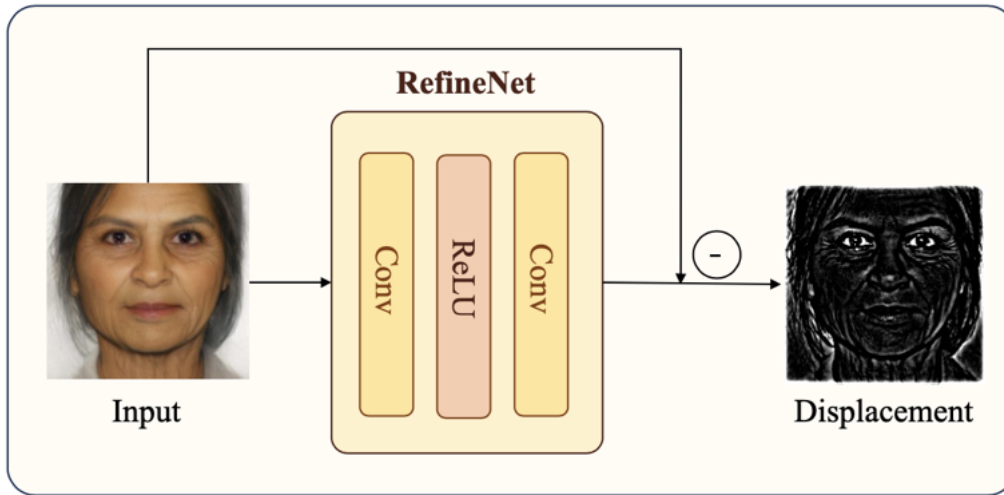


Figure 5.12: **Detailed shape reconstruction.** *RefineNet* extracts the high-frequency information from the input image as a displacement map.

After we extract the displacement maps D from re-aged results, we first convert the coarse geometry G_c to the normal map N_c , then we compute the detailed geometry G_d as

$$G_d = G_c + D \odot N_c \quad (5.6)$$

The detailed geometry G_d can be further converted to the detailed normal map N_d , thus generating the detailed shading maps by $S(T, l, N_d)$ and obtaining the detailed rendering

results from $\mathcal{R}(G_d, S(T, l, N_d), c)$.

Training and Testing Datasets

The datasets for this network are the same image datasets as other networks in our pipeline.

Implementation Details

Having generated the coarse shape using *GeoNet*, we do not consider the vertices in this training process. During the training of *RefineNet*, we focus on the quality of the rendering results in comparison to reference images. To enhance the training efficacy, we incorporate two derivative losses. Following the methodology outlined in [73], we integrate the 2D gradients of both the predictions and the reference images. This approach aims to extract high-frequency information while simultaneously suppressing noise, leading to more accurate reconstructions. The loss function for this network is computed as follows:

$$\mathcal{L}_{shape.d} = \lambda_{pho}\mathcal{L}_{pho} + \lambda_{dx}\mathcal{L}_{dx} + \lambda_{dy}\mathcal{L}_{dy} \quad (5.7)$$

where $\mathcal{L}_{pho}$ is the photometric loss, $\mathcal{L}_{dx}$ and $\mathcal{L}_{dy}$ are derivative losses on the x and y axis, respectively. $\lambda_{pho} = 1.0$, $\lambda_{dx} = 0.1$, and $\lambda_{dy} = 0.1$.

- **Photometric loss:** The equation is the same as that for the coarse shape reconstruction.
- **Derivative loss** is calculated as

$$\mathcal{L}_{dx} = \left\| \frac{\partial R}{\partial x} - \frac{\partial \tilde{R}}{\partial x} \right\|_1, \quad \mathcal{L}_{dy} = \left\| \frac{\partial R}{\partial y} - \frac{\partial \tilde{R}}{\partial y} \right\|_1 \quad (5.8)$$

We employ Adam to optimize the loss function of this network with a learning rate of $1e-5$ and a 0.99 decay rate per epoch.

5.4 Results and Evaluation

In this section, we present the results of our pipeline, displaying both 2D re-aged images and the corresponding 3D coarse and detailed shape reconstructions as detailed in Section 5.4.1. To evaluate the effectiveness of our methods, we conduct qualitative and quantitative comparisons with other state-of-the-art methods, as discussed in Sections 5.4.2 and 5.4.3, respectively. Furthermore, we implement two user studies in Section 5.4.4 to gather subjective feedback on the perceived realism and quality of the generated images and shapes. These evaluations enable a thorough understanding of our method’s capabilities and limitations within the context of facial age transformation and reconstruction.

5.4.1 Results

Figure 5.13 illustrates the 2D re-ageing results of facial images from 14 years old to 80 years old employing both *AgeEditNet* and *SkinNet*. The images in red boxes represent the original photographs of each subject. As can be seen, our method effectively produces realistic facial structural changes through *AgeEditNet*. This network can adapt to all the subjects with different skin types, and at the same time, preserve their identities throughout the aging process.

On the other hand, our approach leverages *SkinNet* to edit biophysical skin tones. The results indicate a gradual decrease in the concentrations of melanin and hemoglobin from ages 14 to 80. Involving this adjustment not only reflects the natural facial ageing process, but also enhances the overall realism of simulating age-related changes in facial appearance.

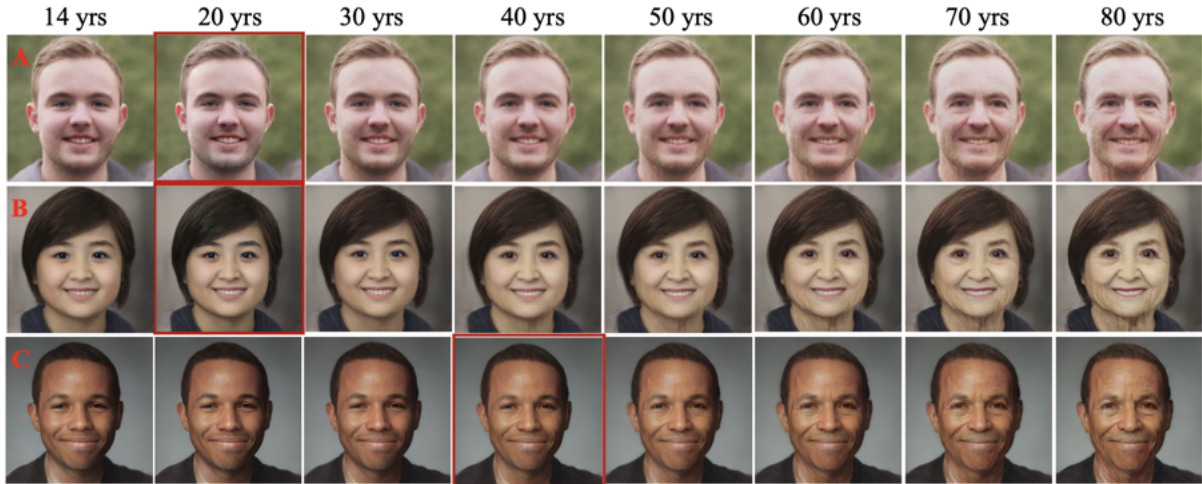


Figure 5.13: Age transformation results from 14 to 80 years old. The input images are outlined with a thin red border. Our methods generate realistic facial structural changes, such as wrinkles, with good identity preservation, as well as skin color changes due to ageing.

As shown in this figure, the first column of Figure 5.14 presents the coarse shape reconstructions of three subjects of 20 years old corresponding to the 2D re-ageing results, generated by the *GeoNet*. These results clearly show that our method can generate high-quality 3D shape reconstructions from a single image.

The second to fourth columns of Figure 5.14 display the corresponding detailed shape reconstructions for ages 40, 60, and 80, produced by *RefineNet*. These results illustrate that our method performs well on 3D shape reconstructions in both coarse and detailed stages. As the subjects age, notable aging effects, such as wrinkles and other structural changes on the facial surface, become increasingly obvious. These details enhance the visual realism of the models while also capturing the subtle features that characterize the aging process.

These examples demonstrate that our method is capable of effectively addressing diverse subjects across different ages, ethnicities, and genders. As a result, it produces realistic re-ageing outcomes while maintaining good identity preservation.

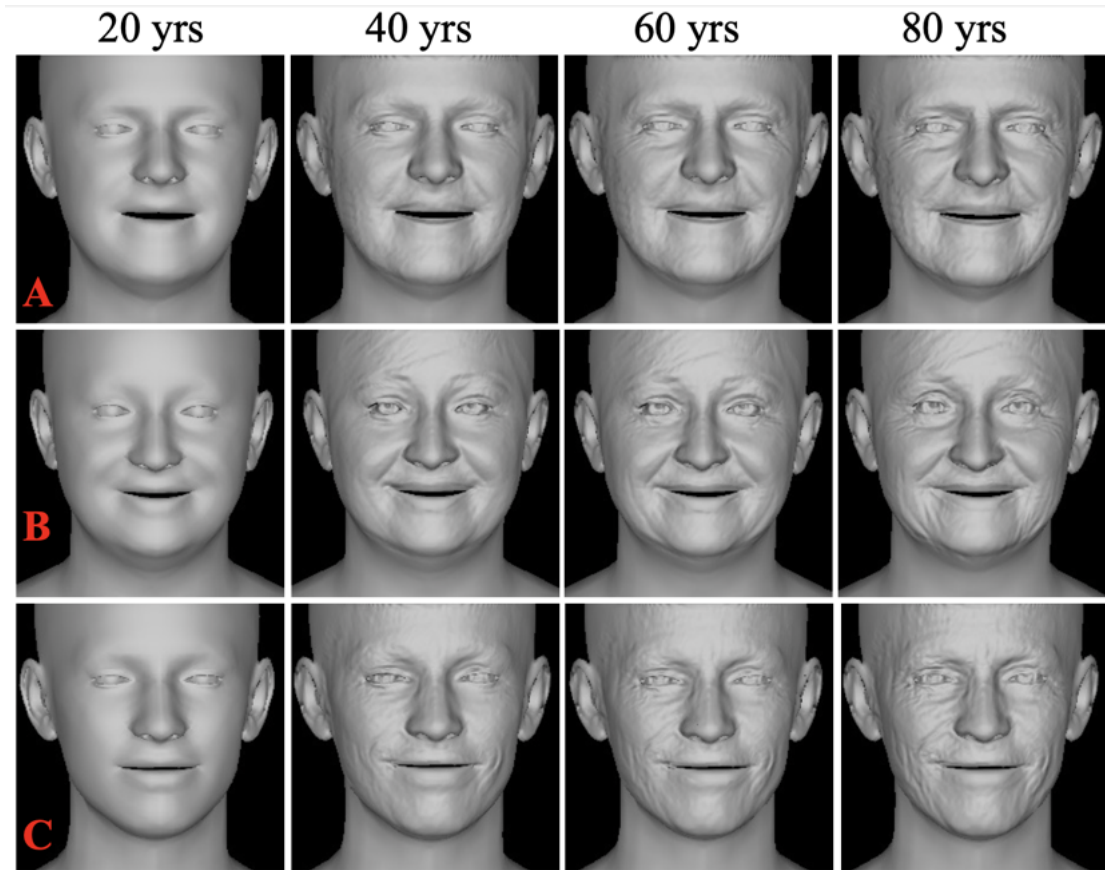


Figure 5.14: Coarse shape reconstructions with 20 years old and detailed shape reconstructions of 40, 60, and 80 years old.

As described before, the input dataset is 10 years old while the reference of the output dataset is 80 years old. This means that the network is trained from young to old incrementally. Here, we present the results of a cyclic aging process to demonstrate the robustness of our method. To be specific, we start with an input image of a 22-year-old subject, apply our method to age it to 62 years old, and subsequently reverse the process to return to the original 22-year-old appearance.

Figure 5.15 illustrates this ageing cycle, which clearly indicates that our method is both robust and consistent throughout the aging and de-aging processes while effectively preserving the identity of the subject across different age representations.

We also present an extreme example of age reversal, demonstrating the transformation

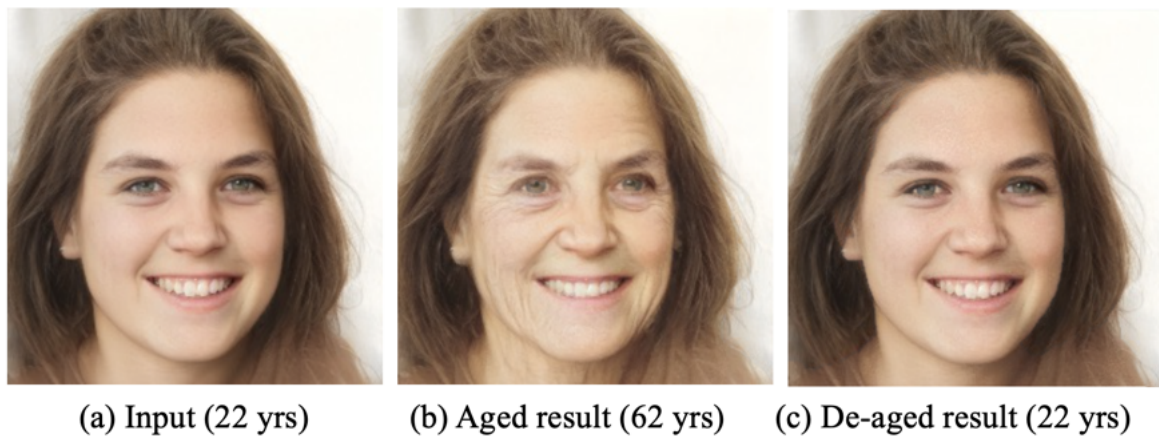


Figure 5.15: An example of a round-trip transformation, aging a face from 22 years old (a) to 62 years old (b), then de-aging back to 22 years old (c).

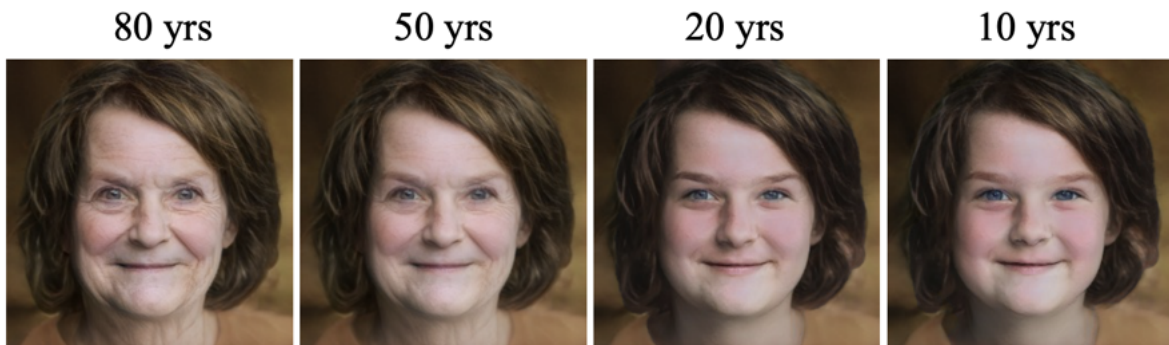


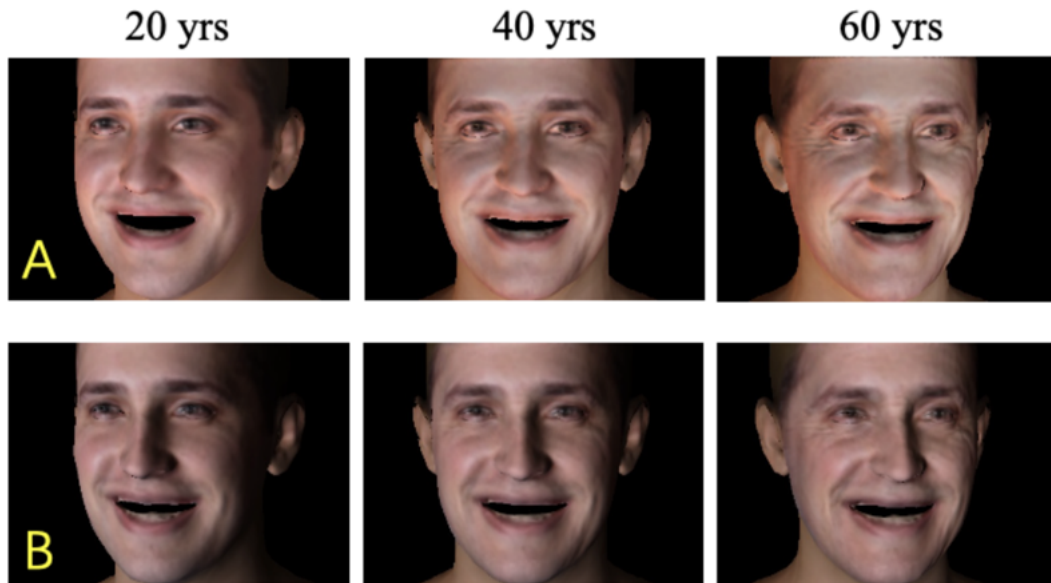
Figure 5.16: An extreme example of a subject from 80 to 10 years old.

from 80 years old to 10 years old, as illustrated in Figure 5.16.

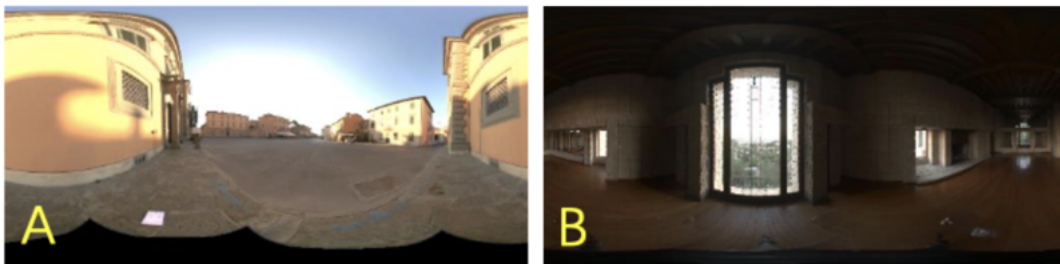
Once both the coarse and detailed shape reconstructions are obtained, along with the re-aged results, we proceed to render these outputs using various environmental maps. For this rendering process, we employ a straightforward normal-based technique based on the Phong reflection model.

Figure 5.17 illustrates the rendering results that integrate the detailed shape and re-aged texture for a subject at 20, 40, and 60 years of age, as shown in (a). These results are further enhanced by applying two different environment maps, displayed in (b). In our rendering model, the specular roughness and specular reflectance are set as empirical constants across the facial surfaces, providing a uniform appearance while maintaining the visual integrity of the aging effects. These results not only show the adaptability of

our method to different lighting conditions but also highlight the realistic rendering of age-related changes.



(a) Rendering results



(b) Environment maps

Figure 5.17: Multi-view rendering results of 20, 40, 60 years old with two environment maps.

5.4.2 Qualitative Evaluation

We are the first to propose a 3D re-aging method in computer graphics and vision. Therefore, in this part, we compare our method against both 2D re-aging approaches and 3D reconstruction approaches with a single input, separately.

In the 2D re-aging comparison, we compare our method with four recent methods: SAM [1], LATS [58], HRFAE [78], and FRAN [82]. As illustrated in Figure 5.18, our approach

achieves realistic 2D re-aging results through the integration of *AgeEditNet* and *SkinNet*, demonstrating good identity preservation across age transformations from 14 to 60 years old. In contrast, while *SAM* produces high-quality re-aging results, it struggles with maintaining identity consistency. This issue is particularly obvious in male subjects, where it becomes challenging to tell whether the re-aged face corresponds to the original input image. Furthermore, *SAM* alters the background of the input image, detracting from the focus on facial changes. Both *HRFAE* and *LATS* have good performance in preserving identity; however, these two methods produce under-aged results. Among these methods, *LATS* is the most effective for re-aging subjects to a younger age. This comparative evaluation underscores the strengths and limitations of existing methods, highlighting the main advantages of our 3D re-aging method in achieving both realism and identity preservation.

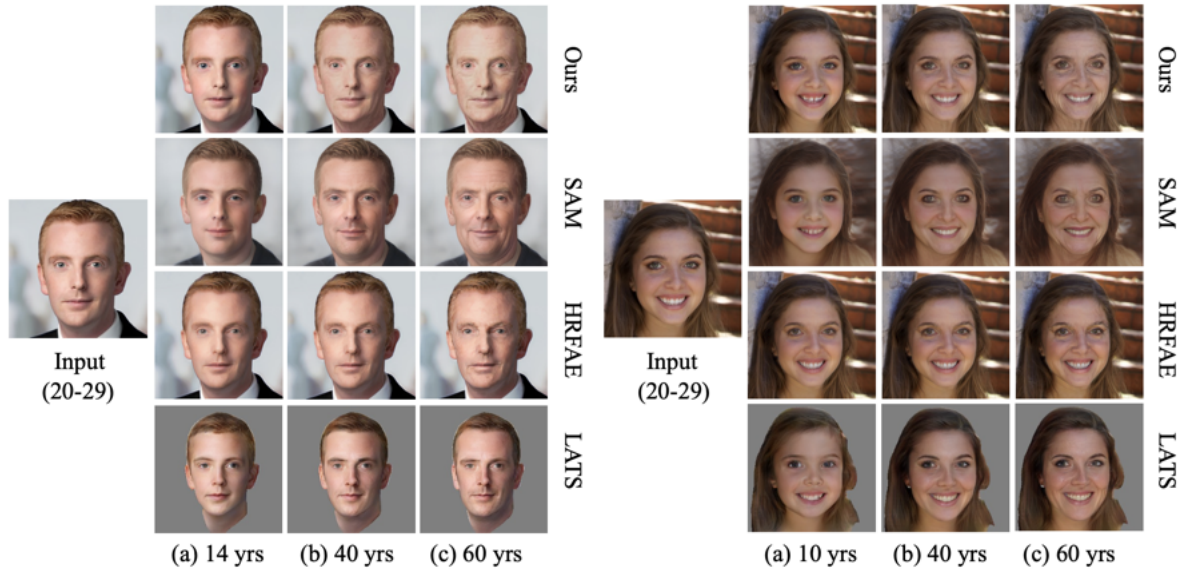


Figure 5.18: Qualitative comparison among our method and three SOTA ageing methods, namely *SAM* [1], *HRFAE* [78], and *LATS* [58]. Notably, *SAM* shows a large identity loss compared to the other three methods, particularly for the male subject. In contrast, *HRFAE* demonstrates improved identity preservation, but this method is constrained by the limited age range (20 – 69 years old). Additionally, challenges in producing significant changes for the oldest age group are observed in both *HRFAE* and *LATS* although *LATS* is the best method among these four to de-age images to the children. Aside from the re-ageing quality, *SAM* and *LATS* can NOT preserve the background in the input image, especially for the right subject with a complex background.

Except for these three methods, we also compare our approach with FRAN, the latest re-aging approach that can produce high-quality results while preserving identity [82]. Due to the unavailability of the source code for FRAN, we obtained sample images from their publication for comparative analysis, as shown in Figure 5.19.

As illustrated, both FRAN and our method effectively generate significant structural changes corresponding to ageing. However, our approach demonstrates superior performance in terms of identity preservation, particularly in some facial regions such as the mouth and eyes. Additionally, our method incorporates biophysical skin editing into our pipeline, which enables more accurate skin tone modifications. Furthermore, our method has a wider age range from 10 to 80 years old, in contrast to FRAN, which does not effectively address the children’s period. This enhances the applicability of our method across various age groups.

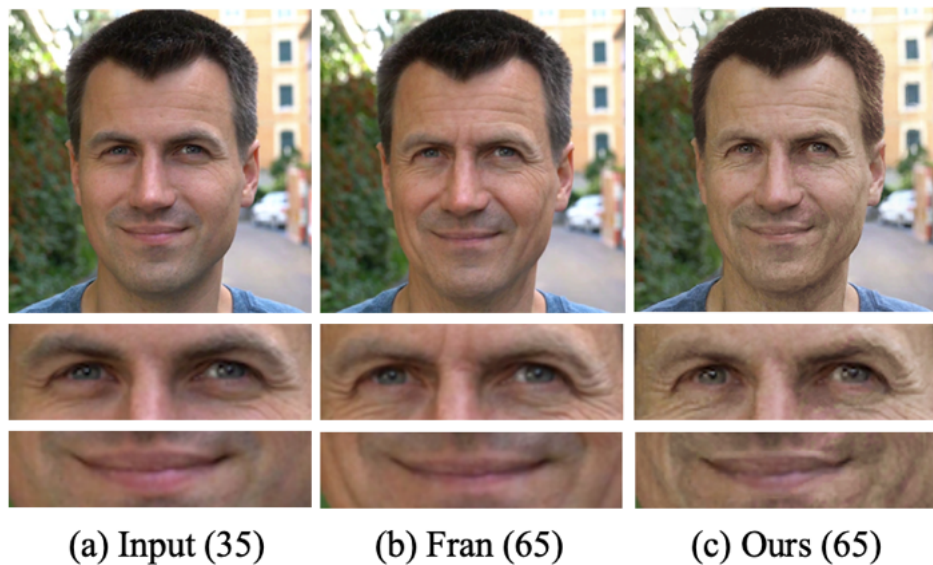


Figure 5.19: Comparison of re-aged results using FRAN [82] and our method. Input image from FRAN [82].

In the 3D reconstruction part, we compare our method with two recent approaches: DECA [22] and EMOCA [16]. While these methods primarily focus on 3D reconstruction, rather than re-aging, we utilize the re-aged results generated from our pipeline as inputs for their methods, while our method directly produces final output results.

Figure 5.20 illustrates the superior performance of our approach in capturing aging effects on geometric features in both young and old ages, such as wrinkles around the eyes and chin, which are important details of aged faces.

Overall, our qualitative evaluation demonstrates the robustness and adaptability of our re-aging framework in producing high-quality 2D and 3D results that accurately reflect age transformations.

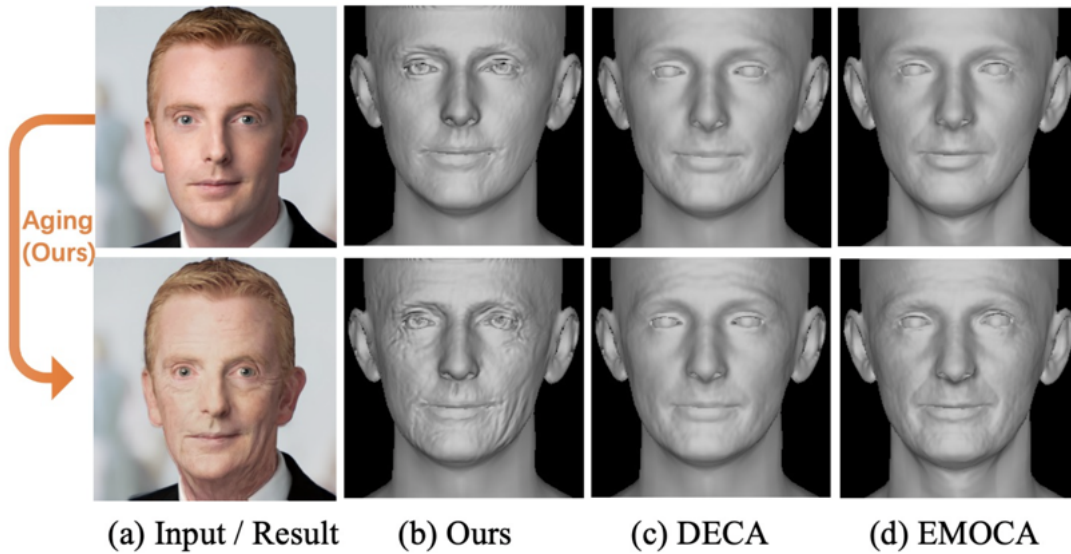


Figure 5.20: Comparison of 3D detailed reconstruction among DECA [22], EMOCA [16], and ours. DECA and EMOCA can NOT alter ages, therefore we feed them the re-aged images using our methods. The input image is taken from the CelebA dataset.

5.4.3 Quantitative Evaluation

We also conducted a quantitative evaluation to compare the performance of our method with three recent re-aging approaches, including SAM [1], LATS [58], and HRFAE [78]. As previously discussed, one of the primary challenges in the re-aging task is the lack of image datasets with paired age labels. To address this, we utilized two key criteria for evaluation: (1) age prediction accuracy and (2) identity preservation. To achieve this, we employed the publicly available face recognition API, InsightFace [18], which estimates the age of an image and generates a corresponding embedding based on [18]. The cosine

similarity between embeddings of images with different ages is then calculated to assess identity preservation.

For the evaluation, we randomly selected 100 images from the CelebA-HQ dataset [40]. This dataset contains "young" and "old" age labels rather than accurate ages. Therefore, we assume the input images to be approximately 25 years old, with a target re-aging age of 60 years old. The results are summarized in Table 5.2, which presents a comparison between the four methods.

In terms of average age prediction, our method and SAM demonstrate are closer to the target age compared to LATS and HRFAE. This result also aligns with the conclusions from our qualitative evaluation, where LATS and HRFAE tend to generate younger-than-expected results. This makes these two methods perform better in identity preservation. However, when both criteria (age prediction accuracy and identity preservation) are taken into account, our method outperforms the other three approaches, showing a better balance between re-aging effects and identity preservation.

Table 5.2: Quantitative evaluation

	avg. age pred.	identity preservation
SAM [1]	56.93	0.6173
LATS [58]	43.5681	0.9230
HRFAE [78]	49.0181	0.9607
Ours	59.17	0.8518

5.4.4 User Study

We implemented two user studies to supplement our qualitative and quantitative evaluations, comparing our method with two recent re-aging approaches: SAM [1] and FRAN [82]. Since the code for FRAN is not publicly available, we employ test images borrowed from their paper. We select 12 pairs of images with both genders and ages ranging from 20 to 80 years old. Each pair appears four times in random order on the screen, resulting in

576 rankings collected for each study.

To ensure consistency and eliminate potential environmental effects, we invited 10 participants for each user study and use the Psychophysics Toolbox extensions (PsychToolbox v3.0.19) [8, 61, 44] to implement these two studies in Matlab. All participants are asked to complete the studies in a dark room using the same screen to minimize visual differences. During the studies, the re-aged images generated by the three methods gradually disappeared once participants made their selections. We aggregate the responses and employ statistical analyses using Fisher’s exact test, to determine the significance of differences in user preferences across methods.

The first user study focuses on evaluating re-aging quality. Participants were asked to rank the images based on which method produced the most realistic aging results relative to the target age. The frequency of each method being chosen as the top option is displayed in Figure 5.21(a). As shown, our method (44.1%) performs slightly better than FRAN (36.3%), while SAM (19.6%) ranked the lowest. Statistical analysis shows significant differences, with p-values of $p = 0.0082$ (FRAN vs. ours), $p < 1 \times 10^{-9}$ (FRAN vs. SAM), and $p < 1 \times 10^{-18}$ (ours vs. SAM).

The second user study evaluates identity preservation, asking participants to rank the methods with the most similar faces to the input faces. Figure 5.21(b) shows that both our method (54.5%) and FRAN (42.7%) significantly outperform SAM (2.8%), a conclusion consistent with our previous evaluations. The statistical significance of the results is also strong, with p-values of $p < 1 \times 10^{-4}$ (FRAN vs. ours), $p < 1 \times 10^{-66}$ (FRAN vs. SAM), and $p < 1 \times 10^{-96}$ (ours vs. SAM).

In summary, these user studies demonstrate that our method achieves superior performance in both re-aging quality and identity preservation when compared to FRAN and SAM. Our approach not only generates more realistic aging effects but also preserves the essential facial identity of the subjects more effectively. These results further solidify our

method’s robustness and applicability in 3D re-aging tasks.

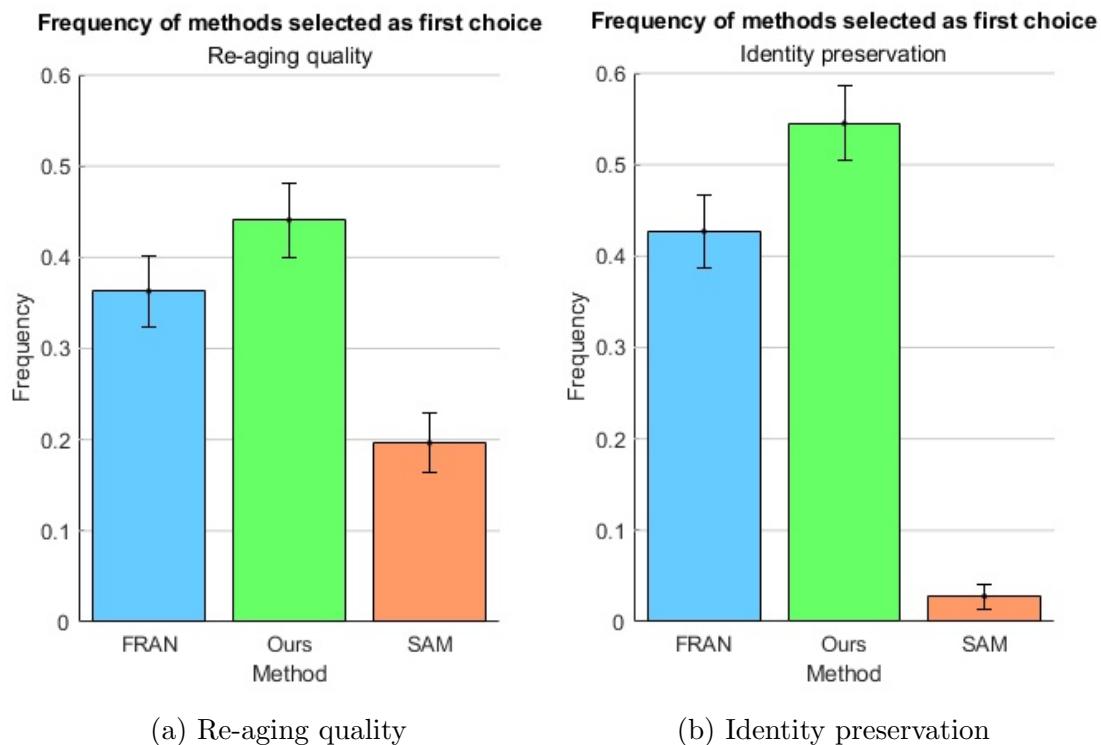


Figure 5.21: Bar plot showing the frequency with which each method was chosen as the first choice, along with 95% confidence intervals.

5.5 Conclusion

To summarize, we proposed a novel 3D facial re-aging framework that simultaneously edits both the appearance and geometry of facial images to simulate aging across a wide age span, from 10 to 80 years. Unlike existing 2D-based methods that focus solely on texture transformation, our method captures biophysical and structural changes while preserving individual identity. The framework integrates three key networks: SkinNet, which adjusts melanin and hemoglobin levels to reflect age-related skin tone variations; AgeEditNet, a diffusion-based network leveraging α -(de)blending to simulate smooth transitions in facial texture and structural changes such as wrinkles; and GeoNet, which adapts the coarse 3D geometry based on the target age. High-frequency geometric details, including fine wrinkles and displacement, are further refined using RefineNet.

This pipeline allows for realistic, controllable age transformation from a single in-the-wild image, shown in Section 5.4, however, it still has some limitations. First, the age range in our current framework is constrained to ages between 10 and 80 years, employing a diffusion-based approach. This excludes the infant and early childhood stages, limiting the method’s applicability in re-aging tasks involving very young subjects. Extending the model to cover these early developmental stages would be an important area for future work.

Second, while our method performs well in identity preservation when compared to other state-of-the-art techniques, it lacks the ability to simulate certain aging characteristics. For instance, it does not account for features such as gray hair and skin sagging, which are critical effects of ageing. Integrating more advanced generative models to capture these features would enhance the realism of our aging simulations.

Another limitation is the presence of a halo effect in some final outputs, particularly around facial boundaries. This artifact is primarily due to the use of a binary facial mask during processing. A potential solution to this issue is to replace the binary mask with alpha matting, which would enable smoother transitions and more natural blending between the face and the surrounding background. Additionally, incorporating an eye mask could help maintain the accuracy of eye color throughout the ageing process, which is sometimes lost in the current implementation.

Moreover, the resolution of our training dataset is currently limited to 256×256 pixels, which restricts the level of detail captured in the generated images. Training on higher-resolution datasets would allow for more detailed and refined 3D reconstructions, particularly in texture quality, wrinkles, and skin features.

Finally, our method does not separate environment maps from the input images, which affects the relightability of the generated 3D faces. In scenarios where accurate relighting is required, this limitation can be problematic. Future work could focus on disentangling

environment lighting from the input, enabling more flexible and realistic relighting of age-transformed 3D faces under varying conditions.

In summary, while our method can generate high-quality 3D re-aging results, addressing these limitations would further improve its performance and widen its applicability in computer vision and graphics.

Chapter 6

Conclusion

This thesis focuses on accurate modeling and editing of facial appearances. It begins with a hyperspectral skin model with a practical measurement method, then explores the approaches for parameter estimation and skin reflectance reconstruction, and further introduces its application in skin appearance editing. Finally, the work extends to a comprehensive framework for realistic facial re-ageing in both appearance and shape.

In detail, we introduced a diffusion-based hyperspectral skin model by incorporating five chromophore parameters which represent the skin reflectance in both invisible and visible ranges from 300nm to 1000nm (Chapter 3). We also presented a practical measurement setup consisting of four LCD monitors surrounding the subject to illuminate the faces with different lighting conditions, and a camera located in the middle to capture photographs. Given these photographs, we proposed a three-stage LUT search for parameter estimation by employing white, blue, and green illuminations to excite different chromophores in each stage. Compared to the conventional LUT search, this approach can predict comparable results while accelerating the search time. However, all the LUT search methods have a common limitation that the sampling density of chromophores results in discrete values, thus affecting the quality of the results.

To overcome this limitation, we introduced a neural network consisting of an Encoder and a Decoder. The Encoder aims to estimate five spectral parameters in our skin model, which are further passed to the Decoder to reconstruct full hyperspectral reflectance using PCA-basis (Chapter 4). The reconstructed reflectance can be visualized in the visible domain with different uniform illuminations and also in the invisible domains, including the UVA and NIR spectrums. We further simplified this complete network with a single input image under D65 illumination. As these images are in the standard color space, this network can adapt to any image after color correction. Our skin model enables having a wide range of applications, such as skin appearance editing. We simulated various skin statuses, including tanned, lightened, flushed, and paled, by modifying different chromophore maps.

Age is an important factor that affects these skin properties. For example, melanin and hemoglobin concentrations decrease with age, leading to lighter and paler skin tones. In Chapter 5, we proposed a framework for facial age transformation in both appearance and shape. Existing re-ageing methods focus on the facial structural changes of two-dimensional images while overlooking the skin color changes. In contrast, our proposed method considered both the skin properties editing and facial structure changes to predict more realistic results. Additionally, we also extended this method to three dimensions. Benefiting from some current 3D shape reconstruction approaches, we integrated a geometric neural network to generate high-quality coarse shapes with head size control. To add age-appropriate details, we involved a lightweight refinement network that extracts the ageing details to help with detailed shape reconstruction. We evaluated our method quantitatively and qualitatively with several recent 2D re-ageing methods and 3D shape reconstruction approaches, validating that our method outperforms the current re-ageing techniques.

Although these methods can generate high-quality results, they still have several limitations. For the skin modeling part, the LUTs are sensitive to the lighting conditions

and capturing setup. Therefore, it is challenging to employ this method to estimate the parameters and reconstruct the full skin reflectance in the in-the-wild images. To overcome this issue, a possible research direction is to eliminate the lighting conditions of the input images. The future work could explore the joint estimation of environment maps and accurate skin reflectance. This would enable generalizing in-the-wild images captured under arbitrary lighting conditions, significantly broadening its applicability in real-world scenarios.

Building on the biophysical skin model, another future research can extend into the field of computational cosmetics, enabling biophysical modeling and realistic simulation of the effects of skincare products and makeup on skin appearance. This could involve adding an additional optical layer to simulate how cosmetic products modify light interaction with the skin surface and subsurface layers. Due to the diverse cosmetic materials and their complex optical properties, this direction would be challenging and requires a large number of experiments. However, it has strong potential for applications such as virtual makeup try-on and personalized skincare recommendation systems, particularly in the context of e-commerce and virtual shopping platforms.

Additionally, our work does not address skin appearance rendering. While rendering is a crucial area in computer graphics, Therefore, another interesting future work is fast photorealistic rendering with spectral properties.

For the re-ageing part, our method mainly focuses on identity preservation while sacrificing some ageing effects, such as skin sagging and gray hair. To overcome this issue, future work can involve more realistic ageing characteristics by incorporating generative Decoders, such as StyleGAN. Moreover, building upon a large 3D face scan dataset with various ages would help with producing more realistic re-ageing results in 3D shapes, thus enabling the production of a re-ageing dataset with paired ages spanning a long time in two and three dimensions.

There are several possible research directions within the topics discussed in this thesis. We hope that our work can contribute to the field of facial appearance modeling and editing, and we believe there will be more promising findings for substantial contributions to the area in the future.

Bibliography

- [1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Trans. Graph.*, 40(4), jul 2021.
- [2] Carlos Aliaga, Mengqi Xia, Hao Xie, Adrian Jarabo, Gustav Braun, and Christophe Hery. A Hyperspectral Space of Skin Tones for Inverse Rendering of Biophysical Skin Properties. *Computer Graphics Forum*, 2023.
- [3] Sarah Alotaibi and William A. P. Smith. A biophysical 3d morphable model of face appearance. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 824–832, 2017.
- [4] Michael Attas, Trevor Posthumus, Bernie Schattka, Michael Sowa, Henry Mantsch, and Shuliang Zhang. Long-wavelength near-infrared spectroscopic imaging for in-vivo skin hydration measurements. *Vibrational Spectroscopy*, 28(1):37–43, 2002. A Collection of Papers Presented at SHEDDING NEW LIGHT ON DISEASE: Optical Diagnostics for the New Millennium, Winnipeg, Canada, June 25-30, 2000.
- [5] A N Bashkatov, E A Genina, V I Kochubey, and V V Tuchin. Optical properties of human skin, subcutaneous and mucous tissues in the wavelength range from 400 to 2000nm. *Journal of Physics D: Applied Physics*, 38(15):2543, jul 2005.
- [6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive*

- Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.
- [7] Laurence Boissieux, Gergo Kiss, Nadia Thalmann, and Prem Kalra. Simulation of skin aging and wrinkles with cosmetics insight. *Computer Animation and Simulation*, 01 2000.
- [8] David H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10(4):433 – 436, 1997.
- [9] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.*, 34(4), jul 2015.
- [10] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.*, 34(4), jul 2015.
- [11] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.
- [12] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9429–9439, 2019.
- [13] Tenn F. Chen, Gladimir V. G. Baranoski, Bradley W. Kimmel, and Erik Miranda. Hyperspectral modeling of skin appearance. *ACM Trans. Graph.*, 34(3), may 2015.
- [14] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units(elus), 2016.
- [15] Symon Cotton, Ela Claridge, and Per Hall. A skin imaging method based on a colour formation model and its application to the diagnosis of pigmented skin lesions. *Proceedings of Medical Image Understanding and Analysis*, 99:49–52, 01 1999.

- [16] Radek Danecek, Michael J. Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20279–20290, 2022.
- [17] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. page 3844–3852, 2016.
- [18] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.
- [19] Craig Donner and Henrik Wann Jensen. Light diffusion in multi-layered translucent materials. *ACM Trans. Graph.*, 24(3):1032–1039, jul 2005.
- [20] Craig Donner and Henrik Wann Jensen. A spectral bssrdf for shading human skin. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques*, EGSR ’06, page 409–417, Goslar, DEU, 2006. Eurographics Association.
- [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 Trans. Graph.*, 27(5), dec 2008.
- [22] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4), jul 2021.
- [23] Y. Gitlina, G. C. Guarnera, D. S. Dhillon, J. Hansen, A. Lattas, D. Pai, and A. Ghosh. Practical Measurement and Reconstruction of Spectral Skin Reflectance. *Computer Graphics Forum*, 39(4):75–89, 2020.
- [24] Guillermo Gomez-Trenado, Stéphane Lathuilière, Pablo Mesejo, and Óscar Cerdón. Custom structure preservation in face aging. 2022.
- [25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020.

- [26] Paulo Gotardo, J  r  my Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. 2018.
- [27] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1294–1307, 2017.
- [28] Chet S. Haase and Gary W. Meyer. Modeling pigmented materials for realistic image synthesis. *ACM Trans. Graph.*, 11(4):305–335, oct 1992.
- [29] Sen He, Wentong Liao, Michael Ying Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Disentangled lifespan face synthesis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3857–3866, 2021.
- [30] Zhenliang He, Meina Kan, S. Shan, and Xilin Chen. S2gan: Share aging factors across ages and share aging trends among individuals. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9439–9448, 2019.
- [31] Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α -(de)blending: a minimalist deterministic diffusion model. 2023.
- [32] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [34] Shuokang Huang, Kaihan Li, Di You, Yichong Chen, Arvin Lin, Siying Liu, Xiaohui Li, and Julie A. McCann. Wimans: A benchmark dataset for wifi-based multi-user activity sensing. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLII*, page 72–91, Berlin, Heidelberg, 2024. Springer-Verlag.

- [35] Jose A. Iglesias-Guitian, Carlos Aliaga, Adrian Jarabo, and Diego Gutierrez. A biophysically-based model of the optical properties of skin aging. *Comput. Graph. Forum*, 34(2):45–55, may 2015.
- [36] Jose A. Iglesias-Guitian, Carlos Aliaga, Adrian Jarabo, and Diego Gutierrez. A biophysically-based model of the optical properties of skin aging. *Computer Graphics Forum (EUROGRAPHICS 2015)*, 34(2), 2015.
- [37] Akira Ishimaru. Wave propagation and scattering in random media. Oxford University Press, 1978.
- [38] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan. A practical model for subsurface light transport. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, pages 511–518, 2001.
- [39] Jorge Jimenez, Timothy Scully, Nuno Barbosa, Craig Donner, Xenxo Alvarez, Teresa Vieira, Paul Matts, Verónica Orvalho, Diego Gutierrez, and Tim Weyrich. A practical appearance model for dynamic facial color. *ACM Transactions on Graphics*, 29(6):1–10, 2010.
- [40] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.
- [41] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, dec 2021.

- [43] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M. Seitz. Illumination-aware age progression. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3334–3341, 2014.
- [44] M Kleiner, D Brainard, Denis Pelli, A Ingling, R Murray, and C Broussard. What’s new in psychtoolbox-3. *Perception*, 36(14):1–16, 2007.
- [45] Aravind Krishnaswamy and Gladimir V.G. Baranoski. A biophysically-based spectral model of light interaction with human skin. *Computer Graphics Forum*, 23(3):331–340, 2004.
- [46] A. Lanitis, C.J. Taylor, and T.F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.
- [47] Alexandros Lattas, Yiming Lin, Jayanth Kannan, Ekin Ozturk, Luca Filipi, Giuseppe Claudio Guarnera, Gaurav Chawla, and Abhijeet Ghosh. Practical and scalable desktop-based high-quality facial capture. pages 522–537. Springer, 2022.
- [48] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction ”in-the-wild”. *CoRR*, abs/2003.13845, 2020.
- [49] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), nov 2017.
- [50] Xiaohui Li, Giuseppe Claudio Guarnera, Arvin Lin, and Abhijeet Ghosh. Practical measurement and neural encoding of hyperspectral skin reflectance. In *2024 International Conference on 3D Vision (3DV)*, pages 1301–1309, 2024.
- [51] Xiaohui Li, Giuseppe Claudio Guarnera, Arvin Lin, and Abhijeet Ghosh. Realistic Facial Age Transformation with 3D Uplifting. *Computer Graphics Forum*, 2024.

- [52] Arvin Lin, Yiming Lin, Xiaohui Li, and Abhijeet Ghosh. High Quality Neural Relighting using Practical Zonal Illumination. In *Eurographics Symposium on Rendering*. The Eurographics Association, 2024.
- [53] Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11869–11878, 2018.
- [54] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [55] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, GRAPHITE '06, page 381–389, New York, NY, USA, 2006. Association for Computing Machinery.
- [56] Hubert Nguyen. *Gpu Gems 3*. Addison-Wesley Professional, first edition, 2007.
- [57] Emilie Nogu  , Arvin Lin, Xiaohui Li, Giuseppe Guarnera, and Abhijeet Ghosh. Practical rgb measurement of fluorescence and blood distributions in skin. *Color and Imaging Conference*, 32:133–141, 10 2024.
- [58] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 739–755, Berlin, Heidelberg, 2020. Springer-Verlag.
- [59] Jong-Il Park, Moon-Hyun Lee, Michael D. Grossberg, and Shree K. Nayar. Multi-spectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.
- [60] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009*

- Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.
- [61] Denis G. Pelli. The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10(4):437 – 442, 1997.
- [62] S.J. Preece and E. Claridge. Spectral filter optimization for the recovery of parameters which describe human skin. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):913–922, 2004.
- [63] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery.
- [64] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018.
- [65] Erik Reinhard, Erum Arif Khan, Ahmet Oguz Akyz, and Garrett M. Johnson. *Color Imaging: Fundamentals and Applications*. A. K. Peters, Ltd., USA, 2008.
- [66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

- [68] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 252–257, 2015.
- [69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [70] Jin-Li Suo, Xilin Chen, S. Shan, Wen Gao, and Qionghai Dai. A concatenational graph evolution aging model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:2083–2096, 2012.
- [71] Yusuke Tazoe, Hiroaki Gohara, Akinobu Maejima, and Shigeo Morishima. Facial aging simulator considering geometry and patch-tiled texture. In *ACM SIGGRAPH 2012 Posters*, SIGGRAPH ’12, New York, NY, USA, 2012. Association for Computing Machinery.
- [72] B. Tiddeman, M. Burt, and D. Perrett. Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, 21(5):42–50, 2001.
- [73] A. Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2017.
- [74] Norimichi Tsumura, Nobutoshi Ojima, Kayoko Sato, Mitsuhiro Shiraishi, Hideto Shimizu, Hirohide Nabeshima, Syuuichi Akazaki, Kimihiko Hori, and Yoichi Miyake. Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH ’03, page 770–779, New York, NY, USA, 2003. Association for Computing Machinery.
- [75] Yuheng Wang, Juan Ye, Xiaohui Li, and David L Borchers. Towards automated animal density estimation with acoustic spatial capture-recapture. *Biometrics*, 80(3):ujae081, 08 2024.

- [76] Kaida Xiao, Julian Yates, Faraedon Zardawi, Suchitra Sueeprasan, N. Liao, Liz Gill, Chelsea Li, and Sophie Wuerger. Characterising the variations in ethnic skin colours: A new calibrated data base for human skin. *Skin Research and Technology*, 23, 06 2016.
- [77] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. pages 598–607, jun 2020.
- [78] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8624–8631, 2020.
- [79] David Zargaran, Florence Zoller, Alexander Zargaran, Tim Weyrich, and Afshin Mosahebi. Facial skin aging: key concepts and overview of processes. *International Journal of Cosmetic Science*, 2022.
- [80] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [81] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. *European Conference on Computer Vision*, 2022.
- [82] Gaspard Zoss, Prashanth Chandran, Eftychios Sifakis, Markus Gross, Paulo Gotardo, and Derek Bradley. Production-ready face re-aging for visual effects. *ACM Trans. Graph.*, 41(6), nov 2022.