De-rendering 3D Objects in the Wild

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

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 \cite{13,14}, recent learning-based ap-
proaches 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.

**Intrinsic 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,
is however surjective. This means that due to the highly
portable lighting for either training or inference is chal-
ing 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],
ear-planar surfaces [28], indoor scenes [26] or other syn-	hetic 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 ques-
tionable 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 [11,7,17,25,55,58], 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 decompo-
sition 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 forma-
tion process, i.e. rendering an image from its intrinsic com-
ponents. Our method learns to invert this process—de-
derendering the image—by extracting the intrinsic compo-
nents from an input image \( I_{in} \in [0, 1]^{3 \times H \times W} \).

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 ren-
dering process. Second, we leave enough flexibility in the
rendering process, such that the model can learn to over-
come 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 nor-
mal map \( N \in [-1, 1]^{3 \times H \times W} \). Given an image, directly
predicting a normal map with a neural network is problem-
atic, 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 nor-
mal map \( N_D \)) and a refinement normal map \( N_{\text{ref}} \) from the
image, and combine them:

\[
N = \frac{N_D + N_{\text{ref}}}{\|N_D + N_{\text{ref}}\|} \tag{1}
\]

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, dif-
fuse, 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, po-
tentially 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 shini-
ness value \( \alpha \in [0, \alpha_{\text{max}}] \) and a global specularity intensity
\( \alpha_{\text{spec}} \in [0, 1] \) for the whole object. Summarizing, we rep-
resent the lighting as \( L = (s_{\text{amb}}, s_{\text{dir}}, l, \alpha, \alpha_{\text{spec}}) \) and the intrinsic
material properties as \( (A, \alpha, \alpha_{\text{spec}}) \).

We obtain an image \( I \) from shape, material, and lighting
through the following rendering equation, where \( u \in \Omega =
\{1, \ldots, H\} \times \{1, \ldots, W\} \) represents a pixel location.

\[
\mathbf{I}_u = \tau(s_{\text{amb}}A_u + s_{\text{dir}}(N_u^T l_A) + a_{\text{spec}}(N_u^T v) \alpha)), \tag{2}
\]

where \(\tau(I_u) = I_u^{1/\gamma}\), \(\gamma = 2.2\) 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., \(||I - \hat{I}||^2\)) 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 \(D\) and its normal map \(N\)) 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 \(L_e = (s_{\text{amb}}, s_{\text{dir}}, l_e)\) such that the aggregated shading corresponds to the brightness of the input image.

\[
\arg\min_{L_e} \sum_{u \in \Omega} (2B_u - (s_{\text{amb}}, s_{\text{dir}}, l_e) N_{c,u} l) \tag{3}
\]

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, \(N_e\) is the coarse normal map. With this light estimate, an initial albedo estimate \(\hat{A}_e\) can be obtained by inverting the shading equation:

\[
\hat{A}_{c,u} = \mathbf{I}_u \left(\frac{s_{\text{amb}}, s_{\text{dir}} N_{c,u} l_e}{c}\right)^{-1} \tag{4}
\]

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 \(\hat{A}_e\), 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):

\[
\arg\min_{A_e} \|\delta_x A_e - \delta_x \hat{A}_e\|_2 + \|\delta_y A_e - \delta_y \hat{A}_e\|_2 + \\
\lambda_{TV} \left(\|\delta_x A_e\|_1 + \|\delta_y A_e\|_1\right) \tag{5}
\]

We use \(\delta_x\) and \(\delta_y\) to signify the computation of image gradients, which can, for example, be obtained by applying the Sobel operator to the image. We obtain \(L_e\) and \(A_e\) by opt-
timizing 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 $I_{in}$. A shape network $\Phi_{\text{shape}}$ predicts both the depth map $D, D_u \in [d_{\text{min}}, d_{\text{max}}]$ and the normal refinement map $N_{\text{ref}}$, which is normalized after prediction and used to obtain the final normal map $N$ with Eq. (1). The albedo network $\Phi_{\text{albedo}}$ predicts the albedo map $A, A_u \in [0, 1]$, and the light network $\Phi_{\text{light}}$ predicts the light parameters $s_{\text{amb}}, s_{\text{dir}} \in [0, 1]$, as well as, shininess $\alpha \in [0, \alpha_{\text{max}}]$ and specular intensity $a_{\text{spec}} \in [a_{\text{min}}, a_{\text{spec}}]$.

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.

$$L_c = \sum_{u \in \Omega} \lambda_D \|D_u - D_{c,u}\|_1 + \lambda_N N_{T_u} N_{c,u} + \lambda_A \|A_u - A_{c,u}\|_1 + \lambda_L \|L - L_c\|_2$$

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 influenced by the direction $\hat{I}$.

$$L_{\text{rec}} = \frac{1}{|\Omega|} \sum_{u \in \Omega} \|I_u - \hat{I}_u\|_1 + \frac{1}{2} \left(1 - \text{SSIM}(I, \hat{I})\right)$$

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 $I'$ (mainly influenced by the direction $\hat{I'}$).

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 $\hat{I}$, which is also used in the reconstruction loss term, and one with randomly sampled lighting conditions, denoted as $\hat{I}'$. 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 $\hat{I}$ instead of the original image $I_{in}$ as positive example when training $\Phi_{\text{disc}}$, 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: $L_{\text{gan}} = (1 - \Phi_{\text{disc}}(\hat{I}'))^2$.

We can then train the whole model end-to-end using

$$L = L_c + \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{gan}} L_{\text{gan}}$$

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 $I_{\text{spec,ref}}(I)$ 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 $64 \times 64$.

**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
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 $L_1$, $L_2$, and SSIM error metrics, we use mean angle deviation in degrees $\text{DIA}(\hat{N}, \hat{N}) = \frac{1}{|\Omega|} \sum_{u \in \Omega} \cos^{-1}(N_u \cdot \hat{N}_u)$ 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, $\mu_A$ 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).

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-

<table>
<thead>
<tr>
<th>Model</th>
<th>Normal $N$ MSE ↓</th>
<th>DIA ↓</th>
<th>SSIM ↑</th>
<th>Albedo $A$ MSE ↓</th>
<th>DIA ↓</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRFS [2]</td>
<td>0.331</td>
<td>52.994</td>
<td>0.113</td>
<td>0.724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ShapeNet-Intr. [49]</td>
<td>N/A</td>
<td>N/A</td>
<td>0.114</td>
<td>0.726</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SISaSVBRDF [20]</td>
<td>0.288</td>
<td>42.801</td>
<td>0.112</td>
<td>0.752</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neur. Rel. [43]</td>
<td>0.228</td>
<td>41.603</td>
<td>0.093</td>
<td>0.723</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>0.173</td>
<td>37.807</td>
<td>0.075</td>
<td>0.760</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Normal $N$ MSE ↓</th>
<th>DIA ↓</th>
<th>SSIM ↑</th>
<th>Albedo $A$ MSE ↓</th>
<th>DIA ↓</th>
<th>SSIM ↑</th>
<th>Specular $I_{spec}$ MSE ↓</th>
<th>DIA ↓</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Albedo</td>
<td>0.162</td>
<td>36.5</td>
<td>0.088</td>
<td>0.750</td>
<td>0.124</td>
<td>0.077</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Shape</td>
<td>0.506</td>
<td>68.7</td>
<td>0.079</td>
<td>0.757</td>
<td>0.108</td>
<td>0.058</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No GAN</td>
<td>0.169</td>
<td>37.2</td>
<td>0.075</td>
<td>0.762</td>
<td>0.123</td>
<td>0.073</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>0.173</td>
<td>37.8</td>
<td>0.075</td>
<td>0.760</td>
<td>0.112</td>
<td>0.059</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
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 $I$, predicted albedo $A$ and normals $N$, diffuse shading map $I_{\text{diff}}$, specular shading map $I_{\text{spec}}$ and reconstructed image $\hat{I}$. Further, we also show the shading maps ($I'_{\text{diff}}, I'_{\text{spec}}$) and reconstructed image $\hat{I}'$ under new lighting conditions. Our model achieves a high level of detail on shape and material reconstruction and convincing relighting results.

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.
<table>
<thead>
<tr>
<th>Input</th>
<th>Albedo / Reflectance</th>
<th>Normal (Diffuse)</th>
<th>(Shading)</th>
<th>Specularity / Roughness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIRFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Image Shape And SVBRDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural Relighting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shapenet Intrinsics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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].

<table>
<thead>
<tr>
<th>Geometry &amp; Albedo Improvement.</th>
<th>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.</th>
</tr>
</thead>
</table>

| 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 on 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


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


[29] Zhengqi 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


