This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Learning Physics-guided Face Relighting under Directional Light

Thomas Nestmeyer*
MPI for Intelligent Systems

Jean-François Lalonde
Université Laval

Iain Matthews
Epic Games

Andreas Lehrmann*
Borealis AI

Abstract

Relighting is an essential step in realistically transferring objects from a captured image into another environment. For example, authentic telepresence in Augmented Reality requires faces to be displayed and relit consistent with the observer's scene lighting. We investigate end-to-end deep learning architectures that both de-light *and* relight *an image of a human face. Our model decomposes the input image into intrinsic components according to a diffuse physics-based image formation model. We enable non-diffuse effects including cast shadows and specular highlights by predicting a residual correction to the diffuse render. To train and evaluate our model, we collected a portrait database of 21 subjects with various expressions and poses. Each sample is captured in a controlled light stage setup with 32 individual light sources. Our method creates precise and believable relighting results and generalizes to complex illumination conditions and challenging poses, including when the subject is not looking straight at the camera.*

Supplementary material can be found on our project page https://lvsn.github.io/face-relighting

1. Introduction

In recent years Augmented Reality (AR) has seen widespread interest across a variety of fields, including gaming, communication, and remote work. For an AR experience to be immersive, the virtual objects inserted in the environment should match the lighting conditions of their observed surroundings, even though they were originally captured under different lighting. This task, known as *relighting*, has a long history in computer vision with many seminal works paving the way for modern AR technologies [2, 3, 21, 34, 39].

Relighting is often represented as a physics-based, two-stage process. First, *de-light* the object in order to recover its intrinsic properties of reflectance, geometry, and lighting. Second, *relight* the object according to a desired target lighting. This implies an exact instantiation of the rendering equation [17] operating on lighting and surface reflectance representations capable of capturing the true nature of the light-material-geometry interactions. In practice, errors occur due to imperfect parametric models or assumptions. One common approximation is to assume diffuse materials [2, 39]. Another approximation is smooth lighting, e.g. modeled as low-order spherical harmonics, which cannot produce hard shadows cast from point light sources like the sun. We consider the hard problem of relighting human faces, which are known for both their complex reflectance properties including subsurface scattering, view-dependent and spatially-varying reflectance, but also for our perceptual sensitivity to inaccurate rendering. Recent image-to-image translation approaches rely on deep learning architectures (e.g. [16]) that make no underlying structural assumption about the (re)lighting problem. Given enough representational capacity an end-to-end system can describe any underlying process, but is prone to large variance due to over-parameterization, and poor generalization due to physically implausible encodings. Test-time manipulation is also difficult with a semantically meaningless internal state. While this could potentially be alleviated with more training data,

(a) ℓ_{src}, input (b) ℓ_{dst}, prediction (c) ℓ_{dst}, ground truth

Figure 1: **Overview.** Given an unseen input image (**a**) from the test set that was lit by the according directional light ℓ_{src} above it, we relight it towards the directional light ℓ_{dst} in (**b**). To judge the performance, we provide the corresponding ground truth image in (**c**).

*The majority of this work was performed while the authors were at Facebook Reality Labs. Corresponding author: T.Nestmeyer@gmail.com

acquiring sufficient amounts is very time consuming.

Recent approaches have demonstrated that explicitly integrating physical processes in neural architectures is beneficial in terms of both robust estimates from limited data and increased generalization [25, 39, 43]. However, these approaches have focused on the *de-lighting* process, and use the simplified physical models for relighting that do not model non-diffuse effects such as cast shadows and specularities.

In this work, we bridge the gap between the expressiveness of a physically unconstrained end-to-end approach and the robustness of a physics-based approach. In particular, we consider relighting as an image-to-image translation problem and divide the relighting task into two distinct stages: a physics-based parametric rendering of estimated intrinsic components, and a physics-guided residual refinement. Our image formation model makes the assumption of directional light and diffuse materials. The subsequent refinement process is *conditioned* on the albedo, normals, and diffuse rendering, and dynamically accounts for shadows and any remaining non-diffuse phenomena.

We describe a neural architecture that combines the strengths of a physics-guided relighting process with the expressive representation of a deep neural network. Notably, our approach is end-to-end trained to simultaneously learn to both de-light *and* relight. We introduce a novel dataset of human faces under varying lighting conditions and poses, and demonstrate our approach can realistically relight complex non-diffuse materials like human faces. Our directional lighting representation does not require assumptions of smooth lighting environments and allows us to generalize to arbitrarily complex output lighting as a simple sum of point lights. To our knowledge, this is the first paper showing realistic relighting effects caused by strong directional lighting, such as sharp cast shadows, from a single input image.

2. Related work

Intrinsic images. Intrinsic image decomposition [3] and the related problem of shape from shading [57] have inspired countless derived works. Of interest, [2] propose to simultaneously recover shape, illumination, reflectance and shading from a single image and rely on extensive priors to guide an inverse rendering optimization procedure. Other methods recover richer lighting representations in the form of environment maps given the known geometry [26]. More recent approaches rely on deep learning for the same task, for example using a combination of CNN and guided/bilateral filtering [31] or a pure end-to-end CNN approach [10] with the common problem of hard to come by training data. Available datasets may include only sparse relative reflectance judgements [4], or sparse shading annotations [20], which limits learning and quantitative evaluation.

While many previous works focus on lighting estimation from objects [2, 12, 26, 29] or even entire images [11, 15, 18, 54, 55], few papers explicitly focus on the *relighting* problem. Notably, [36] use a small number of images as input, and, more recently, [52] learn to determine which set of five light directions is optimal for relighting. Image-to-image translation [16] combined with novel multi-illumination datasets [30] has lately demonstrated promising results in full scene relighting.

The explicit handling of moving hard shadows in [9] and [35] is relevant. While both use multi-view inputs to relight outdoor scenes, our method works on a single input image to relight faces (our multi-view setup is only used to capture training data). Similar to our work, [53] regress to intrinsic components like albedo and normals, but their illumination model is spherical harmonics and therefore does not handle shadows. [38] recently proposed a residual appearance renderer which bears similarities to our learned residual in that it models non-Lambertian effects. Both of the latter works optimize for intrinsic decomposition, whereas we learn end-to-end relighting. Our intrinsic components are only used as a meaningful intermediate representation.

Face relighting. Lighting estimation from face images often focuses on normalization for improving face recognition. For example, [50] use spherical harmonics (SH) to relight a face image, and [47] use a Markov random field to handle sharp shadows not modeled by low-frequency SH models. Other face modeling methods have exploited approximate lighting estimates to reconstruct the geometry [23, 45] or texture [24]. In computer graphics similar ideas have been proposed for face replacement [5, 8]. Low-frequency lighting estimation from a face has been explored in [19, 40, 41]. In contrast, [32] note that eyes reflect our surroundings and can be used to recover high frequency lighting. More closely related to our work, [7] learn the space of outdoor lighting using a deep autoencoder and combine this latent space with an inverse optimization framework to estimate lighting from a face. However, their work is restricted to outdoor lighting and cannot be used explicitly for relighting.

Of particular relevance to our work, neural face editing [43] and the related SfSNet [39] train CNNs to decompose a face image into surface normals, albedo, and SH lighting. These approaches also impose a loss on the intrinsic components, as well as a rendering loss which ensures that the *combination* of these components is similar to the input image. FRADA [22] revisited the idea of relighting for improving face recognition with face-specific 3D morphable models (similar to [43]), while we do not impose any face-specific templates. Single image portrait relighting [58] bypasses the need for decomposition, while still estimating the illumination to allow editing. In a similar line of work, [44] capture faces in a light stage using one light at a time, but then train using smoother illuminations from image based rendering which leads to artifacts when exposed to hard cast shadows or strong specularities. Re-

cently, [28] also used light stage data and train to relight to directional lighting as we do. However, their network expects a pair of images captured under spherical gradient illumination at test time, which can only be captured in a light stage. The portrait lighting transfer approach of [42] directly transfers illumination from a reference portrait to an input photograph to create high-quality relit images, but fails when adding/removing non-diffuse effects.

3. Architecture

The following two sections first introduce an image formation process (Sec. 3.1) and then describe its integration into a physics-based relighting architecture (Sec. 3.2).

3.1. Image formation process

The image formation process describes the physics-inspired operations transforming the intrinsic properties of a 3D surface to a rendered output. The majority of physics-based works are based on specific instantiations of the rendering equation [17],

$$L_o(\omega_o) = \int_{\omega_i \in \Omega} f(\omega_i, \omega_o) L_i(\omega_i) \langle \mathbf{n}, \omega_i \rangle \, \mathrm{d}\omega_i, \quad (1)$$

where ω_i, ω_o are the incoming and outgoing light directions relative to the surface normal $\mathbf{n}$ at the surface point $\mathbf{X}_j$. $L_i(\omega_i)$ and $L_o(\omega_o)$ are the corresponding (ir)radiances, $f(\cdot, \cdot)$ is the BRDF describing the material's reflectance properties, and $\langle \mathbf{n}, \omega_i \rangle$ is the attenuating factor due to Lambert's cosine law.

This model is often simplified further by assuming a *diffuse* decomposition into albedo $a \in \mathbb{R}$ and shading $s \in \mathbb{R}$,

$$a = f(\omega_i, \omega_o), \qquad [\text{const.}] \quad (2)$$

$$s = \int_{\omega_i \in \Omega} L_i(\omega_i) \langle \mathbf{n}, \omega_i \rangle \, \mathrm{d}\omega_i. \quad (3)$$

Non-diffuse effects. A realistic relighting approach must relax modeling assumptions to allow complex reflectance properties such as subsurface scattering, transmission, polarization, etc., and if using (2) specularities. Unfortunately, learning a spatially varying BRDF model $f(\omega_i, \omega_o)$ based on a non-parametric representation is infeasible: assuming an image size of 512×768 and a pixelwise discretization of the local half-angle space [27] would result in 1.7×10^{12} parameters. Learning a low-dimensional representation in terms of semantic parameters [6] seems like a viable alternative but is still prone to overfitting and cannot account for light-material-interactions outside of its parametric space.

We propose a hybrid approach and decompose f into two principled components, a diffuse albedo a and a light-varying residual r:

$$f(\omega_i, \omega_o) = a + r(\omega_i, \omega_o). \quad (4)$$

This turns (1) into

$$L_o(\omega_o) = as + \int_{\omega_i \in \Omega} r(\omega_i, \omega_o) L_i(\omega_i) \langle \mathbf{n}, \omega_i \rangle \, \mathrm{d}\omega_i. \quad (5)$$

For a light source with intensity $I(\omega_i)$, we can identify $L_i(\omega_i) = I(\omega_i) v(\omega_i)$, where $v \in \{0, 1\}$ is the binary visibility of the light source. Under the assumption of a single directional light source from $\widetilde{\omega}_i$, we integrate over one point only, so if we further write $\widetilde{r}(\widetilde{\omega}_i, \omega_o) = r(\widetilde{\omega}_i, \omega_o) I(\widetilde{\omega}_i) \langle \mathbf{n}, \widetilde{\omega}_i \rangle$, we can re-formulate our rendering equation (1) to

$$L_o(\omega_o) = (as + \widetilde{r}(\widetilde{\omega}_i, \omega_o)) \cdot v(\widetilde{\omega}_i). \quad (6)$$

This will be the underlying image formation process in all subsequent sections. While as captures much of the diffuse energy across the image according to an explicit generative model, the residual $\widetilde{r}(\widetilde{\omega}_i, \omega_o)$ accounts for physical effects outside of the space representable by (2) and is modeled as a neural network (akin to [38]). We do not impose any assumptions on $r(\widetilde{\omega}_i, \omega_o)$, even allowing light subtraction, but do enforce a to be close to the ground truth albedo of a diffuse model which we obtain from photometric stereo [51].

Discussion. While directional lights are conceptually simple, they lead to challenging relighting problems. Our combination of an explicit diffuse rendering process and a non-diffuse residual (with implicit shading) serves several purposes: (1) Describing most of the image intensities with a physics-based model means the output image will be more consistent with the laws of physics; (2) Describing specular highlights as residuals alleviates learning with a CNN; (3) Leaving the residual unconstrained (up to ground truth guidance) allows us to model effects that are not explainable by the BRDF, such as subsurface scattering and indirect light; (4) Modeling visibility explicitly helps, because the simple diffuse model does not handle cast shadows. At the same time, expecting the residual to take care of shadow removal by resynthesis is much harder than just masking it.

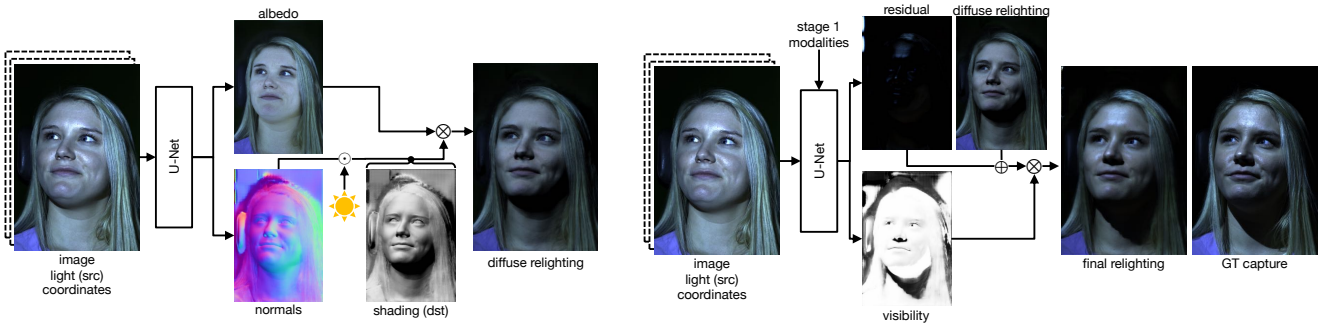
3.2. Physics-guided relighting

Presented with an input image I_{src} that was lit by an input illumination ℓ_{src}, our goal is to learn a generator G, relighting I_{src} according to a desired output illumination ℓ_{dst},

$$G(I_{\text{src}}, \ell_{\text{src}}, \ell_{\text{dst}}) = I_{\text{dst}}. \quad (7)$$

At training time, we assume ℓ_{src} and ℓ_{dst} to be directional lights, which is known to be a particularly challenging instance of relighting and accurately matches our captured data (see Sec. 4). At test time, this is not a limitation, since we can easily fit a set of directional lights to an environment map to perform more complex relighting (see Sec. 6).

Our physics-guided approach to solving the relighting task consists of a recognition model inferring intrinsic components from observed images (de-lighting) and a generative



(a) Stage 1: Diffuse Rendering.

(b) Stage 2: Non-Diffuse Residual.

Figure 2: **Physics-guided relighting with structured generators.** Our generator consists of two stages modeling diffuse and non-diffuse effects. All intrinsic predictions are guided by losses w.r.t. photometric stereo reconstructions. **(a)** We use a U-Net with grouped convolutions to make independent predictions of the intrinsic components. Predicted normals are always re-normalized to unit vectors. Given a desired output lighting, we compute shading from normals and render a diffuse output. **(b)** Conditioned on all modalities inferred in (a), we predict a non-diffuse residual and binary visibility map to model specularities, cast shadows, and other effects not captured by our instance of the rendering equation.

model producing relit images from intrinsic components (relighting). While the recognition model takes the form of a traditional CNN, the generative model follows our image formation process (Sec. 3.1) and is represented by structured layers with clear physical meaning. In line with (6), we implement the latter as a two-stage process: (Stage 1) Using the desired target lighting, we compute shading from predicted normals and multiply the result with our albedo estimate to obtain a diffuse render; (Stage 2) Conditioned on all intrinsic states predicted in stage 1, we infer a residual image and a visibility map, which we combine with the diffuse render according to (6). An illustration of this pipeline is shown in Fig. 2. Since all its operations are differentiable and directly stacked, this allows us to learn the proposed model in an end-to-end fashion from input to relit result.

We introduce losses for all internal predictions, i.e., albedo, normals, shading, diffuse rendering, visibility, and residual. We emphasize the importance of using the right loss function and refer to Sec. 5.1 for a comprehensive study. In order to obtain the corresponding guidance during training, we use standard photometric stereo reconstruction [51].

4. Data

Our data comprises a diverse set of facial expressions captured under various lighting conditions.

4.1. Acquisition

We record our data in a calibrated multi-view light-stage consisting of 6 stationary Sony PMW-F55 camcorders and a total of 32 white LED lights. The cameras record linear HDR images at 2048×1080 / 60 fps and are synchronized with a Blackmagic sync generator that also triggers the LED lights. We flash one LED per frame and instruct our subjects to hold a static expression for the full duration of an LED cycle (32 frames ~ 0.53 s). In order to remove captures with motion, we filter our data based on the difference of two fully lit shots before/after each cycle. For light calibration, we use a chrome sphere to recover directions and intensities in 3D [13] but subsequently express them with respect to each of the 6 cameras, such that we obtain a total of $6 \cdot 32 = 192$ different light directions/intensities for each image.

We record a total of 482 sequences from 21 subjects, resulting in $482 \cdot 6 \cdot 32 \cdot 32 = 2{,}961{,}408$ relighting pairs. Each pair is formed using any one of the 32 lights as input, and any one taken from the same sequence and same camera as output. We split them into 81% (17 subjects) for training, 9.5% (2 subjects) for validation and 9.5% (2 subjects) for testing.[1] We did not ask the subjects to follow any specific protocol of facial expressions, besides being diverse, such that our evaluation on validation/test data is on *both* unseen expressions and unseen subjects.

After extraction of the raw data, we use photometric stereo (PMS) reconstruction [51] to separate the input images I into albedo A, shading S with corresponding normals, and non-diffuse residual images $R = I - A \odot S$ per frame.

4.2. Augmentation

Modern neural architectures are much better at interpolation than extrapolation. It is therefore critical to cover the space of valid light transports as well as possible. To this end, we perform a series of data augmentations steps in an attempt to establish strong correlations throughout the parametric relighting space: (1) We flip all training images along the horizontal and vertical axis, increasing the effec-

[1] The split into training, validation and testing was done manually in an effort to balance the demographics of the subjects. See supplementary material for details.

Table 1: **Loss selection.** We explore the influence of different training losses and evaluation metrics on direct image-to-image translation ("pix2pix") and our structured guidance approach ("ours"). For each class, we show validation scores for all pairwise combinations of 5 training losses (rows) and the same 5 evaluation metrics (columns). The best model for each evaluation metric is shown in bold.

M	Training Loss	Evaluation Metric				
		L_1	L_2	LPIPS	DSSIM	MS-DSSIM
pix2pix	L_1	.0452	.0067	.2564	.1707	.1144
	L_2	.0516	.0082	.2663	.1911	.1369
	LPIPS	.0424	.0062	**.1868**	.1440	.0992
	DSSIM	**.0406**	**.0055**	.2138	**.1378**	.0930
	MS-DSSIM	.0422	.0058	.2358	.1547	**.0913**
ours	L_1	.0406	.0055	.2237	.1484	.0913
	L_2	.0415	.0056	.2302	.1547	.0953
	LPIPS	.0365	.0048	**.1701**	.1308	.0803
	DSSIM	**.0362**	**.0045**	.2008	**.1270**	**.0793**
	MS-DSSIM	.0410	.0055	.2165	.1470	.0910

[M: model; LPIPS: [56]; DSSIM: structured dissimilarity; MS-DSSIM: multi-scale DSSIM]

tive dataset size by a factor of 4. Note that this also requires adaptation of the corresponding light directions and normals; (2) We perform a linear scaling $\mathbf{x}' = s \cdot \mathbf{x}$, $s \sim \mathcal{U}_{[0.6,1.1]}$, of the images, shading, residuals and light intensities. In practice, we did not observe substantial benefits compared to training without scaling; (3) We randomly perturb the light calibration with Gaussian noise $n \sim \mathcal{N}(0, 0.01^2)$ to improve generalization and account for minimal calibration errors; (4) For quantitative results, we perform a spatial rescaling to $\frac{1}{8}$th of the original image resolution (135×256), train on *random crops* of size 128×128 and test on center crops with the same resolution to have comparability with SfSNet. Qualitative results are generated by rescaling to $\frac{1}{2}$ of the original resolution (540×1024), trained on *random crops* of size 512×768 and tested on center crops of that resolution.

5. Experiments

Our models were implemented using PyTorch [33] with a U-Net [37] generator and PatchGAN [14] discriminator (for the final relit image) based on the implementations provided by pix2pix [16]. The images in our dataset are camera RAW, represented as 16-bit linear RGB values nominally in the range $[0, 1]$. There is under- and over-shoot headroom, but for training and evaluation we clamp them into this range and linearly transform to $[-1, 1]$ as input into the network.

5.1. Evaluation metric

Quantitatively comparing the relit prediction $\hat{I}_{\text{dst}}$ of the generator against the ground truth I_{dst} requires an appropriate error measure. We consider the L_1 and L_2 norms but recognize that they do not coincide with human perceptual response. We also consider the "Learned Perceptual Image Patch Similarity" (LPIPS) loss suggested by [56] using the distance of CNN-features pretrained on ImageNet. Another prevailing metric of image quality assessment is structural similarity (SSIM) [48] and its multi-scale variant (MS-SSIM) [49]. In our evaluation, we use the corresponding dissimilarity measure DSSIM $= \frac{1-\text{SSIM}}{2}$, and likewise for MS-SSIM, to consistently report errors.

When defining the loss function during training, the same choices of distance metrics are available. To densely evaluate their performance, we report in Table 1 the results of training all intrinsic layers with the same loss function from the options above. Surprisingly, we conclude that, for our task, using DSSIM for the training loss consistently leads to models which generalize better on the validation set using most of the error metrics. The only exception is evaluation using the LPIPS metric, which is better when also trained using this metric. Therefore, we chose the models trained on DSSIM for computing the final test results.

5.2. Baseline comparisons

We now provide quantitative and qualitative comparisons of our proposed architecture, to related work.

5.2.1 Baselines

Our baselines comprise the following set of related methods.

PMS. To understand the lower error bound of a diffuse model, we take albedo A and shading S from photometric stereo (PMS; [51]) and diffuse render via $A \odot S$. We note that this model has access to all light configurations at the same time, with the desired target illumination amongst them. Since this gives an unfair advantage, we do not highlight results for this model in Table 2.

SfSNet (pretrained). We take the pretrained network of SfSNet [39] and apply it to our data by using their decomposition into albedo and normals, but ignoring the output spherical harmonics estimate. Instead, we compute target shading as the dot product of ℓ_{dst} and normals to have a direct comparison with our assumption of directional light and present the result after diffuse rendering.

SfSNet (retrained). We retrain [39] on the calibrated PMS data and also provide the source illumination as input, to which our model has access as well. Compared to the pretrained model above, this baseline can be seen as a fairer comparison to SfSNet.

Pix2pix. The arguably simplest way to learn (7) from data is to instantiate G as a traditional neural network consisting of a series of generic convolutional layers with no semantic

(a) input image (b) SfSNet [39] (pretrained) (c) SfSNet [39] (retrained) (d) pix2pix [16] (e) ours (f) ground truth

Figure 3: **Qualitative evaluation on unseen subjects and expressions.** We compare relighting (**a**) the input image with (**b/c**) pretrained and retrained variants of SfSNet [39], (**d**) pix2pix [16], and (**e**) our model. In (**f**), we show the ground truth capture of the given target illumination. Notice our model's ability to generate realistic shadows and specular highlights.

meaning and no knowledge of the image formation process.We adapt the pix2pix translation GAN [16] to our use case by conditioning the generator on the input image as well as the source and target illumination. This ensures an objective comparison with our more structured model, which also has access to lighting information.

5.2.2 Evaluation

Qualitative evaluation. We compiled a collection of qualitative results in Fig. 3. While the shading of SfSNet is smooth, it has a bias towards an albedo which probably resembles skin color in their training data and does not distinguish well between different skin tones and hair. As expected, retraining their model on our data leads to more accurate results. Still, due to the diffuse assumption, it looks flat compared to our more expressive model. It misses specularities and surface normals orthogonal to the light direction are missing ambient light from inter-reflections. The pix2pix model generates promising results, but its domain-agnostic architecture often leads to physically implausible artifacts, such as missing shadows. In comparison, the predictions of our proposed architecture are typically the most realistic, mainly due to its need to estimate a consistent albedo, as can be seen for example at the hair in the first row of Fig. 3. The last row shows a hand occluding the face, leading to strong cast shadows that have to be introduced/removed. Our model using intrinsic guidance gracefully handles that case.

While our data allows foreground masking computed from PMS, we show the full image predictions for better judgment. At test time, an off-the-shelf face matting approach, e.g. [46], could be used for cleaning the predictions.

We encourage the reader to look at more qualitative results of this type on our project page[2], where we also show relighting under a moving target illumination.

Quantitative evaluation. In Table 2 (first block), we analyze the quantitative performance of our model in the described scenario with known source illumination ℓ_{src}. The test set comparison with SfSNet and the diffuse pix2pix baseline confirms the importance of our physics-based guid-

[2] https://lvsn.github.io/face-relighting

Table 2: **Quantitative evaluation.** We show a quantitative comparison of our approach to baseline methods. Performance on the test set is reported under the assumption of both known ('with') and unknown ('w/o') source illumination. All models have been trained with the DSSIM loss.

L	Model	Evaluation Metric				
		L_1	L_2	LPIPS	DSSIM	MS-DSSIM
with	SfSNet (R)	.0636	**.0121**	.2508	.1840	.1277
	pix2pix	.0668	.0144	.2430	.1832	.1328
	ours	**.0609**	.0123	**.2144**	**.1618**	**.1138**
w/o	SfSNet (P)	.1359	.0424	.4703	.3221	.3121
	pix2pix	.0815	.0189	.2783	.2076	.1623
	ours	**.0684**	**.0142**	**.2273**	**.1763**	**.1316**
	PMS	.0391	.0047	.1630	.1125	.0561

[L: access to source illumination; LPIPS: [56]; DSSIM: structured dissimilarity, MS-DSSIM: multi-scale DSSIM; PMS: photometric stereo; P/R: (p)retrained]

ance and non-Lambertian residuals. An extension of our model without the assumption of known source illumination (second block) will be discussed in Sec. 6.1. For reference, the PMS reconstruction, restricted to a diffuse model but computed from multiple images, is also shown.

5.3. Additional qualitative comparisons

We provide more qualitative comparisons to the following recent portrait relighting approaches in Fig. 4.

The *mass transport relighting* [42] approach is different in that it defines the target lighting as that of *another* portrait. To match those conditions, we set the desired output to directly be the ground truth reference image. Despite these optimal conditions, [42] fails to generate specular highlights and cast shadows, which are well-captured by our technique.

Single image portrait relighting using an environment map is learned in [44]. Training images are produced by compositing multiple 'one light at a time' captures. As already discussed in [44], the method fails on strong light.

Finally, [58] learn to do *deep portrait relighting* using a spherical harmonics representation, which also handles smooth lighting exclusively, as can be seen in Fig. 4.

6. Extensions

We now demonstrate that our model successfully generalizes to different scenarios, including unknown input illumination, relighting with environment maps, and relighting images captured in the wild.

6.1. Relighting with unknown source illumination

While we cannot remove the need for the target illumination, information about the source illumination is already

(a) input (b) MT (c) SIPR (d) DPR (e) ours

Figure 4: **Additional qualitative comparisons.** We relight the input in the first row to the input in the second row, and vice versa. Results in the Mass Transport (MT) [42] approach and Single Image Portrait Relighting (SIPR) [44] were provided by the authors. For Deep Portrait Relighting (DPR) [58], we use their provided code and approximate the light directions manually.

(a) input (b) with (c) w/o (d) GT

Figure 5: **Relighting with unknown source illumination.** (**a**) Input image. (**b/c**) Our relighting model *with* and *without* access to source illumination. (**d**) Ground truth output.

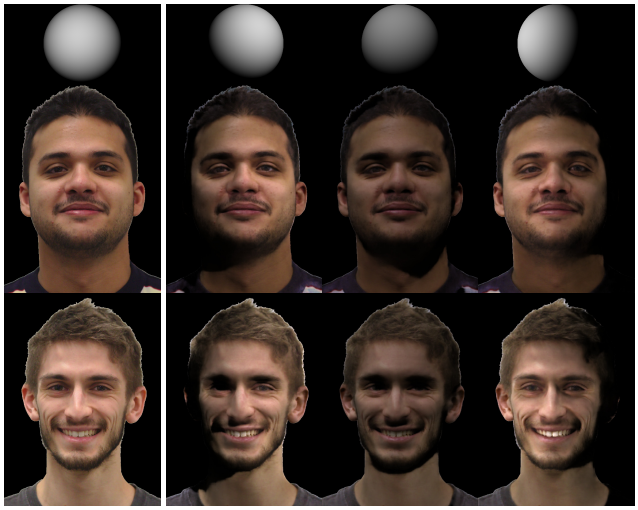
contained in the input image, allowing for implicit learning of ℓ_{src}. To illustrate our model's ability to extract these signals, we trained a version of our architecture without explicit access to the input lighting; these results, as well as a comparison to the corresponding baseline variants, are shown in Table 2 (second block) and Fig. 5. As expected, all models incur a small drop in performance compared to their counterparts with explicit knowledge. Nonetheless, our model *without* access to the source illumination achieves similar (and in some cases better) performance than the pix2pix model *with* access to the source illumination.

6.2. Relighting with environment maps

Directional light sources are a very general representation, and our approach easily allows for relighting with environment maps. While more principled approaches like importance sampling are available [1], for illustrative purposes here we simply sample environment maps by downscaling them to 64×32 pixels and instantiating our relighting prediction with one light direction per pixel. Since light is additive, we then mix the resulting predictions according to their color and intensity. Examples of a relit scene with three environ-

(a) input (b) env. 1 (c) env. 2 (d) env. 3

Figure 6: **Relighting with environment maps.** We consider input images taken under a point light source (**a**) and relight them with respect to 3 different environment maps (**b-d**).



(a) input (b) relit 1 (c) relit 2 (d) relit 3

Figure 7: **Relighting in the wild.** We consider portraits not taken in our capture environment (**a**) and relight them with respect to 3 different target point lights (**b-d**).

ment maps are shown in Fig. 6, more results are contained in the supplementary material on our project page.

6.3. Relighting in the wild

To demonstrate generalization outside the domain of our lab-captured dataset, we conducted experiments using pictures taken in an office environment with a Canon EOS 5D Mark III and visualize results of relighting towards three target lights in Fig. 7.

We emphasize the practical difficulties of relighting those portraits, including unknown discrepancies in the imaging pipeline (camera sensor, illuminant color, image processing etc.) and approximation of the unknown source lighting. Since the portraits are taken under uncontrolled office lighting, this results in images which are diffusely lit by multiple input light sources. Although this violates our model assumption of a single directional light source, we run our model using an input light direction ℓ_{src} in the image that would simply light the portrait centrally. To compensate for different illumination colors in the input, we compute the mean of a 51×76 center patch and apply a linear color transform towards our data distribution. The background is masked by hand, which could be automated [46]. Note that we can show neither quantitative nor qualitative comparisons to ground truth relit images since they do not exist.

7. Conclusion

We propose a method which learns to relight a face with strong directional lighting, accurately reproducing non-diffuse effects such as specularities and hard-cast shadows. We introduce a structured relighting architecture with semantic decomposition into, and subsequent re-rendering from intrinsic components. On our challenging light-stage dataset with directional light, the integration of an explicit generative rendering process and a non-diffuse neural refinement layer within an end-to-end architecture proved to be superior. We found that our model tends to produce shadows and albedo maps that are qualitatively closer to reality than all baselines. A more structured approach also has advantages beyond raw performance, including better interpretability/explainability and the possibility for direct manipulation/extraction of its semantically meaningful intermediate layers for downstream tasks. A comprehensive study of different losses and evaluation metrics highlighted the benefits of training on a perceptual loss over more traditional choices. We therefore believe our model to be useful in a wide range of face-centric and more general applications in, e.g., augmented reality.

Limitations and future work. While our model generally deals well with cast shadows in the input image (see Fig. 4), results get worse when there is so little light that the camera mostly returns noise. Although a crude infilling for those pixels based on context can be learned, an interesting future direction would be to identify these pixels explicitly. A dedicated infilling method, conditioned on the properly relit parts of the image and other intrinsic layers, could be applied to them. To cancel ambient input illumination (as in Fig. 7), future work could experiment with taking a flash photo and subtracting a second non-flash photo (in linear color space).

Acknowledgments. We thank Yannick Hold-Geoffroy and Marc-André Gardner for their valuable feedback, and Zhixin Shu and Tiancheng Sun for running their code on our dataset.

References

[1] Sameer Agarwal, Ravi Ramamoorthi, Serge Belongie, and Henrik Wann Jensen. Structured importance sampling of environment maps. *ACM Transactions on Graphics (TOG)*, 22(3):605–612, 2003. 7

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. 1, 2

[3] Harry Barrow and J Tenenbaum. Recovering intrinsic scene characteristics. *Comput. Vis. Syst., A Hanson & E. Riseman (Eds.)*, pages 3–26, 1978. 1, 2

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 2

[5] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter N Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. *ACM Transactions on Graphics (SIGGRAPH)*, 27(3):39:1–39:8, 2008. 2

[6] Brent Burley. Physically-Based Shading at Disney. In *SIGGRAPH 2012 Courses*, 2012. 3

[7] Dan A Calian, Jean-François Lalonde, Paulo Gotardo, Tomas Simon, Iain Matthews, and Kenny Mitchell. From faces to outdoor light probes. *Computer Graphics Forum*, 37(2):51–61, 2018. 2

[8] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 30(6), Dec. 2011. 2

[9] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. Multiview Intrinsic Images of Outdoors Scenes with an Application to Relighting. *ACM Transactions on Graphics (TOG)*, 34(5):1–16, 2015. 2

[10] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[11] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 9(4), 2017. 2

[12] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1932–1947, 2018. 2

[13] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and spatially-varying BRDFs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1060–1071, 2010. 4

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 5

[15] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 2

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5, 6

[17] James T. Kajiya. The rendering equation. In *Proceedings of ACM SIGGRAPH*, pages 143–150, 1986. 1, 3

[18] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3D object compositing. *ACM Transactions on Graphics (TOG)*, 33(3):1–15, 2014. 2

[19] Sebastian B. Knorr and Daniel Kurz. Real-time illumination estimation from faces for coherent rendering. In *IEEE International Symposium on Mixed and Augmented Reality*, 2014. 2

[20] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[21] Edwin H Land and John J McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971. 1

[22] Ha Le and Ioannis Kakadiaris. Illumination-invariant face recognition with deep relit face images. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 2

[23] Jinho Lee, Raghu Machiraju, Baback Moghaddam, and Hanspeter Pfister. Estimation of 3D faces and illumination from single photographs using a bilinear illumination model. In *Eurographics Conference on Rendering Techniques*, 2005. 2

[24] Chen Li, Kun Zhou, and Stephen Lin. Intrinsic face image decomposition with human face priors. In *European Conference on Computer Vision*, 2014. 2

[25] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 37(6), Nov. 2018. 2

[26] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):129–141, 2016. 2

[27] Wojciech Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003. 3

[28] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (SIGGRAPH)*, 38(4), July 2019. 3

[29] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt,

and Christian Theobalt. LIME: Live intrinsic material estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2

[30] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *IEEE International Conference on Computer Vision*, 2019. 2

[31] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[32] Ko Nishino and Shree K Nayar. Eyes for relighting. *ACM Transactions on Graphics (SIGGRAPH)*, 23(3):704–711, July 2004. 2

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshops*, 2017. 5

[34] Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production facial performance relighting using reflectance transfer. *ACM Transactions on Graphics (TOG)*, 26(3):52, 2007. 1

[35] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A Efros, and George Drettakis. Multi-view Relighting Using a Geometry-aware Network. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[36] Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. Image based relighting using neural networks. *ACM Transactions on Graphics*, 34(4):111:1–111:12, July 2015. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015. 5

[38] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *International Conference on Computer Vision*, 2019. 2, 3

[39] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. SfSNet: Learning shape, refectance and illuminance of faces in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6

[40] Davoud Shahlaei and Volker Blanz. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015. 2

[41] Hyunjung Shim. Faces as light probes for relighting. *Optical Engineering*, 51(7):077002–1, 2012. 2

[42] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)*, 37(1):2, 2018. 3, 7

[43] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[44] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (SIGGRAPH)*, 38(4):79, 2019. 2, 7

[45] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S.M. Seitz. Total moving face reconstruction. In *European Conference on Computer Vision*, 2014. 2

[46] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (TOG)*, 37(4):64, 2018. 6, 8

[47] Yang Wang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and D Samaras. Face re-lighting from a single image under harsh lighting conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[49] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *IEEE Asilomar Conference on Signals, Systems & Computers*, volume 2, 2003. 5

[50] Zhen Wen, Zicheng Liu, and T S Huang. Face relighting with radiance environment maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003. 2

[51] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J. Gortler, David W. Jacobs, and Todd Zickler. From shading to local shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1), 2014. 3, 4, 5

[52] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018. 2

[53] Ye Yu and William AP Smith. InverseRenderNet: Learning Single Image Inverse Rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019. 2

[54] Edward Zhang, Michael F. Cohen, and Brian Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics*, 35(6), 2016. 2

[55] Edward Zhang, Michael F Cohen, and Brian Curless. Discovering point lights with intensity distance fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5, 7

[57] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999. 2

[58] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *IEEE International Conference on Computer Vision*, 2019. 2, 7