De-rendering 3D Objects in the Wild

Felix Wimbauer[1,2] Shangzhe Wu[2] Christian Rupprecht[2]
[1] Technical University Munich, [2] University of Oxford
wimbauer@in.tum.de {szwu, chrisr}@robots.ox.ac.uk

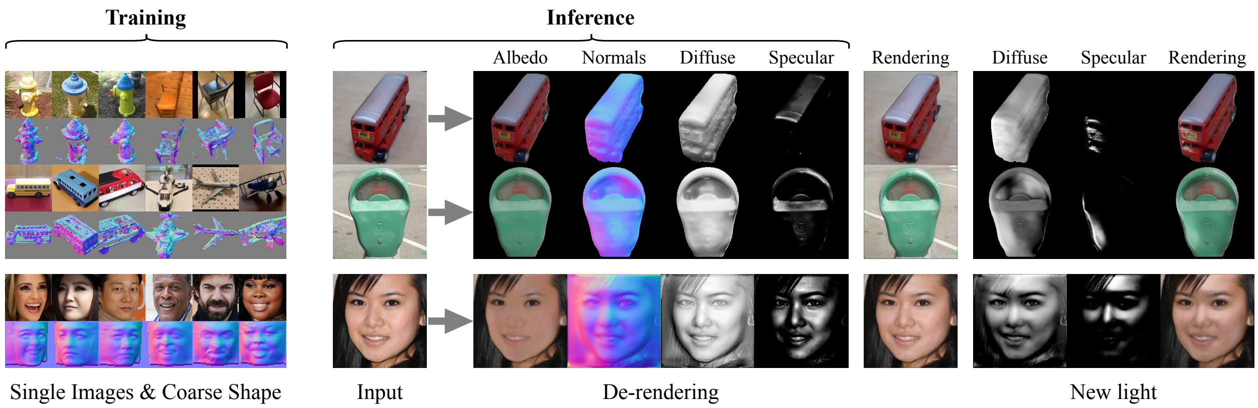

Figure 1. **De-rendering 3D objects from in-the-wild images.** Left: Our method is trained on unconstrained images with coarse shape estimates which can be obtained using off-the-shelf methods or classical structure from motion. Right: Our method de-renders an image into precise shape, material (diffuse albedo, specularity, specular intensity), and lighting (direction, intensity). The image can then be re-rendered and re-lit based on our illumination model. The model generalizes to objects beyond the training categories (*e.g.* parkingmeter).

Abstract

With increasing focus on augmented and virtual reality (XR) applications comes the demand for algorithms that can lift objects from images into representations that are suitable for a wide variety of related 3D tasks. Large-scale deployment of XR devices and applications means that we cannot solely rely on supervised learning, as collecting and annotating data for the unlimited variety of objects in the real world is infeasible. We present a weakly supervised method that is able to decompose a single image of an object into shape (depth and normals), material (albedo, reflectivity and shininess) and global lighting parameters. For training, the method only relies on a rough initial shape estimate of the training objects to bootstrap the learning process. This shape supervision can come for example from a pretrained depth network or—more generically—from a traditional structure-from-motion pipeline. In our experiments, we show that the method can successfully de-render 2D images into a decomposed 3D representation and generalizes to unseen object categories. Since in-the-wild evaluation is difficult due to the lack of ground truth data, we also introduce a photo-realistic synthetic test set that allows for quantitative evaluation. Please find our project page at: https://github.com/Brummi/derender3d

1. Introduction

From a single 2D image, humans can easily reason about the underlying 3D properties of an object, such as the 3D shape, the surface material and its illumination properties. Being able to infer "object intrinsics" from a single image has been a long standing goal in Computer Vision and is often referred to as "inverse rendering" or "*de-rendering*" as it reverses the well-known rendering step of Computer Graphics, where an image is generated from a similar set of object and material descriptors.

De-rendering an image into its physical components, not only plays an important role for general image understanding, but is also key to many applications, such as Augmented/Virtual Reality (XR) and Visual Effects (VFX). In these applications, a decomposed 3D representation can be used to increase the realism by enabling post-processing steps, such as relighting or changing the texture or material properties, which further blurs the line between real and synthetic objects in these environments.

As XR is moving from research and commercial use to consumer devices, a de-rendering method should work on a wide variety of images in the wild to allow a broad adoption of these technologies. While the history of image decomposition literature is long [13,14], recent learning-based approaches have demonstrated this capability on specific categories, such as human portraits [38] and synthetic ShapeNet objects [49], by training on ground-truth data, often obtained using synthetic models or sophisticated light stage capturing systems. However, obtaining large-scale ground-truth material and illumination annotations for general objects "in the wild" is much more challenging and infeasible to collect for *all* objects. Models trained on synthetic data often lack sufficient realism, resulting in poor transfer to real images. Models trained on real data usually focus on a single category (*e.g*. faces or birds [11, 17, 24, 25, 55]) and do not generalize to new classes.

On the other hand, another line of research that has recently gained interest, aims to learn 3D objects in an unsupervised or weakly-supervised fashion, without relying on explicit 3D ground-truth [17, 19, 25, 36, 55]. Although impressive results have been demonstrated in reconstructing 3D shapes of simple objects, few of the methods have considered also recovering specular surface materials as this introduces even more ambiguities to the model. Furthermore, they are generally restricted to a single category.

In this paper, we explore the problem of learning non-Lambertian intrinsic decomposition from in-the-wild images without relying on explicit ground-truth annotations. In particular, we introduce a method that capitalizes on the coarse 3D shape reconstructions obtained from unsupervised methods and learns to predict a refined shape as well as further decomposes the material into albedo and specular components, given a collection of single-view images.

At the core of the method lies an image formation process that renders the image from its individual components. The model then learns to decompose the image through a reconstruction objective. Since this formulation is highly ambiguous, the model relies on several additional cues to enable learning a meaningful decomposition. We bootstrap the training using a coarse estimate of the initial shape. This estimate can come from a variety of sources. For datasets such as Co3D [40], where multi-view information is available, we rely on traditional structure-from-motion pipelines (*e.g*. COLMAP [44]). For specific categories such as faces, existing specialized unsupervised methods can be used to obtain a coarse initial shape estimate. We present a simple method that estimates initial material and light properties using the coarse shape, the input image and a simple lighting model. We can then facilitate learning by using the coarse estimates as initial supervisory signals, which avoids many degenerate solutions that would fulfill the reconstruction objective alone. Finally, to further improve the quality of the decomposition, we introduce a third objective, where the image is rendered with randomized light parameters, and a discriminator helps to ensure realistic reconstructions.

While we do need (pseudo) supervision of the coarse shape during training, the final model can directly decompose an input image without any other input. We show that our model produces accurate and convincing image decompositions that improve the state of the art and even generalizes beyond the categories of objects it was trained on. In our experiments, we show that the model works on a wide variety of objects from different datasets. However, as this is the first method to tackle de-rendering in the wild, there is currently no suitable benchmark to quantitatively evaluate the quality of the decomposition. We thus also introduce a synthetic benchmark dataset, using photo-realistic rendering of 10 objects from several viewpoints. Each image is associated with ground truth per-pixel material properties and lighting information that allows us to directly evaluate the decomposition. The new dataset, code and trained models will be published together with the paper.

2. Related Work

This work studies the problem of learning to de-render images of general objects "in the wild", which lies in the intersection of several fields of Computer Vision and Computer Graphics. In this section, we will first discuss the relevant work on Intrinsic Image Decomposition and Inverse Rendering from multiple images, and direct supervision as well as recent unsupervised approaches.

Instrinsic Image Decomposition. Intrinsic Image Decomposition is a classic task, where the main goal is to factorize an image into a reflectance image and a shading image, *i.e*. separating the true surface color from lighting effects. Since this is a highly ill-posed task, traditional methods often rely on additional heuristics and priors. The classic Retinex algorithm [23] assumes that small variations in image intensity result from shading whereas abrupt changes reveal the true reflectance. Many other priors have also been explored over the past few decades, such as global sparsity constraints on the reflectance [10, 41, 46, 47], and explicit geometric constraints on shading assuming Lambertian surface [2, 21]. Recently, researchers have also studied learning-based approaches, by training on synthetic data [16, 33] or multi-illumination images [27, 31]. In this work, we borrow ideas from this area to constrain the albedo extraction, but aim at decomposing the image into explicit material, shape and lighting factors rather than a single shading map, as this allows for relighting and re-rendering.

Supervised Inverse Rendering. Next, we will focus on inverse rendering methods that recover shape, material and illumination from images. Classical Shape-from-Shading approaches assume Lambertian surface properties [13, 14]. Photometric Stereo techniques [1, 12] recover shape, BRDF material and lighting by solving an optimization problem, given multiple images of a scene captured under various lighting conditions and/or from multiple viewpoints. This has been extended with learning-based approaches [3, 4, 6, 59–61] and recently with implicit (neural) representations [5, 50, 57, 62]. While most of these methods still require multiple images at inference, some have learned priors from multiple views that can be used for single image inference [20, 34]. However, capturing multiple images with controlled lighting for either training or inference is challenging and difficult to apply to objects "in the wild", which is main target of this paper.

The de-rendering task can also be learned with direct supervision, often using synthetic data, like ShapeNet [7], objects [6, 8, 9, 32, 49], synthetic faces/bodies [15, 22, 45], near-planar surfaces [28], indoor scenes [26] or other synthetic objects [16, 29, 43]. However, generating large-scale realistic synthetic data that captures the level of complexity of the real world is challenging, and hence it remains questionable how well these methods generalize to real images. As inverse rendering and relighting of faces and persons is particularly useful, relighting datasets for real faces have been collected using light stage setups [37, 38, 51, 52]. This approach, however, is not feasible for general objects.

Unsupervised Inverse Rendering. Recently, there has been an increasing interest in developing unsupervised or weakly supervised methods for inverse rendering tasks. Several works have attempted to learn 3D shapes of object categories, such as faces and birds, from only single-view image collections [11, 17, 25, 55, 58], with weak supervision such as 2D keypoints, masks, category template shapes or assumptions like symmetry. Most of these focus on shape learning and do not tackle material and lighting decomposition specifically or assume a simple Lambertian shading model. Wu *et al*. [54] recovers shape, shiny material and environment lighting, but focuses only on a single specific type of object—vases—and assumes rotational symmetry.

Unlike all these approaches, this work aims at recovering specular material and illumination on *general objects* from images in the wild, with only coarse geometry estimations during training, that can be obtained from existing methods.

3. Method

In this section we will describe the model and training scheme of our method. Fig. 2 shows an overview of the decomposition, training procedure and losses.

3.1. Rendering - Image Formation Model

Essential to our method is modeling the image formation process, *i.e*. rendering an image from its intrinsic components. Our method learns to invert this process—*de-rendering* the image—by extracting the intrinsic components from an input image $\mathbf{I}_{\text{in}} \in [0,1]^{3\times H\times W}$.

While the rendering process is usually deterministic, it is however surjective. This means that due to the highly complex nature of the image formation process, the inverse is ambiguous and many different combinations of intrinsic materials map to the same image.

We deal with the highly ambiguous inverse rendering step in three ways. First, we make reasonable assumptions about the object's material that drastically simplify the rendering process. Second, we leave enough flexibility in the rendering process, such that the model can *learn* to overcome the approximations used in the first step. Finally, we use coarse shape supervision using traditional methods or existing, object-specific solutions to bootstrap the learning process and to avoid degenerate solutions.

Shape. As the shape of an object has a strong influence on its shading, we will directly link these two components. To compute per-pixel shading, we require a per-pixel normal map $N \in [-1,1]^{3\times H\times W}$. Given an image, directly predicting a normal map with a neural network is problematic, as there is no incentive to adhere to a global shape. Thus, we compute the normal map N_D from a depth map $D \in [d_{\min}, d_{\max}]^{H\times W}$. However, fine geometric details (*e.g*. scratches or small reliefs) have a strong influence on the normals, but little on the global shape. Thus, we predict both a depth map D (and compute the corresponding normal map N_{D}) and a refinement normal map N_{ref} from the image, and combine them:

$$N = \frac{N_{\text{D}} + N_{\text{ref}}}{\|N_{\text{D}} + N_{\text{ref}}\|} \tag{1}$$

Light & Material. Using a very expressive lighting model would allow us to capture highly complex effects for photo-realistic rendering. However, we find that such models add significant difficulty to the learning of inverse tasks without further supervision. To model the lighting, we hence rely on Phong Illumination [39], which considers ambient, diffuse, and specular light components. Additionally, we make the following assumptions. We can observe during training that the shading from the one dominant light source (*e.g*. the sun) is a very important hint for the model. Further, multiple light sources would introduce more ambiguity, potentially harming the correctness of the predicted geometry. Therefore, we model the light as a single directional light source and a global ambient light, both emitting perfectly white light. It is parameterized by ambient and directional strength $s_{\text{amb}}, s_{\text{dir}} \in [0,1]$, and a light direction $l \in SO(3)$. For both terms, we use a combined per-pixel albedo map $A \in [0,1]^{3\times H\times W}$. Specularity is a very complex lighting effect and therefore difficult to extract from a single image. To keep the complexity tractable, we use a global shininess value $\alpha \in [0, \alpha_{\max}]$ and a global specularity intensity $a_{\text{spec}} \in [0,1]$ for the whole object. Summarizing, we represent the lighting as $L = (s_{\text{amb}}, s_{\text{diff}}, l)$ and the intrinsic material properties as $(A, \alpha, a_{\text{spec}})$.

We obtain an image $\hat{\mathbf{I}}$ from shape, material, and lighting through the following rendering equation, where $u \in \Omega =$

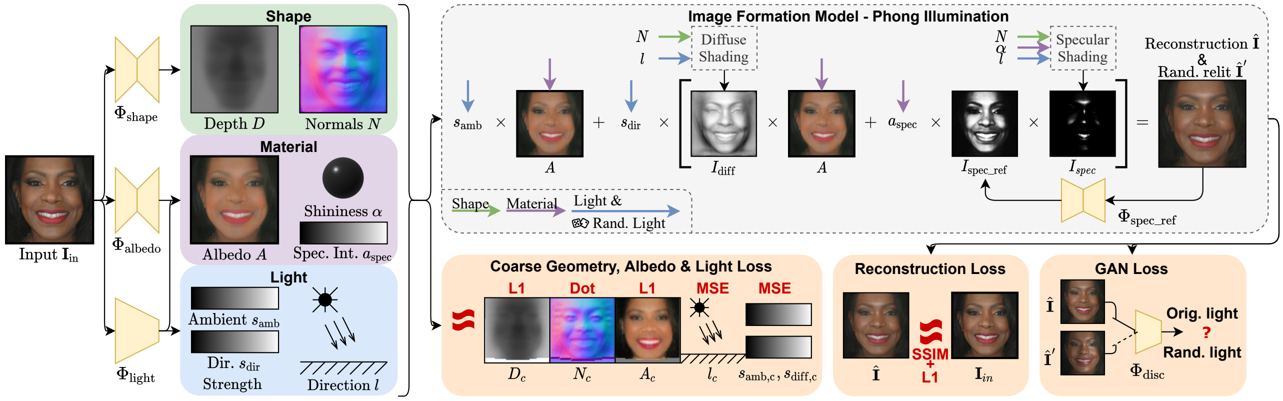

Figure 2. **Model architecture.** Several networks decompose the input image $\mathbf{I}_{\text{in}}$ into shape (depth D and normals N), material (albedo A, shininess α and specular intensity a_{spec}), and light (ambient and directional strength $s_{\text{amb}}, s_{\text{dir}}$ and light direction l). To allow for complex lighting effects, we also introduce a specularity refinement step. Training combines three different loss terms: 1. a loss on the decomposition using coarse estimates 2. a reconstruction loss, and 3. a discriminator loss for a randomly relit image.

$\{1, \ldots, H\} \times \{1, \ldots, W\}$ represents a pixel location.

$$\hat{\mathbf{I}}_u = \tau(\underbrace{s_{\text{amb}} A_u}_{\text{ambient term}} + s_{\text{dir}}(\underbrace{N_u^T l A_u}_{\text{diffuse term}} + \underbrace{a_{\text{spec}} \left(N_u^T v\right)^{\alpha}}_{\text{specular term}})), \quad (2)$$

where $\tau(\mathbf{I}_u) = \mathbf{I}_u^{1/\gamma}$, $\gamma = 2.2$ denotes the approximation of tone-mapping by a gamma function, which is commonly used to ensure a more even brightness distribution.

3.2. De-rendering

Our network architecture is composed of several sub-networks, that predict the different shape, material, and lighting properties of an input image. From these predictions, we can reconstruct the image with the image formation model described above. However, as mentioned before, due to the plethora of ambiguities in the rendering function, a simple reconstruction objective alone (*e.g.* $\|\mathbf{I} - \hat{\mathbf{I}}\|_2^2$) is not sufficient to learn a meaningful decomposition. To overcome this challenge, we propose a training scheme, with two additional objectives that regularize the learning problem and prevent degenerate solutions. As training data, we use an unconstrained set of images with associated coarse geometry estimates. We use the coarse geometry to generate further coarse estimates for the intrinsic components, which we use in auxiliary loss terms. These coarse constraints force the model to predict a semantically correct disentanglement on a global level.

Extracting Coarse Light & Albedo. Because coarse shape (depth map D_c and its normal map N_c) alone does not suffice to constrain the decomposition, we also compute coarse light and albedo estimates from the geometry information through two optimization steps.

As we only need coarse estimates of the intrinsic components, we can make the simplifying assumption that the per-pixel coarse brightness $B \in [0,1]^{H \times W}$ (computed in HSV color space) of the input image is proportional to the combination of ambient and diffuse shading, and discard specular lighting effects. This translates to an albedo map with constant brightness. Given light information, we can obtain the relevant shading map from the coarse geometry. Therefore, we optimize the coarse light parameters $L_c = (s_{\text{amb},c}, s_{\text{dir},c}, l_c)$ such that the aggregated shading map corresponds to the brightness of the input image.

$$\arg\min_{L_c} \sum_{u \in \Omega} \left(2B_u - (s_{\text{amb},c} + s_{\text{dir},c} N_{c,u}^T l_c)\right)^2 \quad (3)$$

We fix the albedo brightness to $\frac{1}{2}$ to avoid color saturation effects and consequently add a scaling factor of 2 for B. Here, N_c is the coarse normal map. With this light estimate, an initial albedo estimate $\tilde{A}_c$ can be obtained by inverting the shading equation:

$$\tilde{A}_{c,u} = \mathbf{I}_u \left(s_{\text{amb},c} + s_{\text{dir},c} N_{c,u}^T l_c\right)^{-1} \quad (4)$$

However, because of the coarseness of the geometry and no modeling of specularity effects, an estimate using this formulation alone will contain many artifacts. To regularize the estimate $\tilde{A}_c$ we refine it using another optimization step. Similar to the constraints used in the intrinsic image decomposition literature [48, 56], we apply total variation regularization (TV) on the albedo as well as a data term that retains the image gradients (*i.e.* edges):

$$\begin{aligned}\arg\min_{A_c} \|\delta_x A_c - \delta_x \tilde{A}_c\|_2 + \|\delta_y A_c - \delta_y \tilde{A}_c\|_2 + \\ \lambda_{\text{TV}} \left(\|\delta_x A_c\|_1 + \|\delta_y A_c\|_1\right) .\end{aligned} \quad (5)$$

We use δ_x and δ_y to signify the computation of image gradients, which can, for example, be obtained by applying the Sobel operator to the image. We obtain L_c and A_c by optimizing Eq. (3) and Eq. (5) respectively using gradient descent, which takes less than a second and can be precomputed for each image (see Fig. 3).

Learning to De-render. We use three different neural networks to predict the intrinsic components from the input image $\mathbf{I}_{\text{in}}$. A shape network Φ_{shape} predicts both the depth map $D, D_u \in [d_{\min}, d_{\max}]$ and the normal refinement map N_{ref}, which is normalized after prediction and used to obtain the final normal map N with Eq. (1). The albedo network Φ_{albedo} predicts the albedo map $A, A_u \in [0, 1]$, and the light network Φ_{light} predicts the light parameters $s_{\text{amb}}, s_{\text{dir}} \in [0, 1]$, as well as, shininess $\alpha \in [0, \alpha_{\max}]$ and specular intensity $a_{\text{spec}} \in [a^{\text{spec}}_{\min}, a^{\text{spec}}_{\max}]$.

We train our model using complementary losses on the decomposition and on the rendered image. This makes the network adhere to globally accurate components, while achieving more detailed reconstructions. The loss is computed using the (precomputed) coarse shape, albedo, and light information as pseudo supervision.

$$\begin{aligned}\mathcal{L}_c = \sum_{u\in\Omega} \lambda_D \|D_u - D_{c,u}\|_1 - \lambda_N N_u^T N_{c,u} + \\ \lambda_A \|A_u - A_{c,u}\|_1 + \lambda_L \|L - L_c\|_2^2\end{aligned} \tag{6}$$

Additionally, there are two losses on the rendered image. First, we apply a reconstruction loss between the rendered and the input image to train our model to capture all local details in the decomposition. Specifically, this loss term is computed from the combination of a per-pixel L1 loss and the patch-based structural similarity score $\text{SSIM}(\mathbf{I}, \hat{\mathbf{I}})$ [53].

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \sum_{u\in\Omega} \|\mathbf{I}_u - \hat{\mathbf{I}}_u\|_1 + \frac{1}{2}\left(1 - \text{SSIM}(\mathbf{I}, \hat{\mathbf{I}})\right) \tag{7}$$

While the reconstruction loss gives a very strong training signal, there often remains some ambiguity, in that, given fixed light, certain details can be modeled either by the material (light independent) or the shape component (light dependent). Such mistakes only become apparent when we render the image under new lighting conditions L' (mainly influenced by the direction l') .

To ensure that we achieve a semantically correct decomposition, we therefore also introduce an adversarial formulation. Concretely, we render two images in each forward pass: one with the predicted lighting conditions, denoted as $\hat{\mathbf{I}}$, which is also used in the reconstruction loss term, and one with randomly sampled lighting conditions, denoted as $\hat{\mathbf{I}}'$. We then train a discriminator network $\Phi_{\text{disc}} \in \mathbb{R}$ to score, whether an image was rendered using the original lighting conditions or whether it was re-lit. For this we use the discriminator from LSGAN [35]. Using the reconstructed image $\hat{\mathbf{I}}$ instead of the original image $\mathbf{I}_{\text{in}}$ as positive example when training Φ_{disc}, has the advantage that the network cannot use artifacts from the image formation models as hints as to whether the image was re-lit or not, as both, real and fake examples come from the same pipeline. The loss term on the relit image is computed as: $\mathcal{L}_{\text{gan}} = (1 - \Phi_{\text{disc}}(\hat{\mathbf{I}}'))^2$. We can then train the whole model end-to-end using

$$\mathcal{L} = \mathcal{L}_c + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{gan}}\mathcal{L}_{\text{gan}}\ , \tag{8}$$

to learn to de-render an image into its intrinsic components.

Refinement. While simplifying the specularity model to two scalars allows for stable training, it can be limiting when there are large differences in material properties across the object. To alleviate this issue, similar to the way we allow the normals to deviate from the underlying shape via a refinement map, we predict a per-pixel specularity refinement map $I_{\text{spec\_ref}}(\hat{\mathbf{I}})$ from the *output* image. We then multiply $I_{\text{spec\_ref}}$ with the specularity term and re-compose the image.

4. Experiments

We conduct extensive experiments to evaluate our method and its individual components.

4.1. Datasets and Metrics

We use three different datasets to cover a wide variety of objects: faces, a collection of common objects in the wild, and a new synthetic and photo-realistic test set with ground truth annotations. Please see the supplement for all details.

CelebA-HQ [18] is a large-scale human face dataset, consisting of 30k high-resolution portrait pictures of celebrities. We roughly crop out the face area and use the corresponding train/val/test split of the CelebA dataset. To obtain the rough initial geometry estimate D_c, we use [55] at a reduced resolution of 64×64.

Co3D [40] is a collection of nearly 19,000 videos capturing objects from 50 MS-COCO [30] categories, that come with per-frame depth, camera pose data, and reconstructed sparse point clouds. First, we use the Point Cloud Library [42] to compute surface normals from the point clouds. The resulting depth and normal maps are very sparse (see Fig. 3). We select a subset of the categories and obtain 23895 training and 2817 testing images.

COSy (Common Objects Synthetic) is a test set we have created to allow for quantitatively evaluation of image decomposition methods. This is necessary as there does not exist a dataset that combines photorealistic images with precise image decomposition ground truth. We hand-select 10 freely available and photorealistic 3D scenes for the Blender 3D modeling software[1] and define 4 different camera views for each. Additionally to the rendered image, we also save the diffuse albedo map, normal map, and foreground mask. We do not use this dataset of 40 images only for testing.

[1] https://blender.org

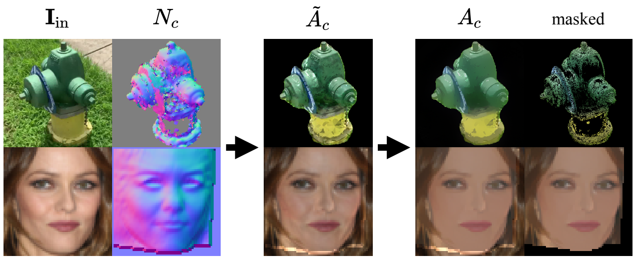

Figure 3. **Coarse Albedo Optimization.** For training, we precompute coarse albedo estimates A_c from the input image $\mathbf{I}_{\text{in}}$ and the coarse shape estimate N_c. We first approximate the light by assuming a Lambertian shading model and inverting it. Subsequently, we apply a smoothing optimization to remove artifacts.

Post-processing for Training. For every image, we compute the normal map from the rough initial depth estimate and optimize a light and albedo approximation (see Fig. 3).

We train one model for the **CelebA-HQ** dataset and one model for the **Co3D** dataset. All hyper-parameters are the same for both models (see supplementary material) except for $\lambda_L = 0$ on **Co3D**, as the geometry is very sparse and the light estimate is often not accurate enough, hindering the convergence of the training.

Metrics. As other methods use different image formation models, and thus obtain a different representation for shading, it is not possible to directly compare shading maps. This constrains us to quantitative evaluation of normal and albedo map only. Aside from the common $\mathcal{L}_1$, $\mathcal{L}_2$, and SSIM error metrics, we use mean angle deviation in degrees $\text{DIA}(N, \hat{N}) = \frac{1}{|\Omega|}\sum_{u\in\Omega}\cos^{-1}(N_u^T\hat{N}_u)$ for normals, and the scale-invariant error $\text{SIE}(A, \hat{A}) = \frac{1}{|\Omega|}\sum_{u\in\Omega}\|A_u - \mu_A - (\hat{A} - \mu_{\hat{A}})\|_2^2$ for albedo, as it can be estimated only up to a constant scale factor. Here, μ_A is the average albedo across the whole image $\mu_A = \frac{1}{|\Omega|}\sum_{u\in\Omega} A_u$.

4.2. Results

Qualitative Evaluation. To demonstrate the capabilities of our method, we first evaluate it on a diverse selection of samples, as shown in Fig. 4. Regardless of the category and background, we obtain globally correct results with a very high fidelity. Critically, even though the COSy dataset object categories are *not* part of the training categories, we observe the same level of detail. This demonstrates our method's generalization capabilities to novel objects and categories. In addition to the decomposition results, we also show that our method produces realistic images and shading maps when changing the light.

Further, we compare our results with state-of-the-art methods for intrinsic image decomposition [2, 29, 43, 49] in Fig. 5. All methods are able to predict reasonable albedo maps, that capture the major color components. However, the albedo maps of [2, 29, 43] still contain color gradients and lighting effects around the edges and corners. [49] is able to remove almost all specular components of the roof but introduces artifacts, for example, at the top of the roof. Our method successfully removes shading effects and does not contain artifacts. For normal prediction, [2] does not capture the shape of the object, nor fine details. Although [29] and [43] predict seemingly detailed normal maps, closer inspection shows that they are not physically grounded (*e.g.* the normals on the windows point upwards). Our normal map is both detailed and it adheres to the global shape. Finally, our shading maps are computed from the normal map and additional material properties. This is why they are similarly detailed and based on physical properties. Diffuse and specular effects are captured correctly. The diffuse shading map predicted by [49] is very detailed, but does not capture the light direction.

	Normal N		Albedo A	
Model	**MSE** ↓	**DIA** ↓	**SIE** ↓	**SSIM** ↑
SIRFS [2]	0.331	52.994	0.113	0.724
ShapeNet-Intr. [49]	N/A	N/A	0.114	0.726
SISaSVBRDF [29]	0.288	42.801	0.112	0.752
Neur. Rel. [43]	0.228	41.603	0.093	0.723
Ours	**0.173**	**37.807**	**0.075**	**0.760**

Table 1. **Comparison with the state of the art.** We show good improvements over previous methods on the COSy dataset. [29] and [43] were trained using flash photographs.

	Normal N		Albedo A		Specular I_{spec}	
Model	**MSE** ↓	**DIA** ↓	**SIE** ↓	**SSIM** ↑	**MSE** ↓	**L1** ↓
No Albedo	0.162	36.5	0.088	0.750	0.124	0.077
No Shape	0.506	68.7	0.079	0.757	0.108	0.058
No GAN	0.169	37.2	0.075	0.762	0.123	0.073
Ours	0.173	37.8	0.075	0.760	0.112	0.059

Table 2. **Ablation.** Results on the COSy dataset when deactivating components of our model. Concretely, we set the $\lambda = 0$ coefficient for the respective loss term and then do a full training run.

Quantitative Evaluation. As neither **CelebA-HQ** nor **Co3D** contain explicit, dense ground truth for the different intrinsic image components, we apply the **Co3D** model on our newly introduced **COSy** dataset to perform quantitative evaluations. We compare against state-of-the-art image decomposition methods [2, 29, 43, 49], as shown in Tab. 1. Across all metrics, we achieve best accuracy on normal and albedo extraction. The fact, that our method (like the others) was not trained on this test set, highlights its strong generalization capabilities.

Single Image Relighting. To demonstrate the usefulness of de-rendering, we perform relighting on the **CelebA-HQ** dataset. Fig. 6 shows comparison of our method with state-of-the-art face relighting methods [63] and [15]. As a re-

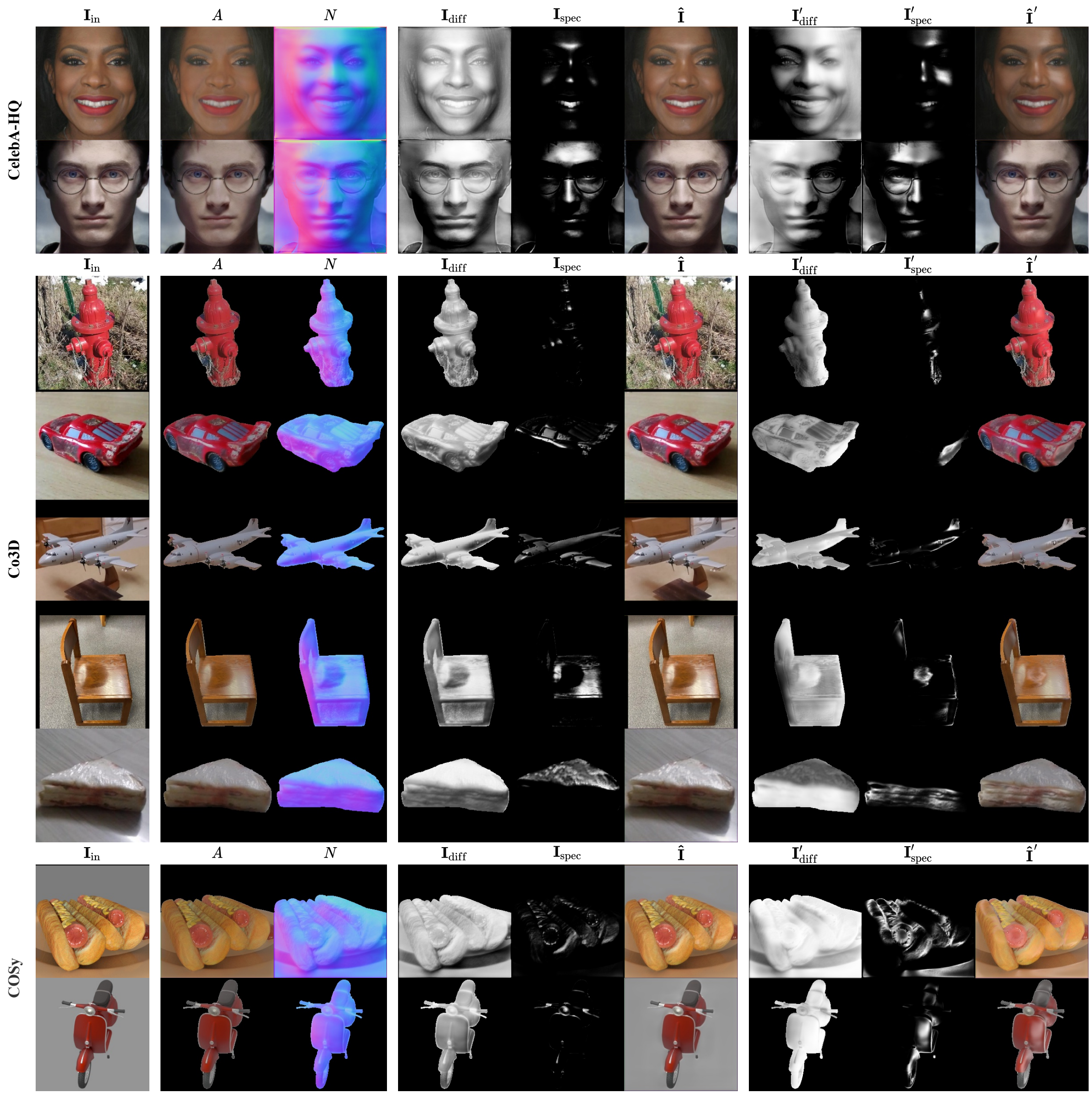

Figure 4. **Qualitative results.** We train a model on each of the two datasets **CelebA-HQ** and **Co3D**, and show the respective decomposition results from the test sets. To highlight the generalization capabilities, we also apply the **Co3D** model to samples from our synthetic test set **COSy**. Every row contains the input image $\mathbf{I}_{\text{in}}$, predicted albedo A and normals N, diffuse shading map $\mathbf{I}_{\text{diff}}$, specular shading map $\mathbf{I}_{\text{spec}}$ and reconstructed image $\hat{\mathbf{I}}$. Further, we also show the shading maps ($\mathbf{I}'_{\text{diff}}$, $\mathbf{I}'_{\text{spec}}$) and reconstructed image $\hat{\mathbf{I}}'$ under new lighting conditions. Our model achieves a high level of detail on shape and material reconstruction and convincing relighting results.

sult of the underlying explicit image formation model, our method produces visually correct relighting results, which are more color accurate than [63] and contain much fewer artifacts than [15]. This demonstrates that our method can not only perform very well in intrinsic image decomposition, but also compete with methods, that were specifically designed for certain sub-tasks.

4.3. Ablation Study and Analysis

We also conduct several ablation studies and further analyses of the impact of the individual model components.

Loss Components. At the heart of our method is the combination of three losses: a coarse loss on the different intrinsic components, a reconstruction loss, and a discriminator loss. We deactivate each component and then evaluate the resulting models on COSy, as shown in Tab. 2.

When deactivating the albedo and shape losses ($\lambda_A = 0$ and $\lambda_D = \lambda_N = 0$ respectively), the predictions of the respective components become significantly worse. The discriminator loss does not have a large influence on the quality of the albedo and normal accuracy, however, it stabilizes the accuracy of the specular shading map.

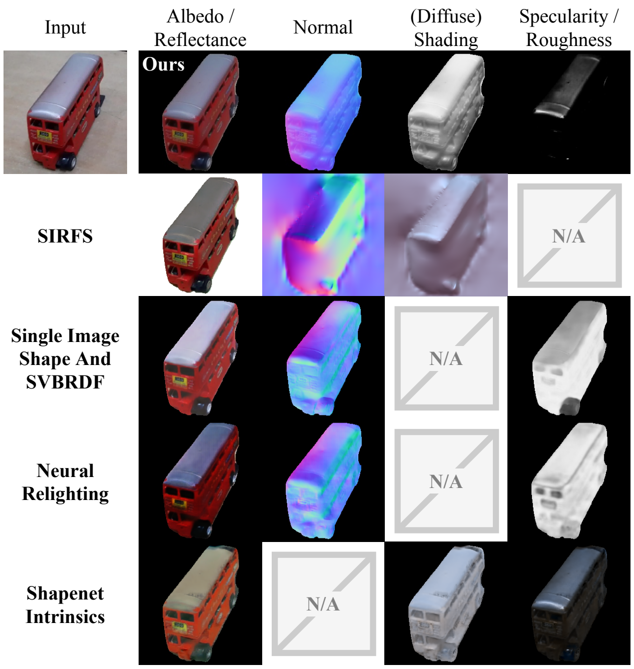

Figure 5. **Qualitative comparison with state of the art.** We show superior image decomposition results compared to SIRFS [2], Single Image Shape and SVBRDF [29], Neural Relighting [43], and ShapeNet Intrinsics [49].

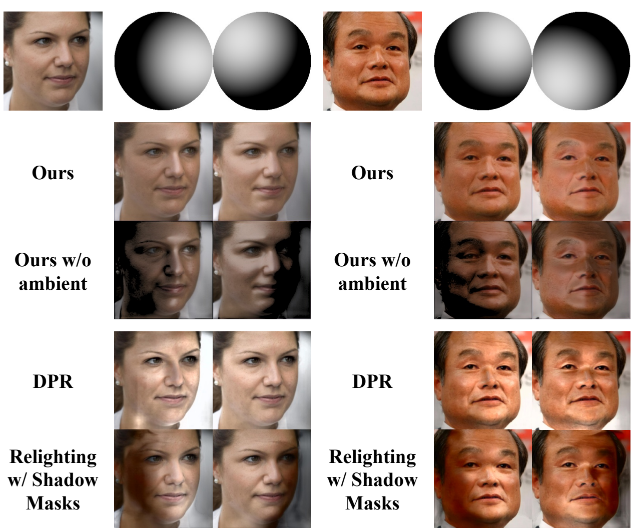

Figure 6. **Qualitative comparison with face relighting methods**. We use our model to relight images from the **CelebA-HQ** test set and compare with state-of-the-art face relighting methods. Our method shows better color accuracy and more robustness.

Geometry & Albedo Improvement. Fig. 3 compares on two (test-set) examples the coarse input during training and the prediction of the trained model. This is to verify that albedo and normal map predictions achieve a significantly higher level of detail and completeness compared to their initial, coarse counterparts that are used to supervise the training process. This is a result of the reconstruction and GAN losses and the explicit image formation model.

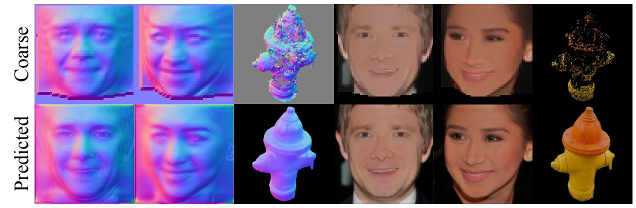

Figure 7. **Coarse estimates vs. prediction.** We compare the rough shape and albedo estimates that we use during training with predictions from the final model and show large improvements. Note that the rough estimates are not available during test time.

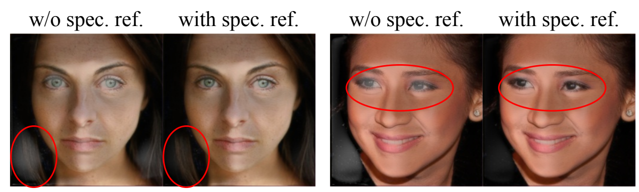

Figure 8. **Specular Refinement.** The specular refinement network improves regions where the complexity of underlying material is difficult to describe explicitly with the lighting model.

Specular Refinement. Fig. 8 demonstrates the specular refinement on two re-lit portrait images. The assumption of shared specular parameters for the entire image, can sometimes lead to specular artifacts in complex regions, which is especially important during relighting. The network effectively removes artifacts both on the hair and around the eyes, leading to a more realistic output.

5. Conclusions

We have presented a method that can factorize in-the-wild images of objects into their intrinsic components shape, material, and lighting. Our proposed learning pipeline does not rely synthetic datasets and only uses sparse geometry estimates during training, which can be obtained using off-the-shelf unsupervised methods. Through a series of ablation studies, we have demonstrated the importance of the different components of our method, particularly the coarse losses. The proposed method achieves high accuracy for all intrinsic components, both on in- and out-of-distribution images, which we measure on our newly introduced synthetic image decomposition test set that we hope will become a new benchmark for de-rendering images in the wild.

Acknowledgments Shangzhe Wu is supported by Meta Research. Christian Rupprecht is supported by Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI) and the Department of Engineering Science at the University of Oxford.

References

[1] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, pages 1–8. IEEE, 2008. 2

[2] Jonathan T. Barron and Jitendra Malik. Shape, Illumination, and Reflectance from Shading. *IEEE TPAMI*, 2015. 2, 6, 8

[3] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV*, pages 294–311. Springer, 2020. 2

[4] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3D Capture: Geometry and Reflectance from Sparse Multi-View Images. In *CVPR*, 2020. 2

[5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 3

[6] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, and Jan Kautz. Two-shot Spatially-varying BRDF and Shape Estimation. In *CVPR*, 2020. 2, 3

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[8] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019. 3

[9] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. Dib-r++: Learning to predict lighting and material with a hybrid differentiable renderer. In *NeurIPS*, 2021. 3

[10] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. *Computer Graphics Forum*, page 1415–1424, 2012. 2

[11] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 2, 3

[12] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE TPAMI*, 2009. 2

[13] Berthold KP Horn. Obtaining Shape from Shading Information. *The Psychology of Computer Vision*, 1975. 1, 2

[14] Berthold K. P. Horn and Michael J. Brooks. *Shape from Shading*. MIT Press, Cambridge Massachusetts, 1989. 1, 2

[15] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *CVPR*, pages 14719–14728, 2021. 3, 6, 7

[16] Michael Janner, Jiajun Wu, Tejas Kulkarni, Ilker Yildirim, and Joshua B Tenenbaum. Self-Supervised Intrinsic Image Decomposition. In *NeurIPS*, 2017. 2, 3

[17] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2, 3

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 5

[19] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *CVPR*, pages 2202–2211, 2019. 2

[20] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep Convolutional Inverse Graphics Network. In *NeurIPS*, 2015. 3

[21] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédo Durand, and George Drettakis. Coherent Intrinsic Images from Photo Collections. *ACM TOG*, 2012. 2

[22] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. Single-image full-body human relighting. In *Eurographics Symposium on Rendering (EGSR)*. The Eurographics Association, 2021. 3

[23] Edwin H. Land and John J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971. 2

[24] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. *NeurIPS*, 33, 2020. 2

[25] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3D reconstruction via semantic consistency. In *ECCV*, 2020. 2, 3

[26] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse Rendering for Complex Indoor Scenes: Shape, Spatially-Varying Lighting and SVBRDF from a Single Image. In *CVPR*, 2020. 3

[27] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018. 2

[28] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *ECCV*, 2018. 3

[29] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to Reconstruct Shape and Spatially-Varying Reflectance from a Single Image. In *SIGGRAPH Asia*, 2018. 3, 6, 8

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5

[31] Andrew Liu, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros, and Noah Snavely. Learning to Factorize and Relight a City. In *ECCV*, 2020. 2

[32] Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, and Jyh-Ming Lien. Material Editing Using a Physically Based Rendering Network. In *ICCV*, 2017. 3

[33] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 2020. 2

[34] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single Image Intrinsic Decomposition without a Single Intrinsic Image. In *ECCV*, 2018. 3

[35] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 5

[36] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Christopher Pal. Unsupervised depth estimation, 3d face rotation and replacement. In *NeurIPS*, 2018. 2

[37] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, pages 5124–5133, 2020. 3

[38] Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM TOG*, 2021. 2, 3

[39] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 3

[40] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, pages 10901–10911, 2021. 2, 5

[41] Carsten Rother, Martin Kiefel, Lumin Zhang, Bernhard Schölkopf, and Peter Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. In *NeurIPS*, 2011. 2

[42] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *ICRA*, Shanghai, China, May 9-13 2011. 5

[43] Shen Sang and Manmohan Chandraker. Single-Shot Neural Relighting and SVBRDF Estimation. In *ECCV*, 2020. 3, 6, 8

[44] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[45] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild. In *CVPR*, 2018. 3

[46] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *CVPR*, 2008. 2

[47] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, 2011. 2

[48] Li Shen, Chuohao Yeo, and Binh-Son Hua. Intrinsic image decomposition using a sparse representation of reflectance. *IEEE TPAMI*, 35(12):2904–2915, 2013. 4

[49] Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. Learning Non-Lambertian Object Intrinsics across ShapeNet Categories. In *CVPR*, 2017. 2, 3, 6, 8

[50] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021. 3

[51] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM TOG*, 38(4):79–1, 2019. 3

[52] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, pages 205–216. Wiley Online Library, 2021. 3

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

[54] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world's revolutionary artefacts. In *CVPR*, 2021. 3

[55] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild. In *CVPR*, 2020. 2, 3, 5

[56] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM TOG*, 31(6):1–10, 2012. 4

[57] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *NeurIPS*, 2020. 3

[58] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021. 3

[59] Ye Yu, Abhimetra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and Will Smith. Self-supervised Outdoor Scene Relighting. In *ECCV*, 2020. 2

[60] Ye Yu and William AP Smith. InverseRenderNet: Learning Single Image Inverse Rendering. In *CVPR*, 2019. 2

[61] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, pages 5453–5462, 2021. 2

[62] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. In *SIGGRAPH Asia*, 2021. 3

[63] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *ICCV*, pages 7194–7202, 2019. 6, 7