Joint Learning of Portrait Intrinsic Decomposition and Relighting

Mona Zehni\textsuperscript{1}, Shaona Ghosh\textsuperscript{2}, Krishna Sridhar\textsuperscript{2}, Sethu Raman\textsuperscript{2}

\textsuperscript{1}Department of ECE and CSL, University of Illinois at Urbana-Champaign
\textsuperscript{2}Apple Inc.

mzehni2@illinois.edu, \{shaona.ghosh, srikrishna.sridhar, sethur\}@apple.com

Abstract

Inverse rendering is the problem of decomposing an image into its intrinsic components, i.e. albedo, normal and lighting. To solve this ill-posed problem from single image, state-of-the-art methods in shape from shading mostly resort to supervised training on all the components on either synthetic or real datasets. Here, we propose a new self-supervised training paradigm that 1) reduces the need for full supervision on the decomposition task and 2) takes into account the relighting task. We introduce new self-supervised loss terms that leverage the consistencies between multi-lit images (images of the same scene under different illuminations). Our approach is applicable to multi-lit datasets. We apply our training approach in two settings: 1) train on a mixture of synthetic and real data, 2) train on real datasets with limited supervision. We showcase the effectiveness of our training paradigm on both intrinsic decomposition and relighting and demonstrate how the model struggles in both tasks without the self-supervised loss terms in limited supervision settings. We provide results of comprehensive experiments on SfSNet, CelebA and Photoface datasets and verify the performance of our approach on images in the wild.

1. Introduction

Image decomposition into its intrinsic components (albedo, normals and lighting), i.e. inverse rendering is a fundamental problem in computer vision \cite{45, 44}. Not only it allows us to understand the geometry of the scene, it also enables us to manipulate different components and generate a desired output image. For example, by changing the illumination while fixing the albedo and the surface normals, a relit image of the same scene is obtained. Inverse rendering has a wide range of applications in image editing, subject relighting and augmented reality.

On a single image specifically, intrinsic decomposition is extremely challenging. Under Lambertian assumption, this originates from an underconstrained system of equations, thus leading to ambiguity and non-uniqueness of the decomposition solution. In such scenarios having strong priors on the individual components help in accurate image decomposition. Deep learning models that are trained on large datasets have proven to form strong priors in various ambiguous settings such as 3D reconstruction from single view image \cite{49, 8}, image inpainting \cite{53} and super-resolution \cite{18}, to name a few.

Here, we also resort to a deep learning based data-driven approach for the single-image intrinsic decomposition task. Previous state-of-the-art methods have tackled this problem through supervised training on all intrinsic components either on synthetic \cite{26, 31}, real datasets \cite{52} or a combination of both \cite{37}. Here, we propose a new semi self-supervised training paradigm that helps in learning the intrinsic decomposition and relighting tasks in regimes where ground truth labels for intrinsic components are not avail-
able. Our approach leverages the images of the same scene lit under different illuminations, i.e. multi-lit images, and their inherent consistencies during training, while operating on single images at inference. For this purpose, we generate a multi-lit dataset from the synthetic SfSNet [37] and real CelebA [27] datasets, called multi-lit CelebA. Compared to self-supervised methods such as [54] applied to scene datasets, our method does not require multi-view image datasets and allows for hybrid training on synthetic and real datasets. We evaluate the effect of our training approach in two settings: 1) hybrid training on real and synthetic data, 2) training on real datasets with limited supervision.

We restate our contributions as: 1) We introduce a new self-supervised cross-relighting loss. Assuming a pair of multi-lit images, cross-relighting term enforces consistency between relit images obtained after swapping illumination vectors between the pair. To facilitate the use of this loss term, we generate a multi-lit dataset by relighting images from CelebA [27] with lighting vectors sampled from the synthetic SfSNet dataset [37]. 2) Through extensive experiments and comparison with multiple baselines, we demonstrate the importance of the cross-relighting loss term in either hybrid training on real and synthetic datasets or training with limited supervision on real datasets. We show that without this loss term, learning the intrinsic decomposition and relighting tasks fail in limited supervision settings.

2. Related Work

Intrinsic decomposition and subject relighting have been extensively studied in the literature. Here, we summarize the existing methods.

Intrinsic Image Decomposition: Early works on intrinsic image decomposition are mainly derived by solving the equations provided by the Lambertian reflectance model [3] [22] [34]. To impose proper constraints for intrinsic decomposition of single images which is a highly ill-posed problem, many papers introduced priors on different components [2]. However, deriving effective priors that model the underlying physics is challenging.

Recently, data driven models surged in the intrinsic decomposition field due to their exceptional performance in learning effective priors from large datasets. Deep belief networks [44] and encoder-decoder architectures [21] [41] [37] [26] [4] have particularly proven effective. In the latter, the image is encoded into a latent representation which is then processed by separate decoders, to achieve different components such as albedo, normals and lighting. These architectures are trained on 1) synthetic [31] [26] [23] [38] [59], 2) unconstrained real datasets [54], or 3) a mix of both [37] [36] [6] [58]. While models trained on purely synthetic data with full supervision obtain reasonable results on real images, they still suffer from the domain gap between the real and synthetic datasets. Training on real datasets addresses this domain gap. However, it introduces a new challenge: lack of ground truth labels for the intrinsic components. To resolve this issue, many papers pursue the generation of pseudo-labels for real portrait datasets by fitting a 3D morphable model (3DMM) [5] [41] [48], using a pre-trained model on synthetic data [37] [36], or photometric stereo [32]. The main drawbacks of estimating these labels are: 1) The estimated components can be erroneous. This error then propagates to the trained models on these datasets, 2) Obtaining labels is often prohibitively expensive and not attainable for commercial use.

Self-supervised learning has gained increasing popularity to allow training on unlabeled datasets. For the intrinsic decomposition task, self-supervison is obtained by dense correspondence between pixels across multiple views [54], training on a sequence of multi-lit images or video streams [25] [42], model-based shape reconstruction [46], or through reconstruction loss (imposing consistency between the original images and the re-rendered one from the estimated intrinsic components) while training on a mix of labeled and unlabeled datasets [37] [36] [16]. Here, we introduce a new self-supervised loss term that 1) reduces the need for pseudo-labels and multi-stage training [37], 2) does not require a sequence of images as input during training [25] [42], 3) does not rely on strong priors posed in [54] [25] [46] for training in limited supervision scenarios (no labels on albedos and normals exist) where the intrinsic decomposition from single image is highly ambiguous. Furthermore, compared to [28] proposing an unsupervised intrinsic decomposition technique given multi-lit images at training, we further disentangle the lighting component from the normals, thus facilitating relighting and light transfer between a source and a target image pair.

Subject Relighting: Relighting an image has been the focus of a multitude of recent works. [52] [29] use multiple input images alongside the target lighting to generate the relit subject. While in [29], these images are recorded under spherical color gradient illumination [12] [11], in [52] the input images correspond to multiple fixed directional lightings. These methods are in line with the idea that a new relit image is obtained by a combination of densely sampled images under one lighting at a time (OLAT) light stage [9]. Encoder-decoder models estimating the source lighting and the image latent vector that is later on fed to the decoder alongside the target lighting to generate the relit image are proposed in [57] [43]. Furthermore, the task of light transfer from a source to a target image has been addressed as a mass transport problem in [40]. While these methods are capable of relighting subjects under different illuminations, they lack the interpretability that is put forward by the intrinsic decomposition-based relighting models. Nestmeyer et al in [32] proposed a physics-based model that relights an image given a target lighting while taking into account non-
diffuse components. Joint relighting and multi-view synthesis from multiple input images has also been discussed in [51] [56] [7]. Opposed to our approach, these models rely on multiple input images at inference to extract the full geometry of the underlying scene.

### 3. Method

Our goal is to decompose a single-view RGB image $I$ into its intrinsic components including albedo $A$, normal $N$ and lighting $L$. The albedo and normals have the same dimensionality as the image, i.e. $n \times n \times 3$. While for the albedo, different channels represent color, for normals, they mark the orientation of the normal along 3D axes. Furthermore, we represent colored lighting by its spherical harmonic (SH) coefficients up to second order, thus $L \in \mathbb{R}^{9 \times 3}$. While SH coefficients up to second order mainly constitute low-frequency terms, they are sufficient to estimate the irradiance up to 1% error [33]. We assume $I$ is generated under Lambertian reflectance as,

$$ I(p) = f_{\text{render}}(A(p), N(p), L) = A(p) \circ (L^T N_{\text{SH}}(p)) $$

where $p$ marks the $p$-th pixel and $f_{\text{render}}$ is the differentiable function that maps the intrinsic components to the image domain. Also, $N_{\text{SH}}$ is the transformed normals in SH domain and $\circ$ denotes element-wise multiplication.

#### 3.1. Architecture

The decomposition architecture that we adopt here is inspired by the SfSNet architecture [37] depicted in Fig. 2, consisting of 1) a shared encoder, 2) two streams of cascaded residual blocks to extract albedo and normal features and 3) three decoders for albedo, normals and lighting. On the encoding path, the image is processed through convolution layers, batch norm [14] and ReLU [50] activations. On the decoding path, to upsample the downsampled features, we use deconvolution layers. To estimate the lighting, we concatenate the image, albedo and normal features and feed it to a trainable light estimator. This is inspired by the fact that given an image, its albedo and normals, lighting is estimated by solving a linear equation defined in (1). The light estimator block recovers the SH representation of lighting up to the second order.

#### 3.2. Training

We adopt a combination of supervised and self-supervised strategies using a mix of real and synthetic datasets. Below, we describe our training methods specific to the different datasets and the available annotations.

**Supervised**: For synthetic datasets, ground truth (GT) albedo, normal and lighting are available. Thus, these ground truth values are used to supervise the estimation of $A$, $N$ and $L$ during training. Further, through the reconstruction loss, we enforce the reconstructed image (combining the estimated components following (1)) to be close to the original input image.

**Self-supervised**: We also train our model on real datasets to improve generalization on unconstrained images in the wild. As ground truth intrinsic components are not often available for these datasets, we use self-supervision. Thus, apart from the reconstruction loss, we introduce a new term called cross-relighting loss.

To explain cross-relighting consistency terms, let us consider two images $I_1$ and $I_2$. These images share the same albedo and normals and only differ in terms of lighting, where $I_1 = f_{\text{render}}(A_1, N_1, L_1)$ and $I_2 = f_{\text{render}}(A_2, N_2, L_2)$, with $A_1 = A_2$ and $N_1 = N_2$. Therefore, we expect the estimated albedos and normals for $I_1$ and $I_2$ to be the same. Furthermore, by swapping the estimated lighting for the two images, we expect $f_{\text{render}}(\hat{A}_1, \hat{N}_1, L_2)$ to be close to $I_2$ and $f_{\text{render}}(\hat{A}_2, \hat{N}_2, L_1)$ to be close to $I_1$. This adds further consistency between the relit images which has shown to be
beneficial during training. Furthermore, we verify that the use of cross-relighting loss term reduces the need for providing pseudo-labels for the real unconstrained images [10]. Imposing the cross-relighting loss term is further illustrated providing pseudo-labels for the real unconstrained images [10].

Finally, the final loss that is used to train our architecture on a mix of synthetic and real datasets is defined as,

\[ L = \lambda_A \| \hat{A} - A \|_1 + \lambda_N \| \hat{N} - N \|_1 + \lambda_L \| \hat{L} - L \|_2 + \lambda_{rec} L_{rec} + \lambda_{relit} L_{relit} \]

(2)

where the first three terms mark the supervision on albedo, normals and lighting and are only effective for samples from the synthetic dataset. Furthermore, \( L_{rec} \) and \( L_{relit} \) are the reconstruction and relighting consistency terms defined as,

\[ L_{rec} = d(f_{encoder}(\hat{A}_1, \hat{N}_1, \hat{L}_1), I_1) + d(f_{encoder}(\hat{A}_2, \hat{N}_2, \hat{L}_2), I_2) \]

\[ L_{relit} = d(f_{encoder}(\hat{A}_1, \hat{N}_1, \hat{L}_1), I_2) + d(f_{encoder}(\hat{A}_2, \hat{N}_2, \hat{L}_2), I_1) \]

with \( d(\hat{x}, x) \) denoting the distance metric between \( \hat{x} \) and \( x \). In our experiments, we choose \( \| . \|_1 \) as \( d \) to better preserve the high frequency details in the reconstructed image and intrinsic components. We compute \( L_{rec} \) and \( L_{relit} \) in LAB color space, as it leads to a higher quality relighting results.

Note that, we only need a multi-lit dataset for the training phase, while inference is performed on a single image. We use existing multi-lit dataset PhotoFace [55] and also generate a multi-lit dataset from single-lit CelebA [27] and synthetic SfSNet [37] dataset. We call this dataset multi-lit CelebA. Details are discussed in Section 4.1.

Non-hybrid training: While we use (2) to train on a mixture of synthetic and multi-lit datasets, we also investigate a setting in which we only train on multi-lit datasets with lighting annotations. As discussed in Section 4.1, our multi-lit CelebA dataset consists of images of the same subjects under different lightings alongside the GT illumination vectors. In this setting, we train the same architecture in Fig. 2 with light supervision, \( L_{rec} \) and \( L_{relit} \) terms. Note that, here there is no longer any form of supervision on albedo and normals, as we are not training on synthetic data anymore. To provide some priors on the estimated albedos and normals, we make slight changes in the architecture. For the albedos, we add a sigmoid non-linearity to bound the estimated albedos between [0, 1]. Furthermore, to ensure that the normals have unit norms, we further normalize \( N \) pixel-wise. In Section 4, we present results on this form of training and argue that without \( L_{relit} \), the model is not able to learn a meaningful decompositions between the albedos and the normals. This highlights the importance of \( L_{relit} \) in regimes with limited available annotations.

4. Experiments

4.1. Generating Multi-Lit CelebA dataset

To generate the multi-lit CelebA dataset, we use the images from CelebA [27] and lighting vectors from synthetic SfSNet dataset [37]. To this end, we train a skip-connection based encoder-decoder architecture (similar to [37]) on the labeled synthetic dataset only. Then, we use this network to generate pseudo-labels for the CelebA dataset, while training another network with a mixture of SfSNet and CelebA datasets. Next, to relight CelebA images, we estimate their albedos and normals using the pretrained network, sample 5 different lighting vectors from the SfSNet dataset and generate relit images by passing the estimated albedo, normal and sampled lightings through the rendering function in (1). During training with \( L_{relit} \), we choose random pairs of the relit images of the same subject. When training on a mix of synthetic and multi-lit CelebA dataset, to ensure that the original samples from the CelebA are used, we always choose the original image from CelebA as one of the images in the multi-lit image pair, while choosing the other image randomly from the relit versions. Samples of the multi-lit CelebA dataset are provided in Fig. 3(a).

4.2. Training details

We examine two training paradigms: 1) train on a mix of synthetic and real datasets using the loss defined in (2), 2) train on multi-lit CelebA dataset with light supervision, \( L_{rec} \) and \( L_{relit} \). We adopt the synthetic dataset provided by [37] which includes GT albedos, normals, SH lighting and segmentation maps alongside each image. A sample of this dataset is visualized in Fig. 3(b). The SfSNet dataset consists of \( \sim 190k \) samples, where we use 70%, 10% and 20% as train, validation and test splits, respectively.

For real unconstrained images, we use CelebA [27] and Photoface [55] datasets. To incorporate \( L_{relit} \) on CelebA during training, we use multi-lit CelebA dataset. The CelebA dataset originally contains \( \sim 200k \) samples, but after generating its multi-lit version, \( \sim 600k \) pairs of multi-lit images are available. We use the same splits provided in [27] for CelebA and multi-lit CelebA datasets. On the other hand, the Photoface dataset already contains images of the same subject and pose under different lightings. It also provides annotations in the forms of albedos and normals computed from shape from shading methods [10]. For each subject and pose, 4 different lightings exist. Thus, during training we use \( \binom{4}{2} = 6 \) pairs of multi-lit images. To
avoid having samples of the same subject in different splits, we separate the splits on a subject basis with 100 identities in the test set (to further adhere to the existing baselines evaluated on this dataset). To preserve the personal identity of the subjects, we present results with redacted faces. We cover the eyes and mouth of the samples from real datasets with black boxes.

To further adapt CelebA images to the SfSNet dataset which only contain the face region, we extract the face region for each image. For this purpose, we create a face mask from the convex hull of the facial keypoints extracted by [19] (outlined by the jaw line and the eyebrows). All images are then resized to 128 × 128. In the rest of this section, unless otherwise stated our models are trained on SfSNet and multi-lit CelebA dataset using the loss in (2).

In our experiments, we used two different encoder-decoder based architectures. In the first architecture (Arch 1), we have skip connections between the encoder and both A and N decoders. We use Arch 1 for training models on SfSNet dataset only. In Arch 2, we no longer have skip-connections. This architecture is used for training models on a mixture of SfSNet and real datasets. We use a similar network architecture as [37] for Arch 1 and 2, following their ablation study and analysis on the effectiveness of each architecture in different settings. The light estimator module, consists of a 1 × 1 convolution layer followed by batch-norm, ReLU and adaptive average pooling that maps the features to a 256-dimensional vector. This vector is then fed to a fully connected layer to output a 27-dimensional lighting vector. In all experiments, we use the Adam optimizer with 5×10^{-4} weight decay and learning rates between 0.0002 and 0.001 depending on the experiment. More de-
4.3. Results

Light transfer: In Fig. 4, we compare the light transfer quality of our model in contrast to several benchmarks on various datasets including MultiPIE [13], DPR [57] and CelebA-HQ [17]. Note that, we did not use samples from MultiPIE, DPR and CelebA-HQ datasets and rather took screenshots of the images provided in [57] and evaluated the performance of our models on these screenshots. We provide a brief description of the baselines as: DPR [57] trains an encoder-decoder architecture on multi-lit image pairs with lighting annotations, Shih et al [39] and Li et al. [24] relight a content image given a style image, Shu et al [40] formulates the light transfer as a mass transport problem. Note that while these baselines have been trained on $512 \times 512$ images, our model and SfSNet [37] are trained on $128 \times 128$ images. To have a fair comparison on our model and SfSNet, we resize the input images to $128 \times 128$ and operate on facial region, while reporting results for the other baselines on $512 \times 512$ images.

Note that, in Fig. 4-(a) and (c), the DPR baseline has the advantage of knowing the target lighting and relighting the source image based on this GT illumination. However, for our method and SfSNet we estimate the lighting from the reference and transfer it to the source image. As our approach and SfSNet use colored light unlike DPR [57] (trained on monochromatic light), the color tone of the reference is transferred as the color of the light, thus the tone of the relit image matches the source. This is due to the
nature of the synthetic dataset we have trained on, which contains colored light and low variety of skin tones. One can use other synthetic datasets with variations in the skin tones and monochromatic lighting. Note that our approach succeeds in accurately estimating the lighting from the reference image and relighting the source image even though we are training on lower resolution images without knowing the target lighting vector. Also, compared to the baselines, the transferred light better matches the reference and the relit image looks more natural with no visible artifacts (compared to the results from [39, 24, 40]). These baselines suffer from accurately understanding the lighting from the reference image and then transferring it to the source image. Also in Fig. 4(a), compared to DPR, the facial shades on the relit images of our model better match the reference. This suggests that our model does a better job in estimating the lighting as it is trained on more natural looking images with a wide variation of lightings compared to the dataset used in DPR [57].

In Fig. 5, we closely examine the quality of the light transfer on CelebA dataset [27]. Note that compared to SfSNet, our model demonstrates more medium-frequency details (e.g. wrinkles) in the shading component (denoted by S_\* in the figure). This in turn leads to better quality of the relit images by our model compared to SfSNet. We presume this is because of the use of pseudo labels during training by SfSNet. While these labels are helpful in stabilizing the training on real dataset, they might be erroneous and thus limit the quality of the learnt components. Furthermore, our method is implicitly trained on the light transfer task through the cross-relighting loss term, thus improving light transfer quality compared to SfSNet. In addition, Fig. 5 exemplifies successful and challenging cases for the light transfer task. For instance, in the presence of cast shadows both our model and SfSNet have a hard time in accurately decomposing the albedo and normals by SfSNet. Even though, it is observed that the decomposition and the relighting task on parts of the face which are under strong shadows or dimly lit is more challenging as they contain less information.

In addition, we assess the quality of the estimated normals by our method against several baselines trained on Photoface in Table 1. The metric used is the angular error (in degrees) between the GT and estimated normals at each pixel. We report on mean (per pixel error) and the percentage of pixels with normal angular errors less than 20\degree, 25\degree and 30\degree. Note that, our method outperforms all previous methods in terms of the quality of the estimated normals. This gain verifies the importance of using \( L_{relit} \) and how it helps in a more accurate decomposition especially for low-light images.

**Decomposition performance:** In Fig. 4, we present the estimated intrinsic components by our model. We also relight Img. 1 with the estimated lighting from Img. 2, hence we expect the relit image to be close to Img. 2. Our results verify the accuracy in the estimated components and highlight challenging scenarios for the decomposition task (last three examples). While the model generalizes well on various poses, it is having difficulty in the intrinsic decomposition task on parts of the images which are partially occluded or contain cast shadows. This is due to the scarcity of such examples with GT labels in the training set.

**Importance of cross-relighting loss:** To highlight the importance of \( L_{relit} \), we train two models only on multi-lit CelebA dataset. In both models we train with light supervision and \( L_{rec} \). We also use \( L_{relit} \) while training the second model. Note that no form of supervision on \( \hat{A} \) and \( \hat{N} \) are provided in training these two models. In Fig. 8, we provide visual results on the quality of the decomposition and relighting. While both models are successful in the lighting estimation, the model trained without \( L_{relit} \) clearly fails in decomposing the albedo and normals. Even though, it still provides reasonable looking relit images in terms of the color tone, it evidently misses the correct shading which is a result of the interaction between lighting and normals. This stems from the ambiguity of the decomposition task, especially if no reference albedos and normals are provided.

![Table 2: Quantitative comparison with DPR and Pix2Pix on reconstructions and relighting tasks.](image)

| Metric/Model | DPR | Pix2Pix | Syn. only | SfSNet | Ours |
|--------------|-----|---------|-----------|--------|------|
| \( L_1 \) recon | 26.7 | 30.9 | 42.2 | 26.3 | **17.8** |
| \( L_1 \) relit | 34.9 | 48.7 | 48.7 | 39.2 | **30.8** |
| \( L_2 \) recon | 20.8 | 25.5 | 32.4 | 19.9 | **12.8** |
| \( L_2 \) relit | 27.3 | 37.0 | 37.4 | 30.4 | **23.5** |
| SSIM recon | 0.94 | 0.95 | 0.92 | 0.95 | **0.97** |
| SSIM relit | 0.9 | 0.89 | 0.89 | 0.92 | **0.93** |

Table 2: Quantitative comparison with DPR and Pix2Pix on reconstruction and relighting tasks.
Table 3: Quantitative evaluation on multi-lit CelebA dataset. The second row contains the results of a model trained only on the synthetic data, while the rest are trained on a mix of SfSNet and multi-lit CelebA datasets. (Pseudo sp: pseudo supervision, CR: cross-relighting).

| Model / Metric     | Decomp. | $L_1$ recon | $L_1$ relit | $L_2$ recon | $L_2$ relit | SSIM recon | SSIM relit |
|--------------------|---------|-------------|-------------|-------------|-------------|------------|------------|
| Syn. + Real + No Pseudo sp | ×       | 12.2        | 128.6       | 14.5        | 92.8        | 0.97       | 0.83       |
| Syn. only          | ✓       | 42.2        | 48.7        | 32.4        | 37.4        | 0.92       | 0.89       |
| Syn. + Real + Pseudo sp | ✓       | 26.3        | 39.2        | 19.9        | 30.4        | 0.94       | 0.92       |
| Syn. + Real + CR   | ✓       | 17.3        | 32.0        | 12.5        | 24.4        | 0.97       | 0.93       |
| Syn. + Real + CR + LAB | ✓       | 17.8        | 30.9        | 12.8        | 23.5        | 0.97       | 0.93       |
| Real + light       | ×       | 19.7        | 46.7        | 15.0        | 37.7        | 0.96       | 0.88       |
| Real + CR + light  | ✓       | 17.8        | 25.8        | 13.8        | 20.4        | 0.96       | 0.93       |

(Ablation study): In Table 3, we present a quantitative ablation study on the test set of the multi-lit CelebA dataset. Our metrics are average per pixel $L_1$ and $L_2$ errors for the reconstructed and relit images alongside structural similarity index measure (SSIM). While there is no GT labels for the albedos and normals for CelebA multi-lit dataset, we evaluate the decomposition performance visually. For training on synthetic data only i.e. Syn. Only, we train a network with Arch 1 using the SfSNet dataset. For the rest of the methods, we train a network with Arch 2. Pseudo sp (pseudo supervision) refers to the use of pseudo-labels obtained from the pre-trained Syn. Only model during training. The results demonstrate that without pseudo-supervision or the use of $L_{relit}$, learning intrinsic decomposition from a mix of synthetic and real datasets fail. Furthermore, the results verify the improvement provided by imposing $L_{rec}$ and $L_{relit}$ in LAB color space.

In the last two rows, we quantify the performance of the model trained only on multi-lit CelebA dataset with light and reconstruction supervision and with/without $L_{relit}$. The results again verify the importance of the cross-relighting loss in limited supervision settings and how a model trained with light, $L_{rec}$ and $L_{relit}$ on multi-lit CelebA dataset has close (or even better on some metrics) performance to the models trained on a hybrid dataset.

Comparison to DPR and Pix2Pix: We compare our model against two benchmarks specialized for the subject relighting task and retrained on multi-lit CelebA. For DPR [57], we train a skip-connection based encoder-decoder architecture which relights an input image given a target illumination. In Pix2Pix [15], we treat the lighting as a style vector and train a generator-discriminator pair adversarially, conditioned on the lighting vector. Note that for these two baselines, the target lighting is required and assumed to be known during training unlike our model. We report the quantitative results in Table 2 where we have retrained SfSNet on multi-lit CelebA. Note that our model outperforms all baselines on both reconstruction and relighting tasks. Compared to DPR and Pix2Pix which only focus on the relighting task and use the lighting vector alongside the image/features to relight an image, our model additionally performs intrinsic decomposition which is helpful in improving the relighting performance.

5. Conclusion

In this paper, we studied the inverse rendering problem, i.e. decomposing a single view image into its intrinsic components such as albedo, normals and illumination. We proposed a new self-supervised training paradigm, leveraging consistencies between images of the same scene under different illuminations, i.e. multi-lit images. We also generated a multi-lit dataset from the CelebA dataset, named multi-lit CelebA. We examined our training approach under two different settings: 1) hybrid training on synthetic and multi-lit CelebA dataset, 2) limited supervision training on multi-lit CelebA. We showed in both settings our model is successful in learning intrinsic decomposition and relighting tasks despite the challenging ambiguities of the
problem. We presented qualitative and quantitative comparisons with various state-of-the-art baselines to demonstrate the strength of our approach.

References

[1] A. Bansal, B. Russell, and A. Gupta. Marr Revisited: 2D-3D model alignment via surface normal prediction. In CVPR, 2016.
[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(8):1670–1687, 2015.
[3] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman. Recovering intrinsic scene characteristics. Comput. Vis. Syst., 2(3-26):2, 1978.
[4] A. S. Baslamisli, T. T. Groenestege, P. Das, H. Le, S. Karaoglu, and T. Gevers. Joint learning of intrinsic images and semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.
[6] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu. Photorealistic facial details synthesis from single image. In Proceedings of the IEEE International Conference on Computer Vision, pages 9429–9439, 2019.
[7] Z. Chen, A. Chen, G. Zhang, C. Wang, Y. Ji, K. N. Kutulakos, and J. Yu. A neural rendering framework for free-viewpoint relighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5599–5610, 2020.
[8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision – ECCV 2016, pages 628–644, Cham, 2016. Springer International Publishing.
[9] P. Debevec, T. Hawkins, C. Tchou, H. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’00, page 145–156, USA, 2000. ACM Press/Addison-Wesley Publishing Co.
[10] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 10(4):439–451, 1988.
[11] G. Fyffe and P. Debevec. Single-shot reflectance measurement from polarized color gradient illumination. In 2015 IEEE International Conference on Computational Photography (ICCP), pages 1–10, 2015.
[12] G. Fyffe, C. A. Wilson, and P. Debevec. Cosine lobe based relighting from gradient illumination photographs. In 2009 Conference for Visual Media Production, pages 100–108, 2009.
[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. 2013.
[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15, page 448–456. JMLR.org, 2015.
[15] P. Isola, J. Zhu, T. Zhou, and A. A Efros. Image-to-image translation with conditional adversarial networks. CVPR, 2017.
[16] M. Janner, J. Wu, T. D Kulkarni, I. Yildirim, and J. Tenenbaum. Self-supervised intrinsic image decomposition. In Advances in Neural Information Processing Systems, pages 5936–5946, 2017.
[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In International Conference on Learning Representations, 2018.
[18] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
[19] D. E. King. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10:1755–1758, 2009.
[20] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.
[21] T. D Kulkarni, W. F Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In Advances in neural information processing systems, pages 2539–2547, 2015.
[22] E. H Land and J. J McCann. Lightness and retinex theory. Josa, 61(1):1–11, 1971.
[23] L. Lettry, K. Vanhoye, and L. van Gool. Darn: A deep adversarial residual network for intrinsic image decomposition. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1359–1367, 2018.
[24] Y. Li, M. Liu, X. Li, M. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
[25] Z. Li and N. Snavely. Learning intrinsic image decomposition from watching the world. In Computer Vision and Pattern Recognition (CVPR), 2018.
[26] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. In SIGGRAPH Asia 2018 Technical Papers, page 269. ACM, 2018.
[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.
[28] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In Proceedings of the European Conference on Computer Vision (ECCV), pages 201–217, 2018.
[29] A. Meka, C. Haene, R. Pandey, M. Zollhoefer, S. Fanello, G. Fyffe, A. Kowdle, X. Yu, J. Busch, J. Dourgarian, P. Denny, S. Bouaziz, P. Lincoln, M. Whalen, G. Harvey, J. Taylor, S. Izadi, A. Tagliasacchi, P. Debevec, C. Theobalt, J. Valentini, and C. Rhemann. Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. volume 38, July 2019.

[30] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. arXiv e-prints, page arXiv:1411.1784, Nov. 2014.

[31] T. Nestmeyer, J. Lalonde, I. Matthews, and A. Lehmann. Learning physics-guided face relighting under directional light. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[32] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’01, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery.

[33] Ruo Zhang, Ping-Sing Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(8):690–706, 1999.

[34] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[35] S. Sengupta, J. Gu, K. Kim, G. Liu, D. W. Jacobs, and J. Kautz. Neural inverse rendering of an indoor scene from a single image. In International Conference on Computer Vision (ICCV), 2019.

[36] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In Computer Vision and Pattern Recognition (CVPR), 2018.

[37] J. Shi, Y. Dong, H. Su, and S. X Yu. Learning non-lambertian object intrinsics across shapeNet categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1685–1694, 2017.

[38] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. Style transfer for headshot portraits. ACM Trans. Graph., 33(4), July 2014.

[39] Z. Shu, S. Hadap, E. Shechtman, K. Sunkavalli, S. Paris, and D. Samaras. Portrait lighting transfer using a mass transport approach. ACM Trans. Graph., 37(1), Oct. 2017.

[40] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[41] G. Song, J. Zheng, J. Cai, and T. Cham. Recovering facial reflectance and geometry from multi-view images. Image and Vision Computing, page 103897, 2020.

[42] T. Sun, J. T. Barron, Y. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. Debevec, and R. Ramamoorthi. Single image portrait relighting. ACM Trans. Graph., 38(4), July 2019.

[43] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. 2012.

[44] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(9):1459–1472, 2005.

[45] A. Tewari, M. Zollhoefer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Oct 2017.

[46] G. Trigeorgis, S. Snape, I. Kokkinos, and S. Zafeiriou. Face normals “in-the-wild” using fully convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 340–349, 2017.

[47] G. Trigeorgis, S. Snape, I. Kokkinos, and S. Zafeiriou. Face normals “in-the-wild” using fully convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 38–47, 2017.

[48] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018.

[49] B. Xu, N. Wang, T. Chen, and Mu Li. Empirical Evaluation of Rectified Activations in Convolutional Network. arXiv e-prints, page arXiv:1505.00853, May 2015.

[50] Z. Xu, S. Bi, K. Sunkavalli, S. Hadap, H. Su, and R. Ramamoorthi. Deep view synthesis from sparse photometric images. ACM Transactions on Graphics (TOG), 38(4):1–13, 2019.

[51] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. Deep image-based relighting from optimal sparse samples. ACM Trans. Graph., 37(4), July 2018.

[52] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[53] Y. Yu and W. AP Smith. Inverserendernet: Learning single image inverse rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[54] S. Zafeiriou, M. Hansen, G. Atkinson, V. Argyriou, M. Petrrou, M. Smith, and L. Smith. The photoface database. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 132–139, 2011.

[55] X. Zhang, S. Fanello, Y. Tsai, T. Sun, T. Xue, R. Pandey, S. Orts-Escolano, P. Davidson, C. Rhemann, P. Debevec, J. T. Barron, R. Ramamoorthi, and W. T. Freeman. Neural light transport for relighting and view synthesis. arXiv preprint arXiv:2008.03806, 2020.

[56] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. Deep single portrait image relighting. In International Conference on Computer Vision (ICCV), 2019.
6. Supplementary Material

This section includes supplementary results and descriptions of the experiments discussed in the main draft.

6.1. Notations

Let $\text{Conv}(c_i, c_o, k, s, p)$ denote a convolution layer with $c_i$ and $c_o$ input and output channels, $k$ kernel size, stride $s$ and padding $p$. Also, $\text{bn}(c_i)$ represents a batch normalization layer with $c_i$ input channels. We define a convolution layer followed by ReLU and batchnorm as a convolution block, denoted by $\text{ConvB}(c_i, c_o, k, s, p) = [\text{Conv}(c_i, c_o, k, s, p) - \text{relu} - \text{bn}(c_o)]$. We also define deconvolution block similarly as $\text{DeconvB}(c_i, c_o, k, s, p) = [\text{ConvTranspose}(c_i, c_o, k, s, p, p_o) - \text{relu} - \text{bn}(c_o)]$ with $p_o$ denoting the output padding. Also, $\text{FC}(n_i, n_o)$ is a fully connected layer with $n_i$ and $n_o$ input and output sizes.

6.2. Our architecture

We use two architectures in our experiments. In Arch 1 which is used for training on the synthetic dataset only, we have a U-net based encoder-decoder. The architecture of the encoder is $[\text{ConvB}(3, 32, 5, 2, 2) - \text{ConvB}(32, 64, 5, 2, 2) - \text{ConvB}(64, 128, 5, 2, 2) - \text{ConvB}(128, 256, 5, 2, 2)]$. The features output by the encoder are then passed through two separate cascade of convolution blocks ([$\text{ConvB}(256, 256, 3, 1, 1) - \text{ConvB}(256, 256, 3, 1, 1) - \text{ConvB}(256, 256, 3, 1, 1)$]), to represent the normal and albedo features. The normal and albedo features are then passed through two separate decoders with the same architectures as $[\text{DeconvB}(256, 128, 5, 2, 2, 1) - \text{DeconvB}(128, 64, 5, 2, 2, 1) - \text{DeconvB}(64, 32, 5, 2, 2, 1) - \text{DeconvB}(32, 16, 5, 2, 2, 1) - \text{ConvB}(16, 3, 5, 2, 1)]$. Note that we have skip-connections from different layers of the encoder to the respective layers of the decoders. For the light decoder, we concatenate the normal and albedo features along with the image features from the encoder and pass it through a $\text{ConvB}(768, 256, 1, 1, 0)$, an average pooling and $\text{FC}(256, 27)$.

In Arch 2, which is used for training on a combination of synthetic and real dataset, we no longer have skip connections. We use the same architecture as [37] for Arch 2.

6.3. Details on the DPR baseline

Our architecture for the DPR baseline is a skip-connection based encoder-decoder similar to [57]. The encoder consists of 5 convolution blocks as $[\text{ConvB}(3, 16, 5, 1, 2) - \text{ConvB}(16, 16, 3, 2, 1) - \text{ConvB}(16, 32, 3, 2, 1) - \text{ConvB}(32, 64, 3, 2, 1) - \text{ConvB}(64, 155, 3, 2, 1)]$. The last 27 channels of the extracted feature from the encoder is then passed through an adaptive average pooling layer followed by two fully connected layers to output a $27 \times 1$ lighting vector. This would be the estimated lighting vector for the input (source) image.

At the decoder side, the target lighting is first passed through two fully connected layers (followed by batch norm and ReLU) to generate a $27 \times 1$ feature vector for the target lighting. This feature vector is then repeated to have the same width and height as the 2D features generated by the encoder. This repeated feature is then concatenated with the first 128 channels of the encoder feature to form a feature with 155 channels. This feature is then passed through the decoder structured as: $[\text{DeconvB}(155, 64, 3, 2, 1, 1) - \text{DeconvB}(64, 32, 3, 2, 1, 1) - \text{DeconvB}(32, 16, 3, 2, 1, 1) - \text{DeconvB}(16, 16, 3, 2, 1, 1) - \text{ConvB}(16, 3, 3, 1, 1)]$. Note that we have 4 skip connections from the encoder to the decoder. The features from different layers of the encoder are passed through convolution layers with kernel size 3 and the same number of input/output channels as the input features. The loss function used to train the DPR baseline is:

$$L(\theta) = \| I_2 - f_\theta(I_1, L_2) \|_1 + \lambda \| L_1 - \hat{L}_1 \|_1 + \| I_1 - f_\theta(I_2, L_1) \|_1 + \lambda \| L_2 - \hat{L}_2 \|_1$$

where $(I_1, I_2)$ are the multi-lit image pair from the multi-lit CelebA dataset, $L_1$ and $L_2$ are the ground truth (GT) lighting vectors for $I_1$ and $I_2$ respectively. Also, $\hat{L}_1$ and $\hat{L}_2$ are the estimated source lighting vectors for $I_1$ and $I_2$ images and $f_\theta$ denotes the DPR encoder-decoder network parameterized by $\theta$. Note that while training the DPR baseline, we have supervision over the relit image and the estimated source lighting by the network.

6.4. Details on Pix2Pix baseline

For the Pix2Pix [15] baseline, we have a conditional GAN [30] model. The generator consists of a U-net based encoder-decoder. The encoder takes as input the image and the target lighting vector. To concatenate these two before feeding it to the encoder, we first repeat the lighting vector to have the same width and height as the input image. Then, we concatenate this repeated vector with the input image along the channel dimension. The encoder consists of 4 convolution layers with output channel sizes of 32, 64, 128 and 256 with ReLU activations and batchnorm, kernel size 5 and stride length 2. The decoder also have 4 deconvolution blocks to upsample the lower resolution feature from
the encoder. Also, there are skip-connections between respective layers of the encoder and decoder to allow the flow of high frequency features directly from the input to the output. The discriminator consists of 3 convolution layers, with 32, 64 and 128 channels and a kernel size of 5 and stride 2. The output of the cascaded convolution blocks is then passed through an adaptive average pooling layer and two fully connected layers with 128 and 1 output sizes. To have the final score from the discriminator between 0 and 1, we pass the final output through a sigmoid. The discriminator takes as input the relit image along the target lighting. We use the following loss to train the Pix2Pix baseline:

$$L(\varphi, \phi) = \lambda \| G_\varphi(I_s, L_t) - I_t \|_1 + \log(D_\phi(I_t)) + \log(1 - D_\phi(G_\varphi(I_s, L_t)))$$

where $G_\varphi$ and $D_\phi$ denote the generator and the discriminator parameterized by $\varphi$ and $\phi$. $I_s$ and $I_t$ are the source and target images and $L_t$ denotes the target lighting. Note that in (4) the first term enforces the relit image to be close to the ground truth target image. We train Pix2Pix baseline on multi-lit CelebA [27] dataset which contains lighting vectors from the SfSNet dataset [37].

6.5. Challenging examples

In Fig. 9 we present some challenging examples. Note that, the model has difficulties in the decomposition task for some non-frontal poses. Also, cast shadows alongside partial occlusions result in less accurate decomposition (although still reasonable).

6.6. Visual results on Photoface

In Fig. 10 we visualize decomposition and relighting results on the Photoface dataset [55] for a model trained from scratch on the Photoface dataset. We provide both successful and failed examples, outlined in green and red. For low-light images or parts of the face under shadow, the recovered albedo has low quality. This then leaks to the relit image, leading to a lower quality relit image and visible artifacts.

6.7. Decomposition results

In Fig. 11 we present more decomposition and relighting results for the weakly supervised model and the model trained on a combination of synthetic and real datasets.

6.8. Visual comparisons to DPR and Pix2Pix

In Fig. 12-13 we visually compare the relighting performance of our model against the DPR and Pix2Pix models we trained on the multi-lit CelebA dataset. Note that our model compared to the baselines generate relit images that better match the target lighting and look natural.

6.9. Light transfer

In Fig. 14-15 we present additional visual results and comparisons with SfSNet [37] on the light transfer task.
(a) The weakly supervised model. During training we have used light supervision, reconstruction and cross-relighting loss terms.

(b) Model trained on a mixture of synthetic and real datasets. We have full supervision on synthetic samples, while only use $L_{\text{rec}}$ and $L_{\text{relit}}$ for samples from multi-lit CelebA.

Figure 11: Decomposition results of different models.

Figure 12: Comparison between the relighting results of our model vs DPR [57] and Pix2Pix [15] baselines trained on multi-lit CelebA dataset. The target lights are depicted in the first row.

Figure 13: Comparison between the relighting results of our model vs DPR [57] and Pix2Pix [15] baselines trained on multi-lit CelebA dataset. The target lights are depicted in the first row.
Figure 14: Comparison between our model and SfSNet [37] for the light transfer task. $S\cdot$ denotes shading.

Figure 15: Comparison between our model and SfSNet [37] for the light transfer task. $S\cdot$ denotes shading.