Relighting Humans in the Wild:
Monocular Full-Body Human Relighting with Domain Adaptation

D. Tajima[1] Y. Kanamori[1] Y. Endo[1]

[1]University of Tsukuba, Japan

Figure 1: *Given a masked human image, our method can relight the subject with more faithful skin highlights than the method by Kanamori and Endo [KE18] (left). We also achieve stable video relighting even under dynamic illuminations (right).*

Abstract

The modern supervised approaches for human image relighting rely on training data generated from 3D human models. However, such datasets are often small (e.g., Light Stage data with a small number of individuals) or limited to diffuse materials (e.g., commercial 3D scanned human models). Thus, the human relighting techniques suffer from the poor generalization capability and synthetic-to-real domain gap. In this paper, we propose a two-stage method for single-image human relighting with domain adaptation. In the first stage, we train a neural network for diffuse-only relighting. In the second stage, we train another network for enhancing non-diffuse reflection by learning residuals between real photos and images reconstructed by the diffuse-only network. Thanks to the second stage, we can achieve higher generalization capability against various cloth textures, while reducing the domain gap. Furthermore, to handle input videos, we integrate illumination-aware deep video prior *to greatly reduce flickering artifacts even with challenging settings under dynamic illuminations.*

CCS Concepts
• *Computing methodologies* → *Image manipulation; Neural networks;*

1. Introduction

Human image relighting allows us to change lighting effects on a photographed person under a new illumination. We have witnessed rapid, substantial progress in the photorealism of human image relighting, thanks to the deep learning approaches. The current mainstream of learning-based relighting techniques relies on su-

pervised training with datasets generated from 3D human models. However, the scalability or quality of the training datasets causes problems. For example, the specialized capturing device called the *Light Stage* [DHT*00] provides high-quality data for relighting, but the Light Stage data are accessible only to self-sufficient research teams. As such, Light Stage datasets are often small (e.g., 70 individuals [PEL*21]). On the other hand, we can construct a larger synthetic dataset by collecting 3D scanned human models from commercial websites, but most such 3D models support diffuse materials only. Consequently, human relighting techniques suffer from poor generalization capability against, in particular, various cloth shapes and textures and synthetic-to-real domain gap.

To overcome the domain gap while obtaining higher generalization capability, we propose a two-stage supervised method for human image relighting with domain adaptation. The first stage trains a neural network to infer diffuse-only reflection from the 3D models. For this, we extend the network of Kanamori and Endo [KE18] with the spatially-frequency (SF) loss [Ich18] to enhance high-frequency details and a skin mask [WYL*20] to emphasize skin regions in estimating albedo and light. The second stage introduces another network to learn non-diffuse reflection; we apply the diffuse-only network to real photos and reconstruct diffuse-only images. The second-stage network then learns the residuals between the real photos and diffuse-only reconstructions. Because we can collect many 3D models and real photos for training (e.g., 521 3D models and 8,900 photos in our case), our networks can achieve higher generalization capability against various input data.

We further extend our method to handle human videos. Naïvely applying our method frame-by-frame suffers from flickering artifacts. To enhance the temporal consistency, we integrate the *deep video prior* [LXC20], the state-of-the-art video stabilizer. However, we found a problem when we applied dynamic illuminations for video relighting; because the deep video prior stabilizes output frames too strongly, resultant videos do not exhibit time-varying shading effects due to dynamic illuminations. By conditioning the deep video prior with the dynamic illuminations, we can synthesize flicker-tolerant relighting videos with dynamic illuminations. Note that our approach is versatile; it can be applied to any frame-by-frame relighting methods with dynamic illuminations to greatly reduce flickering.

We summarize our contributions as follows:

1. A two-stage framework for human image relighting that supports non-diffuse reflection with domain adaptation,
2. An improved diffuse-only inverse rendering with light occlusions, and
3. An illumination-conditioned deep video prior for flicker-tolerant video relighting with dynamic illuminations.

For the source code and pre-trained network models, please visit our project page[†].

[†] http://cgg.cs.tsukuba.ac.jp/~tajima/pub/relighting_in_the_wild/
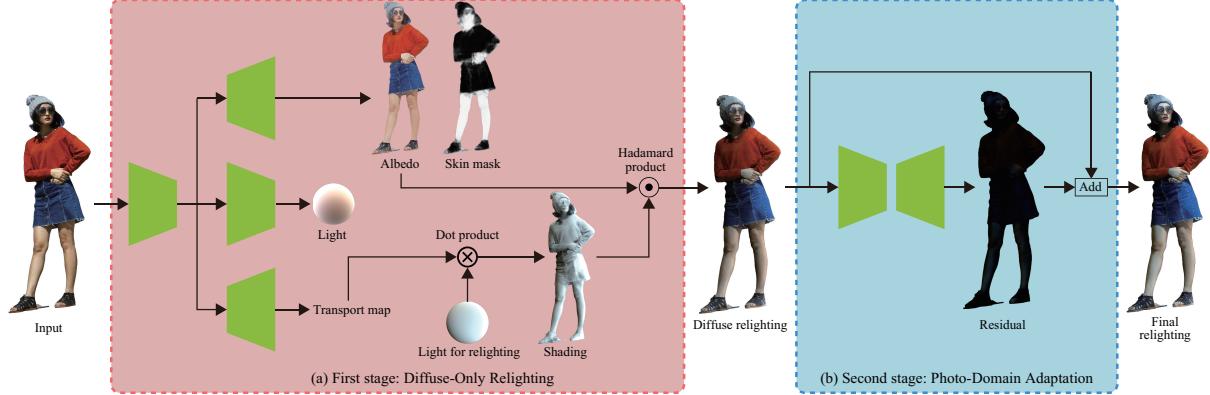
2. Related Work

Because relighting has a long history of research, here we only introduce recent learning-based approaches that target single human images or videos as inputs. The previous studies differ in the targeted body parts (e.g., face only or full-body), training data types (e.g., 3D data or image only), and methodologies (e.g., inverse rendering or GAN inversion [XZY*21]).

Portrait relighting. In the relighting literature, the term "portrait" means that the input images are face-only or headshot (i.e., hair and face with cloth up to chest) photos. Face-only methods [TZK*17; SKCJ18; ZHSJ19] often rely on the fitting of 3D morphable models [BV99; EST*20] to grasp the face geometry. Face relighting is also accomplished by transferring shading effects from reference images [SHS*17; SPB*14]. The recent approaches for headshots [SBT*19; MHP*19; NLML20; ZFT*21; PEL*21] utilize the high-quality data captured using a special device called the *Light Stage* [DHT*00] (or similar ones [WYL*20; ZZW*21]). The Light Stage can capture subjects lit by a single directional light from each light direction, yielding *one-light-at-a-time* (OLAT) images. An OLAT image contains complicated lighting effects (including subsurface scattering and interreflection), and the superposition of OLAT images can represent lighting effects by environmental illuminations. The basic strategy of recent techniques is then to regress OLAT images directly or indirectly from the input headshots [SBT*19; MHP*19; NLML20; ZFT*21; PEL*21] or videos [ZZW*21]. The problem is that Light Stage data are not publicly available but accessible only to the self-sufficient research teams. As such, the Light Stage dataset used in each work is rather small (e.g., only 18 [ZZW*21] or 70 individuals [PEL*21]), resulting in poor generalization capability, particularly in handling cloth texture variations (see, e.g., Section 8 in [PEL*21]). Furthermore, to the best of our knowledge, there are no publicly-available Light Stage datasets for full-body human figures, which are our target in this paper.

Full-body human relighting. Relighting of full-body images or videos has various applications such as human image cut-and-pasting [XADR12] and image-based virtual try-on [Min21] but has not been explored well so far. Kanamori and Endo [KE18] proposed a full-body human image relighting based on inverse rendering using convolutional neural networks (CNNs). The problem is that their method only supports diffuse relighting because their training dataset generated from 3D scanned human figures contains only diffuse textures. Our two-stage method adopts an improved version of their method in the diffuse-only first stage and handles non-diffuse relighting via domain adaptation to real photos in the second stage. We elaborate their method in Section 3.

Domain adaptation. It is well known that supervised methods learning from synthetic data often suffer from the domain gaps between synthetic and real data. A remedy for this is domain adaptation. Domain adaptation has been successfully integrated in various tasks such as 3D human pose estimation [ZWKG19], 3D object reconstruction [PRA19], image dehazing [SLR*20], monocular depth estimation [ZFGT19], semantic segmentation [ZY21], and object detection [SN20]. To the best of our knowledge, our ap-

Figure 2: *Overview of our two-stage relighting approach. The first-stage network infers diffuse-only relighting results, which is then refined in the second-stage network to handle non-diffuse reflection.*

proach is the first attempt of domain adaptation for human image relighting.

GAN inversion. An interesting, promising way of realistic face relighting is to utilize pre-trained GAN networks, namely, *GAN inversion* [XZY*21]. In GAN inversion, an input image is projected into the latent space of a pre-trained GAN network via optimization or an additional encoder, is manipulated in the latent space, and is then re-synthesized as a semantically manipulated yet realistic image. GAN inversion also allows us realistic face relighting [RTD*21; AZMW21] without involving rendering process. Although pre-trained GAN networks for faces are publicly available [Pin21], those for full-body human images are not available and thus we cannot employ GAN inversion for our purpose.

A concurrent work by Lagunas et al. [LSY*21] extends the method by Kanamori and Endo [KE18] by explicitly handling specular reflections with higher-order (i.e., 4th-order) spherical harmonics (SH) and a residual shading term. Their method is orthogonal to ours and could be further enhanced by our domain adaptation technique.

3. Brief Review of Our Baseline [KE18]

We adopt the network by Kanamori and Endo [KE18] as our baseline for the diffuse-only network used in the first stage. Their multi-task network is designed for the inverse rendering of full-body human images, and their three decoder branches extract an albedo map, light coefficients, and a light transport map respectively from each masked image. The light coefficients and light transport maps are represented as coefficient vectors of second-order SH. Unlike the previous methods [SKCJ18; ZHSJ19] that did not consider light occlusions, their method explicitly handles light occlusion at each pixel of the light transport map, yielding realistic shading, particularly in hollowed regions such as armpits, crotches, and garment wrinkles. A drawback is that their method is limited to diffuse reflection because their training dataset is synthesized from 3D scanned human models, which usually contain only diffuse components as textures.

We also point out two more problems in their method, namely, 1) the albedo-light color ambiguity and 2) the output blurriness.

Albedo-light color ambiguity: The image decomposition in reverse rendering holds an intrinsic ambiguity; for example, from a masked photo of blue cloth, we cannot know whether a blue cloth is lit by a white light or a white cloth is lit by a blue light. As a result, although the multiplication of their albedo map and light is close to the input color, each color sometimes becomes strange, particularly if the ground-truth light is dissimilar to any of the training data.

Output blurriness: Although their method uses L1 total variation (TV) losses to enhance image edges, their albedo maps and shading maps (i.e., the multiplication of the light transport map and light) often lose details. For example, the important texture details of faces, e.g., eyes, eyebrows, and lips in albedo maps, are smeared out.

We address these two problems in Section 4.1.
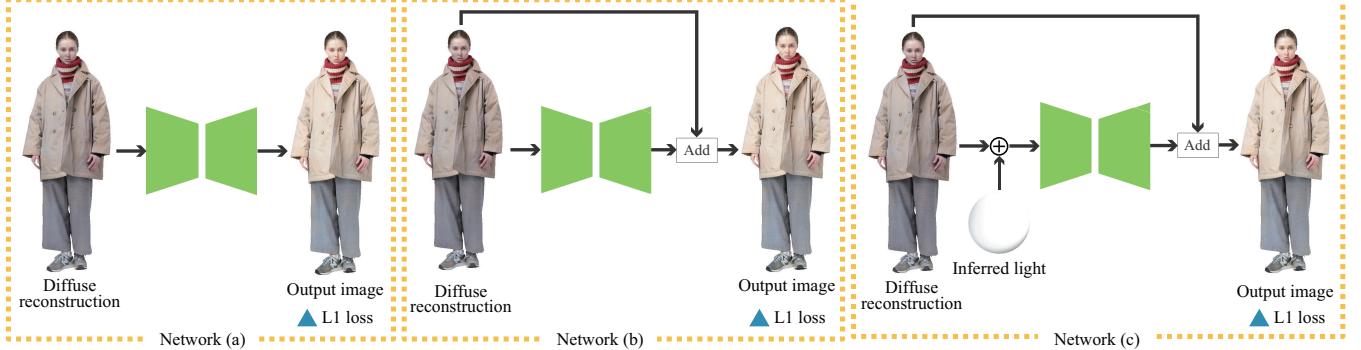
4. Method

Our method takes a single, masked human image as input and applies relighting under a new illumination. We account for diffuse and non-diffuse reflection by a two-stage approach; we synthesize a diffuse-only image in the first stage (Section 4.1) and then compensate non-diffuse components in the second stage (Section 4.2). Regarding the training datasets (Section 5), the first-stage network uses synthetic images rendered from 3D scanned models. In contrast, the second-stage network relies on the outputs of the first stage and real human photos to account for non-diffuse residuals. Figure 2 shows the overview of our method.

Furthermore, we extend our method to human video relighting and demonstrate stable results with much less flickering using our illumination-conditioned deep video prior (Section 4.4).

4.1. First Stage: Diffuse-Only Network

We adopt the baseline network architecture [KE18] with some modifications for the diffuse-only network in the first stage. From

Figure 3: *Three candidate network architectures for the second stage. We adopt (b) because of its best performance.*

a single human image, the diffuse-only network extracts an albedo map, light coefficients, and a light transport map, as explained in Section 3. We can reconstruct a diffuse-only version of the input image by multiplying the three components, i.e., dot products of the light coefficients and light transport map, followed by pixel-wise multiplication of the albedo map. The diffuse-only reconstruction is then fed to the second stage.

Here we propose solutions to address the two problems mentioned in Section 3 as follows:

Solution to albedo-light color ambiguity: The strange output colors in the baseline method partially stem from their small light dataset; they only used 40 lights for training. We thus increase the training data for lights from 40 to 2,760 by carefully selecting naturally-looking lights. Also, we disambiguate the colors of albedo maps and lights based on human-intrinsic color constancy, inspired by Wang et al. [WYL*20]; because skin colors are often stable in albedo maps, we integrate a mechanism to emphasize skin region by inferring an albedo map and a skin mask simultaneously (see Section 4.3 for the implementation detail). Note that Wang et al. extracts face masks to emphasize face regions, but face regions are often too small in our target images (i.e., full-body images). In contrast, our skin mask covers a face and other regions such as arms and legs.

Solution to output blurriness: Instead of TV losses used in the baseline [KE18], we introduce the spatially-frequency (SF) loss [Ich18] for optimizing albedo and shading maps to enhance their details. The SF loss is a weighted sum of L2 losses for images filtered by the normalized Laplacian-of-Gaussians (LoG) with different sigmas. The normalied LoG kernel is defined as follows:

$$G_{\sigma}(\mathbf{x}) = -\frac{1}{2\pi\sigma^2}\left(2 - \frac{\|\mathbf{x}\|^2}{\sigma^2}\right)\exp\left(-\frac{\|\mathbf{x}\|^2}{\sigma^2}\right), \qquad (1)$$

where σ is the scale and $\mathbf{x}$ is the pixel position within the kernel. For each subband i, we use different $\sigma_i \in \{0.6, 1.2, 2.4, 4.8, 9.6, 19.2\}$ and weights $w_i \in \{600, 500, 400, 20, 10, 10\}$ to put higher emphasis on higher-frequency subbands. Let A and S be the ground-truth albedo and shading maps and $\tilde{A}$ and $\tilde{S}$ be their inferred counterparts. The SF losses $\mathcal{L}_{SF}^{albedo}$ and $\mathcal{L}_{SF}^{shading}$ for albedo and shading

Table 1: *Quantitative comparison of three network architecture for the second stage. The best value for each metric is in boldface.*

	RMSE ↓	SSIM ↑	LPIPS ↓
(a)	0.0419	0.990	0.0141
(b)	**0.0417**	**0.996**	**0.0059**
(c)	0.0428	0.990	0.0136

maps are then defined as follows:

$$\mathcal{L}_{SF}^{albedo} = \frac{1}{CWH}\sum_{i}^{6} w_i \|G_{\sigma_i} * A - G_{\sigma_i} * \tilde{A}\|_2^2, \qquad (2)$$

$$\mathcal{L}_{SF}^{shading} = \frac{1}{CWH}\sum_{i}^{6} w_i \|G_{\sigma_i} * S - G_{\sigma_i} * \tilde{S}\|_2^2, \qquad (3)$$

where $*$ denotes convolution while C, W, and H are the numbers of channels, width, and height of the maps, respectively. We observe that the SF loss improves high-frequency details particularly in the output albedo maps. We demonstrate the resultant albedo maps in Section 6.

4.2. Second Stage: Refinement Network

In the second stage, the refinement network learns the differences between the masked photos of human figures and diffuse-only counterparts reconstructed in the first stage. We adopt an encoder-decoder network with skip connections as the network architecture. Specifically, we considered the following three architectures (see Figure 3):

(a) A network that directly generates an output containing both diffuse and non-diffuse reflections (Figure 3, left),

(b) A network that outputs the residuals from the diffuse-only reconstruction (Figure 3, middle), and

(c) A variant of (b) with inferred light information (Figure 3, right).

Among these candidates, (a) is the simplest network architecture to directly output non-diffuse residuals. (b) is inspired by ResNet [HZRS16] so that the network outputs the difference between the ground-truth and the first-stage output. (c) is to account for the light specified for relighting.

Table 1 shows the result of quantitative comparison among the three architectures. We can see network (b) was the best, and thus we adopt (b) for the refinement network. One might consider that network (c) should have been better due to the light integration. However, in reality (c) was inferior, probably due to the erroneous lights inferred by our baseline network; note that there are no ground-truth lights for real photos.

4.3. Network Training

Here we summarize how to train our networks. We first train the diffuse-only network in the first stage using a synthetic dataset. After freezing the network parameters, we apply the pre-trained diffuse-only network to real photos of human figures to calculate diffuse-only reconstructions. We then train the refinement network to learn the residuals between the real photos and diffuse-only counterparts.

We explain our network architectures. For the first stage, we borrowed the encoder-decoder network by Kanamori and Endo [KE18] (see Section 5 in their paper); it has three decoders for albedo, light transport, and light, respectively. Their encoder has six convolutional layers, a residual block, and skip-connections with the decoder branches for albedo and light transport maps. We modified their network as follows. We replaced each transposed convolution in the decoders with upsampling and convolutional layers to avoid checkerboard artifacts. In the albedo decoder, we modified the final layer so that it generates not a three- (i.e., RGB albedo map) but a four-channel output (i.e., albedo map and skin mask). For the second stage, we used a U-net-like architecture whose encoder and single-branch decoder are almost identical to those used in the first stage, except for the numbers of input/output channels.

We used the following loss functions for training. The loss function $\mathcal{L}^{1st}$ for the first stage is defined as:

$$\mathcal{L}^{1st} = \sum_{r \in \mathcal{R}} \mathcal{L}_r^{reconst.} + \sum_{s \in \mathcal{S}} \mathcal{L}_s^{shading} + \mathcal{L}_{SF}^{albedo} + \mathcal{L}_{SF}^{shading} + \mathcal{L}^{focal}, \quad (4)$$

where $\mathcal{L}_r^{reconst.}$ and $\mathcal{L}_s^{shading}$ are the reconstruction and shading losses used in the baseline [KE18]; $\mathcal{R}$ and $\mathcal{S}$ are sets of subscripts for ten types of reconstruction losses (i.e., three unary losses for albedo map, light transport map, and light, and seven patterns of triple products of albedo map, light transport map, and light) and three types of shading losses (i.e., three patterns of double products of light transport map and light), respectively. Please refer to the paper [KE18] for more details. As explained in Section 4.1, we replace their TV losses with our SF losses $\mathcal{L}_{SF}^{albedo}$ and $\mathcal{L}_{SF}^{shading}$. We also added a focal loss $\mathcal{L}^{focal}$ [WYL*20] for the skin mask. Each loss in $\mathcal{L}^{1st}$ has the same weight (i.e., 1). In the second stage, we only used an L1 loss between the final outputs and real photos.

4.4. Flicker-Tolerant Human Video Relighting

We also propose a simple extension of our method to human video relighting. We apply our method for each frame of an input video separately. However, as is often the case with frame-by-frame image processing, this straightforward approach suffers from flickering because we cannot guarantee temporal consistency. To remove

the flickering artifacts, we adopt the *deep video prior* (DVP) proposed by Lei et al. [LXC20]. Inspired by the finding [UVL18] that CNNs learn low-dimensional manifolds faster than noise, Lei et al. extended this idea to the temporal dimension to reduce temporal noise (i.e., flickering), resulting in a simple formulation with excellent noise reduction. Unlike previous approaches [BTS*15; LHW*18], DVP does not need additional training datasets or estimating optical flow. To apply DVP, we train an untrained network (of the same architecture as the second stage) to fit the frames of an input video (i.e., the original video before applying relighting) to those of the relit video containing flickering, only with the perceptual loss. We then terminate the training manually before noise appears in the result.

We confirmed that DVP removes flickering excellently with static illuminations but causes a problem with dynamic illuminations. If we animate an illumination (by, e.g., rotating it), flickering artifacts are removed, but the illumination looks as if it were static; we cannot see time-varying shading effects due to the dynamic illumination. This is probably because some input frames are similar, but their ground-truth outputs are different from each other; in such a case, the network has to infer different outputs from similar inputs by learning one-to-many correspondence, which is difficult to learn. To account for dynamic illumination, we condition the network by feeding light information in addition to each frame. Specifically, we vectorize the SH coefficients of the illumination for each frame and replicate it for each pixel in the frame. We then concatenate the input frame and the tiled SH coefficient vectors to feed the network. The SH coefficients serve as a powerful cue to distinguish similar input frames, enabling time-varying shading effects due to dynamic illuminations.

5. Dataset

We gathered 541 3D scanned models in standing poses from several commercial websites for the first stage and used 521 for training and 20 for test. For each 3D model, we rendered a binary mask, albedo map, and light transport map at the resolution of 1024×1024, as done in the baseline [KE18]. We collected HDR environment maps from HDRI Heaven[‡] and calculated their SH coefficients after rotating them at $36°$ degree intervals. We used the resultant 2,760 lights for training. For test, we used the 10 lights provided by the authors [KE18]. To prepare "ground-truth" skin masks for our synthetic human image dataset, we applied a semantic segmentation network trained with the *ModaNet* [ZYKP18] dataset, which does not contain "skin" labels. We extracted skin masks as the intersections of our binary masks and "background" regions in the ModaNet dataset.
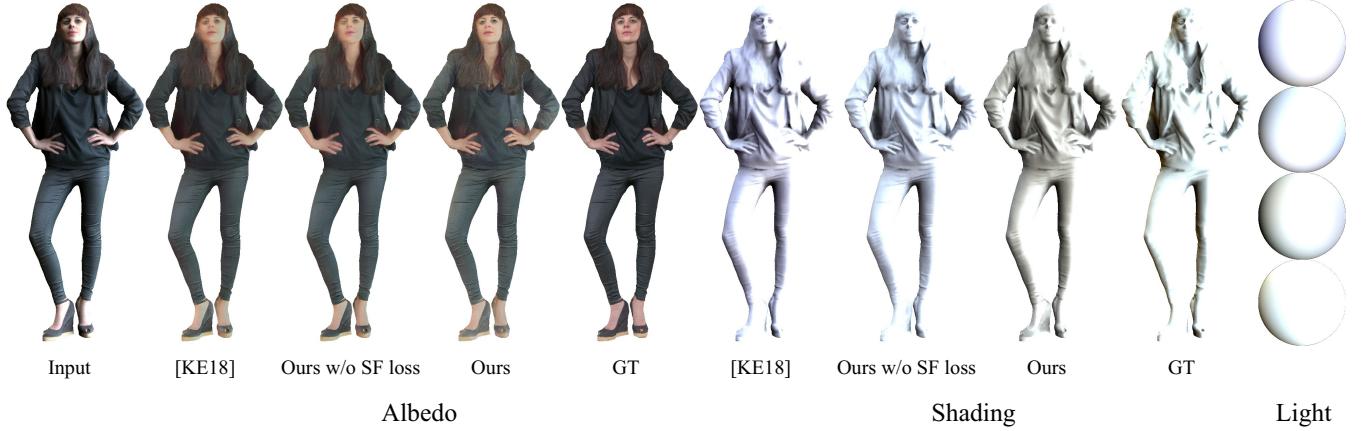
For the second stage, we collected 9,152 photos of standing fashion models from a website of an apparel company and used 8,900 photos for training and 252 for test. We extracted binary masks for the real photos using a commercial service[§]. The photos and binary masks were uniformly resized to 1024×1024 with appropriate paddings.

[‡] https://hdrihaven.com/
[§] https://remove.bg/

Table 2: *Quantitative ablation study for our first stage (plus [SKCJ18]) with synthetic data. The best scores are in boldface.*

	RMSE ↓				SSIM ↑			LPIPS ↓		
	Albedo	Light	Shading	Reconst.	Albedo	Shading	Reconst.	Albedo	Shading	Reconst.
[SKCJ18]	0.101	0.290	0.274	0.0581	0.982	0.813	0.994	0.0280	0.0888	0.0102
[KE18]	0.104	0.265	0.207	0.0600	0.979	0.969	0.992	0.0447	0.0444	0.0177
Ours w/o SF loss	0.0996	0.142	0.120	0.0428	0.977	0.979	0.992	0.0494	0.0413	0.0228
Ours	**0.0932**	**0.131**	**0.112**	**0.0377**	**0.986**	**0.982**	**0.996**	**0.0229**	**0.0292**	**0.00650**

| Albedo | Shading | Light |
| Input   [KE18]   Ours w/o SF loss   Ours   GT   [KE18]   Ours w/o SF loss   Ours   GT |

Figure 4: *Qualitative evaluation for our first stage with synthetic data. Lights are sorted in the same order as albedo and shading maps, from top to bottom.*

Table 3: *Quantitative evaluation of output images reconstructed from real photos. The best scores are in boldface.*

	RMSE ↓	SSIM ↑	LPIPS ↓
[SKCJ18]	0.0749	0.988	0.0130
[ZHSJ19]	0.0813	0.983	0.0201
[KE18]	0.0783	0.989	0.0197
Ours	**0.0417**	**0.996**	**0.0059**

(4) Evaluation of human video relighting.

As the evaluation metrics, we calculated RMSE within binary masks, SSIM and LPIPS [ZIE*18] within bounding boxes of masks to exclude pixels out of masks or bounding boxes. Because real photos do not have ground-truth relighting results, we measured the errors between input images and reconstructed images for quantitative evaluations. Our results in this paper were generated from the test data only.

6. Results

We implemented our method using Python and PyTorch and conducted training and inference on NVIDIA GeForce GTX 1080 Ti. We used Adam as an optimizer for the first stage and RAdam for the second stage, setting the exponential decay rates for the moment estimates as {0.5, 0.999}. We scheduled the learning rates in the range of [0.001, 0.0001] using the cosine annealing within 20 epochs per cycle. Our batch size was one. The computational time for one-epoch training was about 30 minutes when we used one GPU to process 1024×1024 images. We terminated the training at 180 epochs, where the learning curves reached a plateau. The time for testing a 1024×1024 input image was about 0.18 seconds.

To evaluate our method, we conducted the following experiments:

(1) Ablation study of the first stage,
(2) Difference evaluation between the first and second stages,
(3) Comparison with existing relighting methods, and

6.1. Ablation Study of the 1st Stage

We evaluated the effectiveness of our improvements (i.e., integration of the skin mask and SF loss, explained in Section 4.1) over the original diffuse-only baseline [KE18]. Specifically, we compared 1) the baseline [KE18], 2) "Ours w/o SF loss" (i.e., skin mask only), and 3) "Ours" (i.e., with both skin mask and SF loss). We trained the baseline [KE18] with the same training data (i.e., both lights and 3D models) as ours for a fair comparison. Tables 2 and 3 show the results of quantitative evaluations with synthetic data and real photos as inputs, respectively. Both tables demonstrate that "Ours" recoded the best scores in all metrics. Figures 4 and 5 show the qualitative comparisons with synthetic data and real photos as inputs, respectively. We can see that skin masks suppress highlights in skin regions and thus help recovering natural skin colors in the albedo maps. With the SF loss, high-frequency details are extracted in the albedo maps as well as the reconstructed images.

| Input | [KE18] | Ours w/o SF Loss | Ours | [KE18] | Ours w/o SF Loss | Ours |

Albedo | Reconstruction

Figure 5: *Qualitative evaluation for our first stage with real photos.*

6.2. Difference Evaluation between the 1st and 2nd Stages

Figure 6 shows the reconstructed results of the first (i.e., diffuse-only) and second (i.e., refinement) stages as well as their differ-ences. The RMSE scores under the reconstructed images are al-most halved from the first to the second stages. The blueish im-ages visualize per-pixel absolute differences. The first-stage results have large errors particularly in skin regions whereas the errors in

RMSE	0.0506	0.0293			RMSE	0.0514	0.0297			
RMSE	0.0419	0.0234			RMSE	0.0499	0.0257			
Input	First stage reconstruction	Second stage reconstruction	First stage error	Second stage error	Input	First stage reconstruction	Second stage reconstruction	First stage error	Second stage error	

Figure 6: *Comparison of our first- and second-stage outputs. The blueish images encode per-pixel absolute errors.*

the second stage are reduced substantially, which implies that the glossiness (e.g., skin highlights) is recovered in the second stage.

6.3. Comparison with Existing Relighting Methods

We compared our method with three existing relighting methods [SKCJ18; ZHSJ19; KE18]. Both of [SKCJ18; ZHSJ19] were originally proposed for face relighting. We trained [SKCJ18] with our synthetic dataset. In the training of [ZHSJ19], we had to fabricate normal maps for photographed people, which we synthesized using a recent method for single-image 3D human reconstruction [SSSJ20]. We then trained our own implementation of [ZHSJ19] with real photos obtained from an apparel website and lights inferred using the baseline [KE18].

Table 3 demonstrates that our method outperforms the others in all metrics w.r.t. the reconstruction errors. Figure 7 shows the qualitative comparison, where the rectangles emphasize skin highlights (red), occluded regions (blue), and albedo differences (green). Overall the method by Zhou et al. [ZHSJ19] did not work well probably because it could not learn light variations because fashion models in apparel websites are almost always front-lit. As pointed out by Kanamori and Endo [KE18], [SKCJ18] makes occluded regions (blue rectangles) unnaturally bright because it ignores light occlusion. [KE18] generates blurry albedo maps (red rectangles) with possibly wrong colors (green rectangles). Ours yields more natural skin colors with highlights (red rectangles).

To further validate our method, we conducted a user study based on the Friedman test [SC88]. Using six sets of relighting results shown in our paper, we requested 20 subjects to rank four methods (i.e., [ZHSJ19], [SKCJ18], [KE18], and ours) for each set via an online survey using the Google Forms, asking the naturalness of images relit under given lights. The display order of the four methods was shuffled for each set. Consequently, we obtained 120 (= 6×20) samples. We summed up each sample's rank (i.e., 1, 2, 3, or 4) for each method. The rank sums (the smaller the better) of the four methods are 417, 314, 297, and 172, respectively, which means ours is the best. We confirmed that the ranking has a statistically significant difference ($p < 0.001$). Furthermore, post-hoc tests revealed that ours outperforms the others statistically significantly.

6.4. Evaluation of Human Video Relighting

We obtained the input videos of Zablotskaia et al. [ZSZS19], and then applied our method in the following three different ways:

"w/o DVP": Naïvely applying our method frame by frame,
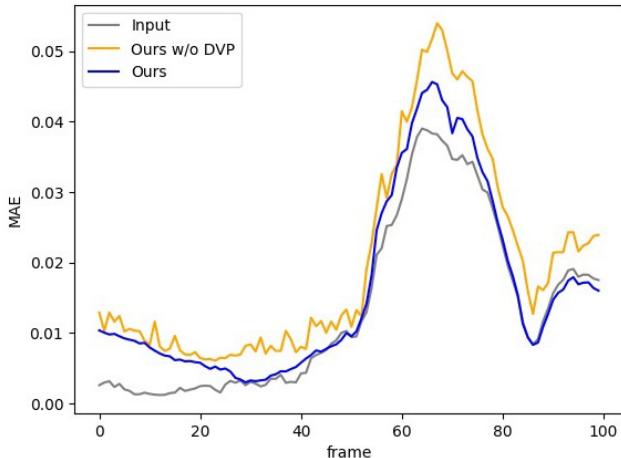
Figure 7: *Comparison of relighting results. The rectangles emphasize skin highlights (red), occluded regions (blue), and albedo differences (green).*

"w/ DVP": Simply applying DVP to our initial results, and
"w/ DVP + light": Applying DVP with light conditioning.

Figure 10 shows the resultant frames with two video sequences. **Please watch the accompanying video to see the video outputs.** When we applied our method without DVP, there were heavy flickering artifacts in the outputs. When we employed DVP, the flickering artifacts were greatly reduced. However, DVP stabilizes the output too strongly that it even cancels the time-varying shading effects by dynamic illuminations. Consequently, the females subjects look as if they were lit by static lights. With our light-aware DVP, we can account for dynamic illumination, while reducing flickering greatly. The graphs in Figure 8 plot the differences between two consecutive frames of the output video sequences (Figure 10, left) with a dynamic illumination. Specifically, the differences are calcu-

Figure 8: *Graphs of temporal differences between two consecutive video frames.*

Input   Albedo map   Shading map   Reconstruction

Figure 9: *Failure case of albedo-shading decomposition in the first stage.*

lated as mean absolute error (MAE) within intersections of masks in two consecutive frames. Although this naïve MAE measurement is not the best way because we do not consider correspondences among frames, we can see the tendency that our light-aware DVP (the orange graph) yields lower errors than without DVP (the blue graph).

7. Limitations

Although our method can handle non-diffuse relighting, it inherits the limitations of the diffuse-only baseline [KE18]. If texture patterns are complicated or lighting conditions are extraordinary, the albedo-light color ambiguity is hard to resolve. Consequently, albedo and shading maps might have wrong colors (see Figure 9 as an example), causing poor relighting. To address this issue, we would like to provide richer information to the networks; the light inference will be improved if we feed not only foreground images

but also background images. For better albedo inference, a larger training dataset will help networks learn richer albedo variations.

Our light-conditioned DVP greatly reduces flickering under dynamic illuminations but does not completely eliminate it. We hypothesize that this is because the light information provided by SH coefficients is insufficient. We would like to explore a better approach in future.

8. Conclusions

In this paper, we have proposed a two-stage supervised method for relighting full-body human images or videos, while enhancing generalization capability and reducing the synthetic-to-real domain gap. The first stage infers diffuse-only relighting better than the baseline [KE18], thanks to skin masks [WYL*20] and the SF loss [Ich18]. The second stage compensates the synthetic-to-real gap by inferring the differences between diffuse-only relighting and real photos. We further accomplished stable human video relighting by conditioning the deep video prior [LXC20] with light information. As a result, we can achieve high temporal coherence even under dynamic illumination.

There are several directions for future work. Although we did not compare our method with Light-Stage-based methods [SBT*19; WYL*20; PEL*21] due to the lack of publicly-available full-body image datasets, we would like to confirm their advantage (i.e., the ability to account for arbitrary illumination effects such as hard shadows, glossy reflections, subsurface scattering, and inter-reflection) in our setting. We also would like to account for the background image for better light inference.

References

[AZMW21] ABDAL, RAMEEN, ZHU, PEIHAO, MITRA, NILOY J., and WONKA, PETER. "StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows". *ACM Trans. Graph.* 40.3 (May 2021) 3.

[BTS*15] BONNEEL, NICOLAS, TOMPKIN, JAMES, SUNKAVALLI, KALYAN, et al. "Blind video temporal consistency". *ACM Trans. Graph.* 34.6 (2015), 196:1–196:9 5.

[BV99] BLANZ, VOLKER and VETTER, THOMAS. "A Morphable Model for the Synthesis of 3D Faces". *Proc. of SIGGRAPH 1999*. Ed. by WAGGENSPACK, WARREN N. ACM, 1999, 187–194 2.

[DHT*00] DEBEVEC, PAUL E., HAWKINS, TIM, TCHOU, CHRIS, et al. "Acquiring the reflectance field of a human face". *Proc. of SIGGRAPH 2000*. Ed. by BROWN, JUDITH R. and AKELEY, KURT. ACM, 2000, 145–156 2.

[EST*20] EGGER, BERNHARD, SMITH, WILLIAM A. P., TEWARI, AYUSH, et al. "3D Morphable Face Models—Past, Present, and Future". *ACM Trans. Graph.* 39.5 (June 2020) 2.

[HZRS16] HE, KAIMING, ZHANG, XIANGYU, REN, SHAOQING, and SUN, JIAN. "Deep Residual Learning for Image Recognition". *CVPR 2016*. June 2016 4.

[Ich18] ICHIMURA, NAOYUKI. "Spatial Frequency Loss for Learning Convolutional Autoencoders". *CoRR* abs/1806.02336 (2018). arXiv: 1806.02336 2, 4, 10.

[KE18] KANAMORI, YOSHIHIRO and ENDO, YUKI. "Relighting humans: occlusion-aware inverse rendering for full-body human images". *ACM Trans. Graphics* 37.6 (Dec. 2018), 1–11 1–6, 8, 10.

[LHW*18] LAI, WEI-SHENG, HUANG, JIA-BIN, WANG, OLIVER, et al. "Learning Blind Video Temporal Consistency". *ECCV 2018*. Vol. 11219. Lecture Notes in Computer Science. Springer, 2018, 179–195 5.

[LSY*21] LAGUNAS, MANUEL, SUN, XIN, YANG, JIMEI, et al. "Single-image Full-body Human Relighting". *32nd Eurographics Symposium on Rendering (EGSR 2021)*. Eurographics Association, 2021, 167–177 3.

[LXC20] LEI, CHENYANG, XING, YAZHOU, and CHEN, QIFENG. "Blind Video Temporal Consistency via Deep Video Prior". *NeurIPS 2020*. 2020 2, 5, 10.

[MHP*19] MEKA, ABHIMITRA, HÄNE, CHRISTIAN, PANDEY, ROHIT, et al. "Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination". *ACM Trans. Graph.* 38.4 (2019), 77:1–77:12 2.

[Min21] MINAR, MATIUR RAHMAN. *A Curated List of Awesome Virtual Try-on (VTON) Research*. https://github.com/minar09/awesome-virtual-try-on. 2021 2.

[NLML20] NESTMEYER, THOMAS, LALONDE, JEAN-FRANÇOIS, MATTHEWS, IAIN A., and LEHRMANN, ANDREAS M. "Learning Physics-Guided Face Relighting Under Directional Light". *CVPR 2020*. IEEE, 2020, 5123–5132 2.

[PEL*21] PANDEY, ROHIT, ESCOLANO, SERGIO ORTS, LEGENDRE, CHLOE, et al. "Total Relighting: Learning to Relight Portraits for Background Replacement". *ACM Trans. Graph.* 40.4 (Jan. 2021) 2, 10.

[Pin21] PINKNEY, JUSTIN. *Awesome Pretrained StyleGAN2*. https://github.com/justinpinkney/awesome-pretrained-stylegan2. 2021 3.

[PRA19] PINHEIRO, PEDRO O., ROSTAMZADEH, NEGAR, and AHN, SUNGJIN. "Domain-Adaptive Single-View 3D Reconstruction". *ICCV 2019*. IEEE, 2019, 7637–7646 2.

[RTD*21] R., MALLIKARJUN B., TEWARI, AYUSH, DIB, ABDALLAH, et al. "PhotoApp: Photorealistic Appearance Editing of Head Portraits". *CoRR* abs/2103.07658 (2021) 3.

[SBT*19] SUN, TIANCHENG, BARRON, JONATHAN T., TSAI, YUN-TA, et al. "Single image portrait relighting". *ACM Trans. Graphics* 38.4 (July 2019), 1–12 2, 10.

[SC88] SIEGEL, S. and CASTELLAN, N.J. *Nonparametric statistics for the behavioral sciences*. Second. McGraw–Hill, Inc., 1988 8.

[SHS*17] SHU, ZHIXIN, HADAP, SUNIL, SHECHTMAN, ELI, et al. "Portrait lighting transfer using a mass transport approach". *ACM Trans. Graph.* 36.4 (2017) 2.

[SKCJ18] SENGUPTA, SOUMYADIP, KANAZAWA, ANGJOO, CASTILLO, CARLOS D., and JACOBS, DAVID W. "SfSNet: Learning Shape, Reflectance and Illuminance of Faces 'in the Wild'". *CVPR 2018*. June 2018 2, 3, 6, 8.

[SLR*20] SHAO, YUANJIE, LI, LERENHAN, REN, WENQI, et al. "Domain Adaptation for Image Dehazing". *CVPR 2020*. IEEE, 2020, 2805–2814 2.

[SN20] SASAGAWA, YUKIHIRO and NAGAHARA, HAJIME. "YOLO in the Dark - Domain Adaptation Method for Merging Multiple Models". *ECCV 2020*. Vol. 12366. Lecture Notes in Computer Science. Springer, 2020, 345–359 2.

[SPB*14] SHIH, YI-CHANG, PARIS, SYLVAIN, BARNES, CONNELLY, et al. "Style transfer for headshot portraits". *ACM Trans. Graph.* 33.4 (2014), 148:1–148:14 2.

[SSSJ20] SAITO, SHUNSUKE, SIMON, TOMAS, SARAGIH, JASON M., and JOO, HANBYUL. "PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization". *CVPR 2020*. IEEE, 2020, 81–90 8.

[TZK*17] TEWARI, AYUSH, ZOLLHÖFER, MICHAEL, KIM, HYEONG-WOO, et al. "MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction". *ICCV 2017*. IEEE Computer Society, 2017, 3735–3744 2.

[UVL18] ULYANOV, DMITRY, VEDALDI, ANDREA, and LEMPITSKY, VICTOR S. "Deep Image Prior". *CVPR 2018*. 2018, 9446–9454 5.

[WYL*20] WANG, ZHIBO, YU, XIN, LU, MING, et al. "Single image portrait relighting via explicit multiple reflectance channel modeling". *ACM Trans. Graph.* 39.6 (2020), 220:1–220:13 2, 4, 5, 10.

[XADR12] XUE, SU, AGARWALA, ASEEM, DORSEY, JULIE, and RUSHMEIER, HOLLY E. "Understanding and improving the realism of image composites". *ACM Trans. Graph.* 31.4 (2012), 84:1–84:10 2.

[XZY*21] XIA, WEIHAO, ZHANG, YULUN, YANG, YUJIU, et al. *GAN Inversion: A Survey*. 2021. arXiv: 2101.05278 [cs.CV] 2, 3.

[ZFGT19] ZHAO, SHANSHAN, FU, HUAN, GONG, MINGMING, and TAO, DACHENG. "Geometry-Aware Symmetric Domain Adaptation for Monocular Depth Estimation". *CVPR 2019*. Computer Vision Foundation / IEEE, 2019, 9788–9798 2.

[ZFT*21] ZHANG, XIUMING, FANELLO, SEAN RYAN, TSAI, YUN-TA, et al. "Neural Light Transport for Relighting and View Synthesis". *ACM Trans. Graph.* 40.1 (Jan. 2021) 2.

[ZHSJ19] ZHOU, HAO, HADAP, SUNIL, SUNKAVALLI, KALYAN, and JACOBS, DAVID W. "Deep Single-Image Portrait Relighting". *ICCV 2019*. Oct. 2019 2, 3, 8.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A., et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". *CVPR 2018*. IEEE Computer Society, 2018, 586–595 6.

[ZSZS19] ZABLOTSKAIA, POLINA, SIAROHIN, ALIAKSANDR, ZHAO, BO, and SIGAL, LEONID. "DwNet: Dense warp-based network for pose-guided human video generation". *30th British Machine Vision Conference 2019 (BMVC 2019)*. BMVA Press, 2019, 51 8.

[ZWKG19] ZHANG, XIHENG, WONG, YONGKANG, KANKANHALLI, MOHAN S., and GENG, WEIDONG. "Unsupervised Domain Adaptation for 3D Human Pose Estimation". *Proc. of the 27th ACM International Conference on Multimedia (MM 2019)*. ACM, 2019, 926–934 2.

[ZY21] ZHENG, ZHEDONG and YANG, YI. "Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation". *Int. J. Comput. Vis.* 129.4 (2021), 1106–1120 2.

[ZYKP18] ZHENG, SHUAI, YANG, FAN, KIAPOUR, M. HADI, and PIRAMUTHU, ROBINSON. "ModaNet: A Large-scale Street Fashion Dataset with Polygon Annotations". *2018 ACM Multimedia Conference on Multimedia Conference*. Ed. by BOLL, SUSANNE, LEE, KYOUNG MU, LUO, JIEBO, et al. ACM, 2018, 1670–1678 5.

[ZZW*21] ZHANG, LONGWEN, ZHANG, QIXUAN, WU, MINYE, et al. "Neural Video Portrait Relighting in Real-time via Consistency Modeling". *CoRR* abs/2104.00484 (2021) 2.

Figure 10: *Relighting results for human videos. Please watch the accompanying video for the details.*