Full-body Human De-lighting with Semi-Supervised Learning

Joshua Weir[1], Junhong Zhao[1], Andrew Chalmers[1,2], and Taehyun Rhee[2]

[1] Victoria University of Wellington, Wellington, New Zealand
{josh.weir,j.zhao,andrew.chalmers}@vuw.ac.nz
https://www.wgtn.ac.nz
[2] University of Melbourne, Melbourne, Australia
taehyun.rhee@unimelb.edu.au
http://www.unimelb.edu.au

Abstract. Removing undesired shading from human images is crucial in supporting various real-world applications. While recent advancements in deep learning-based methods show promise in addressing this challenge, there persists a struggle to accurately separate texture from shading, which often results in unresolved shading artifacts and altered texture patterns. This issue is exacerbated by dataset limitations, such as the lack of diverse real-world clothing styles in realistic datasets and oversimplified assumptions about human reflectance illumination environments. To solve the problem, our paper introduces a novel semi-supervised deep learning method to effectively assemble both real and synthetic data for better disentanglement of texture and shading. We present a global sparsity constraint designed on both labeled and unlabeled data to minimize color variations in the inferred shading map, enhancing texture recovery. By applying this constraint, our method demonstrates improved handling of a broad range of fashion-related textures in the real-world test. Additionally, we address the disparity between real and synthetic data with a novel domain adaptation module to realize effective transfer from synthetic to real images. This module is designed based on the insights of gamma correction, and demonstrates improved shadow removal in real-world images. By integrating these methods, our approach achieves state-of-the-art results, reducing unwanted shading artifacts while maintaining the integrity of underlying textures in real-world scenarios.

Keywords: De-lighting · Full-body · Semi-supervised Learning

1 Introduction

Removing unwanted lighting features from an image to reveal its true albedo, a process known as "de-lighting," is an important step in various computer graphics and vision applications. Recently human relighting [18, 27, 36, 47], virtual try-on [13, 24], and avatar creation [5, 15], have become popular, where human image de-lighting has contributed a critical component in generalizing these downstream tasks to adapt seamlessly to various real-world lighting conditions.

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

Previous studies on human de-lighting have primarily focused on faces and upper-body portraits [25, 41, 46]. Full-body de-lighting exhibits more challenges due to greater color and texture variations, along with more complex shadows and occlusions resulting from diverse body poses and shapes. Prior research has stressed the importance of large datasets with sufficient variations on different poses, clothing, and lighting conditions [16,48] for training reliable lighting-aware models.

Recent works have established ground-truth datasets using specially designed capturing setups such as light stages [26, 27, 36]. Although the captured data is realistic, such setups are costly and come with many on-site constraints, limiting accessibility for the broader research community and posing challenges in capturing human subjects with sufficient diversity. As an alternative, virtual 3D human models from various commercial/non-commercial sources have been employed [16, 18] for creating resembling synthetic datasets. While these data creations are more widely accessible and controllable on shape, pose, and albedo texture diversity, diversity issues still persist due to the limited variability of clothing and fashion accessories. Moreover, the synthetic procedure often oversimplifies human reflectance properties, leading to a domain gap in full-body human representation, especially in accurately conveying complex shading features, such as specular highlights, subsurface scattering, and the effects of multiple light bounces. Data limitations existing in both synthetic and real-world capturing hinder the learning capability of neural networks on real-world images and exacerbate the de-lighting performance in the wild test scenarios, manifesting as unresolved shading artifacts and unintentional alterations of texture patterns.

In this paper, we present a novel semi-supervised deep learning method for human de-lighting (see Fig. 1) . We devise a novel network architecture and training scheme to take advantage of the utility of labeled synthetic data and overcome its limitations by exploiting a vast dataset of unlabeled real-world photos. The rendered images of 3D human models under diverse illumination conditions, paired with their intrinsic ground-truths facilitate learning for the de-lighting task. Simultaneously, a large dataset of unlabeled real-world images serves to regularize texture overfitting and bridge the shading discrepancies between the synthetic and realistic domains in a semi-supervised manner.

We address texture overfitting on both labeled data and unlabeled data by designing a new loss constraint on the ratio between the source image and its estimated albedo. Assuming such ratio images represent the removed shading and should have a sparse color palette influenced by light directions and occlusions, we propose a global sparsity loss that reduces the overall color variation in this shading map for each normal direction, leading to better albedo estimation. We further adapted this loss to accommodate unlabeled data, allowing our model to learn a vast range of garments and fashion styles from an extensive dataset of images in the wild.

Furthermore, we observe that the domain gap between real-world images and rendered 3D models can be narrowed by adjusting the gamma settings. Consequently, to address unresolved shadows in real images, we propose a novel transfer learning paradigm with a domain adaptation module geared towards learning per-image gamma adjustments and a fine-tuning of adjusted images. This approach adapts our pre-trained de-lighting network to better handle real-world images with more complex shading. Our contributions are summarized as:

- A novel semi-supervised human de-lighting method that is trained using a sparse set of synthetic 3D human images with intrinsic ground-truth, and a large set of unlabeled real photos, achieving robustness with regards to diminishing shading artifacts while maintaining non-shading based content.
- A global sparsity loss applicable to both labeled and unlabeled data to minimize color variance in the inferred shading, leading to more stable texture recovery, and enabling the learning of texture patterns from a broad set of unlabeled images.
- A novel synthetic-to-real transfer learning paradigm with a domain adaptation module for per-image gamma adjustments, improving the model's robustness to complex real-world shadows.

2 Related work

Classical methods: Early research on estimating human reflectance from monocular images primarily employed optimization frameworks with the help of statistical priors like morphable face models [4, 9, 11, 38, 44], and simple lighting assumptions such as directional light or spherical harmonics [29]. To tackle this

problem in more general scenes, numerous works utilize image-based priors. For instance, Shen *et al.* [32] implemented a local reflectance constraint on neighboring pixels with similar chromaticity, since shading typically manifests as intensity variation. Other studies [6, 7] propose global-sparsity constraints on the albedo map through entropy minimization, noting that most objects are composed of a limited palette of base colors. Nonetheless, these methods rely on overly simplistic assumptions about the intrinsic characteristics of real-world images, which may not generalize well to complex images such as full-body humans.

Deep supervised Learning: In recent times, CNN-based methods have emerged as highly effective solutions for de-lighting and relighting humans in diverse and uncontrolled environments [16, 27, 31, 34, 39, 43]. The pioneering work by Kanamori *et al.* [18] introduced the first full-body relighting method using a large dataset of human images with ground-truth intrinsics. Since then, further advancements have been made to enhance its capability in generating challenging shading elements like specularities and cast shadows [20, 34, 37]. However, most of these techniques rely on a simple image-to-image translation pipeline for albedo estimation, which may not effectively remove all shading artifacts.

Weir *et al.* [41] utilized residual image learning and a masked loss function to improve the disentanglement between shading and reflectance, while Ji *et al.* [16] suggested that skip-connections in the widely used U-Net architecture [30] were responsible for shading entanglement and proposed an alternative network architecture based on HR-Net [35].

Deep semi-supervised learning: Various studies have delved into semi supervised learning to address the constraints inherent in synthetic datasets. SfS-Net [31] employed a photometric reconstruction loss to capture intricate facial details from real-world photos. Lumos [43] proposed a portrait relighting method with synthetic-to-real domain adaptation, focusing on correcting the albedo image. While these methods are good at preserving original texture details, their reliance on generating a re-lit image constrains them based on the expressive capabilities of the underlying relighting framework.

3 Method

In Fig. 2, we depict the workflow of our de-lighting network. To improve performance, we condition the albedo network on inferred geometry and semantic labels, following the approach of prior works [16, 27, 39]. Our Prior-Net calculates normal and semantic parse maps from the input image. Concurrently, our proposed domain adaptation module assesses gamma adjustment parameters $\gamma 1$ & $\gamma 2$ for both the input and albedo images. The resultant normal, parse map, and gamma ($\gamma 1$)-adjusted input images are then processed through our Alb-Net to infer the albedo image, which is subsequently adjusted by $\gamma 2$.

During the training process, we use a combination of two datasets: synthetic dataset, $\mathcal{D}_S$, which comprises rendered human models with ground-truth albedo, normal, and parsing labels, and real dataset, $\mathcal{D}_R$, which consists of photographs captured in real-world scenarios, and contains parsing labels only. The images

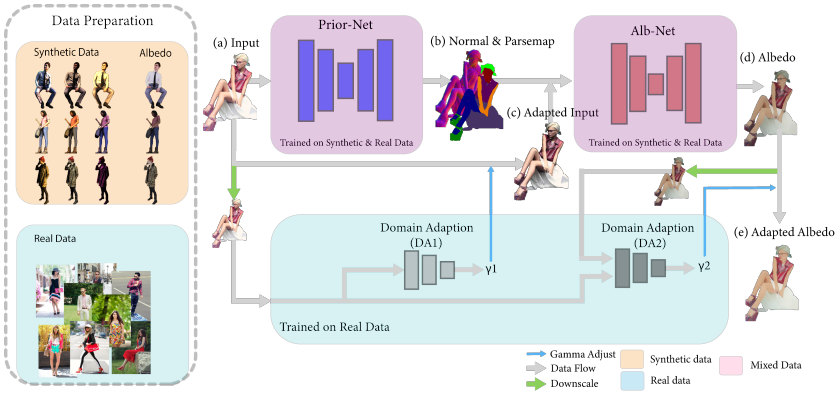

Fig. 2: Our semi-supervised training and inference pipeline for human de-lighting.

from the real dataset, $\mathcal{D}_R$, are used for global-sparsity loss calculation and domain adaptation training, aiming to enhance the model's ability to generalize to diverse texture patterns and real-world shading features. For convenience, we use subscripts $_{\mathbf{S}}$, $_{\mathbf{R}}$ and $_{\mathbf{M}}$ (e.g., $\mathbf{I_S}$, $\mathbf{I_R}$, $\mathbf{I_M}$) to represent batches of images containing only synthetic, only real, and a mix of both, respectively.

3.1 Baseline Model Design

Prior-Net: The Prior-Net network was trained with the loss function $\mathcal{L}_{prior}$ formulated as:

$$\mathcal{L}_{prior} = \lambda_{L1}\|\mathbf{N_S} - \tilde{\mathbf{N}}_\mathbf{S}\|_1 + \lambda_{Vgg}Vgg\left(\mathbf{N_S}, \tilde{\mathbf{N}}_\mathbf{S}\right) + \lambda_F F\left(\mathbf{P_M}, \tilde{\mathbf{P}}_\mathbf{M}\right), \tag{1}$$

where $\mathbf{N_S}$ and $\tilde{\mathbf{N}}_\mathbf{S}$ represent the ground-truth and predicted normal maps, respectively, obtained from the synthetic dataset $\mathcal{D}_S$, Similarly, $\mathbf{P_M}$ and $\tilde{\mathbf{P}}_\mathbf{M}$ represent the ground-truth and predicted parsing maps. Vgg represents the VGG-16 perceptual loss [17,33], and F represents the multi-class focal loss [39]. We assign the values λ_{L1}, λ_{Vgg}, and λ_F as 1, 5, and 100 respectively.

The parsing map identifies up to seven semantic regions: face, hair, hat, bag, arms, legs, and clothes, which are passed to the albedo network (denoted as Alb-Net) Alb-Net in the second stage to support the de-lighting process. The selection of semantic labels is based on two key observations. Firstly, the reflectance of skin regions generally exhibits less variation compared to that of hair and clothes. Secondly, certain objects such as limbs, hats, and hand-held items often cause significant occlusions. By categorizing these regions in an initial stage, we can effectively guide our Alb-Net to achieve more robust de-lighting. An ablation of the parsing map prior is illustrated in Fig. 4 and Tab. 1.

Alb-Net: We use the HR-Net architecture proposed by [16] as the backbone for our Alb-Net network. The baseline model is trained on our synthetic data using the loss function $\mathcal{L}_{bas}$ formulated as:

Fig. 3: The motivation behind our global sparsity loss, where (d) and (e) represent the ratios $\frac{(a)}{(b)}$ and $\frac{(a)}{(c)}$ respectively. The inaccurate color estimation of the shorts in (c) affects the shading in (e), resulting in a visually noticeable error as indicated by the contrast.

$$\mathcal{L}_{bas} = \lambda_{L1} \|\mathbf{A_S} - \tilde{\mathbf{A}}_\mathbf{S}\|_1 + 2\lambda_{Vgg} Vgg\left(\mathbf{A_S}, \tilde{\mathbf{A}}_\mathbf{S}\right), \tag{2}$$

where $\mathbf{A_S}$ and $\tilde{\mathbf{A}}_\mathbf{S}$ denote the ground-truth and predicted albedos from the synthetic dataset $\mathcal{D}_S$. While this baseline loss effectively removes shading artifacts on synthetic testing data, it struggles to generalize well to the diverse texture space of human images in real-world scenarios. Addressing and contributing to this issue will be the focus of the next section.

3.2 Global Sparsity Loss

Given an input image $\mathbf{I}$, its shading $\mathbf{S}$ can be expressed in terms of a ratio image [14, 47] between itself and its albedo, given by $\mathbf{S} = \frac{\mathbf{I}}{\mathbf{A}}$. In a scenario with Lambertian reflectance and directional lighting, the color at any point on this shading map depends on the light intensity, light angle of incidence, and occluding objects in the scene. This means the global color variance should be sparse with respect to each normal direction. This point is illustrated in Fig. 3, where we show the inferred shading map with respect to both the ground-truth albedo $\mathbf{A}$, and a predicted albedo $\tilde{\mathbf{A}}$. Comparing Fig. 3 (e) and (d), we can see that if the color of a certain region is predicted incorrectly, it will result in significant color changes in the ratio image. Therefore, a sparse shading map is linked with more globally consistent albedo inference, this motivates us to limit the color palette of the inferred shading $\tilde{\mathbf{S}}$ to attain more accurate shading removal and avoid unintended removal of non-shading based content.

We achieve this by designing a global sparsity loss based on a minimum entropy. Prior de-lighting works [6, 7, 10] utilized minimum entropy losses on the albedo for shading removal. However, the color variation in human images can be extensive, especially when considering the diversity in fashion. It is thus unreasonable to assume minimal color variation in the albedo. Instead, we focus on the sparsity assumption for the shading map and avoid blurring or miscoloring of vital texture patterns. Our global sparsity $\mathbf{G}(\cdot)$ is defined as follows:

$$\mathbf{G}(S,N) = \mathbb{E}\left[-log\left(\frac{1}{n}\sum_{i=1}^{n} KDE(\mathbf{d}(S,N,i))\right)\right]. \tag{3}$$

Here, KDE is the kernel density estimator [10] to approximate the probability density function. S and N are the shading and normal maps respectively, n denotes the sample size, and $\mathbf{d}$ calculates the distance of each pixel value in the shading map S relative to the pixel S_i, which we defined as:

$$\mathbf{d}(S,N,i) = (S - S_i) \odot \exp\left(-\frac{(N \cdot N_i - 1)^2}{\sigma^2}\right), \tag{4}$$

where $\odot$ represents pixel-wise multiplication. The exponential factor in the equation serves to emphasize the importance of the distance penalty when the normal vectors are similar, effectively preventing unnecessary penalization of shading removal in the presence of non-uniform lighting. The variance parameter σ determines the acceptable cosine angle between the normals. We assign n and σ to 50 and 0.1 respectively.

Then, we incorporate the global sparsity loss using the following equations:

$$\mathcal{L}_{GS1} = \lambda_{GS1} \cdot \mathbf{G}\left(\frac{\mathbf{D_S}}{\tilde{\mathbf{A}}_\mathbf{S}}, \mathbf{N_S}\right), \tag{5}$$

To ensure unbiased global sparsity calculation, we eliminate cast-shadows from our estimated shading maps using the ratio image $\frac{\mathbf{D_S}}{\tilde{\mathbf{A}}_\mathbf{S}}$ where $\mathbf{D_S}$ is the input image rendered without cast-shadows. Here, λ_{GS1} is assigned a value of 800.

Since no ground-truth labels exist for real data $\mathcal{D}_R$, two considerations must be taken into account: Firstly, the true color of any pixel in $\mathbf{I_R}$ with a max value of 1 is unknown. Thus, we drop these pixels from our calculation. Secondly, we cannot isolate cast-shadows from the equation like in $\mathcal{L}_{GS1}$. Instead, we mitigate the shadow bias by operating on the shading's chromaticity value. In general, cast shadows are characterized by variations in image luminance, while changes in chromaticity values are typically attributed to variations in reflectance [32]. By operating on chromaticity, we minimize the penalty for shadow removal while still penalizing texture removal. The final global sparsity loss function to real data $\mathcal{D}_R$, $\mathcal{L}_{GS2}$, is defined as:

$$\mathcal{L}_{GS2} = \lambda_{GS2} \cdot \mathbf{G}(\|\tilde{\mathbf{S}}_\mathbf{R}\|, \tilde{\mathbf{N}}_\mathbf{R}), \tag{6}$$

where $\|\tilde{\mathbf{S}}_\mathbf{R}\|$ denotes the chromaticity values of the inferred shading map $\tilde{\mathbf{S}}_\mathbf{R} = \frac{\mathbf{I_R}}{\tilde{\mathbf{A}}_\mathbf{R}}$, and the parameter λ_{GS2} is assigned a value of 100.

3.3 Domain Adaptation

In real-world data, direct illumination and cast shadows can differ from synthetic data due to physical phenomena like multiple light bounces and realistic reflectance. This leads to visual changes, such as brighter regions under direct

illumination and shadowed areas appearing brighter due to residual lighting effects. This becomes problematic when training on synthetic human models, which only capture Lambertian reflectance, as the resulting domain gap limits the models performance with real-world images.

To improve shading removal, rather than attempting to model or capture these realistic shading effects in the synthetic data, we take advantage of the simplistic shading distribution that the Alb-Net model has already learned from the synthetic dataset, and design a domain adaptation network that modifies the real images such that they are more closely aligned to the distribution of the synthetic images, albeit with greater texture variety. Intuitively, the addition of physically-based light transport properties to a previously Lambertian rendering can only increase the overall brightness of each pixel, and in most cases, decrease the overall contrast between the brightest and darkest pixels in the image as light rays interact with surfaces that would have otherwise been shadowed. For this reason, we assume the synthetic counterpart of each real image will be darker with more pronounced shadows. Hence, our domain adaptation takes the form of a simple gamma adjustment on the input image:

$$\mathbf{I} = \mathbf{I}^{\gamma 1} \quad : \quad 1 < \gamma 1 < 2.2, \tag{7}$$

where $\gamma 1$ is the input gamma coefficient which, since $0 \leq \mathbf{I} \leq 1$, will compress each pixel towards the dark values and increase image contrast (see Fig. 2 (c)). When applied to real images, a higher $\gamma 1$ generally leads to better shadow removal performance due to enhanced shadow clarity, but also has the potential to alter physical appearance, producing a dark or desaturated albedo (see Fig. 2 (d)). For this reason, we train our first domain adaptation network DA1 (see Fig. 2) to estimate the most optimal gamma value $\gamma 1$ for each input image before it's passed to the de-lighting network.

To resolve the loss of texture and enable more flexibility for the values of $\gamma 1$, we design a second gamma adjustment network DA2 (see Fig. 2), which estimates $\gamma 2$ to be applied to the resulting de-lit image:

$$\widehat{\tilde{\mathbf{A}}} = \tilde{\mathbf{A}}^{\gamma 2} \quad : \quad 0 < \gamma 2 \leq 1. \tag{8}$$

Where $\widehat{\tilde{\mathbf{A}}}$ is the adapted albedo image. Here, $\gamma 2$ has the effect of expanding each pixel in $\tilde{\mathbf{A}}$ to higher values, compensating for the lost image brightness resulting from the initial adjustment.

Once the baseline training of the Alb-Net model has converged, we freeze its model parameters and train the domain adaptation networks on our real dataset $\mathcal{D}_R$ by optimizing both DA1 and DA2 with the following loss function:

$$\mathcal{L}_{DA} = \frac{1}{k_3 \nabla\left(\max_{rgb}(\tilde{\mathbf{S}}_{\mathbf{R}})\right)} + k_1 \mathbf{G}(\|\widehat{\tilde{\mathbf{S}}}_{\mathbf{R}}\|, \tilde{\mathbf{N}}_{\mathbf{R}}) + k_2 \mathbf{G}(\widehat{\tilde{\mathbf{S}}}_{\mathbf{R}}, \tilde{\mathbf{N}}_{\mathbf{R}}), \tag{9}$$

where $\widehat{\tilde{\mathbf{S}}}_{\mathbf{R}} = \frac{\mathbf{I}_{\mathbf{R}}}{\widehat{\tilde{\mathbf{A}}}_{\mathbf{R}}}$, ∇ is the total variation loss (TV) and $\max_{rgb}$ returns the maximum of each RGB value in the image. Inverse TV loss is applied to shading

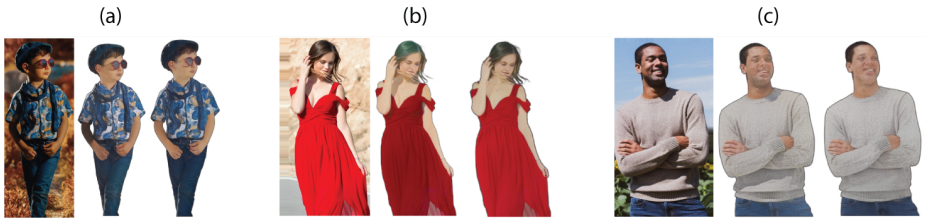

Fig. 4: Parsemap conditioning ablation. Each triplet of images illustrates (left to right) input, output *without* parse, output *with* parse.

intensity to promote the removal of shading artifacts, which are primarily localized to specific regions. Simultaneously, we minimize the shading global sparsity to regularize potential texture distortions in the albedo. The constants $k1$, $k2$, and $k3$ are empirically set to 350, 10, and 25,000 respectively.

4 Data and Implementation

Synthetic dataset: We compiled a synthetic dataset with intrinsic ground-truth labels, denoted as $\mathcal{D}_S$, consisting of 150 3D models of virtual human subjects (134 for training, 16 for testing)[3] obtained from different commercial sources. Each model was posed under three different camera angles and different illumination conditions from which we extracted normal, albedo, foreground-mask, and 154 physically-based renderings in various indoor and outdoor settings [2,3]. To extract ground-truth parse maps, we applied a pre-trained human parsing network [1, 21] to our albedo images which detects up to 18 parsing regions. To enhance the diversity of our dataset and improve the generalization capabilities of the trained model, we apply region-specific data augmentation [42] which introduces random color shifts to various non-skin regions of the training images. Additionally, we insert directional lights into the environment maps to boost robustness to strong shadows. More details can be found in the supplementary document.

Real-world dataset: As for the real dataset, $\mathcal{D}_R$, we utilized the ICCV-15 fashion dataset [22, 23], which comprises a large number of (17,706) full-body portraits captured across a wide range of indoor and outdoor environments but without any albedo ground-truth. Each image is also accompanied by a parse map, which are utilized during Prior-Net training (see Eq. 1). All the training images will be resized to 512×384 resolution for input into the network.

Implementation: We used PyTorch [28] to implement our model, training it on two NVIDIA RTX 6000 graphics cards. We first train our Prior-Net (Sec. 3.1)

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

network until $\mathcal{L}_{prior}$ converges after 5 epochs, froze its parameters, and proceeded to train the Alb-Net network for an additional 5 epochs, aiming for the convergence of the combined losses $\mathcal{L}_{bas} + \mathcal{L}_{GS1} + \mathcal{L}_{GS2}$. During training, each batch consisted of four images from the real dataset $\mathcal{D}_S$ and two images from the unlabeled dataset $\mathcal{D}_R$. After completing this step, we froze the parameters of the initially trained Alb-Net and initiated the training of our Domain Adaptation module with real data (refer to Sec. 3.3). Training for 5 epochs on the entire $\mathcal{D}_R$ dataset. For optimization, we utilized the Adam optimizer [19] with a learning rate of $1e-4$ for Prior-Net and Alb-Net, and $1e-6$ for the domain adaptation module. The inference time for the full model was 154ms on a single graphics card.

5 Results

We evaluated our method against the two recent state-of-the-art methods capable of full-body de-lighting: Total Relighting (TR) [27] and Geometry-aware Single-image Full-body Human Relighting (GSFR) [16]. Two versions of each are tested: The author implemented models, and our implementations of them retrained using our dataset $\mathcal{D}_S$, denoted GSFR (retrained) and TR (retrained). From the qualitative results in Fig. 5, we can see that the original prior works (f & g) can more faithfully preserve texture than our retrained versions (d & e), but struggle under hard shadows (second row) and abnormally colored illumination (bottom row). Our method (c) more effectively resolves harsh shading artifacts and achieves more accurate color inference. This is further illustrated in the wild test (i), where ours clearly achieves more accurate albedo estimation (second and fourth rows) and shadow removal (first and third rows) than prior works. Quantitative results in Tab. 1 show that our method outperforms prior work in terms of MSE, si-MSE [4] [8], PSNR, SSIM [40] and LPIPS [45] on our testing dataset.

5.1 Global Sparsity Evaluation

We assess the advantages of incorporating global sparsity loss in Fig. 6. When employing only the baseline loss $\mathcal{L}_{bas}$, our model tends to eliminate texture patterns and desaturate vibrant colors. The introduction of $\mathcal{L}_{GS1}$ to the synthetic data $\mathcal{D}_S$ yields improved results; however, it falls short in addressing complex textures not present in our synthetic training dataset, evident in the removal of glasses (rows 1 & 3) and the washed-out texture patterns (rows 2, 3 and 5).

The model trained with $\mathcal{L}_{GS2}$ accommodates such features by learning a broader range of textures from our unlabeled dataset $\mathcal{D}_R$. Nevertheless, it tends to miscalculate the brightness of certain regions, as evidenced by the shirt logo in the top row and lightened skin-tone in the fourth row. This discrepancy may be attributed to the fact that $\mathcal{L}_{GS2}$ operates on chromaticity values of the shading map rather than color (refer to Sec. 3.2 for more details).

[4] scale-invariant MSE. The error remains constant regardless of intensity scaling

Table 1: Quantitative results on the synthetic testing dataset.

Method	MSE↓	si-MSE↓	PSNR↑	SSIM↑	LPIPS↓
GSFR	3.719	2.100	13.426	0.783	0.057
TR	5.762	3.084	12.358	0.820	0.093
GSFR (retrained)	0.728	0.463	20.720	0.930	0.036
TR (retrained)	1.012	0.752	19.329	0.917	0.044
w/o GS1 & GS2	0.733	0.461	21.093	0.937	0.032
w/o GS1	0.741	0.407	21.122	0.936	0.031
w/o GS2	0.704	0.381	20.912	0.936	0.030
w/o parse	0.687	**0.355**	20.965	0.936	0.029
Ours Full	**0.646**	0.362	**21.562**	**0.940**	**0.027**

Table 2: Quantitative domain adaptation ablation on the Multi-PIE testing dataset.

Method	MSE↓	si-MSE↓	PSNR↑	SSIM↑	LPIPS×100 ↓
w/o DA	**0.477**	**0.362**	22.542	0.935	0.832
w/ DA	0.520	0.397	**22.960**	**0.942**	**0.764**

Our full model, trained using both $\mathcal{L}_{GS1}$ and $\mathcal{L}_{GS2}$, achieves the most consistent and plausible results in preserving the original texture. This is verified quantitatively in the last row of Tab. 1.

5.2 Domain Adaptation Evaluation

We assess the advantages of domain adaptation qualitatively in Fig. 7. Upon integrating the domain adaptation module, intricate shadows and highlights present in skin and clothing regions are effectively eliminated. By emphasizing shadows and reducing color distortions around shadows, our DA module enhances the performance of a de-lighting network trained using synthetic data.

To quantitatively evaluate our DA module would require real-world images with ground-truth albedos. Since no full-body dataset of this kind is publicly available, we modify the popular face recognition dataset Multi-PIE [12] for our purposes. Multi-PIE comprises a set of portrait images captured from the shoulders up, under 18 directional lighting conditions. We generate input and uniform-lighting images following the method outlined in Weir *et al.* [41]: color-scaled versions of the directional lighting images are used as input to create challenging illuminations, and the average over all directional lighting images is used to approximate a smooth, uniformly-lit image. The ground-truth albedo for this experiment is the uniformly-lit image processed through the respective model to remove any remaining shadows. Our Multi-PIE evaluation dataset consists of 1000 images spanning 140 unique subjects. The results in Tab. 2 indicate that our DA-trained model achieves weaker scores on MSE and si-MSE but

Fig. 5: Comparisons with state-of-the-art methods TR [27] and GSFR [16]. Results on our testing dataset are shown on the left, while in-the-wild results are shown on the right.

stronger scores on PSNR, SSIM, and LPIPS metrics. This suggests that some of the overall color/brightness accuracy was lost due to gamma adjustments, but overall structural accuracy was increased due to enhanced shadow removal, as demonstrated in the qualitative test in Fig. 8.

6 Conclusion

While our method demonstrates significant improvements on real-world shading, failure cases can arise from glossy surfaces such as leather jackets. This is because no such materials were present in our synthetic dataset, so they are indistinguishable from texture patterns even after domain adaptation. Secondly, due to lack of intrinsic ground-truth images for our quantitative evaluation of DA, we depend on de-lit images from our own model, which does not guarantee an accurate ground-truth albedo.

This paper addresses critical challenges for removing undesired shading from real-world human images, a task essential for numerous real-world applications. Our work identifies and tackles issues arising from dataset quality by proposing a novel semi-supervised deep learning method that effectively leverages both synthetic and real-world data for improved disentanglement of texture from shading.

Fig. 6: Illustration of our global sparsity loss: (b) the results of applying only the baseline loss $\mathcal{L}_{bas}$, (c) results without applying $\mathcal{L}_{GS2}$, (d) without applying $\mathcal{L}_{GS1}$, and (e) is the results of the full method trained with both $\mathcal{L}_{GS1}$ and $\mathcal{L}_{GS2}$.

The incorporation of a global sparsity constraint and its adaptation to unlabeled real-world data leads to significant improvements in handling a diverse range of fashion-related textures in real-world tests. Our proposed domain adaptation module, based on our insights into gamma adjustments, effectively narrows the distribution gap between real and synthetic data, thereby enhancing the overall performance of shading removal. Through experimentation, we have demonstrated our approach achieves state-of-the-art results, showcasing advancements in the removal of undesired shading while preserving the integrity of original textures in human images.

Acknowledgements: This work was supported by the Entrepreneurial University Programme from the Tertiary Education Commission in New Zealand, and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Number: RS-2024-00399136).

Fig. 7: Ablation on the domain adaptation module on in-the-wild images. Red arrows in the middle row indicate shadowed regions removed or softened by domain adaptation (bottom row). Best viewed when zoomed in.

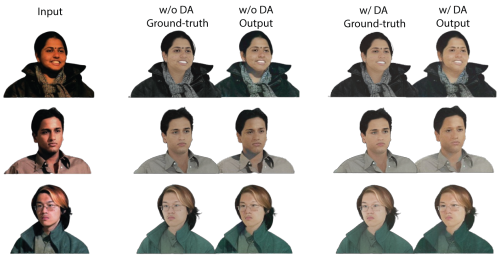

Fig. 8: Ablation on the domain adaptation module on images from the Multi-PIE evaluation dataset. The results w/ DA has less pronounced shadow remaining. Best viewed when zoomed in.

Fig. 9: Limitations: Our method struggles to remove shading from specular surfaces.

References

1. Self-correction-human-parsing. `https://github.com/GoGoDuck912/Self-Correction-Human-Parsing`, accessed: 2023-11-17
2. Laval indoor hdr dataset (2017), `http://indoor.hdrdb.com/`
3. Hdri haven (2020), `https://hdri-haven.com/`
4. Ahmed, A., Farag, A.: A new statistical model combining shape and spherical harmonics illumination for face reconstruction. In: International Symposium on Visual Computing. pp. 531–541. Springer (2007)
5. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2293–2303 (2019)
6. Alldrin, N.G., Mallick, S.P., Kriegman, D.J.: Resolving the generalized bas-relief ambiguity by entropy minimization. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–7. IEEE (2007)
7. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 334–341. IEEE (2012)
8. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE transactions on pattern analysis and machine intelligence **37**(8), 1670–1687 (2014)
9. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
10. Chen, Z., Liu, Z.: Relighting4d: Neural relightable human from videos. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV. pp. 606–623. Springer (2022)
11. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020)
12. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and vision computing **28**(5), 807–813 (2010)
13. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543–7552 (2018)
14. Hou, A., Zhang, Z., Sarkis, M., Bi, N., Tong, Y., Liu, X.: Towards high fidelity face relighting with realistic shadows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14719–14728 (2021)
15. Iqbal, U., Caliskan, A., Nagano, K., Khamis, S., Molchanov, P., Kautz, J.: Rana: Relightable articulated neural avatars. arXiv preprint arXiv:2212.03237 (2022)
16. Ji, C., Yu, T., Guo, K., Liu, J., Liu, Y.: 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. pp. 388–405. Springer (2022)
17. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
18. Kanamori, Y., Endo, Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. arXiv preprint arXiv:1908.02714 (2019)

19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Lagunas, M., Sun, X., Yang, J., Villegas, R., Zhang, J., Shu, Z., Masia, B., Gutierrez, D.: Single-image full-body human relighting. arXiv preprint arXiv:2107.07259 (2021)
21. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). https://doi.org/10.1109/TPAMI.2020.3048039
22. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. Pattern Analysis and Machine Intelligence, IEEE Transactions on **37**(12), 2402–2414 (Dec 2015). https://doi.org/10.1109/TPAMI.2015.2408360
23. Liang, X., Xu, C., Shen, X., Yang, J., Liu, S., Tang, J., Lin, L., Yan, S.: Iccv (2015)
24. Lin, A., Zhao, N., Ning, S., Qiu, Y., Wang, B., Han, X.: Fashiontex: Controllable virtual try-on with text and texture. arXiv preprint arXiv:2305.04451 (2023)
25. Nagano, K., Luo, H., Wang, Z., Seo, J., Xing, J., Hu, L., Wei, L., Li, H.: Deep face normalization. ACM Transactions on Graphics (TOG) **38**(6), 1–16 (2019)
26. Nestmeyer, T., Lalonde, J.F., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5124–5133 (2020)
27. Pandey, R., Escolano, S.O., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: learning to relight portraits for background replacement. ACM Transactions on Graphics (TOG) **40**(4), 1–21 (2021)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
29. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 497–500 (2001)
30. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
31. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In: Computer Vision and Pattern Regognition (CVPR) (2018)
32. Shen, L., Yeo, C., Hua, B.S.: Intrinsic image decomposition using a sparse representation of reflectance. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2904–2915 (2013)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
34. Song, G., Cham, T.J., Cai, J., Zheng, J.: Real-time shadow-aware portrait relighting in virtual backgrounds for realistic telepresence. In: 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 729–738. IEEE (2022)
35. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
36. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P.E., Ramamoorthi, R.: Single image portrait relighting. ACM Trans. Graph. **38**(4), 79–1 (2019)

37. Tajima, D., Kanamori, Y., Endo, Y.: Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In: Computer Graphics Forum. vol. 40, pp. 205–216. Wiley Online Library (2021)
38. Wang, Y., Zhang, L., Liu, Z., Hua, G., Wen, Z., Zhang, Z., Samaras, D.: Face relighting from a single image under arbitrary unknown lighting conditions. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(11), 1968–1984 (2008)
39. Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Transactions on Graphics (TOG) **39**(6), 1–13 (2020)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
41. Weir, J., Zhao, J., Chalmers, A., Rhee, T.: Deep portrait delighting. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
42. Weir, J., Zhao, J., Chalmers, A., Rhee, T.: De-lighting human images using region-specific data augmentation. In: 2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6 (2023)
43. Yeh, Y.Y., Nagano, K., Khamis, S., Kautz, J., Liu, M.Y., Wang, T.C.: 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)
44. Zhang, L., Wang, S., Samaras, D.: Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 209–216. IEEE (2005)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
46. Zhang, X., Barron, J.T., Tsai, Y.T., Pandey, R., Zhang, X., Ng, R., Jacobs, D.E.: Portrait shadow manipulation. ACM Transactions on Graphics (TOG) **39**(4), 78–1 (2020)
47. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7194–7202 (2019)
48. Zhou, T., He, K., Wu, D., Xu, T., Zhang, Q., Shao, K., Chen, W., Xu, L., Yi, J.: Relightable neural human assets from multi-view gradient illuminations. arXiv preprint arXiv:2212.07648 (2022)