Inclusive Portrait Lighting Estimation Model Leveraging Graphic-Based Synthetic Data

Kin Ching Lydia Chau[1] Tao Li[2] Ruowei Jiang[1] Zhi Yu[1] Panagiotis-Alexandros Bokaris[2]
[1]ModiFace Inc. [2]L'Oréal Research & Innovation

Abstract

We propose a synthetic data-based training framework for real-time deep learning models that predict an omnidirectional high-dynamic-range (HDR) environment light map from a single limited field-of-view, low-dynamic-range portrait image. Training lighting estimation models requires paired data of portrait images and the corresponding environment maps. Previous research generates the data by utilizing relightable real-face datasets collected in specialized light stages, and then relighting faces using HDR environment maps. This process is costly and time-consuming, and consequently, the datasets often cover a limited number of subjects and are prone to demographic bias. On the other hand, recent developments in graphic-based synthetic portrait images based on combining a parametric 3D face model with a comprehensive collection of hand-crafted assets, such as skin, hair, and clothing, have shown great advancement in photorealism. Leveraging the ease of collecting diverse synthetic data, we explore their potential in the domain of portrait lighting estimation. Our training framework involves pre-training on synthetic labeled data and fine-tuning on unlabeled real portrait videos. Our model achieves state-of-the-art performance based on the zero-shot evaluation result on the real portrait image benchmark dataset. Furthermore, we conduct a fairness analysis, showing that our model is more robust to demographic differences than the existing state-of-the-art models.

1. Introduction

Understanding environment illumination is crucial in various applications, such as portrait relighting [12, 27], scene simulation [15, 22] and 3D construction [14]. Considering the diverse user base of these applications, it is important to ensure inclusivity and avoid bias towards specific groups. While previous studies have addressed inclusivity in biometric systems [4], face recognition [20], and skin tone recognition [6], the exploration of inclusivity in lighting estimation remains unexplored. To mitigate biometric bias, collecting data from diverse demographic groups is a natural solution. However, collecting data for lighting estimation models is a costly and time-consuming task, as it involves obtaining paired data of portrait images and corresponding light maps.

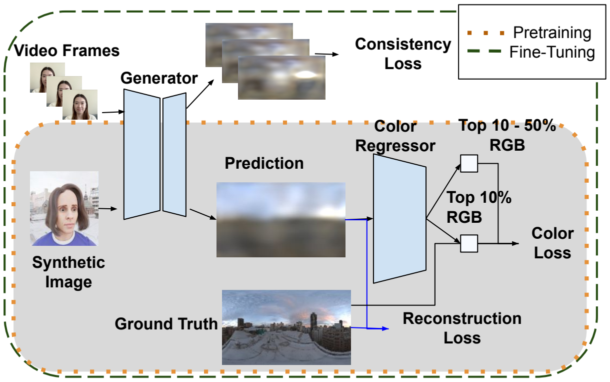

Figure 1. **Model Architecture.** Our approach involves a two-step training process. We pre-train our model using synthetic data with reconstruction loss and color loss. During fine-tuning, we adapt it to the real image domain by introducing a temporal consistency loss based on unlabeled real video frames.

To create a diverse dataset, the Laval Face+Lighting HDR dataset [2] is created through conducting photo sessions involving 9 subjects in 25 distinct lighting environments. To further enhance the variety of lighting and subjects, subsequent researches [5, 12, 13, 19] propose to decouple the collection process of portrait images from that of environment maps. LeGender et al. [12, 13] generates a similar dataset by separately collecting HDR environment maps and recording the reflectance field and alpha matte of 70 diverse subjects using a light stage [3]. They then combine these elements to create paired data using image-based relighting. Similarly, using HDR maps, Sztrajman et al. [19] relights scanned faces from the ICT 3D Relightable Facial Expression Database [19] and Fei et al. [5] relights scanned faces from FaceScape dataset [25] to create the paired data. Due to the requirements of light stages to record facial texture or reflectance information, these datasets only cover a limited number of subjects, potentially leading to models over-fitting faces from certain demographic groups.

Inspired by recent advancements in the photorealism of synthetic face images based on graphics and 3D face models [23], as well as the successful utilization of these synthetic data for training 3D localization tasks [24], we investigate the viability of training our model using these data. Our objective is to achieve performance levels on par with real labeled data and enhance the model's ability to handle subjects from diverse demographics. The main contributions of our study are as follows.

- We propose a training pipeline that first pre-trains the model on graphic-based labeled synthetic data and fine-tunes it on unlabeled videos.
- In contrast to state-of-the-art methods, we show that our model well achieves fairness across gender and ethnicity groups.
- We demonstrate our network's superior performance on both synthetic and real data compared to state-of-the-art methods.

2. Related work

2.1. Lighting Estimation

To achieve a realistic rendering of objects in augmented reality (AR), accurate lighting estimation is crucial. Gardner et al. [7] first introduce an end-to-end deep neural network that directly regresses key lighting locations and intensities from a limited field-of-view photo, without relying on strong assumptions on scene geometry, material properties, or lighting. Subsequent research in this field can be broadly categorized into two approaches: regression models and generative models. The regression models infer low-dimensional geometric and photometric parameters such as sky parameters [11], needlets coefficients [26], and low-dimensional spherical harmonic representations [9, 17, 28]. On the other hand, the generative models directly generate environmental maps [18, 21] or light probes [5]. The regression models have less flexibility and less detailed lighting information, and require additional post-processing steps to convert the lighting parameters to HDR environment maps for image-based relighting. As a result, the generative model approach is a more popular option for mobile or web AR applications [12, 18]. Therefore our model also directly infers the environment maps.

2.2. Graphic-Based Synthetic Face Data

Recently, there has been a growing interest in utilizing graphic-based synthetic face data for training models in facial landmarks and face parsing tasks [23]. These synthetic face data are generated by integrating a parametric 3D face model [1] with a wide range of hand-crafted assets, including skin texture, hair, and clothing, and then relit with HDR maps and rendered using a photo-realistic ray-tracing renderer [23]. Using these graphic-based data offers a high level of photorealism, and studies have shown that models trained on such data have achieved state-of-the-art results in monocular 3D face reconstruction [24]. The graphic-based synthetic data reduces the complexity and cost of collecting lighting data for portrait images and enables the creation of large datasets with diverse subjects and environment maps. Leveraging graphic-based face data, we acquire a comprehensive dataset from the Datagen Platform[1] that consists of over 5,000 actors and 260 environment maps.

2.3. Synthetic and Real Domain Adaption

The existence of a domain-adaptation gap between real and synthetic data is a well-known challenge. Previous studies have addressed this issue by employing generative adversarial networks (GAN) to enhance the realism of the synthetic portrait images [28] or the predicted maps [5]. These methods improve the realism of the predicted maps by encouraging the maps to have high-frequency details. However, since environment maps have a long-tail color intensity distribution [21] and most public datasets[23] only have a few hundreds of indoor and outdoor maps, these methods often result in unstable predictions. In the context of real-time AR lighting estimation models, temporal consistency is an important concern, and flickering in lighting can significantly impact the overall AR experience. Therefore, instead of the GAN approach, we impose a temporal consistent loss based on video data to encourage smooth transitions among the predicted maps of consecutive video frames, thereby enhancing the overall quality and stability of the model.

3. Methods

As illustrated in Fig. 1, we first pre-train our model on synthetic data and then fine-tune the model with synthetic data and unlabelled real video frames.

3.1. Synthetic Training Data

We acquired a diverse synthetic dataset using the Datagen platform. The dataset consists of over 60,000 paired data, comprising 5,000 synthetic actors and over 260 maps. The actors are sampled with an equal distribution of ethnicities (African, Hispanic, Mediterranean, North European, South Asian, Southeast Asian), ages (young, adult, elderly), and genders. We randomize actors' head orientations and add a variety of attributes to them, such as beard, hairstyle, and clothing. For the environment maps, we employ a wide range of HDR lighting maps, representing various times of

[1] https://datagen.tech
[2] https://polyhaven.com
[3] https://hdri-haven.com

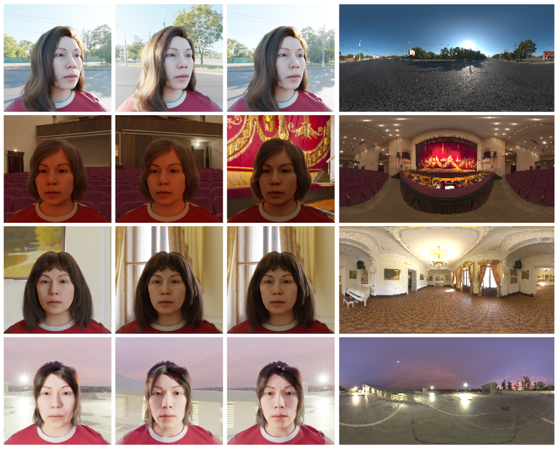

Figure 2. **Illustration of synthetic data.** For each actor, we randomly sample 4 environment maps (right column) and generate 12 images by incorporating different hairstyles, head poses, and other attributes.

day (morning, day, evening, or night) and settings (indoor or outdoor). For each actor, we utilize 4 different HDR maps and 3 orientations per map as illustrated in Fig. 2.

To prioritize the learning of light maps, we crop and center-align heads in the portrait images, and rotate and shift the environment maps to align their centers with the locations of the actors' faces in the portrait images. To introduce more variety in the maps, we incorporate random horizontal flip, exposure, and white balance augmentations.

3.2. Real Portrait Video Data

Our video dataset comprises 4000 portrait videos featuring over 1000 subjects from diverse demographics and lighting environments. Each video maintains a fixed camera position in a stable environment, with the subject positioned at the center, exhibiting different head poses and expressions. By assuming that the environments remain unchanged within short time intervals, the predicted maps of consecutive frames should be invariant to subjects' head poses and expressions.

3.3. Loss Functions

As shown in Fig. 1, we pre-train the generator (G) and the color regressor (D) simultaneously using synthetic data. The total pre-taining loss is the sum of the generator reconstruction, generator color and regressor losses.

$$L_{Pretraining} = L_{Reconstruction} + L_{Color} + L_D \qquad (1)$$

During fine-tuning, we freeze the light color regressor and train the generator using synthetic data and real videos. The total fine-tuning loss is the sum of the reconstruction, color and consistency losses.

$$L_{Finetuning} = L_{Reconstruction} + L_{Color} + L_{Cons} \qquad (2)$$

Generator Reconstruction Loss ($L_{Reconstruction}$) Given an input image x, the reconstruction loss is calculated as a pixel-wise L2 difference between the predicted map $G(x)$ and ground truth map y.

Generator Color Loss (L_{Color}) Pixels with higher intensity in light maps have more impact on rendering as they reveal information regarding key light sources. To account for this, a color loss is added to minimize the average color difference between stronger light sources in the predicted and ground truth maps. Prior research works [8, 21] have demonstrated the effectiveness by utilizing a pixel-wise loss on the top 5-10% highest intensity pixels in the maps. Based on our experiment, the model learns better when this loss is computed using a color regression model (D).

We use LAB color when determining the top k% highest intensity pixels in map y, $M_k(y)$ is defined as the set of pixels P with values greater than $1 - k\%$ percentile in channel L when converting the map to be in LAB color space. We define M_{10} and M_{50} of map y as the following:

$$\begin{aligned} M_{50}\,(y) &= \{P_{RGB} | P_{LAB} \in \{Top\,10\%{-}50\%\,L\,value\}\} \\ M_{10}\,(y) &= \{P_{RGB} | P_{LAB} \in \{Top\,10\%\;L\;value\}\} \end{aligned} \qquad (3)$$

The color regressor is trained in parallel with the generator. It predicts the mean RGB values of the top 10% highest intensity pixels ($M_{10}(y)$) and mean RGB of the top 10%-50% highest intensity pixels ($M_{50}(y)$) from the input maps.

$$L_D = \sum_{k \in \{10,50\}} \mathbb{E}_y \left[\| D_k(y) - \overline{M_k(y)} \|_1 \right] \qquad (4)$$

The light source color loss for the generator is computed as the L1 distance loss between the RGB outputs from the regression model and the average top 10% and 10-50% RGB of the ground truth map.

$$L_{Color} = \sum_{k \in \{10,50\}} \mathbb{E}_x \left[\| D_k(G(x)) - \overline{M_k(y)} \|_1 \right] \qquad (5)$$

Video Frame Consistency Loss (L_{cons}) During fine-tuning, a self-supervised consistency loss is introduced to ensure temporal coherence. Three consecutive video frames (v_t, v_{t+1} and v_{t+2}) are fed into the generator and a pixel-wise L1 distance loss is applied between the predicted map of v_{t+1} and the mix-up of predictions of v_t and v_{t+2}.

$$L_{Cons} = \mathbb{E}_v \left[\left\| (G(v_{t+1}) - \frac{G(v_t) + G(v_{t+2})}{2} \right\|_1 \right] \qquad (6)$$

3.4. Model Architecture

The generator employs an encoder-decoder architecture. It utilizes the first 17 layers of MobileNetv2 [16] to extract features, 5 up-sampling blocks and a convolution output

	Env. Map ↓		Specular ↓		Diffuse ↓	
Model	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor
Zhu et al. [29]	-	-	0.06	0.55	0.119	0.139
Sztrajman et al. [19]	-	0.318	-	0.034	-	0.145
Legendre et al. [13]	0.307	0.258	0.043	0.042	0.121	0.135
Fei et al. [5]	0.268	0.180	0.029	0.018	0.073	**0.069**
Our Pre-trained w/o col loss	0.235	0.171	0.017	**0.015**	0.072	0.071
Our Pre-trained	**0.231**	0.171	0.017	**0.015**	0.067	0.072
Our Fine-tuned	**0.231**	**0.170**	**0.016**	**0.015**	**0.065**	0.071

Table 1. Comparison of siRMSE on LIGHTTEST dataset [5]

layer to generate the maps. Each up-sampling block comprises a convolution layer, a ReLu activation layer, and a bilinear up-sampling layer. Considering the long-tailed intensity distribution of HDR environment maps, the generator predicts maps on a logarithmic scale.

The color regression model consists of six convolution layers and each convolution layer is followed by a Leaky ReLu activate layer. Two fully connected output heads are used to predict the average RGB of the top 10% and top 10-50% highest intensity pixels. See more implementation details in the supplement material.

4. Results and Fairness Study

We benchmark our model against the state-of-the-art models [5, 13, 19, 29] on the LIGHTTEST dataset [5]. The SOTA models are either trained with real portrait images or relit real faces. The LIGHTTEST dataset comprises paired data of real portrait images and environment maps and includes the data from the Laval Face+Light dataset [2]. The subjects in the dataset exhibit a variety of skin tones, and the maps capture both indoor and outdoor backgrounds.

We report the scale-invariant root mean squared error (siRMSE) on all methods in Tab. 1, which is a commonly used evaluation metric for lighting estimation models. The metric is computed directly on the environment maps, and after rendering the maps on a specular sphere and a diffuse sphere. A more detailed ablation study can be found in the supplementary material.

To examine the inclusivity of our model, we conduct a fairness analysis using indoor synthetic data. Synthetic data is used since the real samples in the existing public datasets are predominantly composed of Caucasian male, or Southeast Asian individuals, with limited samples from African, Hispanic, South Asian, or female subjects. As shown in Fig. 3 and Fig. 4, the results show that our model which is trained with a diverse dataset, exhibits minimal performance variation quantitatively and qualitatively across different genders and ethnicities in comparison with SPLiT. A more comprehensive fairness analysis result can be found in supplementary material.

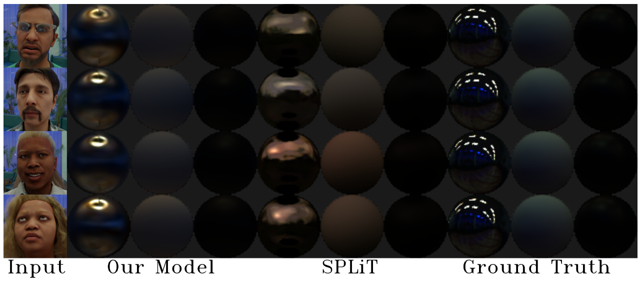

Figure 3. To examine if models are invariant to demographic variation, we utilize synthetic images featuring subjects from diverse ethnicities and genders placed in the same background. We render the predictions on the mirror, specular, and diffused spheres. Predictions of images from the same background should be consistent, that is, the spheres in the same column look identical.

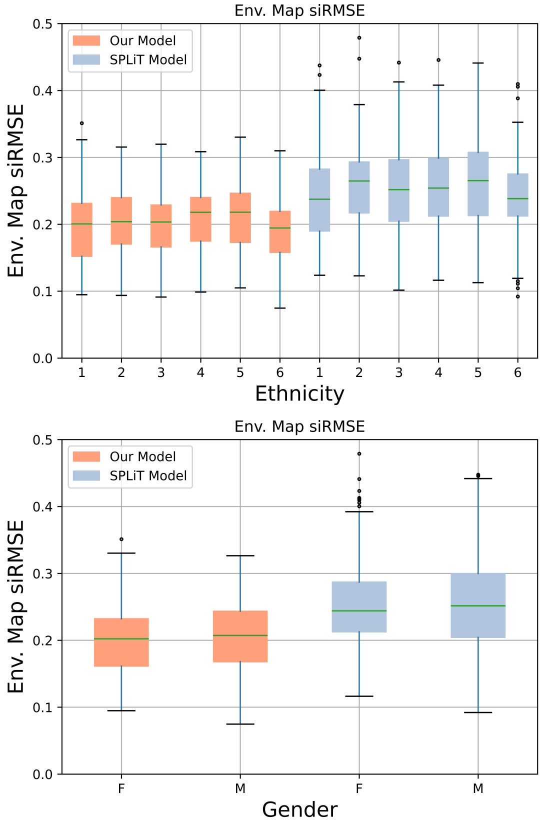

Figure 4. Fairness analysis on ethnicity (#1-6) and gender in comparison with the SPLiT model.

5. Conclusion

Motivated by recent advancements in the photorealism of graphics-based synthetic portrait images, which combine a parametric 3D face model with hand-crafted assets, and the fact that synthetic data simplifies the process of acquiring data from diverse subjects, we propose a training pipeline for portrait lighting estimation models that pre-trains on synthetic data and fine-tunes on unlabeled real videos. Our model achieves state-of-the-art performance in the real portrait image benchmark dataset. Our fairness analysis on gender and ethnicity demonstrates that our model exhibits greater robustness across different demographic groups than the existing state-of-the-art model.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2

[2] Dan A Calian, Jean-François Lalonde, Paulo Gotardo, Tomas Simon, Iain Matthews, and Kenny Mitchell. From faces to outdoor light probes. In *Computer Graphics Forum*, pages 51–61. Wiley Online Library, 2018. 1, 4

[3] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 1

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

[5] Fan Fei, Yean Cheng, Yongjie Zhu, Qian Zheng, Si Li, Gang Pan, and Boxin Shi. Split: Single portrait lighting estimation via a tetrad of face intrinsics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 4

[6] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022. 1

[7] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 2

[8] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7175–7183, 2019. 3

[9] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[11] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7312–7321, 2017. 2

[12] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019. 1, 2

[13] Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. Learning illumination from diverse portraits. In *SIGGRAPH Asia 2020 Technical Communications*, pages 1–4. 2020. 1, 4

[14] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE international conference on computer vision*, pages 3114–3122, 2017. 1

[15] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 1

[16] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[17] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 2

[18] Gowri Somanath and Daniel Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11298–11306, 2021. 2

[19] Alejandro Sztrajman, Alexandros Neophytou, Tim Weyrich, and Eric Sommerlade. High-dynamic-range lighting estimation from face portraits. In *2020 International Conference on 3D Vision (3DV)*, pages 355–363. IEEE, 2020. 1, 4

[20] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021. 1

[21] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European Conference on Computer Vision*, pages 477–492. Springer, 2022. 2, 3

[22] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. *arXiv preprint arXiv:2402.05746*, 2024. 1

[23] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2

[24] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde,

Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 2

[25] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 1

[26] Fangneng Zhan, Changgong Zhang, Wenbo Hu, Shijian Lu, Feiying Ma, Xuansong Xie, and Ling Shao. Sparse needlets for lighting estimation with spherical transport loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12830–12839, 2021. 2

[27] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 802–812, 2021. 1

[28] Hao Zhou, Jin Sun, Yaser Yacoob, and David W Jacobs. Label denoising adversarial network (ldan) for inverse lighting of face images. *arXiv preprint arXiv:1709.01993*, 2017. 2

[29] Yongjie Zhu, Chen Li, Si Li, Boxin Shi, and Yu-Wing Tai. Hybrid face reflectance, illumination, and shape from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5002–5015, 2021. 4