Department of Computer Science

Submitted in part fulfilment for the degree of BEng.

Beyond Racial Bias: Advancing Fairness in In-the-Wild Facial Reconstruction

Harsh Mohan

May 2024

Supervisor: Dr. Patrik Huber

In dedication to all who have faced injustice from biased algorithms, this work aims to make a difference.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Patrik Huber, for his invaluable guidance, support, and patience throughout this research project. His expertise and insights have significantly contributed to the development of this dissertation. I extend my sincere thanks to Dr. Will Smith, James Gardner, and Stephen Da'Prato-Shepar for their invaluable guidance and the many insightful meetings that helped shape this work. Additionally, I am grateful to the faculty and staff of the University of York's Department of Computer Science for providing a conducive learning environment and the resources necessary for my research. I am also grateful to my colleagues and friends for their constant support and the stimulating discussions that enriched my research experience. Their camaraderie and encouragement were vital in overcoming the challenges encountered during this study.

Lastly, I would like to thank my family, especially my sister for their unwavering support and belief in me. Their love and encouragement have been my pillars of strength throughout this journey. This work would not have been possible without the support and contributions of all these individuals and organizations. Thank you all for being an integral part of this achievement.

Contents

List of Figures

List of Tables

Executive Summary

This project addresses the critical challenge of racial bias in 3D face reconstruction, which is an issue of both technical and ethical significance. Our work aims to develop a more equitable and accurate method for 3D facial reconstruction by integrating a learned illumination prior, which is trained independently from any face model. This approach seeks to mitigate the biases inherent in traditional 3D Morphable Models (3DMMs) that often fail to represent diverse facial features accurately under various lighting conditions.

Project Aims and Key Objectives The primary aim of this project is to advance the state of 3D face reconstruction by reducing racial bias and improving accuracy across different skin tones. Key objectives include:

- Developing a novel approach to facial reconstruction that incorporates a learned illumination prior independent of face models.
- Enhancing albedo estimation to accurately separate intrinsic skin color from the effects of lighting, leading to more precise and unbiased reconstructions.
- Evaluating the proposed method against existing benchmarks to validate its effectiveness and fairness.

Motivation The motivation for this project stems from the increasing integration of 3D face reconstruction technologies in everyday applications, including facial recognition, virtual reality, and digital media. These technologies are becoming ubiquitous, and it is critical to ensure they are inclusive and do not perpetuate systemic biases. Achieving fair and accurate 3D reconstructions of all individuals, regardless of skin color, is not only a technical challenge but a moral imperative. Addressing this issue is essential to promote social justice, equity, and trust in technological advancements.

Approach and Methodology The approach taken in this project involves the integration of a learned illumination model trained on a diverse dataset that captures natural lighting variations. This model is decoupled from the facial model to mitigate bias. The methodology includes several key phases:

- Data Preprocessing: Normalization, landmark detection, and segmentation of high-resolution images.

- Model Development: Integration of the FLAME model for dynamic facial structuring, BalancedAlb model for diverse albedo texturing and the RENI++ model for handling complex lighting interactions.
- Rendering and Optimization: Initial rendering using Lambertian reflectance and iterative refinement of model parameters through "Analysis by Synthesis" to closely match synthetic outputs with real-world inputs.

Results Achieved The results of this project demonstrate significant improvements in the accuracy and fairness of 3D facial reconstructions. The learned illumination model successfully mitigates racial bias, providing more consistent and equitable representations across different skin tones. The enhanced albedo estimation effectively distinguishes between skin color and lighting effects, leading to more accurate reconstructions.

Evaluation Against Success Criteria The success of this project was evaluated against the following criteria:

- The bias score, measured as the standard deviation of Individual typology angle (ITA) errors across different skin groups, was significantly reduced, indicating improved fairness and accuracy in facial reconstructions.
- The accuracy of albedo estimation and geometric precision was assessed through qualitative and quantitative comparisons with state-of-the-art methods. Our model consistently outperformed others in these evaluations.
- The model's applicability to various lighting conditions and its flexibility in handling different facial expressions and poses were demonstrated through extensive testing and real-world applications.

Implications for Future Work The implications of our results extend beyond technical improvements, fostering a more inclusive approach to 3D facial reconstruction technology. Future work will focus on:

- Exploring additional methods to further disentangle facial features from lighting effects and expanding the diversity of validation datasets.
- Enhancing the model's adaptability to indoor and artificial lighting conditions, which remain challenging.
- Continuing to address ethical considerations in the development and deployment of 3D reconstruction technologies to ensure equitable representation and application across all demographics.

In conclusion, this project represents a significant step toward creating fairer and more accurate 3D facial reconstruction technologies. By addressing the limitations of existing methods and proposing innovative solutions, we have laid the groundwork for future advancements in this critical area of computer vision and graphics.

1 Introduction

1.1 Overview of 3D Face Reconstruction Challenges

Three-dimensional (3D) face reconstruction is a pivotal technology in the fields of computer vision and graphics, with extensive applications including biometric authentication, animation, virtual reality, and medical procedures. This technology aims to capture the complex geometry of human faces using a combination of imaging techniques and computational algorithms. The importance of accurately reconstructing 3D facial models lies in their ability to provide a realistic and relatable interface for human-computer interaction, which is crucial for applications such as augmented reality (AR), facial recognition systems, and personalized avatars in digital media [1].

Historically, the challenge of 3D face reconstruction has involved capturing the intricate details of the human face, such as skin texture and subtle expressions, under varying lighting conditions. Early techniques relied heavily on multiple images or structured light to create a depth map of the face [2]. While early methods to reconstruct faces from a single image do exist [3], they faced significant difficulties, particularly in separating albedo from illumination and accurately modeling shape. However, advancements in machine learning and 3D imaging have led to the development of more sophisticated methods that can generate high-fidelity 3D models from a single image or in real-time [4], [5].

Recent innovations in deep learning, particularly convolutional neural networks (CNNs), have significantly improved the accuracy and efficiency of 3D face reconstruction. These models are trained on vast datasets of faces to learn depth perception and facial geometry, which enables them to predict 3D shapes from new images with high precision [6], [7]. Furthermore, the integration of generative adversarial networks (GANs) has enhanced the ability to reconstruct faces with realistic textures and fine details, thereby overcoming some of the limitations of earlier reconstruction techniques [8].

Despite these advances, 3D face reconstruction continues to face challenges, particularly in handling diverse facial expressions, orientations, and occlusions. Ongoing research is addressing these issues through more robust algorithms and hybrid approaches that combine traditional geometric modeling with machine learning [9], [10].

1.2 Existing Approaches and Their Limitations

The landscape of 3D face reconstruction is rich with diverse methodologies, each aiming to bridge the gap between two-dimensional imagery and three-dimensional facial models. These approaches range from classical optimization techniques, which meticulously adjust model parameters to match input images, to modern deep learning strategies that leverage vast datasets to learn complex mappings from 2D to 3D. Central to many of these methodologies is the reliance on 3D Morphable Models (3DMMs), a paradigm that encapsulates a wealth of facial information within a manageable set of parameters [3].

Despite the sophistication and advancements these approaches offer, they are intrinsically bound by the limitations of their foundational models and data. A predominant issue arises from the bias that stems from a historical predominance of lighter-skinned subjects in the datasets used to train and validate 3DMMs. As a result, the algorithms learn to associate certain facial features and skin tones with specific 3D structures, skewing their performance when faced with underrepresented groups [11]. Such disparities in algorithmic performance underscore the need for a more inclusive approach to data collection and model training, ensuring that the diversity of the human population is accurately captured and represented.

Furthermore, the challenge of intrinsic image decomposition, a process crucial for separating the innate attributes of a face—such as shape and texture—from external influences like lighting and camera angles [12]—remains a significant hurdle. While multiple images or controlled lighting conditions can aid in disambiguation, these requirements severely limit the applicability of 3D face reconstruction in real-world, "in-the-wild" scenarios, where images taken in uncontrolled settings present diverse lighting conditions and poses, complicating the distinction between intrinsic and extrinsic factors [13]. The reliance on specific, often biased, facial and illumination priors exacerbates the difficulty of achieving accurate reconstructions across diverse populations and conditions [6].

Another critical limitation lies in the static nature of many 3DMMs, which struggle to capture the dynamic range of human facial expressions. This shortcoming not only affects the fidelity of the reconstructed models but also their utility in applications requiring expressive, lifelike avatars or models [14]. Additionally, the computational complexity of some approaches, particularly those employing deep learning, poses barriers to real-time processing and application on resource-constrained devices [13].

1.3 The Problem of Racial Bias

In the domain of 3D face reconstruction, the spectre of racial bias casts a long shadow, compromising the integrity and inclusivity of technological advancements. This bias is not merely a technical oversight but a profound

issue that reflects the underlying data sets and methodologies employed in constructing 3DMMs.

Several studies have demonstrated the demographic biases (sex, age, race) prevalent in biometrics, influencing face verification, age estimation, race classification, and emotion classification [15]. Dedicated benchmark datasets and studies have exposed bias in commercial systems and proposed algorithmic solutions to mitigate these issues [16].

The implications of such bias extend far beyond the realm of computational inaccuracies, venturing into ethical and social territories. In applications where 3D face reconstruction plays a pivotal role, from security and law enforcement to digital entertainment and social media, the repercussions of racial bias are tangible. For individuals with darker skin tones, the misrepresentation and misidentification risk is not just a matter of technological failure but a potent source of systemic discrimination, raising significant concerns about privacy, fairness, and social justice [17]. As these technologies become increasingly embedded in our daily lives, the urgency to address and mitigate racial bias becomes paramount.

It is not merely a technical hurdle but a moral imperative to ensure that the benefits of technological advancements are equitably distributed across all segments of society [18]. The ethical-by-design approach suggested by Brey and Dainow [19] emphasizes integrating ethical considerations during the design phase of algorithm development; this approach is also supported by the European Commission [20]. By advocating for interdisciplinary collaboration, this approach ensures that technological developments are informed by cultural, sociological, and ethical perspectives, thereby promoting a more holistic and responsible approach to innovation [21].

1.4 Our Contribution

Addressing the challenges and limitations entrenched in the domain of 3D face reconstruction, particularly the pervasive issue of racial bias, our project introduces a novel approach: the integration of a learned illumination prior. This initiative represents a significant departure from traditional reliance on biased 3DMMs and moves toward a more equitable and accurate representation of diverse facial features under varying lighting conditions.

Novel Approach to Illumination – At the heart of our contribution is the integration of an illumination model that is learned independently from any face model. This model is trained on a diverse dataset that includes a wide range of natural lighting conditions, aiming to accurately capture the nuances of how different light interacts with various skin colors. By decoupling the illumination model from the facial model, we mitigate the racial bias that stems from the homogeneity of datasets used in traditional 3DMMs.

Enhancing Albedo Estimation – A critical aspect of our method is its focus on improving albedo estimation, a process integral to distinguishing the intrinsic color of the skin from the extrinsic effects of lighting. Albedo estimation has historically been a challenge in 3D face reconstruction, with existing methods often conflating skin tone with shadow or light reflection. Our learned illumination prior allows for a more nuanced disambiguation, leading to more accurate and unbiased reconstructions.

Addressing the Limitations of Existing Methods – Our approach directly confronts the limitations of existing 3D face reconstruction methodologies. By leveraging a learned illumination model, we reduce the dependency on large, biased datasets and controlled imaging conditions. This not only enhances the applicability of our method in diverse and uncontrolled environments but also paves the way for real-time processing capabilities, breaking down barriers to widespread adoption.

Implications and Applications – The implications of our work extend far beyond technical improvements. By fostering a more accurate and fair representation of all individuals, our method holds the promise of advancing applications in facial recognition, virtual reality, and digital media in a socially responsible manner. We envision a future where technologies powered by our approach contribute to reducing systemic biases and promoting inclusivity in digital and real-world spaces alike. We envision applying our model to challenging tasks such as re-illuminating 3D facial models under different poses or expressions to recreate varied depictions of the same face, thus demonstrating the versatility and robustness of our system.

Addressing the Limitations of Existing Methods – In presenting our learned illumination prior, we acknowledge the beginning of an ongoing journey toward eliminating racial bias in 3D face reconstruction. Our work lays the foundation for future research to build upon, encouraging the exploration of additional methods to further disentangle the complex interplay of facial features and lighting. As we move forward, our goal remains steadfast: to ensure that advancements in technology reflect and respect the diversity of the human experience.

2 Related Work

2.1 Foundational Concepts in Computer Graphics

The integration of 3D modeling, projection, and illumination is fundamental in creating a cohesive visual experience in computer graphics. 3D models provide the structural basis, defining the geometry and surface characteristics of objects within a virtual environment. Projection techniques then translate these 3D models into 2D views by simulating how light travels and images are formed through cameras and the human eye, which is crucial for the visualization on digital displays. Finally, illumination models apply lighting effects to these projections, enhancing realism by mimicking how light interacts with different materials and surfaces in the real world. This interplay not only contributes to the aesthetic quality of the final image but also to the practical understanding and manipulation of objects in applications ranging from virtual reality to architectural visualization. Each component is reliant on the others; accurate modeling is necessary for effective projection, and both are incomplete without the nuanced effects provided by advanced lighting techniques, illustrating a symbiotic relationship that drives much of the progress in the field of computer graphics.

2.2 3D Morphable Models

3D Morphable Models (3DMM) have revolutionized the understanding and reconstruction of facial structures in computer vision and graphics. Originally introduced by Blanz and Vetter in 1999 [3], these models employ a statistical approach to represent variations in human faces through a combination of shape and texture data derived from a dataset of 3D face scans. The fundamental premise of 3DMMs lies in the application of Principal Component Analysis (PCA) to these scans, creating a low-dimensional linear subspace that can effectively capture the diversity of facial features in a controlled manner. One can morph between these faces in PCA space, transfer face characteristics from one face to a different face, or generate new faces, which gave the model its name, morphable model [22].

Subsequent research has expanded the capabilities of 3DMMs by integrating more complex models of facial expressions and more diverse datasets. Amberg et al. [23] extended the basic PCA model to include expressions by learning PCA spaces from residual vectors of expressions, thus enhancing the model's sensitivity to facial dynamics. Further contributions, such as by Li et al. [24], combine linear shape spaces with articulated facial

components like jaws and eyeballs, enabling more nuanced animations and realistic portrayals of facial expressions [25]. The introduction of 4D scans has also allowed for the capture of temporal changes in expressions, significantly improving the fidelity and application scope of these models [26].

One of the persistent challenges in the development of 3DMMs is achieving photorealism, particularly when transforming simpler, compact models into high-quality, realistic outputs [27]. For instance, PCA-based models, while efficient and compact, often lack the fine detail needed for high-resolution outputs due to their global nature. The FLAME project [24] addresses this by learning from thousands of accurately aligned 3D scans, aiming to enhance the realism of generic face models used in various multimedia applications. Concurrently, efforts like those by Booth et al. [28], [29] to learn from a linear face model from almost 10,000 diverse facial scans help address the diversity and realism in generated face models by capturing a broader spectrum of human facial variations.

In parallel, significant strides have been made in modeling facial textures and albedos, which is crucial for achieving realistic coloration and illumination in facial reconstructions. Gecer et al. [30] notably advanced this by employing Generative Adversarial Networks (GANs) to learn a nonlinear texture model that captures high-frequency details often lost in traditional linear models. This approach helps mitigate the common problems associated with shading effects, specular reflections, and camera-specific biases that plague earlier models. However, challenges remain, particularly in distinguishing between albedo and illumination effects, a problem exacerbated by the biases towards certain skin colors in existing models [31].

The future development of 3DMMs seems poised to increasingly leverage in-the-wild data, which could democratize the inputs but also complicate the separation of intrinsic facial features from environmental effects. Although these methods still face challenges in accurately modeling non-Lambertian surfaces which are essential for realistic rendering. This approach, while expanding the applicability of 3DMMs, necessitates careful consideration of ethical aspects, particularly in terms of bias and representation. An ongoing concern is the adequate representation of diverse skin tones and facial features across different ethnicities, which is critical for the equitable application of technology in areas such as digital identity verification, animation, and virtual reality [32].

Smith et al. [33] presents a novel statistical model designed to capture and represent the albedo, an inherent color properties of human faces. This model enhances the realism and accuracy of facial reconstructions by separating albedo from lighting effects, using high-resolution face scans and PCA to derive a low-dimensional representation of albedo variations. When integrated with existing 3DMMs, this model significantly improves ap-

plications such as face recognition and photorealistic rendering, particularly under varying lighting conditions.

3DMMs have come a long way since their inception, growing from basic PCA-based models to complex systems capable of rendering high-fidelity, dynamic facial expressions. The integration of diverse data sources and advanced machine learning techniques continues to push the boundaries of what these models can achieve, promising ever more realistic and inclusive applications. However, as the technology advances, it must also confront the challenges of ethical application and representation, ensuring that it serves a global and diverse user base.

2.3 Inverse Rendering of Faces

Inverse rendering addresses the recovery of object and scene properties, such as geometry, reflectance, and illumination, from image data. This process is foundational for understanding and replicating the appearance of objects, particularly human faces, under varied lighting conditions [34]. Traditionally, inverse rendering is complicated due to the challenge of discerning multiple properties simultaneously from a single image. As highlighted by Ramamoorthi and Hanrahan [35], differentiating between texture and lighting without additional information or assumptions remains particularly problematic.

The use of 3DMMs has proven extremely effective in the field of face analysis. These models capture the intrinsic properties of faces, such as shape and texture, independently of extrinsic factors like illumination or camera angle. In recent advancements, various methods have been employed, including the use of depth maps, surface normals, and particularly meshes that ensure dense correspondence across faces. Dense correspondence refers to the method of mapping one set of face data to another in a way that each point on one face has a corresponding point on the other, allowing for precise comparisons and analyses. This approach is crucial for creating accurate simulations of faces and provides essential constraints that aid in solving inverse rendering problems, which would otherwise be underconstrained. [13].

The integration of differentiable rendering into deep neural networks (DNNs) has revolutionized 3D scene understanding. By allowing gradients of the rendering process to be computed, differentiable rendering facilitates the optimization of 3D scene parameters directly from 2D images. This method bridges the gap between 2D image processing and 3D model optimization, enhancing the capability of DNNs to understand and manipulate 3D data efficiently [36].

Humans have a remarkable ability to infer vast information from a single image, relying on cognitive priors developed through experience. NeRFactor by Zhang et al. [37], although not focused on faces, builds on inverse

rendering principles to reconstruct 3D scenes using neural radiance fields (NeRFs). It forms strong priors on geometry, reflectance, and illumination from large datasets, mimicking human perception. Traditionally requiring multiple views or known lighting conditions, recent advances in NeRFactor allow for high-quality reconstructions from limited input, making it practical for everyday imaging scenarios.

Despite significant advancements, inverse rendering of faces using 3DMMs still faces several challenges. The primary issue is the high degree of under-constraint in single-image scenarios, where multiple interpretations of data can lead to ambiguous results. Current research is focusing on overcoming these limitations by integrating more robust data-driven approaches and enhancing the ability to generalize across different face models without overfitting to specific dataset biases. As these technologies continue to evolve, the accuracy and applicability of face rendering in real-world scenarios will likely improve, leading to more realistic and personalized applications in digital media, virtual reality, and automated facial recognition systems.

2.4 Projection Systems in 3D Modeling and Rendering

The concept of projecting a three-dimensional world onto a two-dimensional image plane is pivotal in understanding camera models within the field of 3DMMs. This process, simply known as projection, encompasses several models which vary in their fidelity to real-world camera behaviors. These models, described here in ascending order of accuracy, are essential for tasks such as camera calibration and pose estimation.

The scaled orthographic model simplifies the projection by assuming uniform scaling and orthographic projection, devoid of size, distance, or perspective ambiguities. While it may lack physical realism due to its linear constraints on vertex position, translation, and scale, it serves practical purposes in 3DMM applications, where distance to the camera is considerably larger than the depth variation in the scene. This model is notably used by Bas et al. [38], Blanz et al. [39], and others, who appreciate its simplicity when complex perspective transformations are challenging.

Advancing the concept of the orthographic model, the affine projection allows for arbitrary affine transformations including non-uniform scaling and skew transformations. This model, while still linear, begins to approximate perspective effects more closely. It is favored in scenarios where the estimation of camera parameters needs to remain straightforward but more flexible than the strict orthographic approach, as seen in works by Aldrian and Smith [40] and Huber et al. [41].

At the pinnacle of realism in 3DMM camera models is the perspective projection, epitomized by the pinhole camera model. This model incorporates intrinsic parameters such as the focal length and the principal point location, directly affecting the final image formation. Unlike its simpler counterparts,

the perspective model accounts for the effects of distance on the projected shape, which becomes crucial at closer distances typical of "selfie" scenarios. This model's adoption, particularly in the seminal works of Blanz and Vetter [3] and subsequent studies like Cao et al. [42], underlines its importance in accurately capturing the nuances of real-world projection.

The intricacies of these models underscore the ongoing challenges in camera calibration, especially in deciphering the ambiguities related to shape, scale, and focal length inherent in more complex models. Researchers continue to explore these complexities, as highlighted by recent studies by Bas and Smith [43] and Smith [44], who have critically analyzed the ambiguities associated with perspective projections in practical applications. Through the lens of these models, the dissertation projects forward, aiming to harness these theoretical foundations to address practical challenges in the application of 3DMMs to real-world tasks, where the choice of camera model can significantly influence both the effectiveness and the efficiency of the outcomes.

2.5 Illumination Models

Illumination models in computer graphics are fundamental in simulating how light interacts with surfaces, contributing to the creation of visually realistic environments. The foundation of illumination modeling in computer graphics was laid by the introduction of the Phong reflection model by Bui Tuong Phong in 1975 [45]. This model, pivotal for understanding light simulation, incorporates ambient reflection, which uniformly lights all surfaces; diffuse reflection, which scatters light broadly to create a matte effect; and specular reflection, which creates sharp highlights on shiny surfaces. An advancement of this model, the Blinn-Phong model introduced by Jim Blinn in 1977 [46], optimized the calculation of specular reflections, making the process computationally more efficient by altering the vector calculations used in the original Phong model.

Subsequent developments in illumination models sought greater physical accuracy through techniques like radiosity and ray tracing. These methods, discussed by Cohen et al. [47] and Whitted [48], treat light as a wave that undergoes reflection, refraction, and absorption, thereby enhancing the realism of the scenes but increasing the computational load, which is a significant consideration in real-time graphics.

A major breakthrough in efficient illumination modeling came with the use of Spherical Harmonics (SH), which allowed for the real-time rendering of complex lighting effects by approximating light distribution in a scene. This technique, detailed by Green [49] compresses lighting information into a series of coefficients, reducing the computational demands significantly. SH has been particularly beneficial in video games and virtual reality, offering an efficient means to render diffuse inter-reflections and soft shadows.

The Precomputed Radiance Transfer (PRT) method by Sloan et al. [50] extends the capabilities of SH by facilitating the interaction of light with dynamic objects within a precomputed lighting environment. This adaptation is crucial for applications in dynamic settings such as gaming and virtual reality, where complex lighting environments must be rendered in real-time.

Further enhancements in handling environments with high-frequency lighting variations were achieved through Wavelet Transform methods, which provide a localized spatial frequency analysis, suitable for detailed and contrast-rich visual environments. Additionally, recent research has explored the potential of machine learning in predictive rendering and illumination. These innovative approaches, including the work of Barron and Malik [51], leverage deep learning to predict and manipulate illumination in images, which is useful for tasks like photo relighting and ensuring color consistency across different lighting conditions.

The concept of illumination priors, which utilize pre-existing knowledge about lighting conditions to influence image processing tasks, has proven beneficial in scenarios where direct measurement of illumination is not feasible, especially in inverse-rendering tasks. In these tasks, the goal is to deduce physical properties of a scene—like shape and reflectance—from the observed images. Barron and Malik's method for estimating shape, reflectance, and illumination from a single image exemplifies the use of a low-dimensional parametric model as an illumination prior. This approach simplifies the complex challenge of separating illumination from reflectance in image processing, a key step in inverse rendering.

The exploration of illumination models in computer graphics has transitioned from simple, less computationally demanding models to more sophisticated techniques that offer greater realism and efficiency. The integration of advanced mathematical techniques and machine learning into illumination modeling has opened new avenues for research and application, particularly in the realm of 3DMMs. This evolution reflects a continuous pursuit of more efficient and realistic rendering techniques in the ever-growing field of computer graphics.

2.6 Neural Illumination: Advancements in Lighting Representation

Neural illumination techniques represents a significant shift in the domain of computer graphics and image processing [52]. They are characterized by their adaptive learning capabilities, where the models dynamically learn from a vast range of lighting conditions using deep learning methodologies. This adaptiveness enhances their ability to predict and recreate complex light interactions, significantly enriching the visual quality of virtual environments by decompsing the task into several simpler differentiable sub-tasks [53]. The models excel in dynamic rendering, adjusting in real-time to changing light conditions to produce more realistic scenes. Furthermore,

these approaches are particularly adept at solving inverse lighting problems—deducing lighting conditions from observed images, which is crucial for accurate scene reconstruction and photorealistic rendering. [54]

The Rotation-Equivariant Natural Illumination (RENI) [55] model stands out as a pivotal advancement in the field, integrating generative capabilities with a rotation-equivariant design to efficiently model natural illumination. RENI utilizes a rotation-equivariant neural field representation, which is crucial for maintaining consistent lighting across different orientations, thereby enhancing the rendering of non-Lambertian surfaces while preserving essential high-frequency lighting effects. This model leverages Vector Neurons for adapting spherical images and employs a variational autodecoder, establishing a robust framework for generative tasks related to spherical signals.

RENI has demonstrated remarkable effectiveness in inverse rendering tasks, showing substantial improvements over traditional lighting models. By directly modeling environmental lighting instead of relying on pre-integrated lighting with a fixed Bidirectional Reflectance Distribution Function (BRDF), RENI supports flexible interaction with arbitrary BRDFs during inference. This flexibility is pivotal for extending its application beyond traditional rendering to tasks like shape recovery from specular reflections. Moreover, RENI's handling of High Dynamic Range (HDR) data ensures that the rendering process remains realistic, crucial for applications where natural lighting plays a significant role.

Building upon the foundation laid by RENI, RENI++ [56] introduces several technological enhancements that significantly broaden its application scope and improve its performance in complex lighting environments. The transition to a Transformer-decoder architecture with positional encoding marks a substantial advancement in handling complex data dependencies more effectively. Moreover, the introduction of a scale-free loss and VN-Invariant layer in RENI++ enhances the model's generalization capabilities and computational efficiency, particularly in challenging HDR scenarios.

Neural illumination, particularly through advancements like RENI and RENI++, plays a crucial role in the integration with 3DMMs and differentiable rendering techniques. These models can significantly benefit from the sophisticated lighting simulations provided by neural illumination, enhancing the realism and accuracy of rendered images, especially in 'in-the-wild' scenarios where real-world lighting conditions are unpredictable and complex [57]. As neural models continue to evolve, they promise to further revolutionize the field by providing even more detailed and responsive lighting solutions, pushing the boundaries of what is possible in computer graphics and beyond.

2.7 Benchmarking Racial Bias

Feng et al. [58] highlights a significant gap in the existing resources for evaluating facial albedo estimation: the absence of a comprehensive dataset that accurately represents diverse skin tones and real-world lighting conditions. To address this, the researchers developed a novel dataset featuring high-quality 3D scans of heads from a balanced range of skin colors and ages, rendered under various lighting scenarios using HDR environment maps. This dataset was designed to serve as a benchmark, referred to as FAIR (Facial Albedo Independent of Race), for evaluating both the accuracy and fairness of facial albedo estimation methods.

A central methodological contribution of the research is the introduction of a set of refined metrics for evaluating albedo estimation. The use of the Individual Typology Angle (ITA) [59], calculated from the CIE Lab* color space, represents an innovative approach to quantifying skin tone objectively and uniformly. The higher the ITA, the lighter the skin and this allows for categorization into six skin types, facilitating a more nuanced analysis of performance across different skin colors [60].

Using the new dataset and metrics, the researchers conducted a comprehensive analysis of existing facial albedo estimation methods. The findings revealed a consistent bias towards lighter skin tones across all methods tested. This benchmark enabled the researchers for the first time to quantify the bias of these methods, presenting a clearer picture of the challenges and limitations inherent in current technologies.

To further address these challenges, Feng et al. [58] leverages the insight that the entire scene image, rather than just a cropped image of the face, contains valuable information about lighting. This aids in the disambiguation of lighting and albedo contributions. By conditioning on both the face region and a global illumination signal derived from the scene image, TRUST effectively regresses facial albedo, resulting in more accurate and fair outcomes on their benchmark.

Da'Prato-Shepard et al. [61] built upon similar insights by integrating an RENI, illumination prior learned independently from the 3DMM model representation, providing a constrained search space of illumination and thus, more accurate albedo estimation. Their ablation study justified the model component's choices, demonstrating less bias compared to popular methods like SH. However, even though they managed to reduce bias for darker skin, they had visibly increased bias for lighter skin. Additionally, their implementation showed that given a scene, their inferred environment map was different for different face subjects, and the environment map looked nothing like the scenic lighting and was generally dark, which could have explained their low bias for darker skin.

3 Method

3.1 Architectural Overview

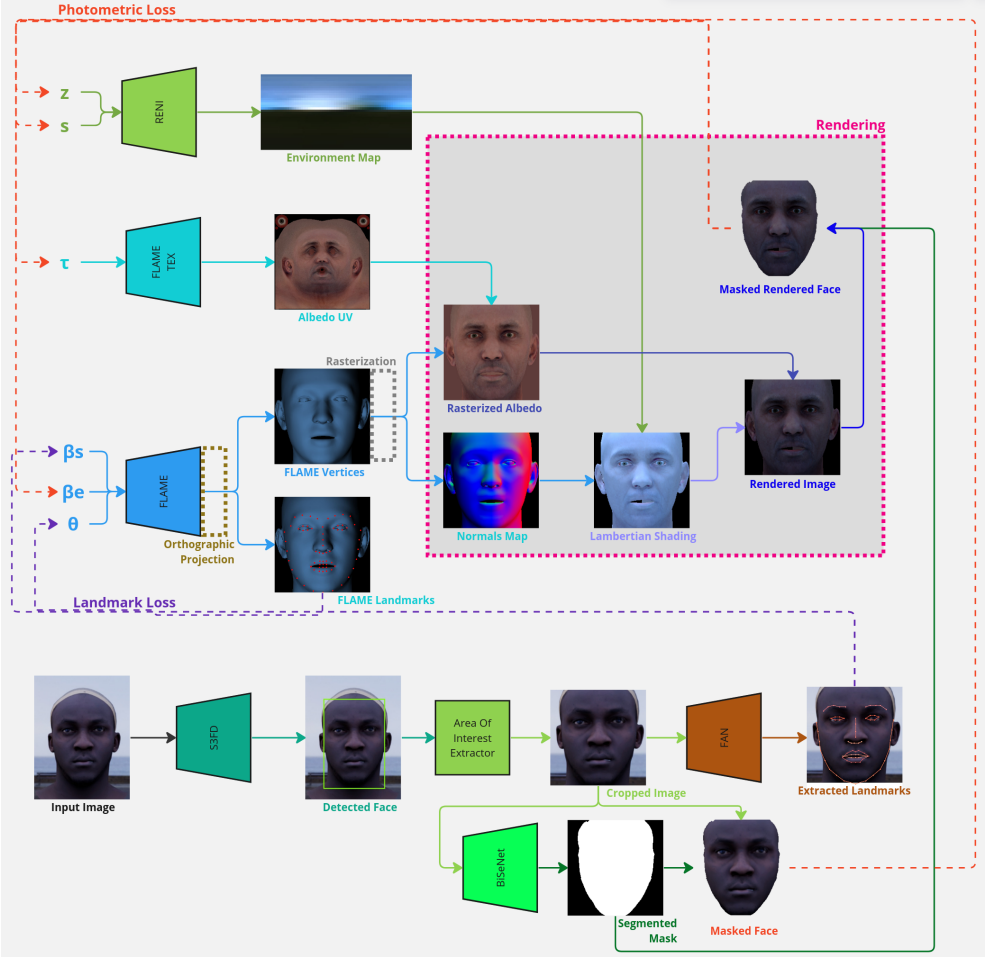


Figure 3.1: Schematic Representation of the 3D Facial Reconstruction Workflow. This diagram provides a visual overview of the system's structured process, showing the sequential phases involved in transforming high-resolution 2D images into photorealistic 3D facial renderings.

Our architecture is designed to encompass various phases that collectively contribute to the robust and efficient 3D facial reconstruction process. Each phase plays a crucial role, and their interactions are pivotal in achieving high-quality outcomes. The process employs an "Analysis by Synthesis" approach in the optimization phase to iteratively refine the model parameters, ensuring the synthetic outputs closely resemble the real-world inputs.

Here's how each phase functions and interconnects within the architecture: **Data Preprocessing**: This foundational phase deals with the collection and preparation of input data. High-resolution images undergo several preprocessing steps, including normalization, landmark detection, and segmentation. This ensures that the data is uniform and standardized, setting a solid base for accurate synthesis and modeling. **Model Development**: Central to our architecture, this phase features the integration of advanced modeling techniques using the FLAME and FLAMETex for dynamic facial structuring and texturing, respectively. Additionally, the RENI++ is incorporated to handle complex lighting interactions, essential for achieving realism in the outputs. This phase generates the initial 3D and illumination models based on the template data. **Rendering**: Before optimization, an initial rendering of the synthesized 3D models is performed to convert them into photorealistic 2D images. This rendering uses techniques such as z-buffering and Lambertian reflectance to closely mimic real-world textures and lighting conditions. The rendered images serve as a baseline to compare against actual images, facilitating the identification of areas needing refinement. **Optimization**: Utilizing the rendered images, this phase applies the "Analysis by Synthesis" methodology to refine the models. It involves adaptive learning rate scheduling to enhance parameter convergence and model stability during training. By continuously analyzing the discrepancies between the rendered outputs and real inputs, the system iteratively adjusts the synthesis to create more accurate and realistic models.

This structured approach, as illustrated in the architectural overview Figure 3.1, ensures a streamlined and effective workflow from data handling to the final realistic rendering of models. Subsequent sections will delve deeper into each phase, outlining their specific roles and the technical specifics underlying their operations.

3.2 Face Detection and Bounding Box Extraction

The Single Shot Scale-invariant Face Detector (S³FD) [62], a CNN, excels in detecting faces under varied conditions and scales, leveraging its architecture that incorporates multiple detection layers tailored for scale invariance. The initialization of S³FD with pre-trained weights from the Large-scale 3D Face Dataset (LS3D-W Balanced) [63] is essential for achieving accurate results across diverse face images, characterized by different poses and lighting conditions. This dataset is crucial for adapting the model to real-world variability in facial recognition tasks. To standardize inputs and enhance detection reliability, the image $\mathbf{I}$ undergoes normalization by subtracting a mean vector $\mu = [104, 117, 123]^T$. S³FD processes the normalized image at multiple scales, identifying potential face regions, each defined by a bounding box and a corresponding confidence score. The set of detected bounding boxes $\mathbf{B}$ is represented as:

$$\mathbf{B} = \{b_i = (x_{\min,i}, y_{\min,i}, x_{\max,i}, y_{\max,i}, c_i) | c_i > \tau, i = 1, 2, \ldots, n\} \quad (3.1)$$

Here, b_i delineates the bounding box coordinates, c_i represents the confidence score, and τ is the threshold that filters out bounding boxes with low confidence. Non-Maximum Suppression (NMS) [64] is then applied to refine these detections by eliminating redundant detections. NMS reduces overlaps and retains the most representative bounding box for each detected face based on a overlap threshold value. This extraction of bounding boxes is paramount, as it ensures accurate localization of facial regions.

3.3 Facial Landmark Detection and Alignment

Following face detection and bounding box extraction, the next step is the extraction of facial landmarks. This process involves pinpointing key anatomical points on the face, such as eye corners, the tip of the nose, and mouth edges. The Face Alignment Network (FAN) [63] is employed for its exceptional accuracy and robustness under various conditions. This CNN, renowned for its capability to capture spatial relationships between facial features, utilizes a deep learning architecture trained on the LS3D-W Balanced dataset, which includes diverse facial types and expressions. Each face image $\mathbf{I}_{\text{crop},i}$ identified by the bounding box is resized to a standard input size for FAN to ensure consistency. FAN incorporates several stacked Hourglass networks [65], adept at detecting facial keypoints by processing feature maps through sequential downsampling and upsampling, producing a series of heatmaps $\mathbf{H}_i$. Landmarks (x_j, y_j) are pinpointed by locating the peak value in each heatmap. The coordinates extracted from the heatmaps are refined to adjust for the scale changes due to resizing, ensuring they match the original image dimensions. The landmarks are outputted as $\mathbf{L} = \{(x_1, y_1), (x_2, y_2), ..., (x_{68}, y_{68})\}$, here $\mathbf{L}$ denotes the set of coordinates for 68 detected landmarks, each representing a specific facial feature. These landmarks are essential for aligning and adapting 3DMMs to the unique contours of an individual's face, as they guide the reconstruction process.

3.4 Facial Segmentation for Detailed Texture Mapping

Following the detection of facial landmarks, the subsequent stage in our 3D facial model involves facial segmentation. The Bilateral Segmentation Network (BiSeNet) [66] architecture features a dual-path design, combining a spatial path that preserves high-resolution details and a context path that captures expansive contextual information. This structure facilitates the efficient processing of both detailed spatial and contextual data, ensuring accurate segmentation of complex facial features. BiSeNet, when trained on the CelebAMask-HQ [67] dataset which consists of high-quality images with detailed facial region annotations, has demonstrated robust performance across diverse imaging conditions. For an input image I, BiSeNet produces a segmentation map S, where each pixel is assigned a label denoting specific facial regions such as the eyes, nose, and mouth. Segmented facial regions are essential for applying texture maps and fitting morphable models with high precision to the contours of the face.

This segmentation ensures that textures and lighting effects are precisely aligned, thereby enhancing the realism of the 3D output. The efficacy of segmentation directly impacts the fidelity of the 3D model, facilitating targeted modifications and enhancements that are vital for authentically rendering facial expressions and characteristics.

3.5 3D Face Model

The FLAME model (Face Landmark-based Articulated Morphable Model) [24] is a computational framework in our pipeline that allows for detailed reconstruction of human faces from images. It parameterizes shape, expression, and pose variations, enabling nuanced control over facial dynamics for realistic adaptations. The model functions through a differentiable scheme that produces a 3D mesh and corresponding facial landmarks, supporting diverse facial geometries and dynamics by manipulating input parameters grouped as follows:

Shape Parameters ($\boldsymbol{\beta}_s \in \mathbb{R}^{100}$) control the baseline facial structure. These are the coefficients that, when multiplied by a shape basis matrix $\mathbf{S} \in \mathbb{R}^{3N\times 100}$, produce the shape deformation:

$$B_{shape}(\boldsymbol{\beta}_s) = \mathbf{S} \cdot \boldsymbol{\beta}_s \tag{3.2}$$

Expression Parameters ($\boldsymbol{\beta}_e \in \mathbb{R}^{50}$) adjust the mesh to reflect specific facial expressions like smiling or frowning. Similar to shape parameters, expression parameters are coefficients for an expression basis matrix $\mathbf{E} \in \mathbb{R}^{3N\times 50}$:

$$B_{exp}(\boldsymbol{\beta}_e) = \mathbf{E} \cdot \boldsymbol{\beta}_e \tag{3.3}$$

Pose Parameters ($\boldsymbol{\theta} \in \mathbb{R}^{6}$) steer head orientation, accommodating rotations and translations. Pose deformations utilize a skinning method with joint-based rotations, described as follows:

$$B_{pose}(\boldsymbol{\theta}) = W(D_{pose}(\boldsymbol{\theta}), \mathbf{J}) \tag{3.4}$$

Here, W is the blend skinning function, D_{pose} is the pose-dependent deformation, and $\mathbf{J} \in \mathbb{R}^{K\times 3}$ is the joint regressor matrix for K skeletal joints.

The vertex positions of the face mesh V are computed by combining these deformations with a neutral face template:

$$V(\boldsymbol{\beta}_s, \boldsymbol{\beta}_e, \boldsymbol{\theta}) = \mathbf{V}_{template} + B_{shape}(\boldsymbol{\beta}_s) + B_{exp}(\boldsymbol{\beta}_e) + B_{pose}(\boldsymbol{\theta}) \tag{3.5}$$

Here, $\mathbf{V}_{template} \in \mathbb{R}^{3N}$ is the neutral face vertex template, where N is the number of vertices in the mesh. Deformations due to shape and expression are linear combinations of basis shapes from $\mathbf{S}$ and $\mathbf{E}$, controlled by Eqs. 3.2 and 3.3 respectively.

Finally, the model computes facial landmarks by selecting specific vertex indices:

$$L = \mathbf{V}_{landmarks} \tag{3.6}$$

where $\mathbf{V}_{landmarks}$ are the vertex positions corresponding to predefined landmark indices on the mesh. This allows landmarks to adapt dynamically to expressions and poses, ensuring accurate alignment and representation across various conditions.

Dynamic memory allocation and hardware acceleration enhance processing efficiency by enabling multiple instances to run concurrently. Parameter initialization sets all parameters—shape ($\boldsymbol{\beta}_s \in \mathbb{R}^{100}$), expression ($\boldsymbol{\beta}_e \in \mathbb{R}^{50}$), and pose ($\boldsymbol{\theta} \in \mathbb{R}^6$)—to zero vectors to begin from a neutral baseline, allowing unbiased parameter exploration during optimization.

The integration of detailed parametrization and dynamic control makes FLAME a robust tool for synthesizing realistic 3D facial models, and being an integral part of the Analysis-by-Synthesis approach.

3.6 Texture Model

Building on the detailed facial geometry provided by the FLAME model, the FLAMETex model extends these capabilities to include high-fidelity texture mapping, necessary for achieving photorealistic 3D reconstructions. The integration of texture not only enhances visual realism but also plays a pivotal role in accurate skin tone representation under varied lighting conditions. FLAMETex utilizes the Balanced Albedo (BalancedAlb) TRUST (Towards Racially Unbiased Skin Tone Estimation) texture model, chosen for its focus on eliminating racial biases in skin tone estimation. This advancement is significant for ensuring fair representation across diverse racial backgrounds, a critical aspect often overlooked in traditional texture models. Texture in FLAME is manipulated through parameters derived from a texture space constructed from a comprehensive dataset. This setup ensures broad generalizability across different ethnicities. Texture parameters (τ) define the skin's coloration and details as:

$$T(\tau) = \mu_t + \sum_{i=1}^{N} \tau_i \phi_i \tag{3.7}$$

Here, $\mu_t \in \mathbb{R}^{3\times512\times512}$ represents the mean texture, $\phi_i \in \mathbb{R}^{3\times512\times512}$ the principal texture components, N (= 50) the number of components utilized, adjustable based on fidelity and resources, and $\tau_i \in \mathbb{R}^{50}$ are the texture coefficients. The texture coefficients are initially set to zero. This initialization reflects starting from the mean texture, which serves as a neutral baseline. Adjustments to these coefficients during model fitting enable the texture to be personalized to match specific facial features more accurately. The output of this is albedo UV map of $512 \times 512 \times 3$ pixels.

We use the BalancedAlb TRUST model in the FLAMETex framework in the hopes that it will reduce racial bias in our facial reconstruction pipeline. Being the most diverse existing model, BalancedAlb, combined with RENI++, ensures accurate and equitable skin tone representation across all racial

groups by incorporating scene disambiguation and a balanced training dataset. This approach not only enhances the visual quality of the models but also promotes fairness and inclusivity in automated systems.

3.7 Camera Model

Following the integration of texture mapping through the FLAMETex model, the next step in our pipeline involves the use of orthographic projection for aligning and scaling the 3D facial models. This projection is applied to simplify the rendering of 3D facial models by neglecting depth information, thus ensuring consistent scale across the image. Unlike perspective projection, orthographic projection does not converge lines, simplifying the calculations and ensuring a consistent scale across the image. In the context of our modeling, camera parameters are structured to support the transformation of 3D vertices and landmarks effectively: **Scale (s)** adjusts the model's size within the viewing frame, ensuring preservation of aspect ratio irrespective of original dimensions. **Translation (t_x, t_y)** moves the model within the 2D plane, positioning the face appropriately within the frame irrespective of pose.

$$X' = s \cdot (X_{\text{proj}} + T) \qquad (3.8) \qquad\qquad X''_{\text{y, z}} = -X'_{\text{y, z}} \qquad (3.9)$$

The camera transformation is encapsulated as $\mathbf{c} = \begin{bmatrix} s & t_x & t_y \end{bmatrix}^T$. where each component captures the relative position and scale of the 3D model with respect to the camera. The transformation from 3D points to their 2D orthographic projection is given by Eq. 3.8, where X_{proj} is the x and y coordinates matrix after projection, and T is the translation vector $[t_x, t_y]$, and s is the scale factor. Depth information is retained in the z-coordinate for rendering depth-dependent effects. This transformation is uniformly applied to both vertex coordinates and facial landmarks from the FLAME model, ensuring consistency in alignment and scaling. Additionally, a reflection transformation adjusts the y and z coordinates post-projection, see Eq. 3.9. This step aligns the model with typical image or scene coordinates, where the y-coordinate increases downwards, essential for correct orientation in Rasterisation process. These techniques guarantee that the spatial dimensions, orientation, and scaling of the 3D facial models are precisely controlled, enhancing both their utility and fidelity for further processing or visualization.

3.8 Illumination Model

Following the alignment and scaling achieved through projection, the next step in our 3D facial reconstruction pipeline involves the integration of advanced neural illumination techniques using RENI++ (Rotation-Equivariant Neural Illumination) model. This extension of the foundational RENI model is tailored for dynamic lighting and intricate light interactions that are characteristic of in-the-wild scenarios, enhancing the photorealism and accuracy of our 3D reconstructions.

In computer graphics, a Lambertian surface refers to an idealized diffuse surface that reflects light uniformly in all directions, resulting in a matte appearance. However, human skin and many real-world materials exhibit non-Lambertian properties, where light reflection varies with viewing and lighting angles, leading to more complex visual effects. RENI++ consists of a sophisticated Transformer-decoder architecture to generate nuanced environmental lighting effects, crucial for realistic rendering of non-Lambertian surfaces like human skin. It operates on a rotation-equivariant neural field, maintaining consistent illumination effects across different orientations and ensuring detailed rendering of high-frequency textures.

The interaction between light and the 3D surface is modeled as:

$$I(x) = \int_{\omega} L(\omega, x) \cdot \rho(\omega, x, v)\, d\omega \tag{3.10}$$

Here, $I(x)$ represents the light intensity at a surface point x, calculated by integrating light contributions from all directions. $L(\omega, x)$ indicates the light arriving at x from direction ω, reflecting the effects of various light sources and environmental interactions. $\rho(\omega, x, v)$ is the Bidirectional Reflectance Distribution Function (BRDF), describing how light is reflected based on its incident direction ω, surface point x, and the viewer's perspective v. The integral over ω sums up these effects from all possible directions, ensuring comprehensive consideration of both direct and indirect lighting.

Within our pipeline, RENI++ is configured with specific environmental map widths and ray sampling numbers to adaptively adjust the lighting based on scene specifics. The model consists of two main parameter vectors: **Latent Space Representation ($\mathbf{z} \in \mathbb{R}^{100\times3}$):** Encapsulates complex lighting variations. **Scale Parameter ($s \in \mathbb{R}$):** Modulates the intensity and distribution of predicted illumination. These dynamically modulate the light intensity and distribution, allowing for real-time adaptation to changing lighting conditions. The output I from RENI++ translates rotation, latent illumination codes, and scale into a coherent illumination map.

Integration of RENI++ within the 3DMM ensures dynamic rendering adjustments in real-time to variable lighting conditions, crucial for inverse rendering tasks where realistic lighting reconstruction is paramount. This holistic approach advances our capability to render realistic human faces under diverse lighting conditions, pushing the boundaries of photorealistic rendering and lighting simulation in dynamic environments.

3.9 Rasterization and Rendering

The final stage in our 3DMM pipeline involves rasterizing and rendering the 3D face model to produce a photorealistic 2D image. This process entails converting geometric data into pixel representations, integrating texture, and applying lighting effects to achieve photorealistic reconstructions. Rasterization involves several critical steps, each governed by precise mathematical operations: **Vertex Transformation:** Eq. 3.11 is used to transform the vertex coordinates $\mathbf{v}$ from world space into normalized device coordinates (NDC) for rendering on a 2D screen. The view matrix M_{view} adjusts coordinates from world space to the camera's perspective, and the projection matrix M_{proj} maps these 3D coordinates into the 2D coordinate system of the screen. **Triangle Setup and Edge Function Evaluation:** Eq. 3.12 determines whether a pixel at coordinates x, y is inside a triangle. The coefficients a, b, c are calculated based on the triangle's vertices, and the sign of $f(x, y)$ indicates if the pixel is within the triangle's boundaries. **Z-Buffering (Depth Buffering):** Eq. 3.15 updates the depth buffer to maintain the correct visibility of surfaces. It compares the current depth z at a pixel location with the existing value in the buffer and retains the minimum value, ensuring that closer objects occlude those farther away. **Attribute Interpolation:** Eq. 3.13 is used for interpolating vertex attributes (like normals and UV coordinates) at pixel p using the barycentric coordinates w_1, w_2, w_3 of the triangle. It effectively blends the attributes of the triangle's vertices based on their proximity to the pixel, where a_i are vertex attributes, and a_p is the interpolated attribute at pixel p.

$$\mathbf{v}' = M_{\text{proj}} M_{\text{view}} \mathbf{v} \tag{3.11}$$

$$f(x, y) = a \cdot x + b \cdot y + c \tag{3.12}$$

$$a_p = \frac{w_1 a_1 + w_2 a_2 + w_3 a_3}{w_1 + w_2 + w_3} \tag{3.13}$$

$$I_p = \rho \sum_{i=1}^{n} (\mathbf{l}_i \cdot \mathbf{n}_p) \mathbf{L}_i \tag{3.14}$$

$$z_{\text{buffer}}[x, y] = \min(z_{\text{buffer}}[x, y], z) \tag{3.15}$$

$$C_p = \mathbf{T}_p \odot I_p \tag{3.16}$$

Each fragment generated contains interpolated values that are passed to the shading stage for further processing: **Lambertian Reflectance:** Eq. 3.14 calculates the light intensity at pixel p using the Lambertian model, which is ideal for simulating diffuse reflections. It incorporates the cosine of the angle between each light source's direction $\mathbf{l}_i$ and the normal at the pixel $\mathbf{n}_p$, multiplied by the light's intensity $\mathbf{L}_i$ and the surface albedo ρ. **Final Image Composition:** This final Eq. 3.16 combines the texture color $\mathbf{T}_p$ and the computed shading intensity I_p at each pixel p to produce the final color C_p. It uses element-wise multiplication to modulate the texture with the shading, reflecting realistic interactions between the material properties and lighting. This rigorous approach ensures that our 3D facial reconstructions are not only geometrically accurate but also exhibit high visual fidelity. By

managing the intricate interplay between geometry, texture, and lighting, the pipeline achieves a realistic rendering of 3D facial models, closely mimicking the complexities of human features under diverse lighting conditions.

3.10 Regularization and Loss Functions

Defining sensible regularization and loss functions is essential to achieve realistic and accurate reconstructions. Regularization helps in maintaining the generalizability of the model across diverse facial images and prevents overfitting, while specific loss functions guide the fitting process towards realistic outcomes. In particular, regularization and landmark-loss are crucial to resolve ambiguities and constrain an otherwise very ambiguous and hard-to-solve problem.

Regularization Techniques: Shape Regularization (3.17) prevents significant deviations in shape parameters by controlling deviation from neutral face shapes, with β_s as shape coefficients. Expression Regularization (3.18) maintains realistic expressions by moderating deformations and regulating β_e, the expression coefficients. Pose Regularization (3.19) ensures natural head orientations; θ are the pose parameters. Latent Codes Regularization (3.20) smooths rapid changes in lighting and texture, with $\mathbf{z}$ as latent illumination codes.

$$L_{\text{shape}} = \frac{1}{2}\sum \beta_s^2 \qquad (3.17) \qquad L_{\text{expr}} = \frac{1}{2}\sum \beta_e^2 \qquad (3.18)$$

$$L_{\text{pose}} = \frac{1}{2}\sum \theta^2 \qquad (3.19) \qquad L_{\text{latent}} = \mathbb{E}[\|\mathbf{z}\|^2] \qquad (3.20)$$

$$L_{landmark} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{2}(l_{ij} - g_{ij})^2} \qquad (3.21)$$

$$\text{smooth}_{L1}(x, y) = \begin{cases} 0.5(x-y)^2 & \text{if } |x-y| < 1 \\ |x-y| - 0.5 & \text{otherwise} \end{cases} \qquad (3.22)$$

$$L_{photo} = \text{smooth}_{L1}(\mathbf{M} \odot \mathbf{I}_{\text{pred}}, \mathbf{M} \odot \mathbf{I}_{\text{true}}) \qquad (3.23)$$

Loss Functions: Landmark Loss (3.21) ensures accurate positioning of facial features by aligning 3D model with 2D facial landmarks. It measures Euclidean distance between predicted landmarks l_{ij} and ground truth g_{ij}. Photometric Texture Loss (3.23) matches the texture of the reconstructed face to the observed image, using a robust Smooth L1 loss (3.22) to handle occlusions and pixel anomalies effectively. These techniques ensure that the fitting process optimizes a balance between fidelity and generalizability, achieving realistic and plausible outputs under diverse conditions.

The overall loss function is defined as:

$$L = L_{photo} \cdot \omega_{photo} + L_{landmark} \cdot \omega_{landmark} + L_{shape} \cdot \omega_{shape} + L_{expr} \cdot \omega_{expr} + L_{pose} \cdot \omega_{pose} + L_{latent} \cdot \omega_{latent} \quad (3.24)$$

3.11 Adaptive Learning Rate Scheduling

Adaptive learning rate scheduling is pivotal for optimizing the 3D face reconstruction model, enhancing convergence and output quality by fine-tuning the learning rates across various model parameters:

Parameter Groupings and Rates: **Geometric Parameters:** $L_r^{\text{default}} = \eta$ is for shape, expression, pose, and camera settings. **Texture Parameters:** $L_r^{\text{tex}} = \eta_{\text{albedo}}$ is specifically for albedo adjustments. **Latent Illumination Codes:** $L_r^{\text{latent}} = \eta_{\text{latent}}$ targets lighting and shading effects.

$$\theta_{t+1} = \theta_t - \frac{\eta_t}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (3.25) \qquad \text{LR}_i(t) = \text{LR}_{i,\text{initial}} \cdot \gamma_i(t) \quad (3.26)$$

The Adam optimizer is utilized for its adaptive moment estimation, crucial for handling large datasets and varied parameter spaces as described by Eq. 3.25 where θ_t are parameters at timestep t, η_t, $\hat{m}_t$, and $\hat{v}_t$ are learning rate, and bias-corrected first and second moment estimates, respectively, ϵ is added for numerical stability.

The optimization process is structured into distinct phases: **Rigid Fitting Phase:** Initial alignment using rigid transformations. **Non-rigid Fitting Phase:** Detailed adjustments to expressions and subtle shape refinements. Learning rates are dynamically adjusted during these phases using Eq. 3.26. where $\text{LR}_i(t)$ is the learning rate for parameter group i at iteration t, and $\gamma_i(t)$ is a scaling factor derived from the current phase. A custom scheduler manages the transition between phases based on predefined criteria, ensuring optimal learning conditions throughout the training process.

Each iteration involves: 1. Gradient calculation via backpropagation. 2. Parameter updates using the current learning rates provided by Adam. 3. Learning rate adjustments by the scheduler, reflective of progress in the fitting phases. This approach ensures that the model efficiently adapts to the complexities of the optimization landscape, resulting in high-quality, realistic 3D facial reconstructions. By intelligently managing learning rates and phases, the system expertly balances exploration and exploitation, leading to effective convergence tailored to the unique dynamics of each parameter set.

4 Evaluation

We compare our model qualitatively and quantitatively with several state-of-the-art (SOTA) methods. MGCNet [68], Deng et al. [69], INORig [70] and DECA [10] use the Basel Face Model (BFM) [71] for albedo estimation; TRUST use the BalancedAlb [58] model; GANFIT [30] uses its own GAN-based appearance model; and CEST [72] is a model-free approach. We conducted analysis using a system equipped with 32 GB of RAM, an NVIDIA GeForce RTX 2070 GPU, and a Ryzen 5 3600X CPU. The software environment included PyTorch 1.10.0, PyTorch3D 0.7.1, and CUDA 12.4.

4.1 Quantitative Analysis

The FAIR benchmark [58] for evaluating the accuracy of albedo estimation on skin color adopts the ITA as an objective metric because of its objectivity, ease of computation from images, and significant correlation with skin pigmentation. See Appx. A to know more about ITA.

The benchmark evaluates the following metrics: **ITA Error**: The mean error in ITA degrees between the predicted albedo UV and the ground-truth albedo UV, calculated over the skin region. **Bias Score**: The standard deviation of the per skin group ITA errors, quantifying bias by showing the variability in performance across different skin groups. **Total Score**: The sum of the ITA error and the bias score, providing an overall performance measure. To facilitate future evaluations using this benchmark, Feng et al. [58] released their validation set constructed using 206 high-quality 3D head scans. It contains synthesized images, cropped images for each head, pre-masked ground-truth albedo UV maps, and ground-truth coordinates of the 68 landmarks. This comprehensive dataset enables robust benchmarking and comparison across different albedo estimation methods. See Appx. B for example images of the dataset and Appx. C for our performance on it.

The comparative analysis of different albedo estimation methods, as detailed in Table 4.1, reveals nuanced insights into the performance of our model relative to established benchmarks. Our model demonstrates strong performance across lighter and tanner skin types (I-IV), though it exhibits a modest decrease in accuracy for darker skin types (V-VI). The bias score of 7.36 for our model signifies a balanced performance across various skin types, significantly lower than that of other methods like GANFIT [30] (29.04) and DECA [10] (26.69). It is also worth noting that Da'Prato-Shepard et al. [61] utilized RENI as their illumination model with an optimization-based

Method	Avg. ITA ↓	Bias ↓	Score ↓	ITA per skin type					
				I	II	III	IV	V	VI
Deng et al. [69]	22.57	22.31	44.89	**8.92**	9.08	8.15	10.90	28.48	69.90
GANFIT [30]	62.29	29.04	91.34	94.80	87.83	76.25	65.05	38.24	11.59
MGCNet [68]	21.41	16.04	37.46	19.98	12.76	8.53	9.21	22.66	55.34
DECA [10]	28.74	26.69	55.43	9.34	11.66	11.58	16.69	39.10	84.06
INORig [70]	27.68	25.73	53.40	23.25	11.88	**4.86**	9.75	35.78	80.54
CEST [72]	35.18	11.08	46.26	50.98	38.77	29.22	23.62	21.92	46.57
TRUST (BFM [71])	16.19	13.99	30.18	12.44	**6.48**	5.69	9.47	16.67	46.37
TRUST (AlbedoMM [33])	17.72	13.95	31.67	15.50	10.48	8.42	**7.86**	**15.96**	48.11
TRUST (BalancedAlb)	13.87	**2.55**	**16.43**	11.90	11.87	11.20	13.92	16.15	18.21
Da'Prato-Shepard et al. [61]	29.21	9.48	38.69	36.40	35.62	28.37	36.70	20.20	**11.45**
Ours	26.53	7.36	33.89	21.94	18.87	20.11	25.65	33.79	38.81

Table 4.1: Comparison to state-of-the-art methods on the FAIR benchmark [58]. Excluding TRUST [58], our model excels in minimizing bias and ITA error, as well as accurate skin colour predictions.

fitting approach similar to ours. However, our iterative model, RENI++, achieved better results, highlighting significant improvements in accuracy and bias reduction over the baseline RENI model. This demonstrates the effectiveness of our refinements in handling complex illumination conditions.

Notably, the TRUST (BalancedAlb) [58] method exhibits the best overall performance in the benchmark, achieving an ITA of 13.87 and a bias score of just 2.55. This model employs a sophisticated approach involving the prediction of global illumination and segment-specific facial analysis, which likely contributes to its superior performance. The comparatively higher bias score of our model suggests a need for incorporating similar techniques to enhance our handling of scene variability and illumination conditions, potentially elevating our performance to match or exceed that of TRUST (BalancedAlb) [58]. We will explore the similar idea in the following section. When excluding TRUST [58] from the comparison, to ensure a more equitable assessment among models that analyze facial images without the influence of advanced environmental pre-processing, our model excels. It outperforms others in terms of bias score, reflecting a robust capability to manage diversity across different skin types with minimal disparity. Our method not only excels in reducing bias but also maintains competitive accuracy across various skin types. This dual achievement underscores the exceptional performance of our model, illustrating its potential as a fair and effective tool.

4.2 Qualitative Evaluation

To evaluate our method qualitatively, we conducted visual comparisons with various SOTA methods. These comparisons were essential in highlighting differences in both the geometric accuracy of facial reconstructions and the fidelity of skin tone representations. Shading mainly affects local geometry, whereas skin tone, a global property of albedo, presents distinct challenges. Accurately estimating shape does not ensure precise skin tone representation, and vice versa. For example, while DECA [10] excels in

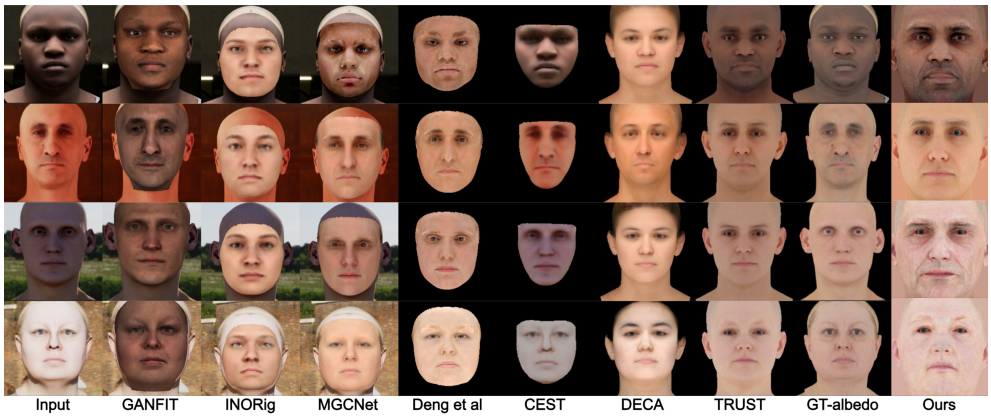


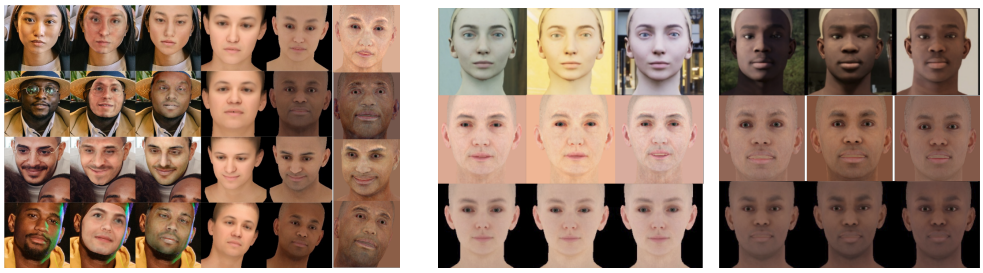
Figure 4.1: Comparing recent face reconstruction methods on synthesised facial images from TRUST benchmark. From left to right: the input image, GANFIT [30], INORig [70], MGCNet [68], Deng et al. [69], CEST [72], DECA [10], TRUST [58], the ground-truth albedo rendering and in the end, our method.

shape reconstruction, it often struggles to capture accurate skin tones due to the constraints imposed by albedo regularization.

In our observations, methods such as GANFIT [30], INORig [70], and DECA [10] tend to produce albedo maps with limited variety. These methods show low ITA error values for certain skin types but exhibit high ITA error values for others, indicating a notable bias toward specific skin tones. The model-free approach adopted by CEST [72], while ambitious, often fails to disentangle lighting effects from albedo accurately, resulting in albedo maps that carry significant remnants of lighting conditions, thereby compromising skin tone accuracy. MGCNet [68] and, to a lesser extent, Deng et al. [69] manage to generate more diverse albedo outputs. However, given that the BFM [71] lacks dark skin tones in its training dataset, representations of these skin tones rely heavily on extrapolation. This extrapolation tends to introduce noise and diminishes accuracy, particularly for darker skin tones.

Our model distinguishes itself by striking an optimal balance between geometric precision and skin tone fidelity. It achieves this by ranking second only to the TRUST [58] model in terms of albedo accuracy, producing consistent and realistic skin tones across a broad range of conditions. Moreover, our method showcases geometric accuracy comparable to that of DECA [10], ensuring that facial reconstructions are both detailed and accurate. Further demonstrating the robustness of our method, we applied it to images of the same subjects taken under various lighting conditions and against differing backgrounds. As illustrated in Figure 4.2, the estimated albedo for each subject remains consistent across these changes, affirming that our method not only captures but also faithfully reproduces skin tones accurately and robustly, irrespective of external lighting or background variations. Although there is a subtle change in albedo under indoor lighting conditions, which we will explore in subsequent sections, but the

reconstructed facial shape remains remarkably consistent.



(a) (b) (c)

Figure 4.2: (a) Qualitative comparisons on real world images. From left to right: the input image, INORig [70], MGCNet [68], DECA [10], TRUST [58], and our method. **(b) & (c)** Given synthesised facial images from FAIR benchmark [58] under varying lighting (indoor and outdoor), our model (2nd row) and TRUST (3rd row) outputs similar albedo and face shape.

4.3 Effect of Illumination Conditions

The FAIR benchmark [58] encompasses a diverse array of lighting scenarios, spanning both natural and artificial conditions. But our illumination model, RENI++ [56], was specifically trained with a focus on outdoor natural lighting conditions. Based on this training specialization, we anticipated superior performance in outdoor settings compared to indoor environments, which typically feature artificial lighting. This hypothesis was substantiated through a detailed analysis of the benchmark data, which we manually categorized into outdoor and indoor classes based on the predominant lighting conditions. The dataset had 45.3% outdoor and 54.7% indoor scenarios, and our findings revealed a significant 49.3% decrease in the bias score (std) from indoor to outdoor scenarios. This marked discrepancy underscores the sensitivity of our model to changes in lighting conditions and highlights a critical area for future optimization.

4.4 Ablation Study

Method	Avg. ITA ↓	Bias ↓	Score ↓	ITA per skin type					
				I	II	III	IV	V	VI
BFM [71] & SH [49]	30.81	27.78	58.58	9.62	6.41	12.50	20.42	54.10	81.79
BalancedAlb [58] & SH [49]	17.02	8.48	25.50	5.49	10.36	17.07	22.17	15.05	31.98
BalancedAlb [58] & RENI++ [56]	26.53	7.36	33.89	21.94	18.87	20.11	25.65	33.79	38.81

Table 4.2: Comparison of different iteration of our model, to verify the contribution of different enhancement to the overall performance.

An ablation study is pivotal for elucidating the influence of various components and configurations within our model. It delineates the enhancements achieved through successive iterations of the model development process, facilitating a deep understanding of each element's contribution to the overall performance. In this study, we evaluated three configurations, as

detailed in Table 4.2 & Appx. D: the baseline model utilizing the BFM [71] with SH, an intermediate model that integrates BalancedAlb model [58] while retaining SH, and our final model which combines BalancedAlb with RENI++ [56].

BFM & SH: This configuration, serving as our baseline, exhibits a high bias score of 27.78 , indicating significant variability in performance across different skin types. It particularly struggles with darker skin tones, as evidenced by elevated ITA errors for Types V and VI. The shortcomings in handling diverse albedos highlight the limitations of the BFM when used without modifications tailored to enhance albedo balance. **BalancedAlb & SH**: By incorporating BalancedAlb, this model configuration achieves a notable reduction in bias score to 8.48. This adjustment yields more consistent performance across skin types, with marked improvements particularly for darker skin tones (Type V: 15.05, Type VI: 31.98). The introduction of BalancedAlb plays a crucial role in moderating the bias inherent in the baseline model, thereby improving fairness and accuracy in albedo estimation. **BalancedAlb & RENI++**: Our final model configuration, which includes RENI++ in addition to BalancedAlb, further lowers the bias score to 7.36. This demonstrates the efficacy of RENI++ in enhancing model stability and consistency across a diverse range of skin types. It's worth noticing that integration of RENI++ caused a spike in avg. ITA error. The major reason behind this could be due to us limiting the resolution of the environment map to 32×64 pixels because of high resource demands of our experiments. The limited resolution may have led to a loss of detail and accuracy in capturing the nuances of outdoor lighting, thereby increasing the ITA error in these scenarios. Additionally, since RENI++ was not optimized for indoor lighting, the error increased further.

In our study, we identified a "baked-in" illumination effect when evaluating indoor lighting conditions, as illustrated in Figure 4.2. This issue arose due to our initial optimization strategy, which optimized albedo for only 250 iterations before adjusting for illumination. This approach was intended to anchor the optimization process with a texture that approximates actual skin color, aiming to capture a more naturalistic appearance from the outset. However, this method proved less effective under the simpler and more uniform indoor lighting, inadvertently causing some lighting features to be embedded directly into the albedo. This was particularly noticeable around specular highlights, where illuminated regions distorted the skin tone estimation. This preliminary strategy also influenced our benchmark evaluation, particularly affecting the ITA error metrics under indoor scenarios and for darker skin tone. In response, we are considering enhanced normalization techniques during the albedo-only optimization phase to prevent albedo from capturing transient lighting effects. By refining our approach to better balance the albedo and illumination components from the initial stages of model training, we aim to mitigate the baked-in effect and improve the

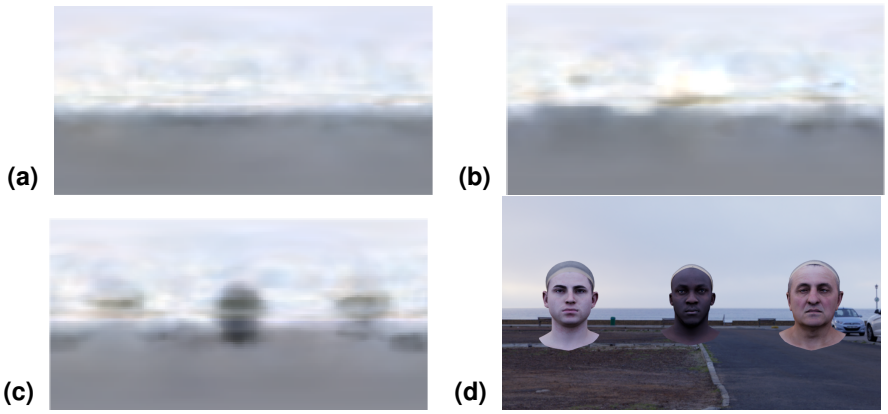


Figure 4.3: Environment maps generated by RENI++ under different preprocessing conditions: (a) Manual face removal, yielding the most accurate background illumination cues. (b) Masking faces using RENI++'s inbuilt feature, introducing some deviations. (c) Raw, unprocessed image, resulting in visible albedo imprints and the least accurate environment map. (d) Original scene image for comparison.

model's accuracy and consistency across diverse lighting conditions. This ablation study not only confirms the validity of our incremental improvements but also sets a clear path for future enhancements to bolster the model's performance even further.

4.5 Investigating RENI++ and Scene Disambiguation

Scene disambiguation involves the process of accurately interpreting and separating different elements within a scene to understand the context and relationships among them. Our current method for estimating scene illumination relies solely on the face region. During the fitting process, the environment map is integrated as an illumination model by the renderer. The optimization is performed on the face crop region, where the RENI++ [56] parameters are adjusted to generate an environment map, which is then used to render the scene illumination onto the face. However, it is intuitive that there is a significant amount of illumination information present in the background of the image, which our current approach overlooks. FLAME [24], utilizes both global scenic and local facial information in a two-step process to estimate illumination. In contrast, our method has so far only employed the local facial information. To investigate how incorporating the background information could improve our model, we conducted a series of tests.

Instead of providing RENI++, with only a cropped facial image, we used whole scene images from the benchmark dataset, as shown in Figure 4.3. We preprocessed these images in three different ways: (a): In this scenario, we manually removed faces from the scene, allowing the model to focus solely on the background illumination cues. This approach yielded the most

Figure 4.4: Demonstrating the versatile application of our model in dynamic and realistic 3D facial rendering. (a) From left to right: an input image, reconstructed render, environment map for relighting, and three novel viewpoints of the face with applied illumination. (b) Shows an input image on the left, and reconstructed render, followed by modification and animation of facial expressions.

accurate environment map generation, as it eliminated any direct influence from the facial regions. (b): We utilized RENI++'s inbuilt mask feature to obscure the facial area. Although this method was effective, it introduced a degree of randomness. The model's attempts to inpaint the masked region resulted in deviations from accurate environmental cues. (c): Providing the model with the raw, unprocessed image resulted in the least favorable outcome. The direct inclusion of faces led to visible albedo imprints in the environment map, demonstrating the significant impact facial features have on the model's ability to accurately generalize the lighting conditions.

Our findings suggest that future work should incorporate background information to estimate the RENI coefficients and environment map more accurately. The manual face removal process could be automated using techniques such as Stable Diffusion Inpainting [73], thereby improving efficiency and consistency in preprocessing the images.

4.6 Applications

Our method aims to create a versatile model with broad applications across various industries. We conducted experiments to showcase the model's capabilities in realistic 3D facial rendering (Figure 4.4). In the first experiment (a), we reconstructed 3D faces from 2D images and re-rendered them under various lighting conditions, demonstrating its use in film, video games, and virtual reality. We also manipulated poses and camera settings to show its flexibility in generating realistic renderings from multiple viewpoints. The second experiment (b) highlighted the model's ability to modify and animate facial expressions and texture, covering diverse skin tones, beneficial for animation, digital marketing, and social media. These experiments demonstrate the model's technical proficiency and its potential to transform digital facial interactions across media and communication platforms, enabling innovative applications and enhancing user experiences.

5 Conclusion and Further Work

This project successfully addresses racial bias in 3D face reconstruction by integrating a learned illumination prior independent of face models. This approach significantly improves accuracy and fairness across various skin tones, outperforming traditional 3D Morphable Models (3DMMs). The BalancedAlb TRUST model enhances albedo estimation, resulting in precise and unbiased reconstructions. Our model shows competitive performance in diverse lighting conditions and facial expressions, marking a substantial step towards fairer and more accurate 3D facial reconstruction technologies. This project underscores the potential for more inclusive and equitable applications in facial recognition, virtual reality, and digital media by leveraging advanced machine learning techniques.

Building on the successes of this project, several avenues for future research and development have been identified:

- Future work will focus on improving the model's adaptability to indoor and artificial lighting conditions. Current models show increased errors under these scenarios, and targeted enhancements could further reduce these inaccuracies.
- Continuing to address ethical considerations in the development and deployment of 3D reconstruction technologies is paramount. This includes ensuring equitable representation and application across all demographics, as well as ongoing monitoring and mitigation of potential biases.
- Further refining techniques for inverse rendering and scene disambiguation to enhance the accuracy of environmental context interpretations. This will improve the model's performance in dynamically changing light conditions and complex scenes.
- Future work should include re-running and evaluating the model without the current resource constraints. This will allow us to fully explore and realize the model's true potential, and may even lead to further improvements in accuracy and fairness.

By continuing to explore these areas, future advancements can build upon the foundation laid by this project, striving towards the goal of creating fairer, more accurate, and inclusive 3D facial reconstruction technologies.

A Individual typology angle

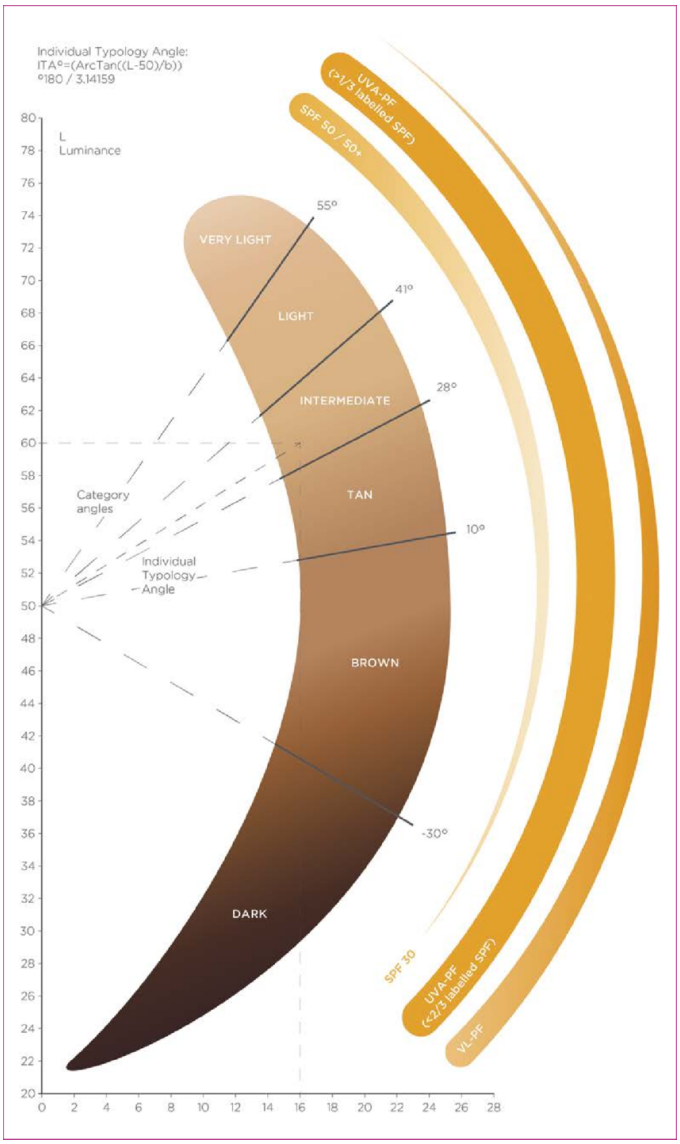


Figure A.1: Individual Typology Angle (ITA) classification of skin types. The ITA is measured in degrees and categorizes skin from very light (ITA $>$ 55°) to dark (ITA $<$ -30°). Each range of ITA values corresponds to a specific skin type, with lower ITA values indicating darker skin phototypes. This figure illustrates the six categories and their respective ITA ranges.

B Evaluation Dataset Examples

In this chapter, we present a variety of examples from evaluation dataset. This dataset is diverse, containing images that showcase different ethnicities, varying lighting conditions, both indoor and outdoor environments, and even some complex shadow patterns.

FAIR [58] evaluation dataset comprises 234 data points. Each data point is structured and includes the following components: **Full Scene Image**: This is a complete image that includes three synthesized heads. The scene is rendered with specific lighting conditions and background settings to mimic realistic scenarios. **Head Crops**: Each of the three heads in the full scene image is cropped and presented individually. This provides a closer view of the head for detailed analysis. **Ground-Truth Albedo UV Maps**: These are pre-masked albedo maps for each head, serving as the ground truth for evaluating the texture and color information predicted by our model. **Ground-Truth Landmark Coordinates**: For each head, a file containing the ground-truth coordinates of 68 facial landmarks is provided. These coordinates are crucial for assessing the accuracy of the facial geometry predicted by any model.

To illustrate the dataset structure, each example below follows this layout: **First Row**: The first row displays the complete raw scenic image. This image includes the three synthesized heads with applied illumination and background settings. **Second Row**: The second row presents the cropped images of each head. These crops are extracted from the full scene image and provide a detailed view of each head individually. **Third Row**: The third row shows the ground-truth masked albedo UV maps for each head. These maps are used as the reference for comparing the predicted albedo from our model.

By providing these detailed examples, we aim to offer a clear understanding of the dataset used for evaluation and the stringent standards against which our model was compared.

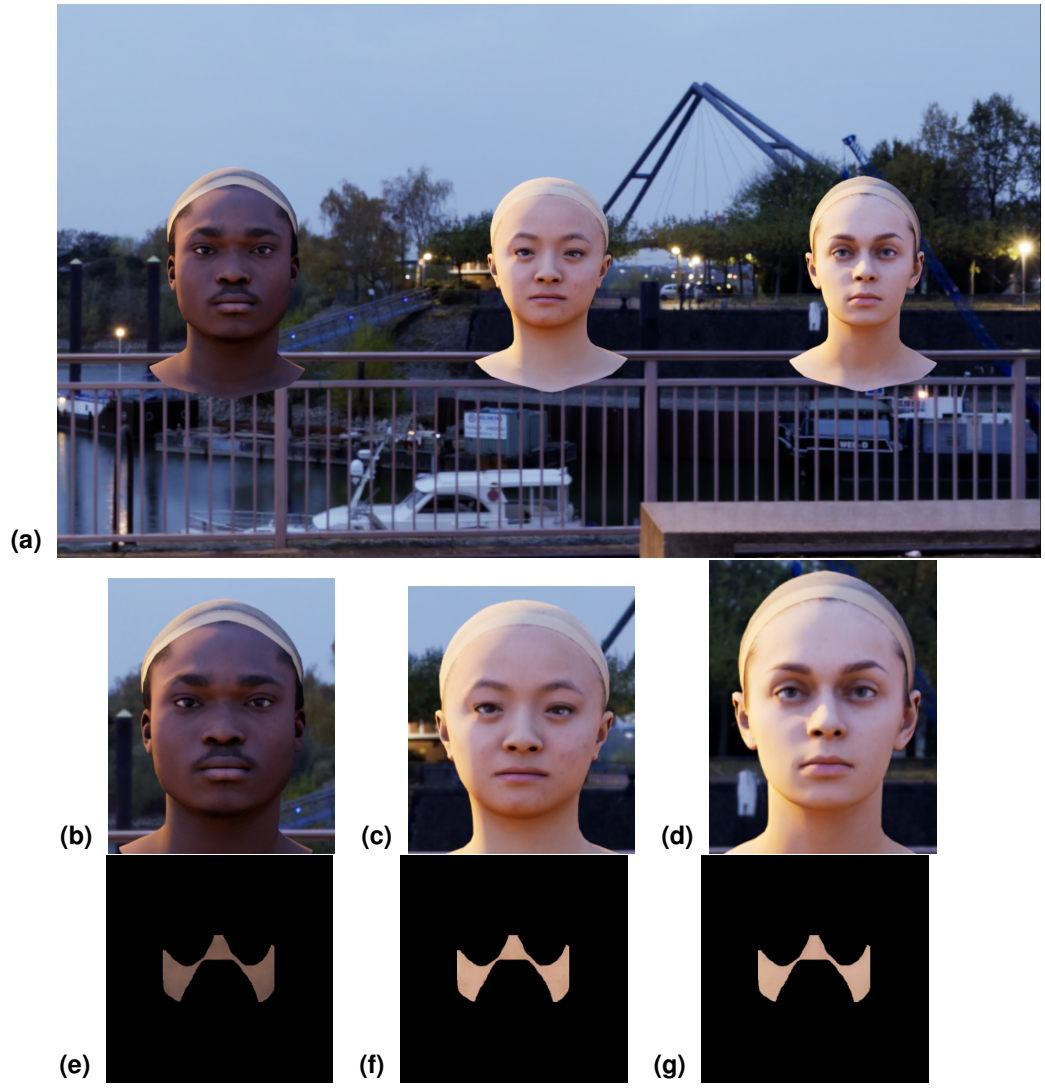


Figure B.1: The image captures an outdoor scene by a canal during the evening. The three synthesized heads are illuminated by a street lamp, casting a warm light on their faces, which adds to the realism of the scene. The ambient evening light combined with the artificial street lamp creates a dynamic lighting condition, ideal for evaluating the model's performance in low-light and mixed lighting scenarios.

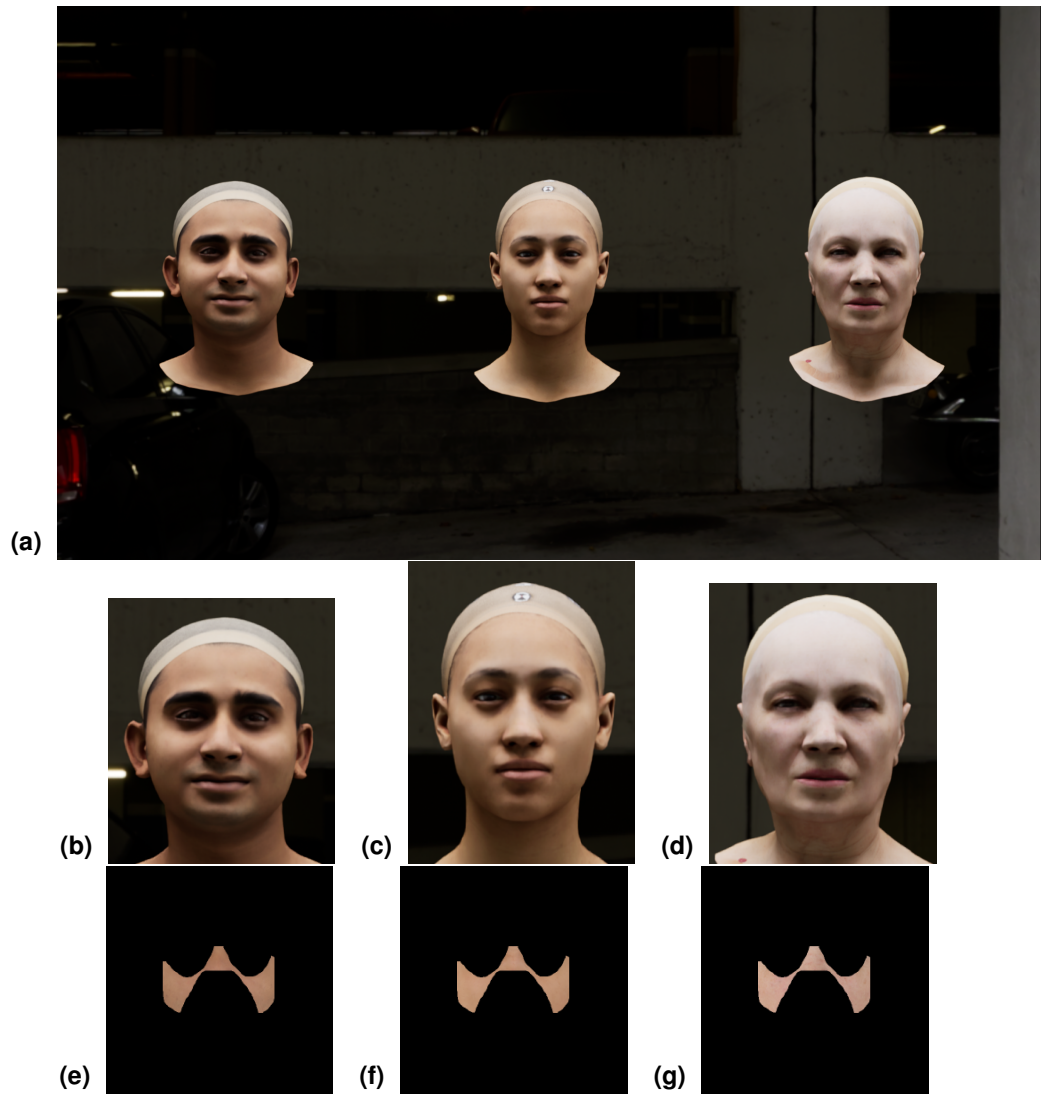


Figure B.2: This example features an indoor scene in a car park at night. The lighting is uniform and white, providing consistent illumination across the three synthesized heads. This setting tests the model's ability to handle uniform artificial lighting, which is common in controlled indoor environments.

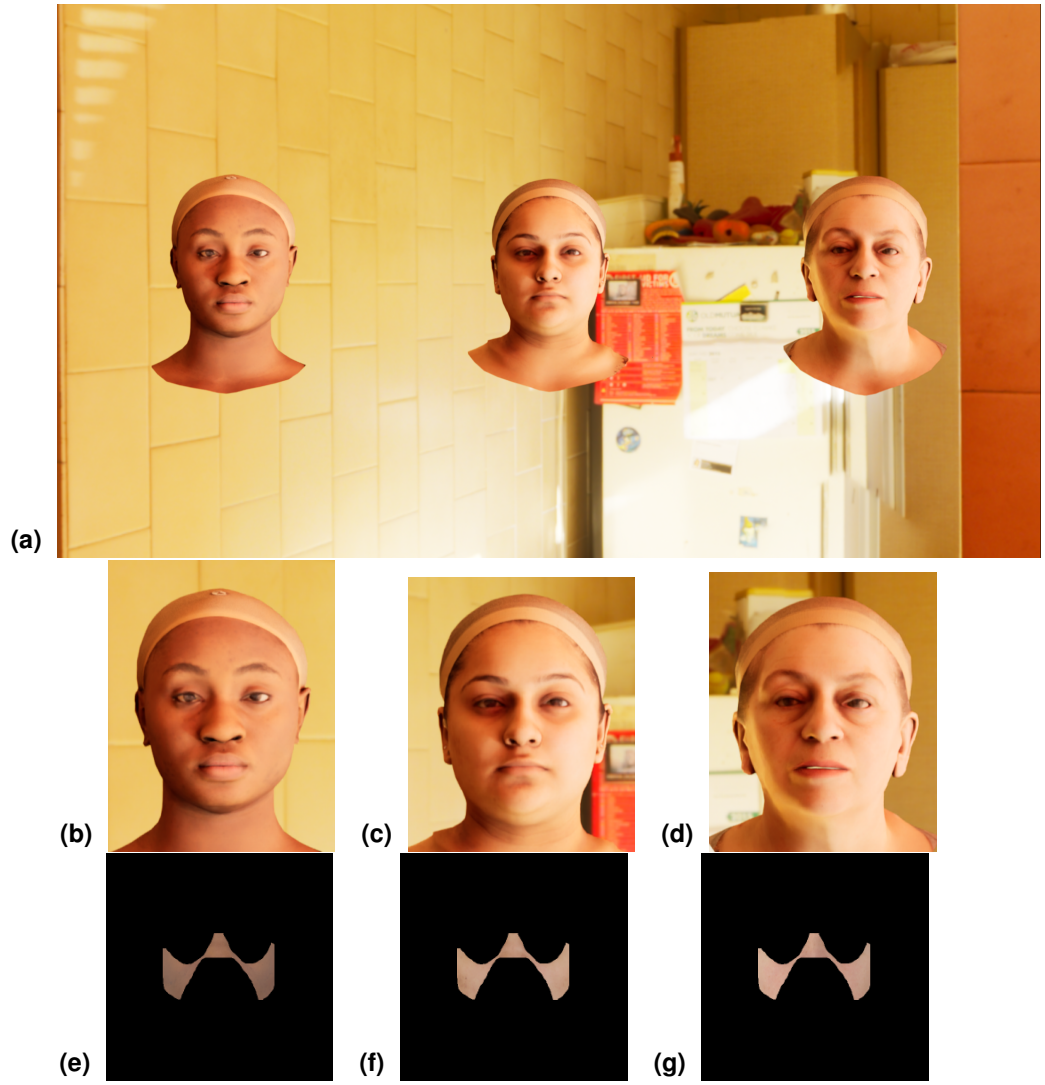


Figure B.3: The scene takes place in a kitchen with extremely bright yellow light illuminating the faces. This light could be natural sunlight or a result of high camera exposure. The intense and warm lighting condition challenges the model to accurately capture and reproduce the texture and color of the faces under high exposure and vibrant lighting.

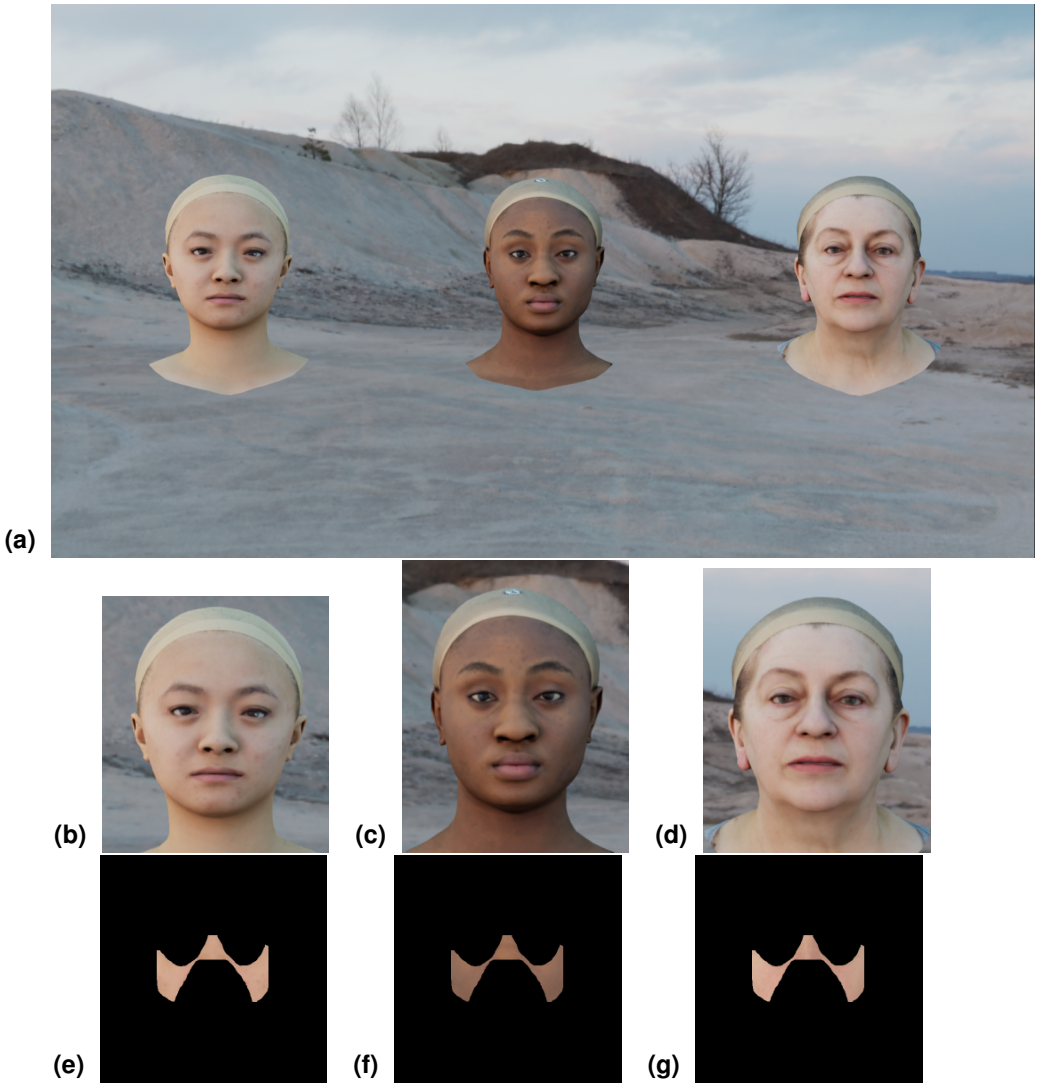


Figure B.4: Set in a desert environment, this outdoor scene features cold color tones with very uniform natural lighting on the faces. The consistent and even natural light allows for the evaluation of the model's performance in depicting faces in a clear, daylight setting with minimal shadows or lighting variations, highlighting the texture and details effectively.

C Model Results

In this chapter, we present the detailed results obtained from our model on the evaluation dataset. The structure of the presented results mirrors the layout of our dataset examples to facilitate a clear and direct comparison.

For each example, the results are organized as follows: **First Row**: The complete raw scenic image containing the three synthesized heads with applied illumination and background. This row provides the original context for the evaluation. **Second Row**: Cropped images of each head from the full scene. These head crops allow for a closer inspection of the facial details and model performance. **Third Row**: The ground-truth masked albedo UV maps for each head. These serve as the reference for evaluating the accuracy of our model's albedo predictions. **Fourth Row**: The model-predicted albedo UV maps. These maps are generated by our model and are compared against the ground-truth albedo maps to assess performance. **Fifth Row**: The rendered reconstructed images. These images are synthesized based on the model's predictions and provide a visual representation of how well the model can recreate the original scene. **Sixth Row**: The illumination maps inferred from the illumination cues on the faces. These maps show the lighting conditions as interpreted by our model. Each column in the results corresponds to one of the three faces in the scene, allowing for an easy comparison across different heads within the same scene.

One significant observation is the consistency of the inferred environment map across all three heads in the same scene demonstrate the effectiveness of our model in maintaining environmental coherence. This consistency validates the robustness of our model in accurately synthesizing and reconstructing scenes with multiple faces under varying conditions.

These results provide comprehensive insights into the performance of our model, showcasing its ability to handle diverse lighting conditions and accurately reproduce facial details and environmental contexts.

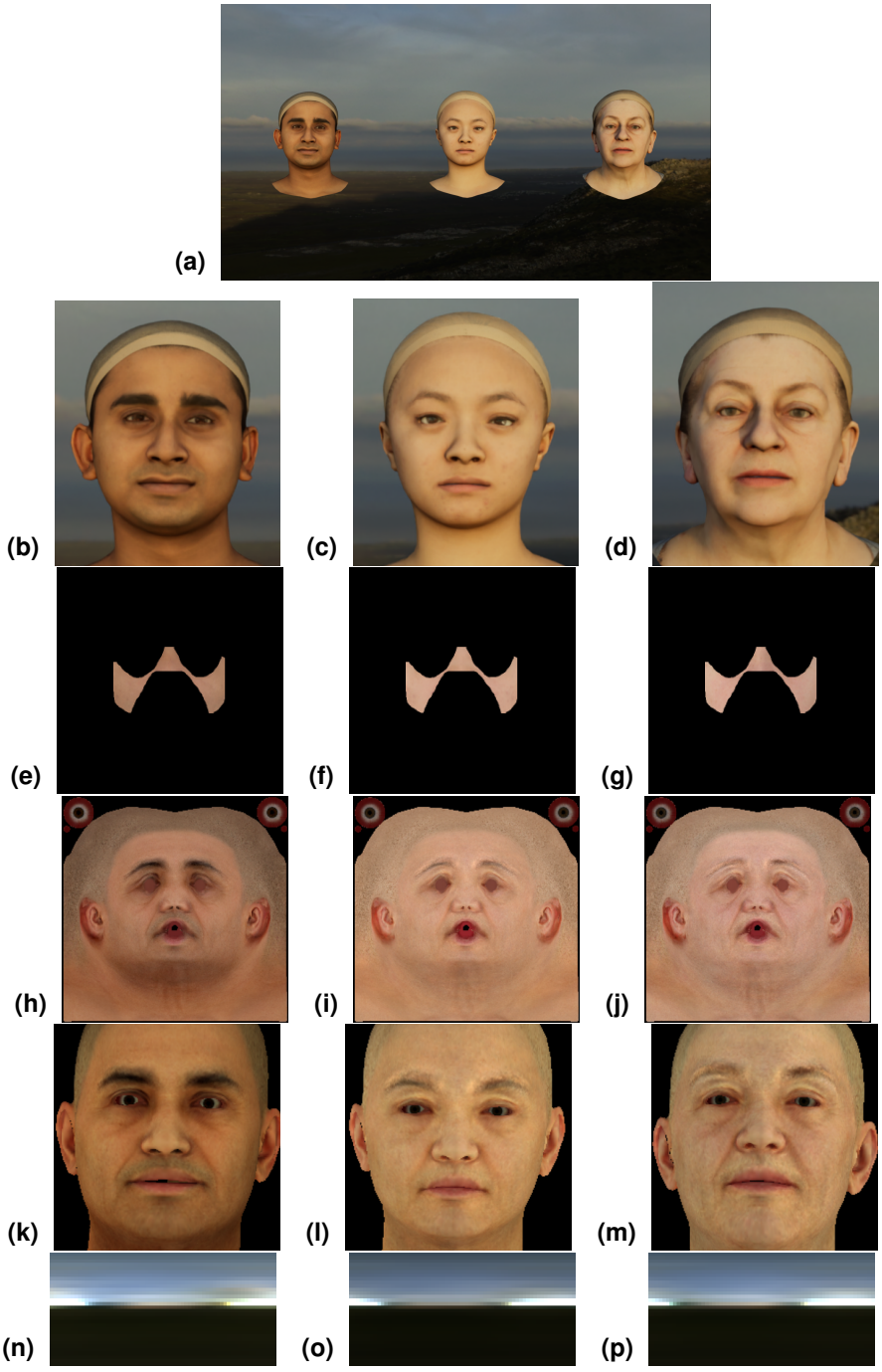


Figure C.1: The scene captures an outdoor environment during dawn, featuring a hilly backdrop. The three synthesized heads are uniformly illuminated by the early sunlight. The ITA errors for the faces are as follows: 0.85, 2.83, 0.72. This setting tests the model's ability to handle natural, soft lighting conditions and maintain consistency across multiple faces.

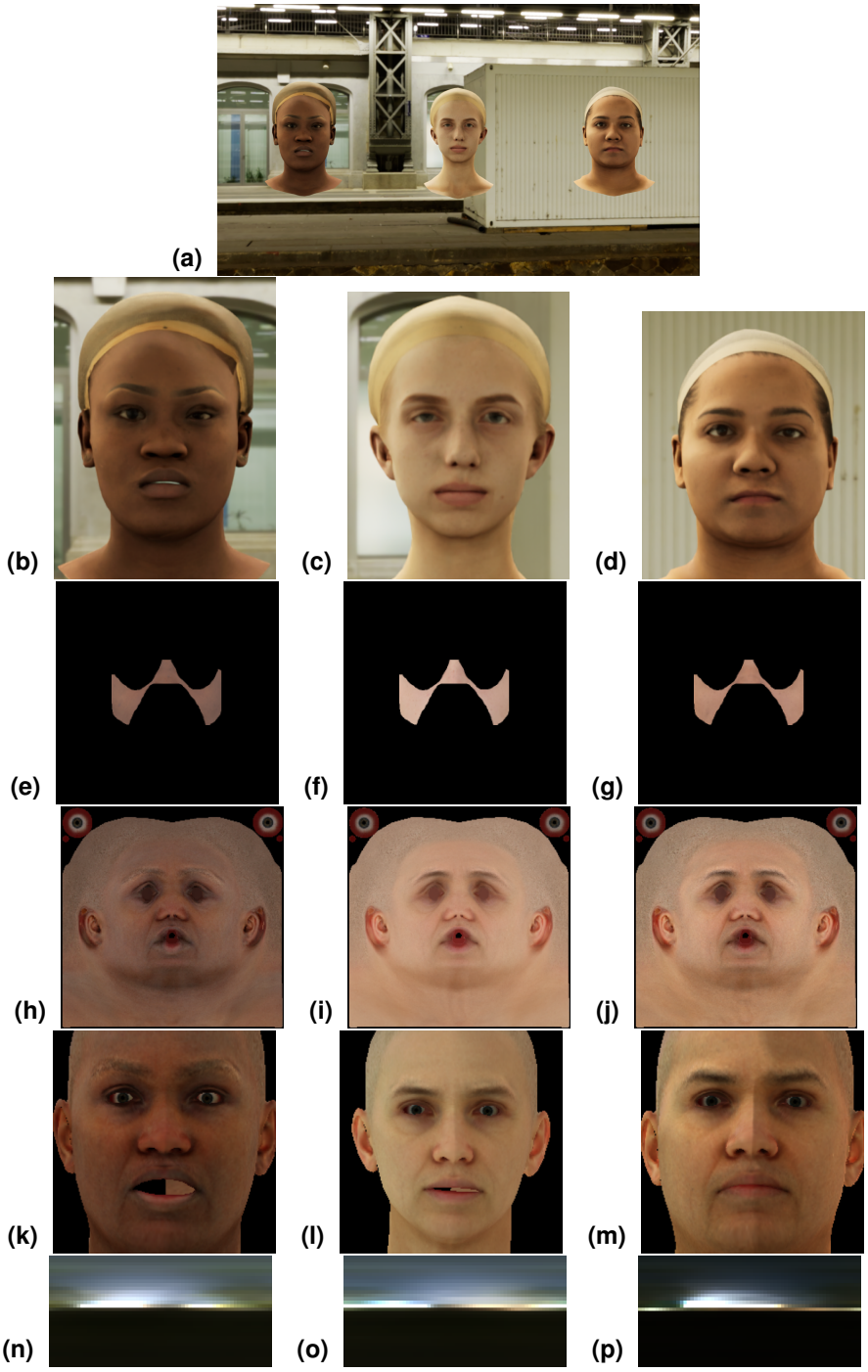


Figure C.2: This example depicts an indoor scene, possibly in a railway station or warehouse, illuminated by an overhead light tube. The lighting is artificial and uniform. The ITA errors for the faces are: 0.95, 11.86, 9.83. This scenario challenges the model's performance under consistent artificial lighting and highlights its robustness in controlled indoor environments.

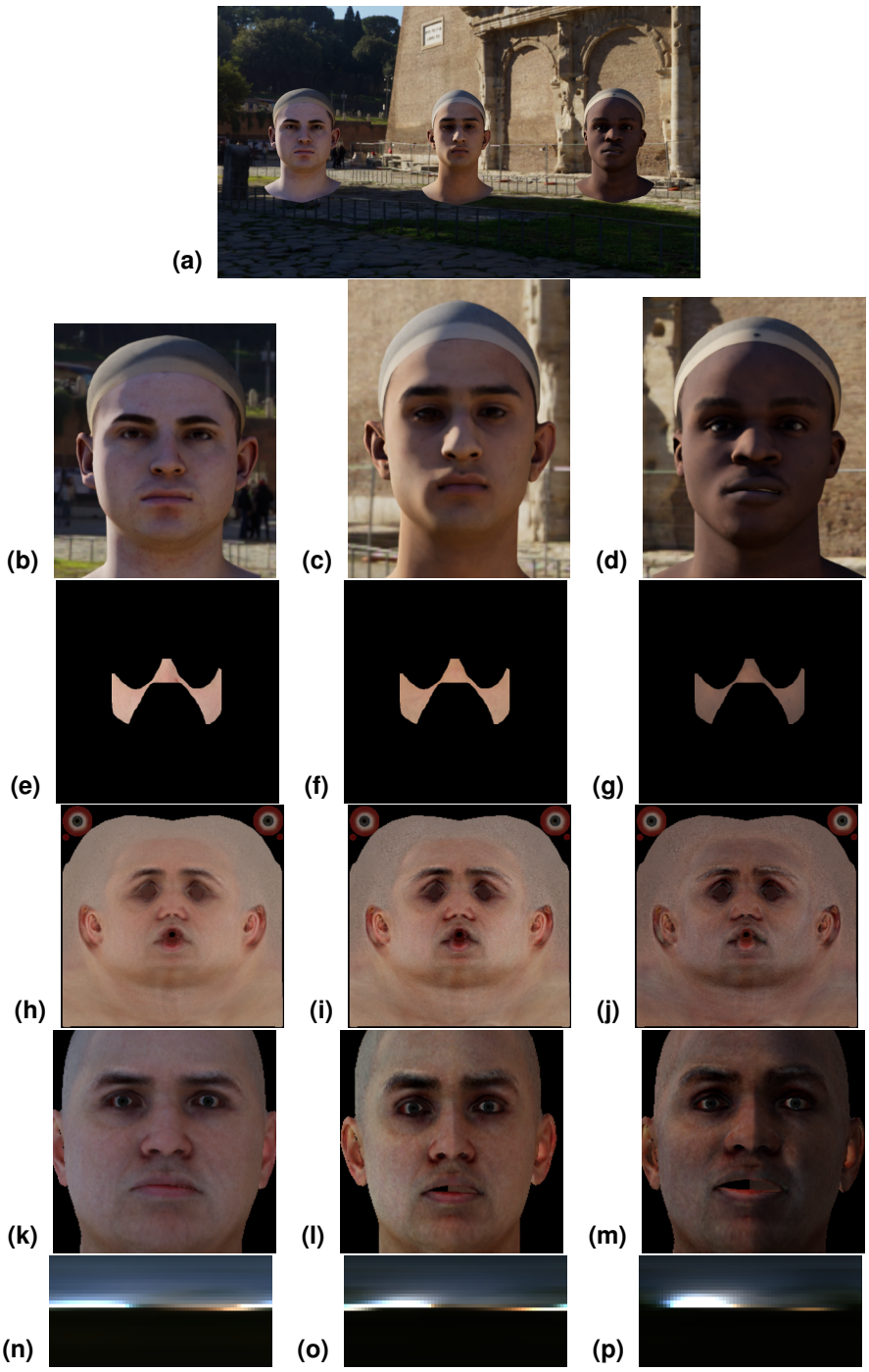


Figure C.3: The image shows an outdoor scene in the afternoon, where a wall's shadow covers the faces, with some sunlight still hitting them. The contrast between the shadowed and sunlit areas tests the model's handling of complex lighting conditions. The ITA errors for the faces are: 27.13, 10.09, 6.56. This setting evaluates the model's ability to manage sharp lighting contrasts and partial occlusions.

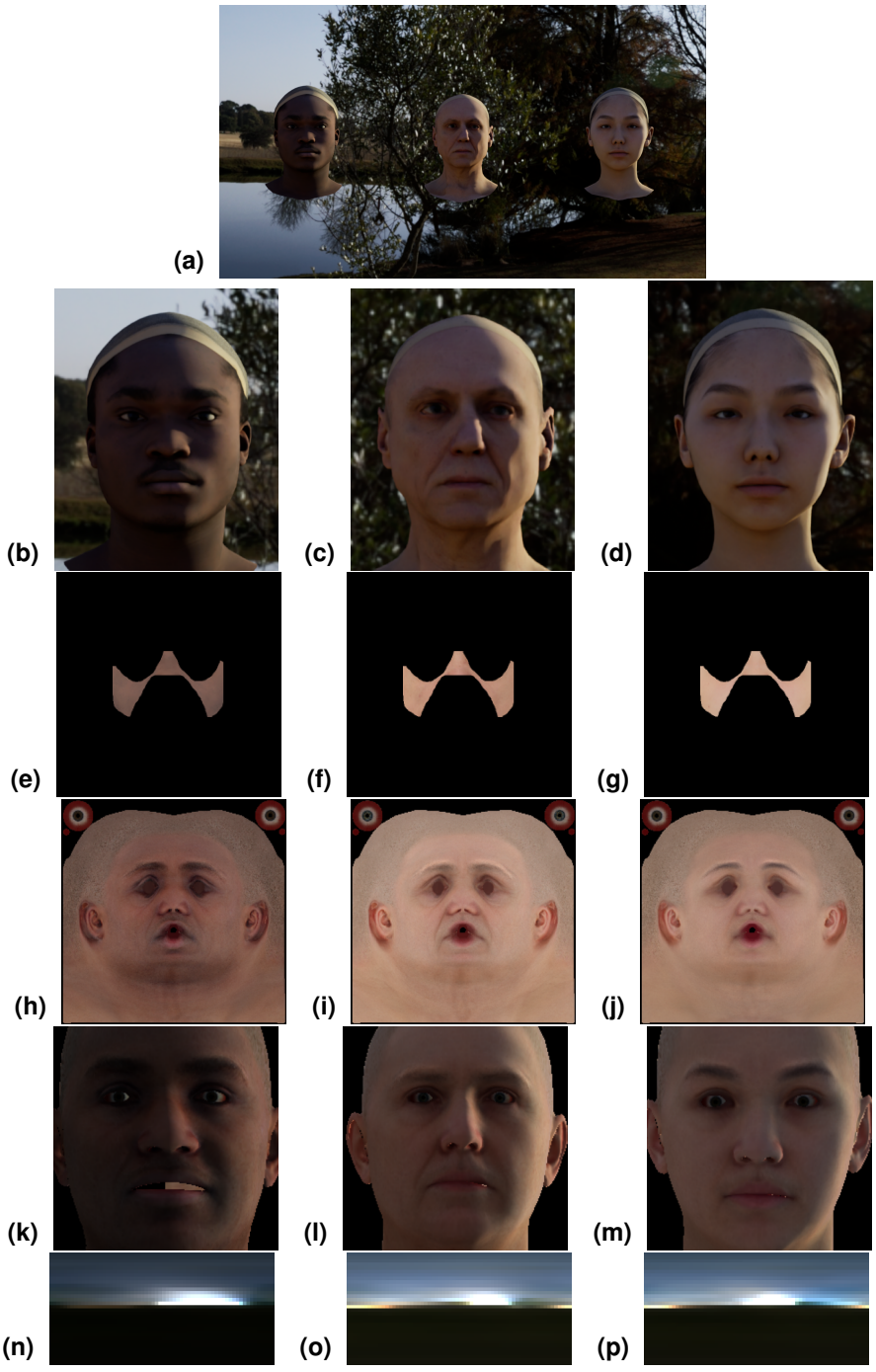


Figure C.4: This scene is set outdoors by a lake, with visible greenery in the background and dim lighting suggesting sunset. The soft, fading light tests the model's performance in low-light conditions. The ITA errors for the faces are: 10.57, 10.21, 10.44. This example highlights the model's capability to handle low-light environments and maintain facial detail accuracy.

D Model Comparison

In this chapter, we present a detailed comparison of our current model with two previous iterations. This comparison is crucial for understanding the improvements and enhancements achieved through successive model developments.

We compare the following three models: **BFM & Spherical Harmonics**: This model uses the Basel Face Model (BFM) combined with Spherical Harmonics (SH) for illumination. **BalancedAlb & Spherical Harmonics**: This intermediate model integrates the BalancedAlbedo (BalancedAlb) technique with Spherical Harmonics. **BalancedAlb & RENI++**: Our current model, which combines BalancedAlb with RENI++, representing the latest advancements in our research.

For each comparison example, the results are organized in the following rows: **First Row**: The input cropped image along with its true albedo UV map. This serves as the ground truth for evaluating the model's performance. **Second Row**: The predicted albedo UV maps generated by each of the three models. This row illustrates how each model interprets and estimates the albedo information from the input image. **Third Row**: The rasterized albedo maps. These maps provide a visual representation of the albedo as processed by the rasterization step, showing how each model translates the predicted albedo UV maps into a final texture map. **Fourth Row**: The rendered images with illumination. These images depict the final output after applying the predicted illumination, offering a comprehensive view of how each model reconstructs the face under varying lighting conditions.

BFM & SH, baseline model demonstrates significant variability in performance, especially across different skin types. Incorporating BalancedAlb reduces the bias score significantly, resulting in more consistent performance across diverse skin types. Our final model configuration further lowers the bias score, showcasing the enhanced stability and consistency achieved with RENI++.

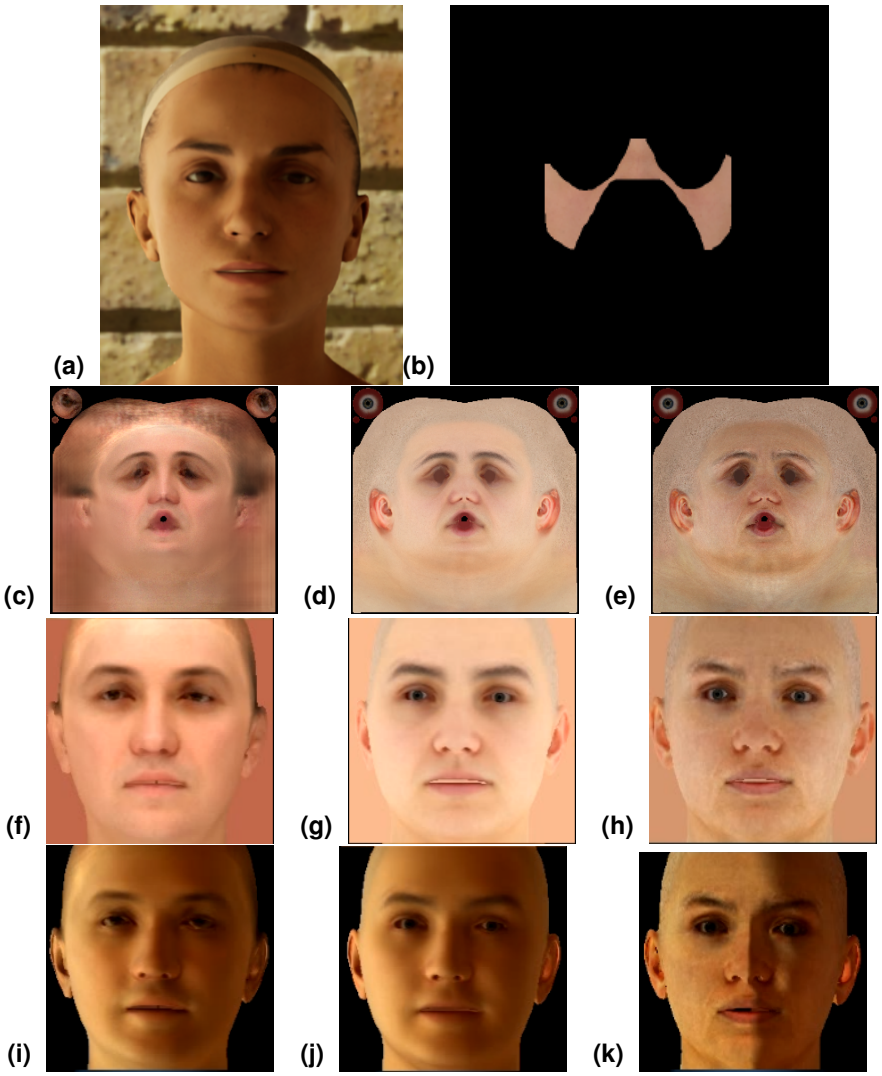


Figure D.1: True ITA: 20.48, Predicted ITA from left to right: 32.27 (BFM & SH), 52.82 (BalancedAlb & SH), and 32.07 (BalancedAlb & RENI++). The oldest model and the current model have similar predictions, whereas the intermediate model shows a significant deviation from the true ITA.

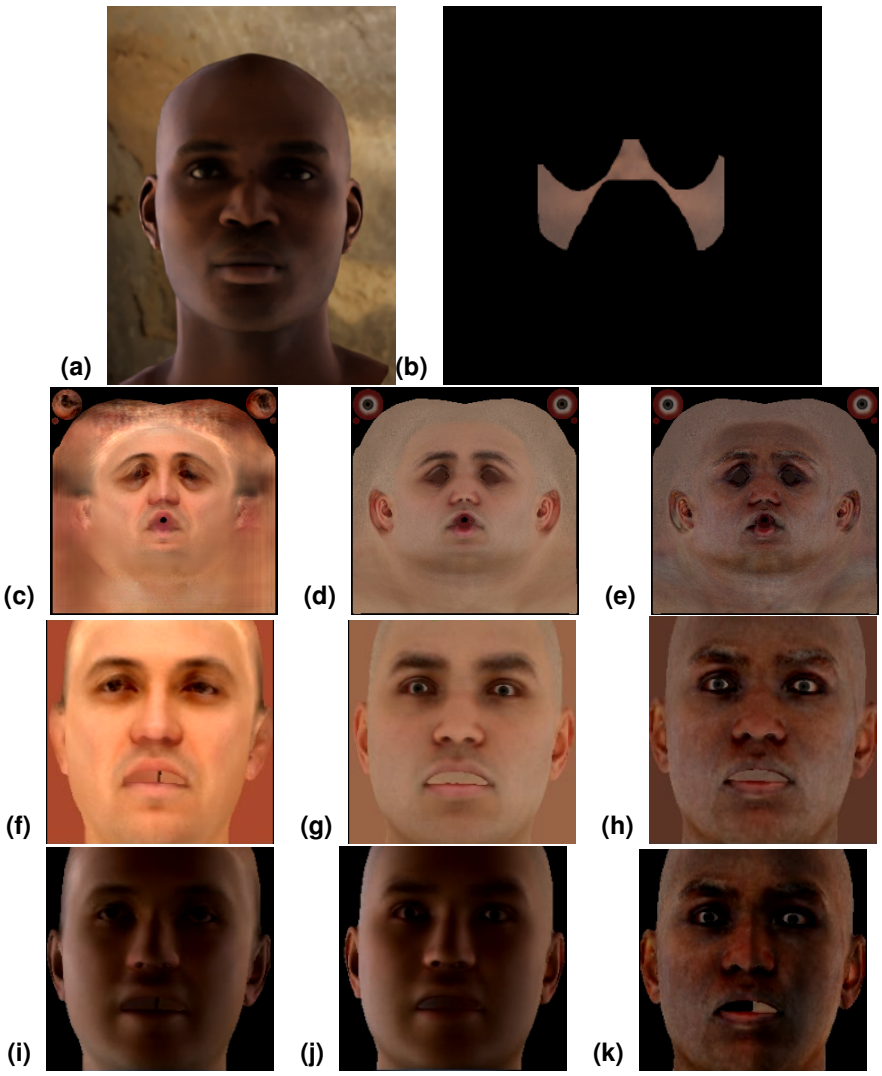


Figure D.2: True ITA: -32.85, Predicted ITA from left to right: 25.50 (BFM & SH), 5.87 (BalancedAlb & SH), and -47.07 (BalancedAlb & RENI++). In this example, the true ITA is negative, indicating darker skin tones. The oldest and intermediate models show a positive bias, while the current model provides a more accurate prediction, closely aligning with the true ITA.

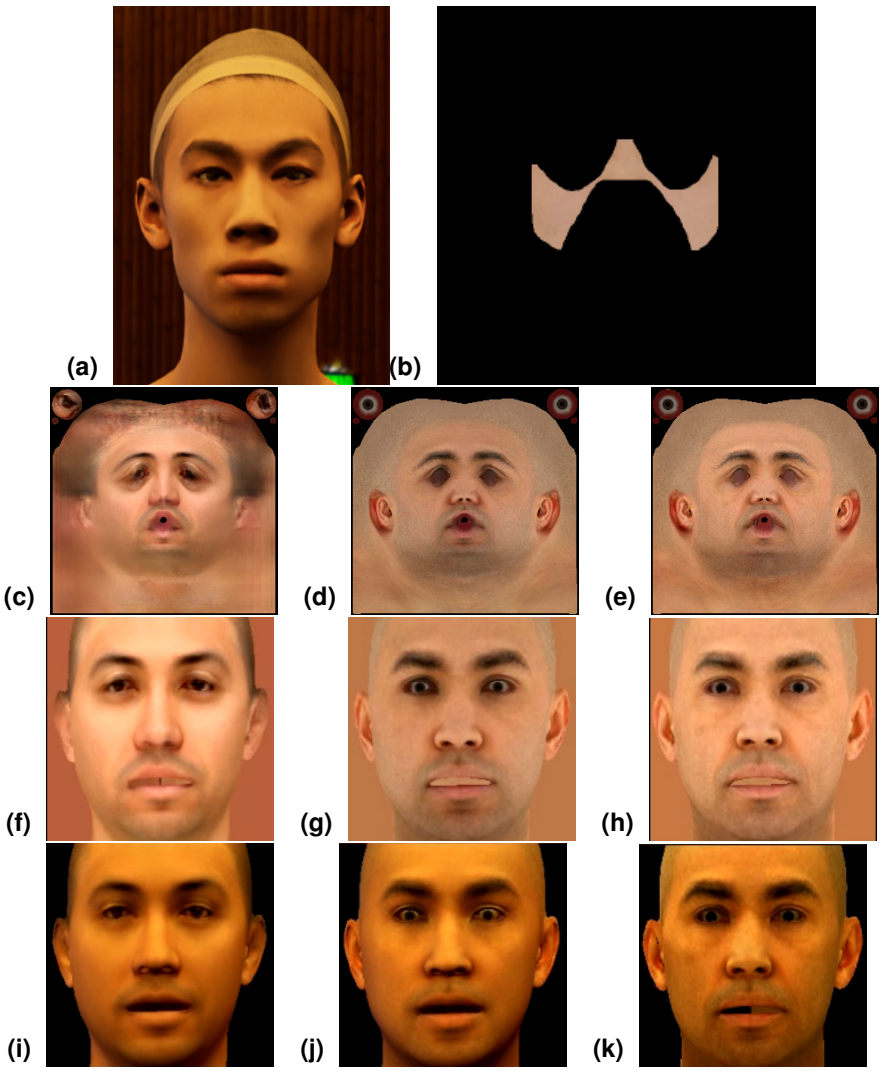


Figure D.3: True ITA: 31.61, Predicted ITA from left to right: 32.11 (BFM & SH), 24.85 (BalancedAlb & SH), and 30.79 (BalancedAlb & RENI++). This example shows a true ITA with a close match to the predictions of both the oldest and current models. The intermediate model, however, has a noticeable deviation.

Bibliography

[1] M. Zollhöfer, J. Thies, P. Garrido *et al.*, 'State of the art on monocular 3d face reconstruction, tracking, and applications,' *Computer Graphics Forum*, vol. 37, no. 2, pp. 523–550, 2018. DOI: https://doi.org/10.1111/cgf.13382. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13382. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13382.

[2] Z. Liu, Z. Zhang, C. Jacobs and M. Cohen, 'Rapid modeling of animated faces from video images,' in *Proceedings of the Eighth ACM International Conference on Multimedia*, ser. MULTIMEDIA '00, Marina del Rey, California, USA: Association for Computing Machinery, 2000, pp. 475–476, ISBN: 1581131984. DOI: 10.1145/354384.376389. [Online]. Available: https://doi.org/10.1145/354384.376389.

[3] V. Blanz and T. Vetter, 'A morphable model for the synthesis of 3d faces,' in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194, ISBN: 0201485605. DOI: 10.1145/311535.311556. [Online]. Available: https://doi.org/10.1145/311535.311556.

[4] E. Richardson, M. Sela and R. Kimmel, '3d face reconstruction by learning from synthetic data,' in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 460–469. DOI: 10.1109/3DV.2016.56.

[5] M. Sela, E. Richardson and R. Kimmel, 'Unrestricted facial geometry reconstruction using image-to-image translation,' in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1576–1585.

[6] A. Tewari, M. Zollhöfer, P. Garrido *et al.*, 'Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2549–2559.

[7] P. Dou, S. K. Shah and I. A. Kakadiaris, 'End-to-end 3d face reconstruction with deep neural networks,' in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5908–5917.

[8] R. Slossberg, G. Shamai and R. Kimmel, 'High quality facial surface and texture synthesis via generative adversarial networks,' *CoRR*, vol. abs/1808.08281, 2018. arXiv: 1808.08281. [Online]. Available: http://arxiv.org/abs/1808.08281.

[9] W. Zielonka, T. Bolkart and J. Thies, 'Towards metrical reconstruction of human faces,' in *European Conference on Computer Vision*, Springer, 2022, pp. 250–269.

[10] Y. Feng, H. Feng, M. J. Black and T. Bolkart, 'Learning an animatable detailed 3d face model from in-the-wild images,' *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[11] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge and A. K. Jain, 'Face recognition performance: Role of demographic information,' *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012. DOI: 10.1109/TIFS.2012.2214212.

[12] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt and C. Theobalt, 'Inversefacenet: Deep single-shot inverse face rendering from a single image,' *arXiv preprint arXiv:1703.10956*, 2017.

[13] S. Sengupta, A. Kanazawa, C. D. Castillo and D. W. Jacobs, 'Sfsnet: Learning shape, reflectance and illuminance of facesin the wild',' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6296–6305.

[14] C. Cao, Q. Hou and K. Zhou, 'Displaced dynamic expression regression for real-time facial tracking and animation,' *ACM Trans. Graph.*, vol. 33, no. 4, Jul. 2014, ISSN: 0730-0301. DOI: 10.1145/2601097.2601204. [Online]. Available: https://doi.org/10.1145/2601097.2601204.

[15] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer and C. Busch, 'Demographic bias in biometrics: A survey on an emerging challenge,' *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[16] M. Wang, W. Deng, J. Hu, X. Tao and Y. Huang, 'Racial faces in the wild: Reducing racial bias by information maximization adaptation network,' in *Proceedings of the ieee/cvf international conference on computer vision*, 2019, pp. 692–702.

[17] J. Buolamwini and T. Gebru, 'Gender shades: Intersectional accuracy disparities in commercial gender classification,' in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, PMLR, 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html.

[18] D. Leslie, 'Understanding bias in facial recognition technologies,' *arXiv e-prints*, arXiv:2010.07023, arXiv:2010.07023, Oct. 2020. DOI: 10.48550/arXiv.2010.07023. arXiv: 2010.07023 [cs.CY].

[19] P. Brey and B. Dainow, 'Ethics by design for artificial intelligence,' *AI and Ethics*, 2023, ISSN: 2730-5961. DOI: 10.1007/s43681-023-00330-4. [Online]. Available: https://doi.org/10.1007/s43681-023-00330-4.

[20] Research Ethics and Integrity Sector (European Commission), *Ethics by design and ethics of use approaches for artificial intelligence*, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf, Nov. 2021.

[21] M. Hildebrandt, 'Algorithmic regulation and the rule of law,' *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20 170 355, 2018. DOI: 10.1098/rsta.2017.0355. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2017.0355. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0355.

[22] P. Huber, 'Real-time 3d morphable shape model fitting to monocular in-the-wild videos.,' Ph.D. dissertation, University of Surrey, 2017.

[23] B. Amberg, R. Knothe and T. Vetter, 'Expression invariant 3d face recognition with a morphable model,' in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1–6. DOI: 10.1109/AFGR.2008.4813376.

[24] T. Li, T. Bolkart, M. J. Black, H. Li and J. Romero, 'Learning a model of facial shape and expression from 4d scans,' *ACM Trans. Graph.*, vol. 36, no. 6, Nov. 2017, ISSN: 0730-0301. DOI: 10.1145/3130800.3130813. [Online]. Available: https://doi.org/10.1145/3130800.3130813.

[25] H. Li, J. Yu, Y. Ye and C. Bregler, 'Realtime facial animation with on-the-fly correctives,' *ACM Trans. Graph.*, vol. 32, no. 4, Jul. 2013, ISSN: 0730-0301. DOI: 10.1145/2461912.2462019. [Online]. Available: https://doi.org/10.1145/2461912.2462019.

[26] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger and M. Niessner, 'Headon: Real-time reenactment of human portrait videos,' *ACM Trans. Graph.*, vol. 37, no. 4, Jul. 2018, ISSN: 0730-0301. DOI: 10.1145/3197517.3201350. [Online]. Available: https://doi.org/10.1145/3197517.3201350.

[27] O. Alexander, M. Rogers, W. Lambeth, M. Chiang and P. Debevec, 'The digital emily project: Photoreal facial modeling and animation,' in *ACM SIGGRAPH 2009 Courses*, ser. SIGGRAPH '09, New Orleans, Louisiana: Association for Computing Machinery, 2009, ISBN: 9781450379380. DOI: 10.1145/1667239.1667251. [Online]. Available: https://doi.org/10.1145/1667239.1667251.

[28] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah and D. Dunaway, 'A 3d morphable model learnt from 10,000 faces,' in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5543–5552. DOI: 10.1109/CVPR.2016.598.

[29] J. Booth, A. Roussos, A. Ponniah, D. Dunaway and S. Zafeiriou, 'Large scale 3d morphable models,' English, *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 8th Apr. 2017, ISSN: 0920-5691. DOI: 10.1007/s11263-017-1009-7. [Online]. Available: https://doi.org/10.1007/s11263-017-1009-7.

[30] B. Gecer, S. Ploumpis, I. Kotsia and S. Zafeiriou, 'Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1155–1164.

[31] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster and T. Vetter, 'Analyzing and reducing the damage of dataset bias to face recognition with synthetic data,' in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2261–2268. DOI: 10.1109/CVPRW.2019.00279.

[32] A. Lattas, S. Moschoglou, B. Gecer *et al.*, 'Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild",' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 760–769.

[33] W. A. Smith, A. Seck, H. Dee, B. Tiddeman, J. B. Tenenbaum and B. Egger, 'A morphable face albedo model,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5011–5020.

[34] M. M. Loper and M. J. Black, 'OpenDR: An approximate differentiable renderer,' in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8695, Springer International Publishing, Sep. 2014, pp. 154–169. DOI: 10.1007/978-3-319-10584-0_11.

[35] R. Ramamoorthi and P. Hanrahan, 'An efficient representation for irradiance environment maps,' *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2001)*, vol. 20, no. 3, pp. 497–500, 2001. [Online]. Available: http://graphics.cs.berkeley.edu/papers/Ramamoorthi-ERI-2001-00/.

[36] H. Kato, D. Beker, M. Morariu *et al.*, 'Differentiable rendering: A survey,' *CoRR*, vol. abs/2006.12057, 2020. arXiv: 2006.12057. [Online]. Available: https://arxiv.org/abs/2006.12057.

[37] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman and J. T. Barron, 'Nerfactor: Neural factorization of shape and reflectance under an unknown illumination,' *ACM Transactions on Graphics (ToG)*, vol. 40, no. 6, pp. 1–18, 2021.

[38] A. Bas, W. A. Smith, T. Bolkart and S. Wuhrer, 'Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences,' in *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, Springer, 2017, pp. 377–391.

[39] V. Blanz, A. Mehl, T. Vetter and H.-P. Seidel, 'A statistical method for robust 3d surface reconstruction from sparse data,' in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, 2004, pp. 293–300. DOI: 10.1109/TDPVT.2004.1335212.

[40] O. Aldrian and W. A. Smith, 'Inverse rendering of faces with a 3d morphable model,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1080–1093, May 2013, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2012.206.

[41] P. Huber., G. Hu., R. Tena. *et al.*, 'A multiresolution 3d morphable face model and fitting framework,' in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, INSTICC, SciTePress, 2016, pp. 79–86, ISBN: 978-989-758-175-5. DOI: 10.5220/0005669500790086.

[42] C. Cao, Y. Weng, S. Lin and K. Zhou, '3d shape regression for real-time facial animation,' *ACM Trans. Graph.*, vol. 32, no. 4, Jul. 2013, ISSN: 0730-0301. DOI: 10.1145/2461912.2462012. [Online]. Available: https://doi.org/10.1145/2461912.2462012.

[43] A. Bas and W. A. Smith, 'What does 2d geometric information really tell us about 3d face shape?' *International Journal of Computer Vision*, vol. 127, pp. 1455–1473, 2019.

[44] W. A. Smith, 'The perspective face shape ambiguity,' in *Perspectives in shape analysis*, Springer, 2016, pp. 299–319.

[45] B. T. Phong, 'Illumination for computer generated pictures,' *Commun. ACM*, vol. 18, no. 6, pp. 311–317, Jun. 1975, ISSN: 0001-0782. DOI: 10.1145/360825.360839. [Online]. Available: https://doi.org/10.1145/360825.360839.

[46] J. F. Blinn, 'Models of light reflection for computer synthesized pictures,' *SIGGRAPH Comput. Graph.*, vol. 11, no. 2, pp. 192–198, Jul. 1977, ISSN: 0097-8930. DOI: 10.1145/965141.563893. [Online]. Available: https://doi.org/10.1145/965141.563893.

[47] D. P. Greenberg, M. F. Cohen and K. E. Torrance, 'Radiosity: A method for computing global illumination,' *The Visual Computer*, vol. 2, no. 5, pp. 291–297, 1st Sep. 1986, ISSN: 1432-2315. DOI: 10.1007/BF02020429. [Online]. Available: https://doi.org/10.1007/BF02020429.

[48] T. Whitted, 'An improved illumination model for shaded display,' *Commun. ACM*, vol. 23, no. 6, pp. 343–349, Jun. 1980, ISSN: 0001-0782. DOI: 10.1145/358876.358882. [Online]. Available: https://doi.org/10.1145/358876.358882.

[49] R. Green, 'Spherical harmonic lighting: The gritty details,' in *Archives of the game developers conference*, vol. 56, 2003, p. 4.

[50] P.-P. Sloan, J. Kautz and J. Snyder, 'Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments,' in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2023, ISBN: 9798400708978. [Online]. Available: https://doi.org/10.1145/3596711.3596749.

[51] J. T. Barron and J. Malik, 'Color constancy, intrinsic images, and shape estimation,' in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato and C. Schmid, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 57–70, ISBN: 978-3-642-33765-9.

[52] S. Song and T. Funkhouser, 'Neural illumination: Lighting prediction for indoor environments,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6918–6926.

[53] L. Wang and K.-J. Yoon, 'Deep learning for hdr imaging: State-of-the-art and future trends,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8874–8895, 2022. DOI: 10.1109/TPAMI.2021.3123686.

[54] Z. Wang, J. Philion, S. Fidler and J. Kautz, 'Learning indoor inverse rendering with 3d spatially-varying lighting,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 538–12 547.

[55] J. Gardner, B. Egger and W. Smith, 'Rotation-equivariant conditional spherical neural fields for learning a natural illumination prior,' *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 309–26 323, 2022.

[56] J. A. Gardner, B. Egger and W. A. Smith, 'Reni++ a rotation-equivariant, scale-invariant, natural illumination prior,' *arXiv preprint arXiv:2311.09361*, 2023.

[57] B. Egger, S. Schönborn, A. Schneider *et al.*, 'Occlusion-aware 3d morphable models and an illumination prior for face image analysis,' *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1269–1287, 1st Dec. 2018, ISSN: 1573-1405. DOI: 10.1007/s11263-018-1064-8. [Online]. Available: https://doi.org/10.1007/s11263-018-1064-8.

[58] H. Feng, T. Bolkart, J. Tesch, M. J. Black and V. Abrevaya, 'Towards racially unbiased skin tone estimation via scene disambiguation,' in *European Conference on Computer Vision*, Springer, 2022, pp. 72–90.

[59] A. CHARDON, I. CRETOIS and C. HOURSEAU, 'Skin colour typology and suntanning pathways,' *International Journal of Cosmetic Science*, vol. 13, no. 4, pp. 191–208, 1991. DOI: https://doi.org/10.1111/j.1467-2494.1991.tb00561.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-2494.1991.tb00561.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-2494.1991.tb00561.x.

[60] M. Merler, N. K. Ratha, R. S. Feris and J. R. Smith, 'Diversity in faces,' *CoRR*, vol. abs/1901.10436, 2019. arXiv: 1901.10436. [Online]. Available: http://arxiv.org/abs/1901.10436.

[61] S. Da'Prato-Shepard, 'Reducing racial bias for in-the-wild 3d face reconstruction using a learned illumination prior,' Submitted in part fulfilment for the degree of MMath in Mathematics and Computer Science., MMath Thesis, Department of Computer Science, Department of Computer Science, May 2023.

[62] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang and S. Z. Li, 'S3fd: Single shot scale-invariant face detector,' in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 192–201.

[63] A. Bulat and G. Tzimiropoulos, 'How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks),' in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

[64] J. Hosang, R. Benenson and B. Schiele, 'Learning non-maximum suppression,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4507–4515.

[65] A. Newell, K. Yang and J. Deng, 'Stacked hourglass networks for human pose estimation,' in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, Springer, 2016, pp. 483–499.

[66] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu and N. Sang, 'Bisenet: Bilateral segmentation network for real-time semantic segmentation,' in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[67] C.-H. Lee, Z. Liu, L. Wu and P. Luo, 'Maskgan: Towards diverse and interactive facial image manipulation,' in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[68] J. Shang, T. Shen, S. Li *et al.*, 'Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency,' in *European Conference on Computer Vision*, Springer, 2020, pp. 53–70.

[69] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia and X. Tong, 'Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[70] Z. Bai, Z. Cui, X. Liu and P. Tan, 'Riggable 3d face reconstruction via in-network optimization,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6216–6225.

[71] P. Paysan, R. Knothe, B. Amberg, S. Romdhani and T. 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*, 2009, pp. 296–301. DOI: 10.1109/AVSS.2009.58.

[72] Y. Wen, W. Liu, B. Raj and R. Singh, 'Self-supervised 3d face reconstruction via conditional estimation,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 289–13 298.

[73] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte and L. Van Gool, 'Repaint: Inpainting using denoising diffusion probabilistic models,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.