De-lighting Human Images Using Region-Specific Data Augmentation

Joshua Weir[†], Junhong Zhao, Andrew Chalmers[†] and Taehyun Rhee[†]
[†]*Computational Media Innovation Centre (CMIC)*
Te Herenga Waka - Victoria University of Wellington
Wellington, New Zealand
{josh.weir, j.zhao, andrew.chalmers, taehyun.rhee}@vuw.ac.nz

***Abstract*—This study aims to use deep learning to recover the diffuse albedo of human images captured under a wide range of real-world lighting conditions. A key challenge here is the wide variety of textures found in full-body human images. While some aspects like skin color have a limited color range, clothing and accessories display a broad spectrum of colors and textures. As a result, creating a comprehensive dataset with accurate labels is unfeasible. To address this, we propose a data augmentation method that involves applying color-shifts to various semantic regions within our training images, all while maintaining realistic appearance. This process is accomplished by initially segmenting the ground-truth albedos into their respective components (e.g., pants, shirt, hair, etc.) using a pre-trained human parsing network. Then, we adjust their hue and intensity channels using randomly chosen values from a carefully defined distribution. Our results show significant improvements in albedo recovery, especially in clothing areas, and better performance with underrepresented skin tones.**

***Index Terms*—human, de-lighting, albedo, data augmentation, deep learning**

I. Introduction

Lighting is essential for defining an image's character through color, reflections, and cast shadows, communicating the subject's context within its environment. In studio setups, photographers exercise precise control over lighting to convey mood and heighten structural clarity. However, in everyday scenarios like handheld photography or home video streaming, individuals often lack the ability to manage scene lighting. Consequently, unwanted shading elements such as shadows, saturation issues, and glare can lead to visual disturbances. These disruptions can undermine the photographer's artistic intent, the immersive perception of a composite, and even the performance of downstream computer vision applications like face recognition, avatar creation, and virtual try-on.

The task of *de-lighting*, as illustrated in Figure 1, involves removing all shading features from a subject, recovering its underlying texture (albedo), and is often considered a crucial initial step for *relighting* [14], [23], [24], [35]. De-lighting presents considerable challenges due to the complex interplay between unknown scene illumination, geometry, and reflectance. Previous research has attempted to address this uncertainty by relying on assumptions derived from prior knowledge in the target domain. For instance, in the case of faces, these assumptions include a similar convex shape and a limited distribution of skin colors. While these assumptions have been successfully utilized in face reconstruction works [5], [11], [27], [33], they are not applicable to images containing non-face attributes such as hair, glasses, and clothing.

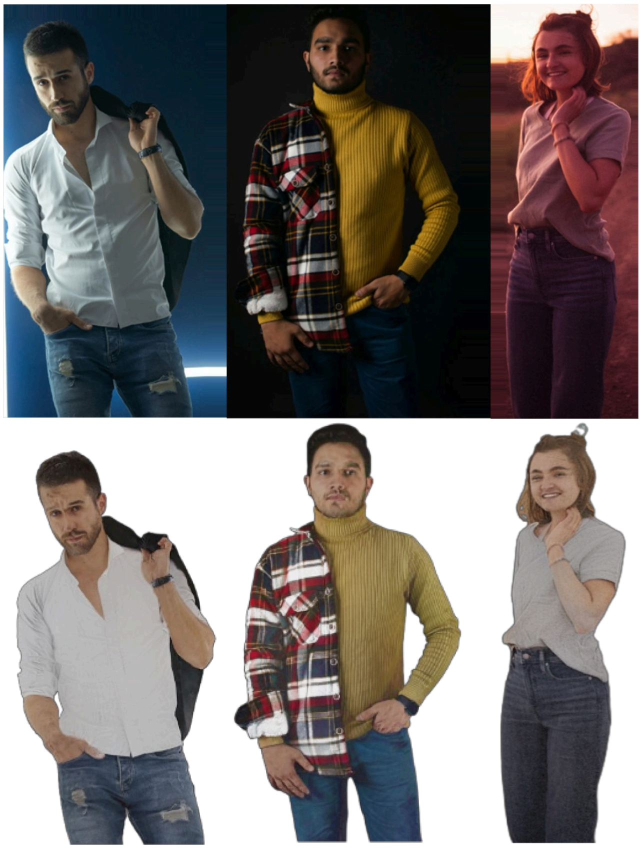

Fig. 1. Given an input, image (top row), our method estimates its albedo image (bottom row), removing shading artifacts while preserving texture.

To overcome these limitations, recent efforts have focused on deep-learning-based approaches for inferring the geometry and albedo of human subjects. However, these methods often require ground-truth supervision labels which are difficult to acquire, making them susceptible to overfitting to specific

979-8-3503-7051-5/23/$31.00 ©2023 IEEE

modalities of the training data. This can result in the inadvertent modification of non-shading-based content like clothing patterns and hair styles if these aspects are poorly represented in the training data.

In our work, we tackle the overfitting problem by designing a data augmentation strategy tailored for full-body human images. Our method stems from the observation that areas like clothing, hair, and accessories exhibit the highest levels of color and texture variability in human images. Conversely, the skin, eyes, and mouth regions tend to have a narrower color distribution, which offers useful constraints for inferring shading under diverse illuminations. Thus, we isolate just clothing and hair regions of our training images and modify them independently without augmenting skin regions. This way we can inject a significant amount of diversity into our dataset while maintaining plausibility of human appearance. We use the most recent state-of-the-art human albedo estimation network as our baseline model [14], and provide a qualitative and quantitative assessment of the improvements gained from our augmentation method

II. Related Work

Face de-lighting: Considerable efforts have been dedicated to the removal of disruptive light-related features, such as dark shadows and specularities, from faces. The primary objectives of these endeavors were to improve face recognition systems [7], [12] and enhance overall image quality [6], [22], [40]. Early works in this domain proposed optimization methods employing Morphable Face models [4], [5], [8], [11], [34]. However, their reliance on parametric models restricted their ability to capture non-facial details and high-frequency intricacies. More recent deep learning methods have addressed this limitation by utilizing feature-wise perceptual losses [20] or employing closed-loop Generative Adversarial Networks (GANs) [13] to recover finer facial details. Nevertheless, a drawback common to these methods is that they tend to focus solely on facial regions, address only specific aspects of the de-lighting process (e.g. shadow removal [40], camera-flash removal [6]), or employ front-facing illumination as the ground-truth, overlooking the sharp reflections it introduces.

An alternative approach involves GAN inversion methods [3], [10], [21], [38]. These methods facilitate face editing by projecting images into the latent space of a pre-trained GAN, thereby disentangling lighting, identity, pose, and expression attributes for independent manipulation. While effective in removing sharp shadows and specular reflections, these methods often struggle to preserve image content that is not explicitly tied to the editing attributes, particularly high-frequency details like freckles and non-facial components such as clothing.

Portrait and full-body de-lighting: Light manipulation of human images using deep learning has been extensively studied in recent years [14], [16], [18], [23], [24], [28], [29], [31], [32], [35], [37], [39], [41]. These approaches typically employ convolutional neural network (CNN) architectures to directly transform images from one lighting condition to another, all while preserving the non-lighting-related content present in the original image. A critical challenge in this domain is the acquisition of datasets that consist of paired images of the same subject under different lighting conditions while maintaining identical poses.

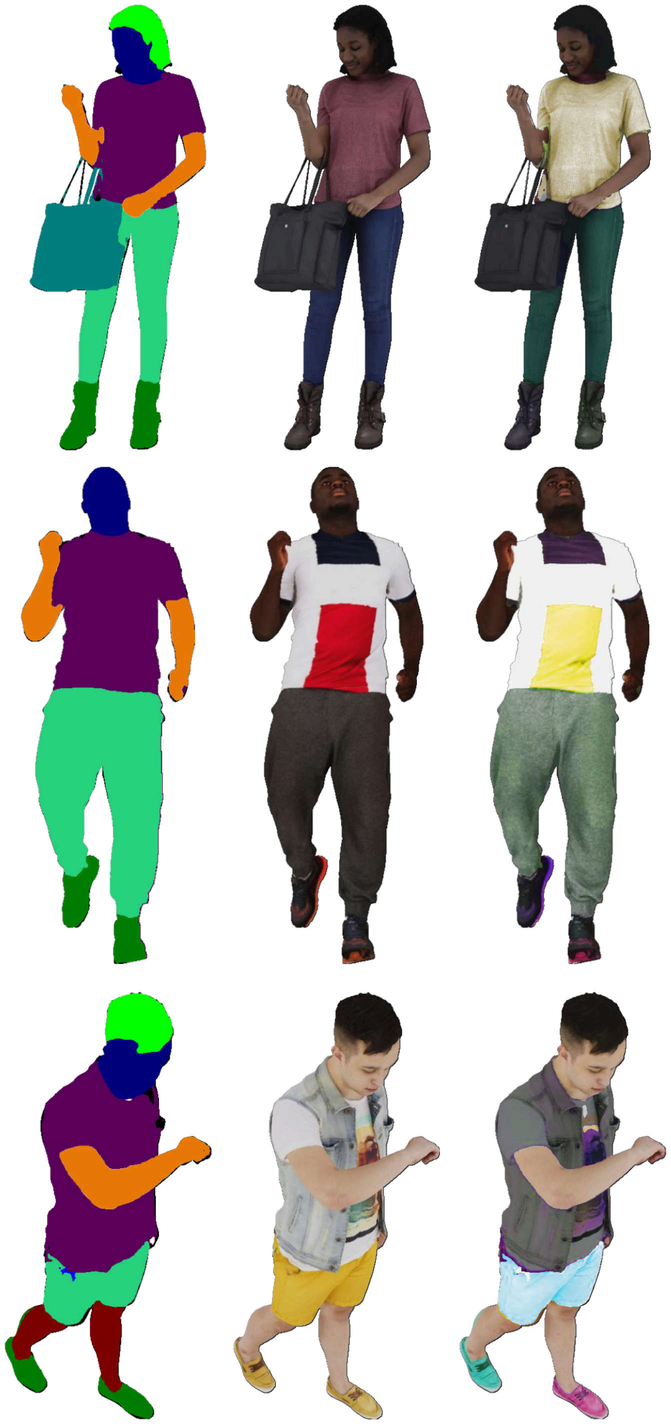

Fig. 2. Illustration of our regional data augmentation. From left to right, parsing map, original albedo, augmented albedo.

To address this challenge, some researchers have resorted to creating their ground-truth datasets using specialized equipment like a light stage [9], [23], [24], [31]. A light stage is a device equipped with an array of LED lights, enabling comprehensive capture of a human subject's reflectance from all directions. Unfortunately, these datasets are often not publicly available due to concerns related to licensing and privacy. Additionally, the high cost associated with acquiring data in this manner often results in datasets with a limited number of subjects, which can lead to overfitting to certain modalities such as skin-tone and clothing colour. To avoid such issues, many researchers have chosen to obtain training data by purchasing scanned 3D models from commercial sources [14], [16], [18], [29], [32]. However, even creating a large and diverse dataset using this method can be prohibitively expensive.

To address the limitations of supervised learning on datasets with sparse labels, some researchers have devised semi-supervised frameworks that enable learning from extensive collections of unlabeled images. For instance, SfSNet [28] introduced a pipeline for inferring geometry, albedo, and illumination from a single face image. Their approach includes an initialization stage where each component is acquired from 3D models, followed by a refinement stage where these components are learned from unlabeled real images through self-reconstruction loss. More recently, Lumos [39] introduced a portrait relighting pipeline with an albedo refinement stage facilitated by adversarial training. One drawback associated with these methods is that their refinement stages do not create a distinct separation between texture and shading, often resulting in shading being embedded into the estimated albedo.

III. Problem Formulation

We represent the input image as $I \in \mathbb{R}^{W \times H \times 3}$ with width W and height H. This image is produced by performing pixel-wise multiplication of its albedo A and shading S, expressed as:

$$I = A \odot S. \tag{1}$$

In line with Ji *et al*'s approach [14], our goal is to derive a set of parameters x for the function $\Omega(I, x) \in \mathbb{R}^{W \times H \times 3}$ which minimizes the following equation:

$$\lambda_{L1} \|A - \Omega(I, x)\|_1 + Vgg\left(A, \Omega(I, x)\right), \tag{2}$$

where the term Vgg represents the VGG-16 perceptual loss function [15] between A and $\Omega(I, x)$, and the constant λ_{L1} is configured to 5.

IV. Data Augmentation

We first utilize the human parsing method of [19] to segment A into a maximum of 14 regions, encompassing segments like hat, hair, sunglasses, upper-clothes, lower-clothes, belt, dress, left-shoe, right-shoe, face, legs, arms, bag, and scarf. In the context of data augmentation, we intentionally avoid altering the regions encompassing the face, arms, and legs. This is to prevent any changes that could lead to an unnatural appearance of skin.

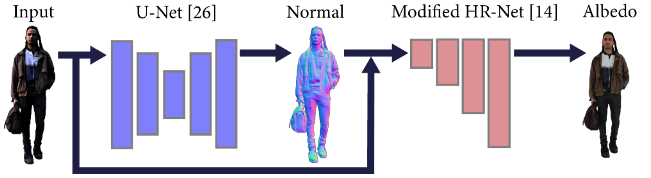

Fig. 3. Illustration of our De-lighting network, following the method outlined by Ji *et al.* [14].

In summary, we partition our albedos A into 11 regions designated as $P = \{A_1, ..., A_{11}\}$. The modification is applied to each region A_i in terms of the hue (A_i^H) and lightness (A_i^L) channels in HSL colour space:

$$\begin{aligned} A_i^H &= A_i^H + H_i & \forall A_i \in P \\ A_i^L &= \max(0.1, \min(0.95, L_i A_i^L)) & \forall A_i \in P. \end{aligned} \tag{3}$$

here, the random variables H_i and L_i are sampled from the uniform distributions of $[0, 2\pi]$ and $[0.1, 2.5]$ respectively. Examples of our segmentations and augmented albedos are shown in Fig. 2. After augmentation, our input images I are created using equation 1.

V. Implementation

We design our de-lighting function $\Omega(I, x)$ using a convolutional neural network (CNN) based image-to-image translation framework. The design of our neural architecture is illustrated in Figure 3, and follows the state of the art method introduced by Ji *et al.* [14].

First, a U-Net style architecture [26] estimates the geometry of the input image in the form of a normal map. Then, this normal map is concatenated with the input image, and passed through a modified HR-Net [30] to infer the final albedo. Conditioning the albedo estimation network with a normal map in this manner has been a design choice utilized in prior works to enhance performance [14], [24], as there exists a significant correlation between geometry and shading.

To ensure high-quality ground-truth albedo labels for our training data, we assemble a dataset comprising of scanned 3D models. Our training data consists of 150 models (134 for training, 16 for testing)[1] obtained from commercial sources. For each model, we produced 154 physically based renderings in various indoor and outdoor settings [1], [2].

We implemented our model using PyTorch [25] and conducted training on two NVIDIA RTX 6000 graphics cards. First, we trained our normal estimation network (see Figure 3) for 5 epochs until convergence. Our normal estimation network is optimized using the following equation:

$$|N - \tilde{N}\|_1, \tag{4}$$

where N and $\tilde{N}$ represent the ground-truth and predicted normals respectively.

[1] Our dataset contained multiple instances of individuals captured with various clothing and poses. We ensured that no identities were repeated between the train and test sets.

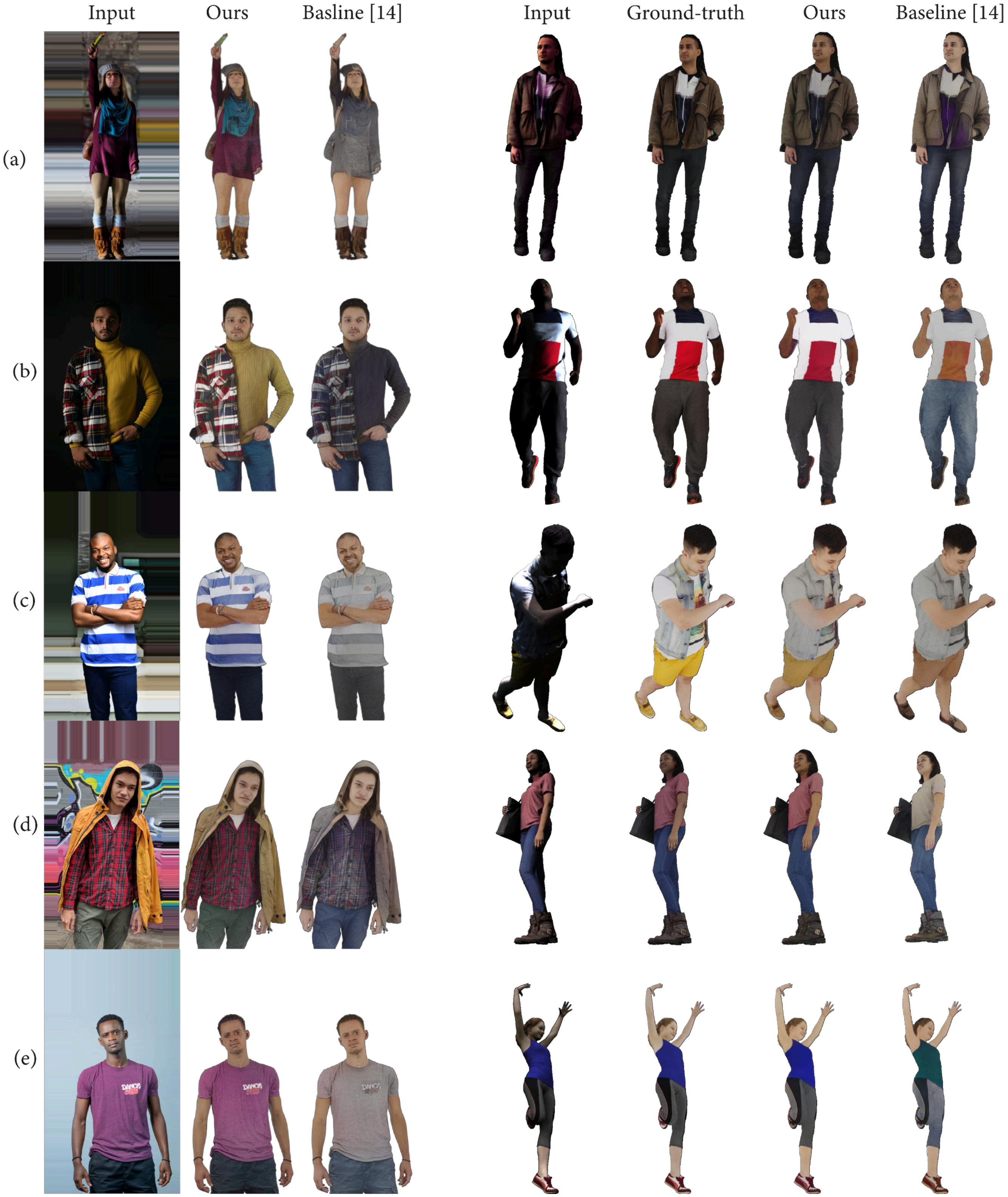

Fig. 4. Qualitative evaluation of our de-lighting with data augmentation (Ours) and without data augmentation (Baseline [14]). Images on the left were captured in the wild, while images on the right are sampled from our testing dataset with available ground-truth.

Then, we freeze the parameters of our normal network, and proceeded to train our albedo estimation network for 5 epochs, optimizing the loss function in equation 2.

Data augmentation (refer to equation 3) was applied every two iterations during training of both normal and albedo networks. Typically, the whole training process was completed within approximately two days. For optimization, we employed the Adam optimizer [17] with a learning rate set to $1e-4$. In terms of inference time, our model performed at 154ms on a single graphics card.

VI. Results

We evaluate the benefits of our data augmentation technique by comparing it with a baseline model. This baseline model is based on our implementation detailed in section V, with the key distinction being the absence of data augmentation. Additionally, this baseline model aligns with the implementation of Ji *et al.* [14] retrained using our dataset. Qualitative results demonstrate significant improvements as illustrated in Figure 4. Our full model trained with data augmentation effectively preserves intricate texture patterns in clothing regions. Notably, our approach also displays reduced bias towards lighter skin tones (see rows a, b, and d in the right column), even though no augmentations were applied to the skin regions. This indicates that the modification of non-skin regions alone introduces enough variability into each albedo image, including the relatively smaller subset with darker skin tones.

Additionally, we compiled a dataset for quantitative evaluation, encompassing 651 images taken from 16 distinct subjects, each exhibiting diverse lighting conditions and poses. Each image in this dataset was paired with its corresponding ground-truth albedo. The outcomes are detailed in Table I. Here, our full method outperforms the baseline on three image similarity metrics: mean squared error (MSE), peak signal to noise ratio (PSNR) and structural similarity index (SSIM) [36].

TABLE I
Quantitative results (mean ±std) on our testing dataset.

Method	MSE($\times$100)$\downarrow$	PSNR$\uparrow$	SSIM$\uparrow$
Ours	1.247 ±1.101	18.600 ±4.490	0.923 ±0.036
Baseline [14]	2.695 ±2.257	15.292 ±4.757	0.887 ±0.052

VII. Conclusion

In summary, we introduce a novel data augmentation method for training human image de-lighting networks, focusing on color adjustments in non-skin regions. Our results demonstrate significant improvements in both qualitative and quantitative aspects, underscoring the critical role of appearance diversification in this context.

Limitations and Future Work: Although our approach showcases enhancements compared to the baseline trained on our dataset, it still exhibits a bias towards lighter skin tones. This highlights the essential requirement for fairness in future datasets to ensure balanced representation. Our method also encounters challenges when it comes to preserving finer details like facial hair and accessories such as sunglasses. Hence, future research could explore more comprehensive approaches to human albedo augmentation, potentially involving the synthesis of intricate clothing textures, facial hair, and tattoos in a manner that looks natural-looking. These advancements hold considerable promise for enhancing human image de-lighting capabilities, ultimately reducing the reliance on expensive acquisition of real-world ground-truth data.

Acknowledgment

This work was supported by the Entrepreneurial University Programme from the Tertiary Education Commission of New Zealand.

References

[1] Laval indoor hdr dataset. *http://indoor.hdrdb.com/*, 2017.

[2] Hdri haven. *https://hdri-haven.com/*, 2020.

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021.

[4] Abdelrehim Ahmed and Aly Farag. A new statistical model combining shape and spherical harmonics illumination for face reconstruction. In *International Symposium on Visual Computing*, pages 531–541. Springer, 2007.

[5] 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*, pages 187–194, 1999.

[6] Nicola Capece, Francesco Banterle, Paolo Cignoni, Fabio Ganovelli, Roberto Scopigno, and Ugo Erra. Deepflash: Turning a flash selfie into a studio portrait. *Signal Processing: Image Communication*, 77:28–39, 2019.

[7] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015.

[8] Xiaowu Chen, Hongyu Wu, Xin Jin, and Qinping Zhao. Face illumination manipulation using a single reference image by adaptive layer decomposition. *IEEE Transactions on Image Processing*, 22(11):4249–4259, 2013.

[9] 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.

[10] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.

[11] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[12] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.

[13] Xianjun Han, Hongyu Yang, Guanyu Xing, and Yanli Liu. Asymmetric joint gans for normalizing face illumination from a single image. *IEEE Transactions on Multimedia*, 22(6):1619–1633, 2019.

[14] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 388–405. Springer, 2022.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[16] Yoshihiro Kanamori and Yuki Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *arXiv preprint arXiv:1908.02714*, 2019.

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

[18] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. Single-image full-body human relighting. *arXiv preprint arXiv:2107.07259*, 2021.

[19] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[20] Shenggui Ling, Ye Lin, Keren Fu, Di You, and Peng Cheng. A high-performance face illumination processing method via multi-stage feature maps. *Sensors*, 20(17):4869, 2020.

[21] BR Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, et al. Photoapp: Photorealistic appearance editing of head portraits. *ACM Transactions on Graphics*, 40(4):1–16, 2021.

[22] Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. Deep face normalization. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019.

[23] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020.

[24] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[27] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5144–5153, 2017.

[28] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[29] Guoxian Song, Tat-Jen Cham, Jianfei Cai, and Jianmin Zheng. Real-time shadow-aware portrait relighting in virtual backgrounds for realistic telepresence. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 729–738. IEEE, 2022.

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[31] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019.

[32] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, volume 40, pages 205–216. Wiley Online Library, 2021.

[33] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017.

[34] Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2008.

[35] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[37] Joshua Weir, Junhong Zhao, Andrew Chalmers, and Taehyun Rhee. Deep portrait delighting. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[38] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021.

[39] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6):1–21, 2022.

[40] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020.

[41] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019.