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

ACCV 2024
ORAL PAPER

Joshua Weir
josh.weir@vuw.ac.nz
Victoria University of Wellington

Junhong Zhao
j.zhao@vuw.ac.nz
Victoria University of Wellington

Andrew Chalmers
andrew.chalmers@vuw.ac.nz
Victoria University of Wellington

Taehyun Rhee*
taehyun.rhee@unimelb.edu.au
The University of Melbourne

1 Introduction

Fig. 1: Our method estimates albedo (bottom) from input (top), removing shading while preserving texture

1.1 Motivation

- De-lighting is the process of removing unwanted lighting artifacts to extract an image's albedo (Fig. 1). This enables applications such as avatar creation, virtual try-on, and relighting to function effetively under various lighting conditions.

1.2 Challenges

- Full-body de-lighting presents challenges due to varied clothing textures, materials, shadows, and body poses.
- Deep learning solutions require large and diverse datasets, but the high cost of creating labelled datasets limits their size and realism, reducing model robustness to real-world scenarios.

1.3 Contributions

- **Semi-supervised learning pipline** (Fig. 2): Integrates both labelled synthetic data and a vast dataset of unlabelled real-world images to effectively diminish shading artifacts while maintaining non-shading based content.
- **Global Sparsity Loss:** Regularizes texture removal. Facilitates learning a broad range of textures outside the scope of the synthetic dataset.
- **Domain Adaptation:** Aligns real-world images with synthetic renderings. Improving model robustness to real-world shading.

2 Method

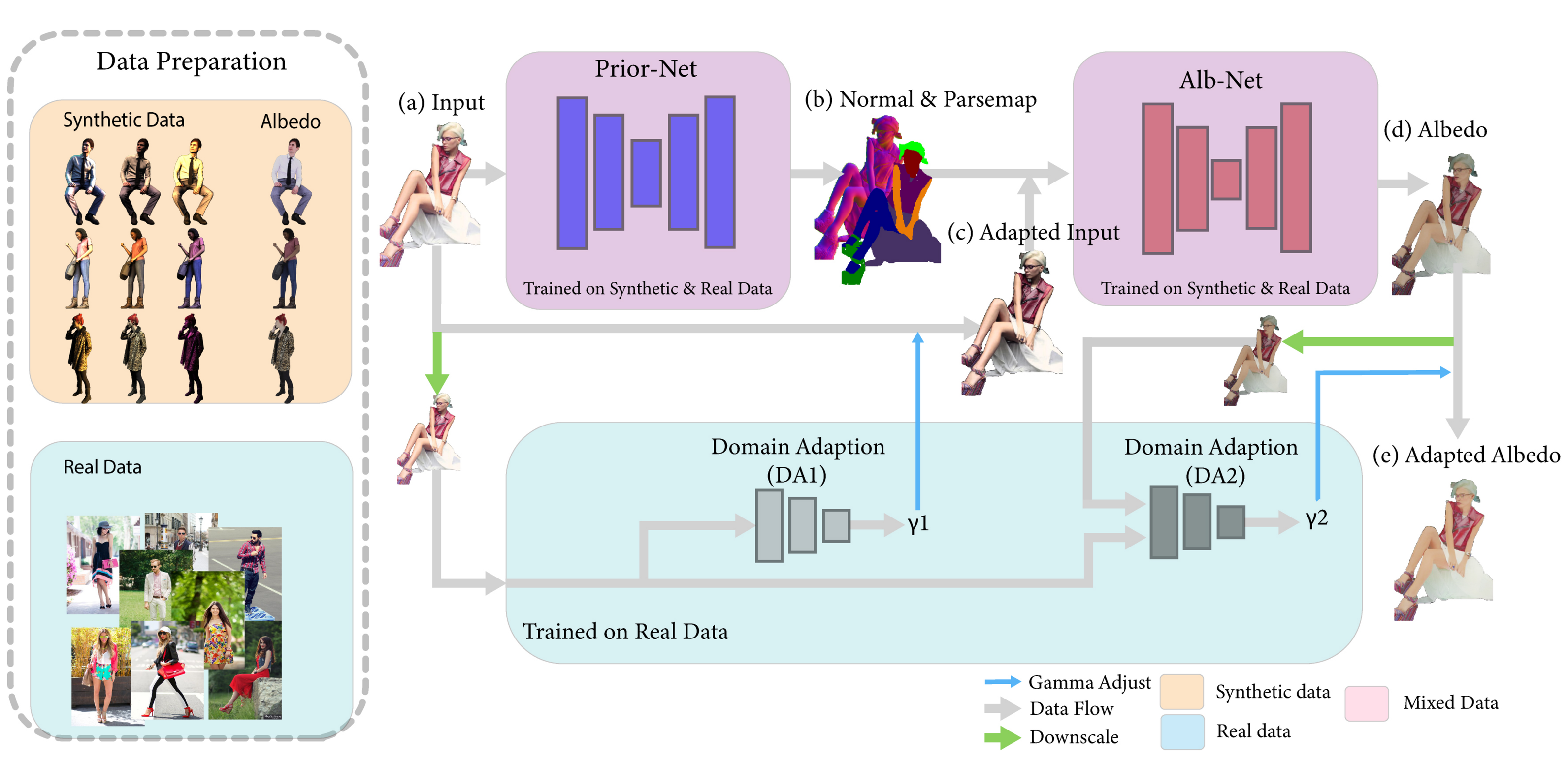

Fig. 2: Our semi-supervised training and inference pipeline for human de-lighting

2.1 Pipeline Design: Our semi-supervised pipeline is illustrated in Fig. 2

- Synthetic data built from 150 3D models (134 for training), each rendered under 154 lighting conditions. Lambertian reflectance only.
- Real data built from a semantic segmentation dataset of 17,706 photographs with semantic parsemap labels.
- Prior-Net and Alb-Net are trained using combined batches of real and synthetic data. L1 loss and VGG-perceptual loss are applied wherever ground-truth labels are available.

2.2 Global Sparsity Loss: For a predicted albedo $\tilde{\mathbf{A}}$ of input image $\mathbf{I}$, the shading map $\frac{\mathbf{I}}{\tilde{\mathbf{A}}}$ should have a sparse colour pallette determined by environment lighting, normals $\mathbf{N}$, and cast shadows. If the color of a certain region is predicted incorrectly, it will result in a significant contrast in the shading map (Fig. 3).

Fig. 3: Global sparsity loss motivation

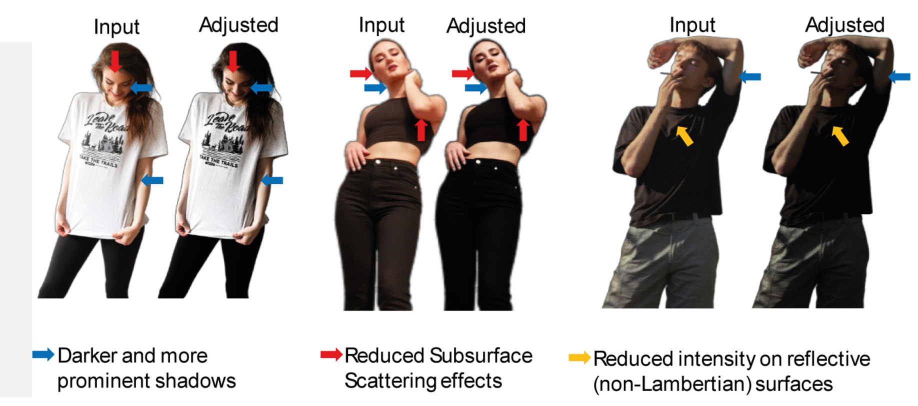

Fig. 4: The results of applying the first gamma adjustment γ1 to real data

To avoid unintended removal of non-shading based content, we enforce shading sparsity with a global sparsity loss $\mathbf{G}(S, N)$ on both labelled and unlabelled when training Alb-Net. This applies minimum entropy to shading S with pixel differences scaled by the cosine angles of the corresponding normals N.

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

Applied to labelled synthetic data

$$\mathcal{L}_{GS2} = \lambda_{GS2} \cdot \mathbf{G}\left(\left\|\frac{\mathbf{I_R}}{\tilde{\mathbf{A}}_\mathbf{R}}\right\|, \tilde{\mathbf{N}}_\mathbf{R}\right)$$

Applied to unlabelled real-world data

Subscripts $\mathbf{S}$ and $\mathbf{R}$ indicate synthetic and real data respectively.
$\mathbf{D_S}$ denotes input $\mathbf{I_S}$ without cast shadows.

2.3 Domain Adaptation: The domain gap between real-world images and rendered 3D models can be narrowed by adjusting the gamma (Fig. 4). Consequently, we train DA1 (Fig. 2) to learn optimal gamma adjustments (γ1) for each input, and DA2 to learn a secondary adjustment (γ2) compensating for brightness adjustments from DA1.

DA1 and DA2 are trained end-to-end with Prior-Net and Alb-Net parameters frozen. The training maximizes total variation loss in the inferred shading to encourage shadow removal, while minimizing global sparsity to prevent texture distortions.

3 Results

3.1 Evaluation: We evaluated our method against the two SOTA methods capable of full-body de-lighting: Total Relighting **(TR)** and Geometry-aware Single-image Full-body Human Relighting **(GSFR)**. Fig. 5 shows ours achieves more accurate shadow removal and texture preservation than prior works.

Fig. 5: Evaluiation against state-of-the-art methods TR and GSFR

3.2 Ablations

- **Global Sparsity:** As shown in Fig. 6, Our full model, trained using both GS1 and GS2 most effectively preserves the original texture.
- **Domain Adaptation:** As shown in Fig. 7, upon integrating the domain adaptation module, intricate shadows and highlights present in skin and clothing regions are effectively eliminated.

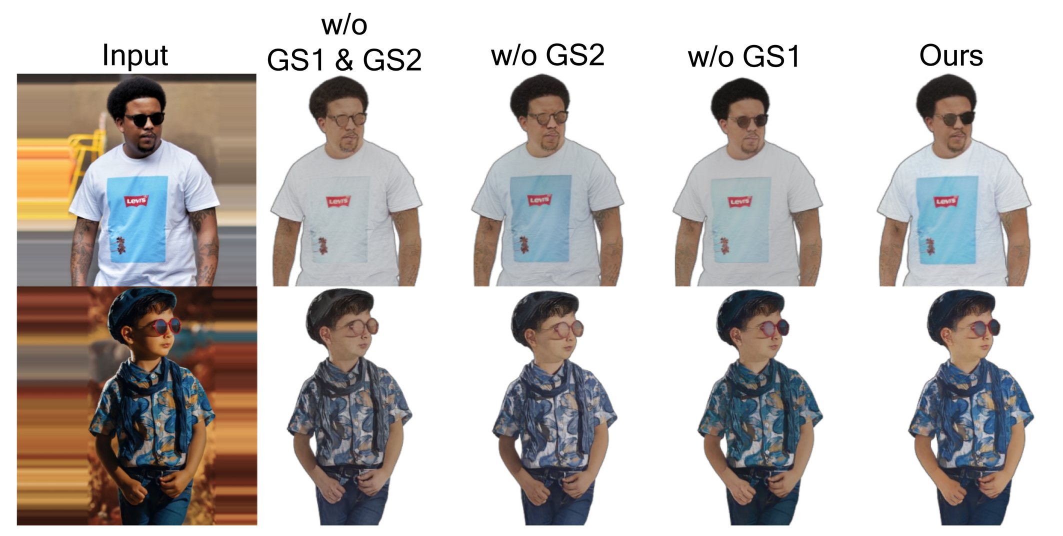

Fig. 6: Global Sparsity Loss ablation

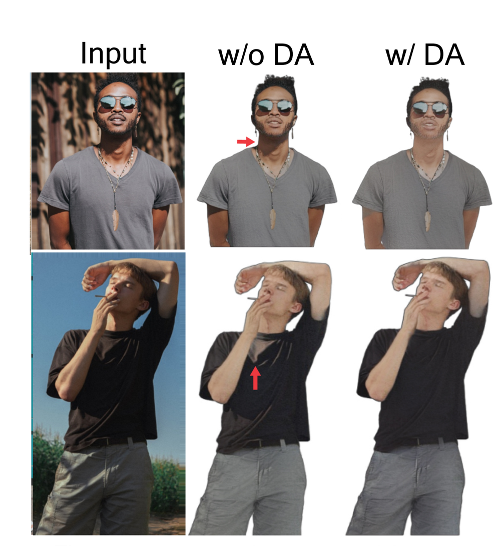

Fig. 7: Domain Adaptation ablation

4 Conclusion

We propose a semi-supervised pipeline to address dataset quality limitations in full-body de-lighting. Our method employs a global sparsity constraint to learn diverse textures and a domain adaptation module to align real and synthetic data, improving shading removal.
Limitations and Future Works: Our method struggles with specular surfaces like leather jackets, absent in the synthetic dataset and indistinguishable from textures despite domain adaptation and global sparsity loss. Future work could explore the separate removal of diffuse and specular lighting artifacts, enabling more precise control over the targeted artifacts using self-supervised loss functions.