Learning Non-Lambertian Object Intrinsics across ShapeNet Categories

Jian Shi
SKLCS, Institute of Software, Chinese Academy of Sciences
University of Chinese Academy of Sciences
shij@ios.ac.cn

Yue Dong
Microsoft Research Asia
yuedong@microsoft.com

Hao Su
Stanford University
haosu@cs.stanford.edu

Stella X. Yu
UC Berkeley / ICSI
stellayu@berkeley.edu

Abstract

We focus on the non-Lambertian object-level intrinsic problem of recovering diffuse albedo, shading, and specular highlights from a single image of an object. Based on existing 3D models in the ShapeNet database, a large-scale object intrinsics database is rendered with HDR environment maps. Millions of synthetic images of objects and their corresponding albedo, shading, and specular ground-truth images are used to train an encoder-decoder CNN, which can decompose an image into the product of albedo and shading components along with an additive specular component. Our CNN delivers accurate and sharp results in this classical inverse problem of computer vision. Evaluated on our realistically synthetic dataset, our method consistently outperforms the state-of-the-art by a large margin.

We train and test our CNN across different object categories. Perhaps surprising especially from the CNN classification perspective, our intrinsics CNN generalizes very well across categories. Our analysis shows that feature learning at the encoder stage is more crucial for developing a universal representation across categories. We apply our model to real images and videos from Internet, and observe robust and realistic intrinsics results. Quality non-Lambertian intrinsics could open up many interesting applications such as realistic product search based on material properties and image-based albedo / specular editing.

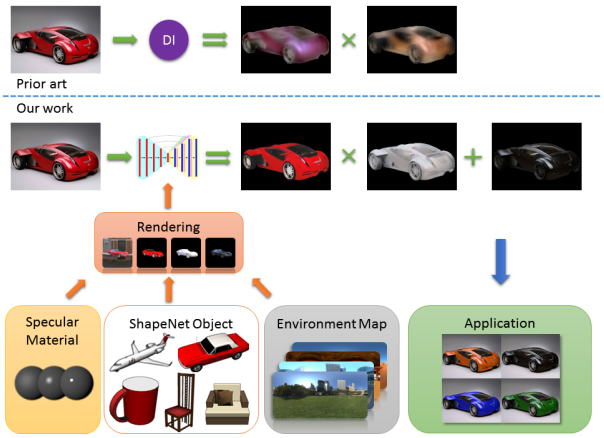

Figure 1: Specularity is everywhere on objects around us and is essential for our material perception. Our task is to decompose an image of a single object into its non-Lambertian intrinsics components that include not only albedo and shading, but also specular highlights. We build a large-scale object non-Lambertian intrinsics database based on the *ShapeNet* dataset, and render millions of synthetic images with specular materials and environment maps. We train an encoder-decoder CNN that delivers much sharper and more accurate results than the prior art of direct intrinsics (DI). Our work enables realistic applications of intrinsics to image-based albedo and specular editing.

1. Introduction

Specular reflection is common to objects encountered in our daily life. However, existing intrinsic image decomposition algorithms, e.g. SIRFS [4] and Direct Intrinsics (DI) [22], only deal with Lambertian or diffuse reflection. Such mismatching between the reality of images and the model assumption often leads to large errors in the intrinsic image decomposition of real images (Fig. 1).

In this paper, our target is to handle specular reflection and solve non-Lambertian object intrinsics from a single image. According to optical imaging physics, we extend the old Lambertian model to a non-Lambertian model with the specular component as an additive residue term:

$$\text{old}: \text{image } I = \text{albedo } A \times \text{shading } S \quad (1)$$

$$\text{new}: \text{image } I = \text{albedo } A \times \text{shading } S + \text{specular } R \quad (2)$$

Inspired by DI [22], we employ a data-driven deep learning approach to capture the associations between the image and its albedo, shading and specular components.

The immediate challenge of our non-Lambertian object intrinsics task is the lack of ground-truth data, especially for our non-Lambertian case, and human annotation appears to be infeasible. Existing intrinsics datasets are not only Lambertian in nature, with only albedo and shading components, but also have their own individual caveates. The widely used MIT Intrinsic Images dataset [13] is very small by today's standard, with only 20 object instances under 11 illumination conditions. MPI Sintel [9] intrinsics dataset, used by *Direct Intrinsics*, is too artificial, with 18 cartoon-like scenes at 50 frames each. Intrinsics in the Wild (IIW) [6] is the first large-scale intrinsics dataset of real world images, but it provides sparse pairwise human ranking judgements on albedo only, inadequate for benchmarking full image intrinsic image decompositions.

Another major challenge is how to learn multiple image regression tasks at both pixel- and intensity- accurate levels. Deep learning has been tremendously successful for image classification and somewhat successful for semantic segmentation and depth regression. The main differences lie in the spatial and tonal resolutions demanded by the output. The state-of-the-art DI CNN model [22] is adapted from a depth regression CNN with a coarse native spatial resolution. Their results are not only blurry, but also with false structures – there could be variations in the intrinsics predicted over a completely flat region of the input image. While benchmark scores for many CNN intrinsics models [23, 37, 38, 22, 24] are improving, the visual quality of these results remains poor, compared with those from traditional approaches based on hand-crafted features and multitudes of priors [7].

Our work addresses these challenges and makes the following contributions.

1. A new non-Lambertian object intrinsics dataset. We develop a new rendering-based object-centric intrinsics dataset with specular reflection based on *ShapeNet*, a large-scale 3D shape dataset.

2. A new CNN model with accurate and sharp results. Our approach not only significantly outperforms the state-of-the-art on multiple error metrics, but also produces much sharper and detailed visual results.

3. Analysis on cross-category generalization. Surprising from deep learning perspective on classification and segmentation, our intrinsics CNN shows remarkable generalization across categories: networks trained only on *chairs* also obtain reasonable performance on other categories such as *cars*. Our analysis on cross-category training and testing results reveals that features learned at the encoder stage are the key for developing a universal representation across categories.

Our model delivers solid non-Lambertian intrinsics results on real images and videos, closing the gap between intrinsic image algorithm development and practical applications.

2. Related Work

Intrinsic Image Decomposition. Much effort has been devoted to this long standing ill-posed problem [5] of decomposing an image into a reflectance layer and a shading layer. Land and McCann [20] observe that large gradients in images usually correspond to changes in reflectance and small gradients to smooth shading variations. To tackle this ill-posed problem where two outputs are sought out of a single input, many priors that constrain the solution space have been explored, such as reflectance sparsity [30, 32], non-local texture [31], shape and illumination [4], *etc*. Another line of approaches explores additional input information, such as image sequences [35], depth [3, 11] and user strokes [8]. A major challenge in intrinsics research is the lack of datasets with ground-truth intrinsics. Grosse *et al*. [13] capture the first real image dataset in a lab setting, with limited variations. Bell *et al*. [6] use crowdsourcing to obtain sparse human judgements on sampled pairs of pixels.

Deep Learning. Narihira *et al*. [23] is the first to use deep learning to learn albedo from IIW's sparse human judgement data. Zhou *et al*. [37] and Zoran *et al*. [38] extend the IIW-CRF model with a CNN learning component. Direct Intrinsics [22] is the first entirely deep learning model that outputs intrinsics predictions, based on a depth regression CNN model [12] and trained on the synthetic Sintel intrinsics dataset. Their results are blurry due to downsampling and convolution followed by deconvolution, and poor due to training on artificial scenes. To improve prediction accuracies and retain sharp details, we build our model upon the success of skip layer connections used in CNNs for classification [15], segmentation [29] and interpolation [27].

Reflectance Estimation. Multiple images are usually required for an accurate estimation of surface albedo. Aittala *et al*. [2] propose a learning based method for single image inputs, assuming that the surface only contains stochastic textures and is lit from known lighting directions. Most methods work on homogeneous objects lit by distant light sources, with surface reflectance and environment lighting estimated via blind deconvolution [28] or trained regression networks [27]. Our work aims at general intrinsic image

decomposition from a single image, without constraints on material or lighting distributions. Our model predicts spatially varying albedo maps and supports general lighting conditions.

Learning from Rendered Images. Images rendered from 3D models are widely used in deep learning, *e.g.* for training object detectors and viewpoint classifiers [33, 21, 14, 25]. Su *et al.* [34] obtain state-of-the-art results for viewpoint estimation by adapting CNNs trained from synthetic images to real ones. ShapeNet [10] provides 330,000 annotated models from over 4,000 categories, with rich texture information from artists. We build our non-Lambertian intrinsics dataset and algorithms based on ShapeNet, rendering and learning from photorealistic images on many varieties of common objects.

3. Intrinsic Image with Specular Reflectance

We derive our non-Lambertian intrinsic decomposition equation based on physics-based rendering. Given an input image, the observed outgoing radiance I at each pixel can be formulated as the product integral between incident lighting L and surface reflectance ρ via this rendering equation [17]:

$$I = \int_{\Omega_+} \rho(\omega_i, \omega_o)(N \cdot \omega_i) L(\omega_i)\, \mathrm{d}\omega_i. \tag{3}$$

Here, ω_o is the viewing direction, ω_i is the lighting direction from the upper hemisphere domain Ω_+, and N is the surface normal direction of the object.

Surface reflectance ρ is a 4D function usually defined as the bi-directional reflectance distribution function (BRDF). Various BRDF models have been proposed, all sharing a similar structure with a diffuse term ρ_d and a specular term ρ_s, and corresponding coefficients α_d, α_s:

$$\rho = \alpha_d \cdot \rho_d(\omega_i, \omega_o) + \alpha_s \cdot \rho_s(\omega_i, \omega_o) \tag{4}$$

For the diffuse component, lights scatter and produce view-independent and low-frequency smooth appearance. By contrast, for the specular component, lights bounce off the surface point only once and produce shinny appearance. The scope of reflection is modeled by diffuse albedo α_d and specular albedo α_s.

Combining reflection equation (4) and rendering equation (3), we have the following image formation model:

$$\begin{aligned} I = \alpha_d \int_{\Omega_+} \rho_d(\omega_i, \omega_o) L(\omega_i)\, \mathrm{d}\omega_i \\ + \alpha_s \int_{\Omega_+} \rho_s(\omega_i, \omega_o) L(\omega_i)\, \mathrm{d}\omega_i = \alpha_d s_d + \alpha_s s_s, \end{aligned} \tag{5}$$

where s_d and s_s are the diffuse and specular shading respectively. Traditional intrinsics models consider diffuse shading only, by decomposing the input image I as a product of diffuse albedo A and shading S. However, it is only proper to model diffuse and specular components separately, since their albedos have different values and spatial distributions. The usual decomposition of $I = A \times S$ is only a crude approximation.

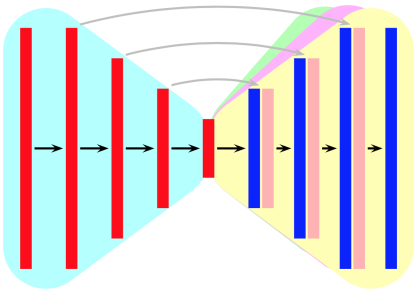

Figure 2: Our mirror-link CNN architecture has one shared encoder and three decoders for albedo, shading, specular components separately. Mirror links connect the encoder and decoder layers of the same spatial resolution, providing visual details. The height of layers in this figure indicates the spatial resolution.

Specular reflectance $\alpha_s s_s$ has characteristics very different from diffuse reflectance $\alpha_d s_d$: Both specular albedo and specular shading have high-frequency spatial distributions and color variations, making decomposition more ambiguous. We thus choose to model specular reflectance as a single residual term R, resulting in the non-Lambertian extension: $I = A \times S + R$, where input image I is decomposed into diffuse albedo A, diffuse shading S, and specular residual R respectively.

Although our image formation model is developed based on physics based rendering and physical properties of diffuse and specular reflection, it does not assume any specific BRDF model. Simple BRDF models (*e.g.* Phong) can be employed for rendering efficiency while complex models (*e.g.* Cook-Torrance) can lead to higher photo-realism.

4. Learning Intrinsics

We develop our CNN model and training procedure for non-Lambertian intrinsics.

Mirror-Link CNN. Fig. 2 illustrates our encoder-decoder CNN architecture. The encoder progressively extracts and down-samples features, while the decoder up-samples and combines them to construct the output intrinsic components. The sizes of feature maps (including input/output) are exactly **mirrored** in our network. We **link** early encoder features to the corresponding decoder layers at the same spatial resolution, in order to obtain pixel-accurate sharp details preserved in early encoder layers. Since output components are closely related to each other, we share the same encoder and use separate decoders for A, S, R.

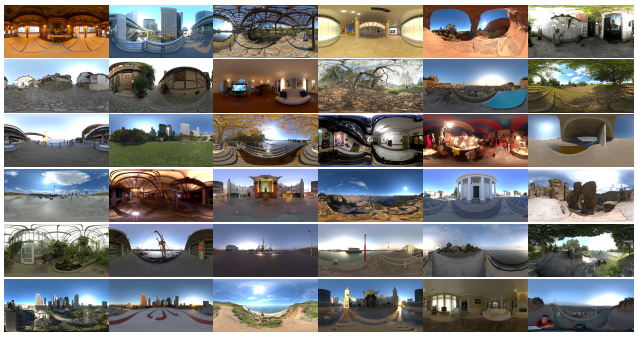

Figure 3: Environment maps are employed in our rendering for realistic appearance, both outdoor and indoor scenes are included. The environment map not only represents the dominate light sources in the scene (*e.g.* sun, lamp and window) but also includes correct information on the surroundings (*e.g.* sky, wall and building). Although a dominate light might be sufficient for shading a Lambertian surface, detailed surroundings provide the details in the specular.

Similar structures have been used in Deep Reflectance Map (DRM) [27] and U-Net [29]. DRM solves an interpolation problem from high resolution sparse inputs to low resolution dense map outputs in the geometry space, ignoring the spatial inhomogeneity of reflectance. U-Net deals with image segmentation. We use multiple decoders with a shared encoder for multiple image regression outputs.

Scale invariant Loss. There is an inherent scale ambiguity between albedo and shading, as only their product matters in the intrinsic image decomposition. DI [22] employs a weighted combination of MSE loss and scale-invariant MSE loss for training their intrinsic networks. Similar to their work, we apply the same loss functions to albedo and shading, while simply take MSE loss for specular. Since we focus on object-level intrinsics, only pixels in the object mask have been used for calculating the loss function and its gradients.

ShapeNet-Intrinsics Dataset. We obtain the geometry and albedo texture of 3D shapes from ShapeNet, a large-scale richly-annotated, 3D shape repository [10]. We pick 31,072 models from several common categories: car, chair, bus, sofa, airplane, bench, container, vessel, *etc*. These objects often have specular reflections.

Environment maps. To generate photo-realistic images, we collect 98 HDR environment maps from online public resources [1]. Indoor and outdoor scenes with various illumination conditions are included, as shown in Fig. 3.

Rendering. We use an open-source renderer Mitsuba [16] to render object models with various environment maps and random viewpoints sampled from the upper hemisphere. A modified Phong reflectance model [26, 19] is assigned to objects to generate photo-realistic shading and specular effects. Since original models in ShapeNet are only provided with reliable diffuse albedo, for each object we randomly pick a Phong material with uniform distribution of specular coefficient $k_s \in (0, 0.3)$ and shininess $N_s \in (0, 300)$, which covers the range from pure diffuse to high specular appearance (Fig. 1). We render albedo, shading and specular layers, and then synthesize images according to Equation 5.

Training. We split our dataset at the object level in order to avoid images of the same object appearing in both training and testing sets. We use $80/20$ split, resulting in $24,932$ models for training and $6,240$ for testing. All the 98 environment maps are used to rendering $2,443,336$ images for the training set. For the testing set, we randomly pick 1 image per testing model. More implementation details can be found in the supplementary material.

5. Evaluation

Our method is evaluated and compared with SIRFS [4], IIW [6], and Direct Intrinsics (DI) [22]. We also train DI using our ShapeNet intrinsics dataset and denote the model as DI*. We adopt the usual metrics, MSE, LMSE and DSSIM, for quantitative evaluation. We also include a simple baseline for shading, which is a constant, and another baseline for albedo, which is the input image itself.

5.1. ShapeNet Intrinsics Dataset

Table 1 shows benchmark scores on our ShapeNet intrinsics test set. Our algorithm consistently outperforms existing approaches. Compared to off-the-shelf solutions, our method provides 40-50% performance gain according to the DSSIM error. Also note that, DI*, i.e. DI trained with our dataset, produces second best results across almost all the error metrics, demonstrating the advantage of our ShapeNet intrinsics dataset.

Numerical error metrics may not be fully indicative of visual qualities, *e.g.* the naive baseline also produces low errors for some cases. Figure 4 provides visual comparisons against ground-truths.

For objects with strong specular reflectance, *e.g.* cars, specular reflection violates the Lambertian condition assumed by traditional intrinsics algorithms. These algo-

ShapeNet intrinsics	MSE		LMSE		DSSIM	
	albedo	shading	albedo	shading	albedo	shading
Baseline	0.0232	0.0153	0.0789	0.0231	0.2273	0.2341
SIRFS	0.0211	0.0227	0.0693	0.0324	0.2038	0.1356
IIW	0.0147	0.0149	0.0481	0.0228	0.1649	0.1367
DI	0.0252	0.0245	0.0711	0.0275	0.1984	0.1454
DI*	0.0115	0.0066	0.0470	0.0115	0.1655	0.0996
Ours	**0.0083**	**0.0055**	**0.0353**	**0.0097**	**0.0939**	**0.0622**
specular	0.0042		0.0578		0.0831	

Table 1: Evaluation on our synthetic dataset. For the baseline, we set its albedo to be the input image and its shading to be 1.0. The last row lists our specular error.

Figure 4: Results on the ShapeNet Intrinsics dataset. Our baselines include SIRFS, IIW, Direct-Intrinsics with released model by the author (DI), and model trained by ourselves on our synthetic dataset (DI*). The top row of each group is albedo, and the bottom is shading. The *Specular* column shows the ground-truth (top) and our result (bottom). We observe that specularity has basically been removed from albedo/shading, especially for cars. Even for the sofa (last row) with little specular reflection, our method still produces good visual results. See more results in our supplementary material.

rithms, *e.g.* SIRFS and IIW, simply cannot handle such specular components. Learning-based approaches, DI, DI*, or our method, could still learn from the data and perform better in these cases. For DI, the network trained on our dataset also has significantly better visual quality, compared with their released model trained on the Sintel dataset. However, their results are blurry, as a consequence from their deep convolution and deconvolution network architecture without our mirrored skip-layer connections.

Our model produces sharper images preserving many visual details, such as boundaries in the albedo and specular images. Large specular areas on the body of cars are also extracted well in the specular residue component, revealing the environment illumination. Such specular areas would confuse earlier algorithms and bring serious artifacts to albedo/shading predictions.

MIT	MSE		LMSE		DSSIM	
intrinsic	albedo	shading	albedo	shading	albedo	shading
SIRFS	**0.0147**	**0.0083**	**0.0416**	**0.0168**	**0.1238**	**0.0985**
DI	0.0277	0.0154	0.0585	0.0295	0.1526	0.1328
Ours	0.0468	0.0194	0.0752	0.0318	0.1825	0.1667
Ours*	0.0278	0.0126	0.0503	0.0240	0.1465	0.1200

Table 2: Evaluation on MIT intrinsics dataset.

Figure 5: Results on the MIT dataset. *Ours** is our ShapeNet trained model fine-tuned on MIT, with data generated by the GenMIT approach used in DI [22].

5.2. MIT Intrinsics Dataset

We also run our network on the MIT intrinsics dataset [13]. While our environment light model is colored and designed for common real-world images, the MIT-intrinsics dataset uses a single grayscale directional light model in a lab capture setting, a scenario that is not included in our ShapeNet intrinsics dataset. The light model differences lead to dramatic visual differences and cause domain shift problems for learning based approaches [22]. We also follow [22] to fine tune our network on the MIT dataset.

Table 2 lists benchmark errors and Fig 5 provides sample results for visual comparisons. SIRFS produces the best numerical results, since the pure Lambertian surface reflection

and grayscale lighting setup best fits the assumption of such prior-based intrinsics algorithms. Direct intrinsics [22] requires fine tuning to reach similar performance. Our model fine tuned on the MIT dataset produces comparable results as SIRFS and better than DI finetuned on MIT; in addition, our results preserve more details compared to DI.

5.3. Real-World Images

Figure 6 shows our results on real-world images, most of them are product images. Although our model is trained entirely on synthetic data, it provides more realistic results than other algorithms, due to photo-realistic rendering that simulates the physical effects of diffuse and specular reflection and to the generalization properties of our model.

Our model also produces surprisingly good results on objects never included in our dataset, *e.g.* the mouse, toy and tomato. In these results such as the car, mouse and potato, specular highlights are correctly estimated, and the corresponding albedo maps are recovered with correct colors. Note that highlight pixels are particularly challenging, since they could be so bright that no diffuse colors are left in the input pixels, essentially invalidating many chroma based intrinsics solutions.

Finally, we apply our model to videos frame by frame, and obtain coherent results without using any constraints on temporal consistency. Please see our supplementary materials.

6. Cross-category generalization

ShapeNet provides semantic category information for each object, allowing in-depth analysis for cross-category performance analysis of our learning based intrinsic image decomposition task. We conduct category-specific training of our network on 4 individual categories which have more than 3,000 objects each: car, chair, airplane and sofa. We evaluate the network on the entire dataset as well as these 4 categories. All these networks are trained with the same number of iterations regardless the number of training data.

Table 3 lists cross-category testing errors. For almost all the categories, training on the specific dataset produces the best decomposition results on that category. This result is not surprising, as the network always performs the best at what it is trained for. Training with all the datasets leads to a small prediction error increase, at less than 0.02 in the DSSIM error.

What is surprisingly is that, on an image of an object category (*e.g.* car) that has never been seen during training (*e.g.* chairs), our network still produces reasonable results, with the DSSIM error on-par or better than existing works that are designed for general intrinsic tasks (Table 1). Figure 7 shows sample cross-category training and testing results: All of our models produce reasonable results, demonstrating cross-category generalization.

Analysis on generalization. Our image-to-image regression network always produces the same physical components: albedo, shading and specular maps, unlike classification networks with semantic labels. Although objects in different categories have dramatically different shapes, textures and appearances, those components have the same physical definitions and share similar structures. Many of those commonalities are widely used in previous intrinsics algorithms,*e.g.* shading is usually smooth and grayscale; albedo contains more color variations and specular is sparse and of higher contrast.

When two categories share some properties, their individually learned networks apply well to the other. For example, the Chair and Sofa categories share similar textures (textile and wood), albedo, and shapes, thus their predictions on all three output channels transfer well to the other category.

We also observe non-symmetry in Table 3: *e.g.* the network trained on Car produces good results on Airplane, while the network trained with Airplane has relative larger error on Car. This difference could be explained by the amount of within-category variations: The car category has more variations in both shapes and textures, and richer variations lead to better generalization. This result can also be observed in the benchmarks on the ALL dataset, where the Car-category network produces the best results except the

Albedo					
	ALL	Car	Chair	Airplane	Sofa
ALL	**0.0939**	0.1014	0.0988	0.0893	0.0716
Car	0.1134	**0.0808**	0.1379	0.1057	0.1002
Chair	0.1181	0.1578	**0.0911**	0.1166	0.0835
Airplane	0.1201	0.1410	0.1338	**0.0757**	0.0954
Sofa	0.1131	0.1348	0.1101	0.1067	**0.0663**
Shading					
	ALL	Car	Chair	Airplane	Sofa
ALL	**0.0622**	0.0685	**0.0549**	0.0596	0.0491
Car	0.0687	**0.0579**	0.0692	0.0683	0.0592
Chair	0.0772	0.1008	0.0561	0.0740	0.0548
Airplane	0.0776	0.0936	0.0738	**0.0481**	0.0629
Sofa	0.0721	0.0877	0.0594	0.0697	**0.0460**
Specular					
	ALL	Car	Chair	Airplane	Sofa
ALL	**0.0831**	0.0866	0.0714	0.1021	0.0730
Car	0.0953	**0.0745**	0.0962	0.1214	0.0854
Chair	0.0982	0.1162	**0.0719**	0.1205	0.0800
Airplane	0.1019	0.1115	0.0980	**0.0871**	0.0939
Sofa	0.0984	0.1115	0.0800	0.1238	**0.0673**

Table 3: Cross-category DSSIM scores. Each row corresponds to a model trained on the specific category, and each column corresponds to the result evaluated on the specific category. Not surprisingly, the lowest errors are mostly on the diagonal when the training and testing sets are the same, except for the shading on Chairs. While category-specific training gives better results on its own category, the results in other categories are surprisingly only slightly worse, demonstrating good generalization.

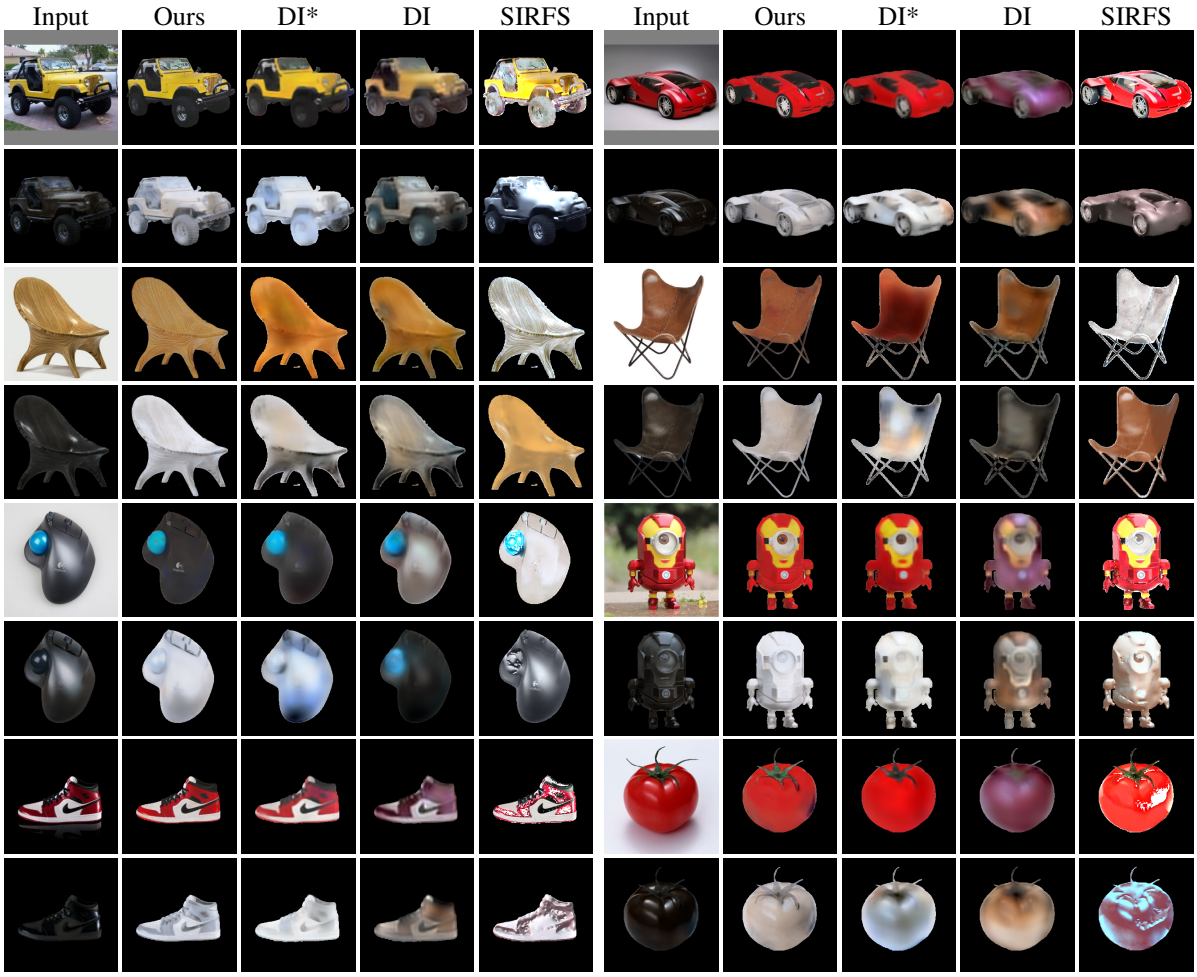

Figure 6: Evaluation on real world images. The first column contains input images (top) and our specular predictions (bottom). For the group of results of an image, the top row gives the predicted albedo and the bottom row gives the shading. We observe that: 1) DI* trained on our dataset produces better results than the publicly released model DI; however, they are still blurry without fine details. 2) SIRFS produces erroneous albedo prediction for strong specular cases, as it does not assume specular reflectance.

	Albedo				Shading				Specular			
	Car	Chair	Airplane	Sofa	Car	Chair	Airplane	Sofa	Car	Chair	Airplane	Sofa
Car	**0.0808**	0.1379	0.1057	0.1002	**0.0579**	0.0692	0.0683	0.0592	**0.0745**	0.0962	0.1214	0.0854
Car-Chair	0.1157	0.1303	0.1182	0.0954	0.0769	0.0678	0.0743	0.0598	0.0833	0.0907	0.1215	0.0882
Chair-Car	0.1311	0.1111	0.1125	0.0929	0.0873	0.0582	0.0711	0.0573	0.1089	0.0736	0.1235	0.0810
Chair	0.1578	**0.0911**	0.1166	0.0835	0.1008	**0.0561**	0.0740	0.0548	0.1162	**0.0719**	0.1205	0.0800
Airplane	0.1410	0.1338	**0.0757**	0.0954	0.0936	0.0738	**0.0481**	0.0629	0.1115	0.0980	**0.0871**	0.0939
Airplane-Sofa	0.1502	0.1324	0.0855	0.0938	0.0940	0.0719	0.0546	0.0609	0.1104	0.0932	0.0916	0.0894
Sofa-Airplane	0.1349	0.1149	0.1032	0.0723	0.0954	0.0628	0.0703	0.0510	0.1129	0.0829	0.1151	0.0763
Sofa	0.1348	0.1101	0.1067	**0.0663**	0.0877	0.0594	0.0697	**0.0460**	0.1115	0.0800	0.1238	**0.0673**

Table 4: Cross-category decoder fine-tuning results. We freeze the encoder component and fine-tune the decoder components in a cross-category setting. Car-Chair denotes the model first trained on cars and then fine-tuned on chairs. Our results show that fine-tuning the decoder would not bring much performance improvement, if the encoder is already biased towards another category. We also observe that cross-category fine-tuning makes little difference when the model is evaluated on a third category, e.g. Car-Chair on Sofa performs similarly to Car on Sofa. These results together indicate the dominating importance of the encoder over the decoder.

Figure 7: Cross-category comparison. Training on one specific category produces the best result when tested on objects from the same category. Categories with similar appearance also share similar results, *e.g.* sofas tested on the model trained on chairs. Dissimilar categories might produce results with artifacts, *e.g.* chairs tested on the model trained on airplanes.

general ALL-network.

We test the role of encoder/decoder in our image-to-image regression task, and verify which is more critical for cross-category generalization. After training on a specific category, we freeze the encoder and fine tune the decoder on another category, e.g. we fine tune the car-trained model on chairs, with the encoder fixed. The encoder features cannot be changed and we can only modify the way the decoder composes them. Table 4 shows the results on fine-tuned models. We observe that finetuning the decoder brings very limited improvement on the dataset it is fine tuned on, indicating that the encoder features are crucial for learning the decomposition. That the model trained on ALL categories produces similar errors to category-specific models is most likely due to the encoder of our model capturing both category-dependent and category-independent features.

(a) Albedo recoloring.

(b) Specular editing.

Figure 8: Image based appearance editing through intrinsic layers. For specular editing, the first row shows scaling specular reflectance intensity by 1.0, 0.5 and 0; the second row shows specular editing by user interaction. The first column shows the original images.

7. Application

Decomposing images into their intrinsic components would benefit many applications. Figure 8 shows image-based material editing [18, 36] examples based on our intrinsics results. We can recolor the diffuse albedo map to simulate a different color paint on the car, while preserving the shading and specular highlights. With our approach, specular highlights can also be edited by simple processing, *e.g.* scaling, or complex user interaction.

8. Conclusion

We extend the intrinsic image problem by introducing a specular term and solve this non-Lambertian intrinsics problem with a deep learning approach. A large scale training dataset with realistic images is generated using physically based rendering on the ShapeNet object repository. Our CNN approach consistently outperforms the state-of-the-art both visually and numerically. Non-Lambertian intrinsics greatly extends Lambertian intrinsics to a much wider range of real images and real applications such as albedo and specular editing.

References

[1] sIBL Archive. http://www.hdrlabs.com/sibl/archive.html.

[2] M. Aittala, T. Aila, and J. Lehtinen. Reflectance modeling by neural texture synthesis. *ACM Trans. on Graphics*, 35(4):65:1–65:13, July 2016.

[3] J. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013.

[4] 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.

[5] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26, 1978.

[6] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics*, 33(4):159, 2014.

[7] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[8] A. Bousseau, S. Paris, and F. Durand. User-assisted intrinsic images. In *ACM Trans. on Graphics*, volume 28, page 130. ACM, 2009.

[9] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conference on Computer Vision*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.

[10] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[11] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *International Conference on Computer Vision*, pages 241–248, 2013.

[12] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *International Conference on Computer Vision*, pages 2650–2658, 2015.

[13] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009.

[14] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Inferring 3d object pose in rgb-d images. *arXiv preprint arXiv:1502.04652*, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, abs/1512.03385, 2016.

[16] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org.

[17] J. T. Kajiya. The rendering equation. In *ACM Siggraph Computer Graphics*, volume 20, pages 143–150. ACM, 1986.

[18] E. A. Khan, E. Reinhard, R. W. Fleming, and H. H. Bülthoff. Image-based material editing. In *ACM Trans. on Graphics*, SIGGRAPH '06, pages 654–663, New York, NY, USA, 2006. ACM.

[19] E. P. Lafortune and Y. D. Willems. *Using the modified phong reflectance model for physically based rendering*. Citeseer, 1994.

[20] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of Optical Society of America*, 61(1):1–11, 1971.

[21] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1688–1695. IEEE, 2010.

[22] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *International Conference on Computer Vision*, pages 2992–2992, 2015.

[23] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 8-10 June 2015.

[24] D. Pathak, P. Kraehenbuehl, S. X. Yu, and T. Darrell. Constrained structured regression with convolutional neural networks. In *http://arxiv.org/abs/1511.07497*, 2016.

[25] X. Peng, B. Sun, K. Ali, and K. Saenko. Exploring invariances in deep convolutional neural networks using synthetic images. *arXiv preprint arXiv:1412.7122*, 2014.

[26] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[27] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars. Deep reflectance maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[28] F. Romeiro and T. Zickler. Blind reflectometry. In *European Conference on Computer Vision*, ECCV'10, pages 45–58, Berlin, Heidelberg, 2010. Springer-Verlag.

[29] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015.

[30] C. Rother, M. Kiefel, L. Zhang, B. Schölkopf, and P. V. Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. In *Advances in neural information processing systems*, pages 765–773, 2011.

[31] L. Shen, P. Tan, and S. Lin. Intrinsic image decomposition with non-local texture cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[32] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 697–704. IEEE, 2011.

[33] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In *BMVC*, volume 2, page 5, 2010.

[34] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *International Conference on Computer Vision*, pages 2686–2694, 2015.

[35] Y. Weiss. Deriving intrinsic images from image sequences. In *International Conference on Computer Vision*, volume 2, pages 68–75. IEEE, 2001.

[36] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez. Intrinsic video and applications. *ACM Trans. on Graphics*, 33(4):80, 2014.

[37] T. Zhou, P. Krähenbühl, and A. A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. *International Conference on Computer Vision*, abs/1510.02413, 2015.

[38] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *International Conference on Computer Vision*, 2015.