Colorful Diffuse Intrinsic Image Decomposition in the Wild

CHRIS CAREAGA and YAĞIZ AKSOY, Simon Fraser University, Canada

Fig. 1. We present a method that can represent in-the-wild photographs in terms of albedo, diffuse shading, and non-diffuse residual components. Our shading layer reflects the colorful nature of multiple illuminants and secondary reflections in the scene, while our residual layer models the specularities and visible light sources.
Image from Unsplash by Rebe Adelaida.

Intrinsic image decomposition aims to separate the surface reflectance and the effects from the illumination given a single photograph. Due to the complexity of the problem, most prior works assume a single-color illumination and a Lambertian world, which limits their use in illumination-aware image editing applications. In this work, we separate an input image into its diffuse albedo, colorful diffuse shading, and specular residual components. We arrive at our result by gradually removing first the single-color illumination and then the Lambertian-world assumptions. We show that by dividing the problem into easier sub-problems, in-the-wild colorful diffuse shading estimation can be achieved despite the limited ground-truth datasets. Our extended intrinsic model enables illumination-aware analysis of photographs and can be used for image editing applications such as specularity removal and per-pixel white balancing.

CCS Concepts: • **Computing methodologies** → **Image representations**; **Image manipulation**.

Additional Key Words and Phrases: intrinsic decomposition, inverse rendering, mid-level vision, shading and reflectance estimation, image manipulation

ACM Reference Format:
Chris Careaga and Yağız Aksoy. 2024. Colorful Diffuse Intrinsic Image Decomposition in the Wild. *ACM Trans. Graph.* 43, 6, Article 178 (December 2024), 12 pages. https://doi.org/10.1145/3687984

Authors' address: Chris Careaga; Yağız Aksoy, Simon Fraser University, Burnaby, BC, Canada.

0730-0301/2024/12-ART178 $15.00
https://doi.org/10.1145/3687984

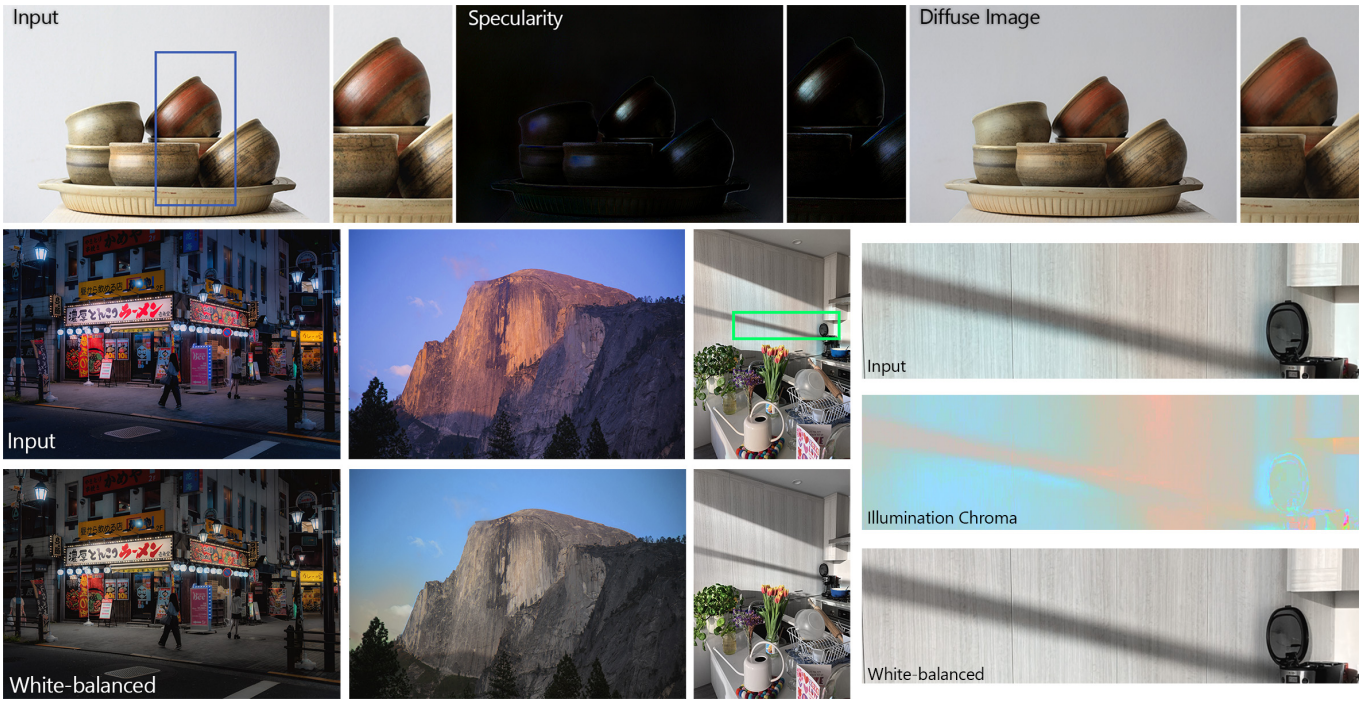

Fig. 2. In this work, we extend the in-the-wild intrinsic decomposition formulations to include a colorful shading component as well as a non-diffuse residual component. This extended image formation enables illumination-aware image editing applications, such as specularity removal as shown at the top, and per-pixel white balancing. Images from Unsplash by NorWood Themes (pots), mos design (street) and Josh Carter (mountain).

1 INTRODUCTION

Intrinsic image decomposition is a long-standing mid-level vision problem that aims to separate the surface reflectance and the effects of the illumination from a single photograph.

Due to the complex interactions between the illumination and the geometry during image formation, it is a highly under-constrained task that requires high-level reasoning about the scene. The lack of real-world training data and the large domain gap between synthetic data and real-world photographs further complicate the task.

Data-driven approaches to this problem have shown recent success, but prior works predominantly rely on the *grayscale intrinsic diffuse model*, $I = A * S$, where I is the input image in linear RGB, A is the 3-channel albedo, and S is the single-channel grayscale shading. Although this model is shown to be useful in making the problem more tractable, it relies on two major assumptions that limit its applicability in real-world scenes.

The first main assumption is the Lambertian world assumption that allows for the two-component multiplicative representation of the image by modeling all surfaces as diffuse. However, by ignoring specular surfaces, this model does not allow for separate editing of diffuse and non-diffuse illumination effects. The second assumption is the single-color shading that limits the model's representation of colorful illumination effects that are common in real scenes such as multiple light sources and inter-reflections. This results in color effects being embedded in the albedo layer as shown in Figure 4, limiting effectiveness in terms of color editing applications.

Few works in the literature attempt to further decompose illumination into diffuse shading and a non-diffuse residual, using the *intrinsic residual model*, $I = A*S+R$, that extends the intrinsic diffuse model with an additive component R that represents non-diffuse lighting effects such as specularities and visible light sources and defines S as an RGB map that reflects the color of illumination. This enhanced capability to model real-world scenes comes at the cost of complexity, increasing the number of unknown variables from 4 to 9 per pixel, exacerbating the under-constrained nature of the problem. Coupled with a lack of diverse ground truth, prior methods that adopt the intrinsic residual model have been constrained to narrow contexts such as objects [Meka et al. 2018; Shi et al. 2017] or faces [Shah et al. 2023; Zhang et al. 2022].

In this paper, we introduce a method that can generate decompositions under the intrinsic residual model for in-the-wild photographs. We start from a decomposition that uses the intrinsic diffuse model and gradually remove the single-color shading and the Lambertian world assumptions to estimate the diffuse albedo and the colorful diffuse shading at high resolutions. As summarized in Figure 3, we first estimate the chroma of the shading using the global context present in the scene that is then used to create a sparse diffuse albedo. Given the diffuse albedo, we further decompose the shading into diffuse and specular components. We show that by breaking this highly under-constrained task into multiple conceptually simpler sub-problems, our method is able to generalize to complex in-the-wild scenes.

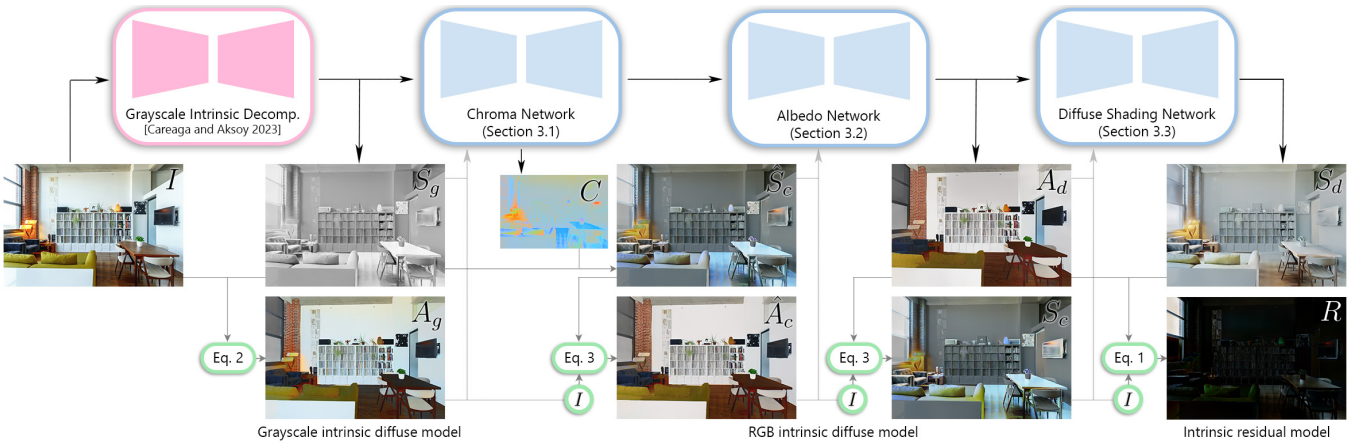

Fig. 3. Our pipeline starts with an input image and a shading/albedo pair generated within the simplified grayscale intrinsic diffuse model generated via an off-the-shelf method. We first extend the image formation model to include colorful shading, and estimate the shading color using our chroma network. This color information is used as input in the second step where we estimate the high-resolution diffuse albedo. In the final step, we remove the Lambertian-world assumption and estimate a colorful diffuse shading component and a non-diffuse residual layer. A single variable is estimated at each step, S_g, C, A_d, and S_d, respectively, and other intrinsic components are computed using the corresponding intrinsic image formation model with increasing representative power.
Image from Unsplash by Nathan Van Egmond.

We extensively evaluate our method's formulation and performance both qualitatively and quantitatively on common benchmarks as well as in-the-wild. We further demonstrate in Figure 2 and Section 5 several illumination-aware image editing applications including per-pixel white-balancing and specularity removal that are made possible by the intrinsic residual model.

2 RELATED WORK

Given the usefulness of intrinsic components in solving challenges in computational photography and image editing, the literature in this domain is extensive, covering multiple interrelated tasks. This section provides a summary of the field focusing on formulations and assumptions made by prior works in the context of our proposed approach. We refer the reader to the survey by Garces et al. [2022] for an in-depth discussion of the intrinsic decomposition literature.

2.1 Intrinsic Decomposition

Grayscale Diffuse Model. The grayscale diffuse intrinsic model has been the predominant assumption since the earliest methods of intrinsic decomposition [Shen et al. 2008; Tappen et al. 2005]. This model has shown continued use due to its simplicity, creating a more tractable problem for both optimization-based [Bell et al. 2014; Garces et al. 2012; Shen et al. 2011; Zhao et al. 2012] and data-driven [Baslamisli et al. 2018a,b; Careaga and Aksoy 2023; Das et al. 2022; Janner et al. 2017; Li and Snavely 2018a; Ma et al. 2018] approaches. As algorithms advance, this simplified model becomes more and more restrictive, causing inferred intrinsic components to stray further from physically accurate quantities.

RGB Diffuse Model. Due to the shortcomings of the grayscale assumption, a few prior works explicitly model a colorful shading component. Li and Snavely [2018b] propose an unsupervised method for learning intrinsic components via time-lapse data, they parameterize their shading component as a grayscale map multiplied by a global RGB color cast. Lettry et al. [2018] propose a similar unsupervised training strategy but take it a step further by estimating an unconstrained RGB shading component. Meka et al. [2021] model an RGB shading layer and further decompose shading into separate light sources, but their method relies on low-level assumptions and user input, making it only suitable for simple scenes. Other works implicitly account for colorful shading effects by directly estimating albedo [Das et al. 2022; Luo et al. 2020, 2023] but these works typically constrain the albedo via an image reconstruction loss using the grayscale diffuse model, therefore lack in ability to accurately model colorful lighting effects.

Residual Model. Extending beyond the well-known intrinsic diffuse model is not common in the literature. Given the difficulty of the problem and lack of real-world ground-truth supervision, prior works have only been able to estimate specularity in specific scenarios. Shi et al. [2017] propose a method for estimating decompositions for singular objects, limiting the real-world applicability of their method. Zhang et al. [2022] use the residual model to estimate intrinsic components, but their method is specifically designed for human faces. Shah et al. [2023] also adopt the residual model. Although they evaluate their model on faces, material images, and simple scenes, they train separate networks for each task. Kim et al. [2013] introduces an optimization formulation to infer the specularities without aiming full intrinsic decomposition. However, their low-level priors often lead to color edges being mislabeled.

Our method learns to estimate unconstrained RGB shading, both specular and diffuse, in the wild without the need for explicit assumptions or constraints. Despite being trained on indoor scenes, our diffuse shading network can generate accurate estimations for out-of-distribution images, as shown in Figure 1.

Fig. 4. The initial albedo map that we use as input contains significant color shifts due to the grayscale shading assumption. Using the shading chroma estimated by our first network (Sec. 3.1), these color shifts are corrected but it fails to remove fine details coming from complex illumination. Our albedo estimation network (Sec. 3.2) is able to remove the effects of the illumination and estimate a sparse albedo map. Image from Unsplash by Holly Stratton.

2.2 Inverse Rendering

Inverse rendering methods tackle the broader task of estimating all intrinsic scene parameters necessary to re-render an image. These methods typically estimate an albedo component explicitly and render shading via inferred geometry and an illumination model. Although this is a slightly different task formulation, inverse rendering methods are generally comparable to intrinsic decomposition methods as they still produce intrinsic components.

One of the earliest approaches by Barron and Malik [2015] uses low-level priors to jointly recover scene intrinsics for simple scenes and isolated objects. Karsch et al. [2014] propose a method for indoor scenes that uses off-the-shelf albedo and depth estimation methods and infers illumination by optimizing for image reconstruction. With the advancement of rendering capabilities, multiple data-driven methods have emerged [Li et al. 2020, 2022; Sengupta et al. 2019; Wang et al. 2021; Zhu et al. 2022b,a]. Given the limited availability of diverse training data, these methods focus on indoor scenes.

Several recent works leverage diffusion-based image generation models to generate plausible intrinsic components conditioned on a given input image [Chen et al. 2024; Kocsis et al. 2024; Zeng et al. 2024]. They model the problem as probabilistic, stemming from the under-constrained nature of the task. Chen et al. [2024] focus on close-up object images and point to the ambiguity between the albedo and illumination colors. Kocsis et al. [2024] focus on indoor images and point to different rendering engines and 3D models in CGI pipelines that occasionally embed several lighting effects in reflectance. They compensate for the random nature of their outputs by averaging over multiple estimations, which results in a loss of details. Zeng et al. [2024] can directly estimate high-resolution intrinsic components, but suffer from being constrained to the latent space of the diffusion model as shown in Figure 6. Due to their fully generative modeling, these methods learn the appearance characteristics of the intrinsic components and work in a similar fashion to style transfer. Zeng et al. [2024] make use of this aspect to demonstrate physically-guided image generation applications. In this work, we focus on the deterministic nature of real-world image formation and show that material and color ambiguities can be resolved through the context present in the scene.

3 METHOD

We aim to decompose an image I into its diffuse albedo A_d and colorful diffuse shading S_d layers with a residual layer R containing non-diffuse illumination effects using the intrinsic residual image formation model:

$$I = A_d * S_d + R. \tag{1}$$

This highly under-constrained problem requires a network to reason about high-level contextual cues about scene geometry, global and local illumination conditions, and material properties. The scarce high-resolution ground truth and the lack of real-world datasets for the diffuse shading component make it challenging for neural networks to statistically model the image formation in the wild.

In order to achieve in-the-wild generalization, we divide the problem into simpler, physically-motivated sub-problems that are convenient for neural networks to model. We start from an existing intrinsic decomposition of the image that relies on the simplified Lambertian intrinsic model with a grayscale shading component S_g:

$$I = A_g * S_g. \tag{2}$$

We use the method by Careaga and Aksoy [2023] to generate an A_g-S_g pair that provides an initial starting point for our method. We gradually remove the grayscale shading assumption, and then the Lambertian-world assumption, to arrive at our extended model in Equation 1. Figure 3 gives an overview of our approach.

3.1 Shading Chroma Estimation

One of the main reasons the grayscale shading assumption is adopted in the literature is that it significantly simplifies the problem by setting the albedo chromaticity to that of the input image. We begin our pipeline by abandoning the grayscale assumption and extend to the RGB intrinsic diffuse model:

$$I = A_c * S_c, \tag{3}$$

that requires inferring the per-pixel chromaticity of the shading layer. For this purpose, we take the input grayscale shading S_g as the luminance of S_c, and estimate the per-pixel chromaticities in our *chroma network*. Borrowing ideas from color constancy literature [Barron 2015; Kim et al. 2021; Murmann et al. 2019], we define the

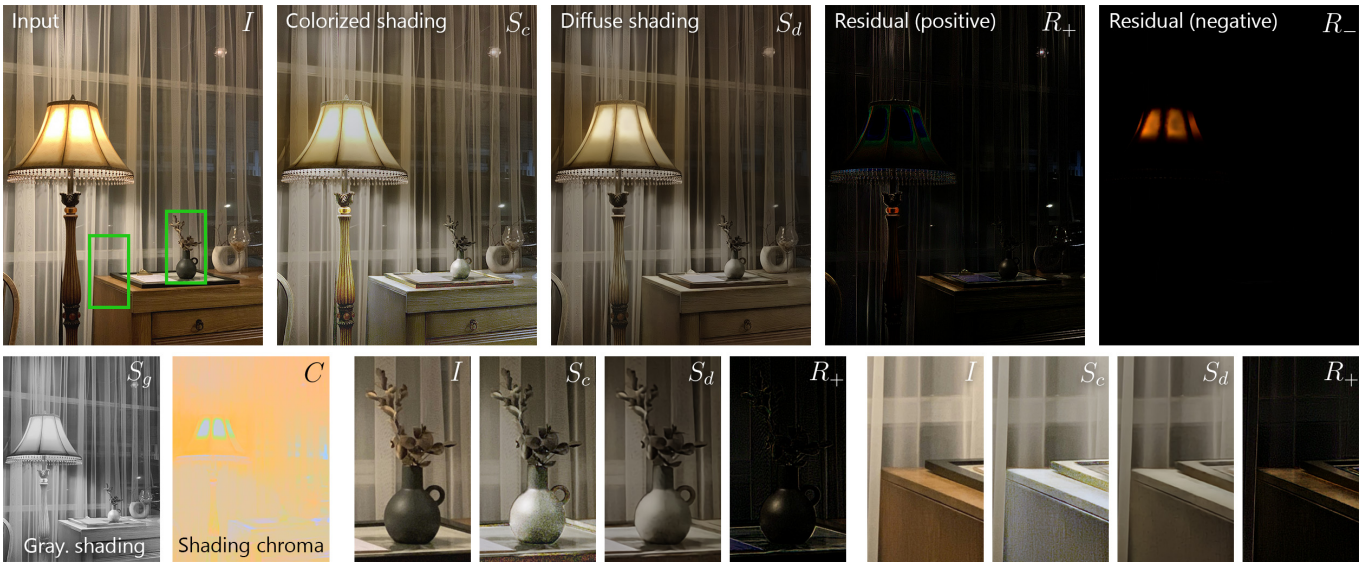

Fig. 5. Starting from a grayscale shading estimation, we first estimate the shading chroma (Sec. 3.1) and create a colorized shading map. In the final step of our pipeline (Sec. 3.3), we further separate the illumination into diffuse shading and non-diffuse residual components. The positive part of the residual represents the specularities in the scene, while the negative part shows the over-exposed regions. Image from Unsplash by Jiwoo Park.

chromaticity as color channel ratios:

$$U = S_c^r / S_c^g \quad V = S_c^b / S_c^g. \tag{4}$$

Given that color channel ratios are unbounded variables, it is challenging to train neural networks with a direct loss on them. Hence, we use a simple mapping to the $[0 - 1]$ range following Careaga and Aksoy [2023] and define our 2-channel target variable C:

$$C = \left[\frac{1}{U+1}, \frac{1}{V+1} \right]. \tag{5}$$

Our chroma network takes the grayscale decomposition (S_g, A_g) and the input image as a concatenated 7-channel input and estimates the 2-channel C. We train this network using the standard mean-squared error and the multi-scale gradient loss commonly utilized in the literature for mid-level vision tasks [Careaga and Aksoy 2023; Li and Snavely 2018a,c; Miangoleh et al. 2024; Ranftl et al. 2020]:

$$\mathcal{L}_{mse}(C) = \frac{1}{N} \sum_{i=1}^{N} (C_i - C_i^*)^2 \tag{6a}$$

$$\mathcal{L}_{msg}(C) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{l=1}^{M} (\nabla C_{i,l} - \nabla C_{i,l}^*)^2, \tag{6b}$$

where C^* is the ground-truth color component image, and $\nabla C_{i,l}$ is the gradient of C at scale l.

The shading chromaticity estimation requires an understanding of the global context present in the scene. It is also a low-frequency variable as discussed by Lettry et al. [2018], making a low-resolution estimation viable. As a result, we utilize a convolutional architecture as detailed in Section 3.4 and estimate C at the receptive field-size resolution. We then combine this low-resolution C with its luminance S_g to construct the RGB shading layer $\hat{S}_c$. Figure 5 shows an example of our estimated chroma.

3.2 Albedo Estimation

The albedo channel, when defined under the grayscale diffuse model, contains strong color shifts coming from colored illumination. The colorized shading $\hat{S}_c$ from our chroma network can be used to compute an approximation to the correct albedo, $\hat{A}_c$, using the RGB diffuse model in Equation 3. However, due to the low-resolution chroma estimation and the lack of enforcement of sparse albedo values up to this point, $\hat{A}_c$ still exhibits illumination-related artifacts as Figure 4 demonstrates.

In order to estimate our final diffuse albedo layer, we define our *albedo network* that takes $\hat{A}_c$ and $\hat{S}_c$ as input together with the input image concatenated to be a 9-channel input and outputs the diffuse albedo A_d. With the global context on illumination color readily provided in its input, the task of our albedo network is to take advantage of the sparse nature of albedo and generate an accurate 3-channel diffuse albedo map. Similar to our chroma network, we use the mean-squared error $\mathcal{L}_{mse}(A)$ and the multi-scale gradient $\mathcal{L}_{msg}(A)$ losses on the albedo to train this network. As Figure 4 shows, this results in a flat albedo without illumination artifacts.

3.2.1 Training datasets. Intrinsic decomposition methods are typically trained with synthetic ground truth. Most synthetic intrinsic datasets readily provide the ground-truth albedo. Furthermore, real-world training data for albedo can be extracted from multi-illumination datasets [Careaga and Aksoy 2023], greatly aiding the in-the-wild generalization. We train our chroma and albedo networks, as well as the ordinal network of Careaga and Aksoy [2023], using 8 synthetic datasets [Krahenbuhl 2018; Le et al. 2021; Li et al. 2023; Roberts et al. 2021; Wang et al. 2022; Yeh et al. 2022; Zheng et al. 2020; Zhu et al. 2022b] and the multi-illumination dataset by Murmann et al. [2019] to provide a good variety of images during training, allowing our albedo estimation to generalize in-the-wild.

Fig. 6. Given the large number of recently proposed diffusion-based methods, we provide a focused qualitative evaluation against these models. These examples show some of the shortcomings of utilizing generative modeling to address the problem of intrinsic decomposition. Since these methods learn a mapping in the latent space of large pre-trained generative models, their outputs can have unintended side-effects like warped faces, and illegible text. These alterations can have a negative impact on downstream editing applications. Additionally, although these methods can achieve the high-level appearance of albedo, they are highly dependent on their training data distribution which can cause effects such as large color shifts and baked-in shading.
Images from Unsplash (from top to bottom) by Mert Kahveci, Joel Muniz, Dollar Gill, and Annie Spratt

3.3 Diffuse Shading Estimation

With the diffuse albedo A_d estimated, we are finally ready to abandon the Lambertian world assumption and estimate the colorful diffuse shading and non-diffuse illumination components in the intrinsic residual model in Equation 1. Given that diffuse shading is highly correlated with the scene geometry, our *diffuse shading network* needs to make use of the geometric cues in the scene to separate the diffuse effects from non-diffuse irradiance such as specularities and visible light sources. This problem can also be seen as the decomposition of $S_c = I/A_d$ in the RGB diffuse model in Equation 3 into diffuse and non-diffuse components.

Our diffuse shading network takes the diffuse albedo A_d, colorized shading from the diffuse model S_c, and the input image as a concatenated 9-channel input. We define the output in the inverse shading space [Careaga and Aksoy 2023] as a three-channel variable $D = 1/(S_d + 1)$ and use the mean-squared error $\mathcal{L}_{mse}(D)$ and the multi-scale gradient $\mathcal{L}_{msg}(D)$ losses during training.

Given the estimated diffuse shading S_d and albedo A_d, we compute the residual non-diffuse layer using the intrinsic residual model in Equation 1:

$$R = I - (A_d * S_d). \tag{7}$$

It should be noted that our estimated diffuse shading is unbounded, and therefore the diffuse image $(A_d * S_d)$ can exceed the input's $[0-$

Fig. 7. Prior grayscale shading works leave residual lighting such as interreflections in the albedo due to their formulation. Some works have attempted to model colorful lighting but the difficulty of the problem is exacerbated by the lack of ground truth data. Due to our formulation, our method is able to remove the colorful lighting effects from the albedo even for in-the-wild images. When compared to our single network baseline, our method generates a sparse albedo with colors that are true to the input image. Image from Unsplash by Eco Warrior Princess.

1] range. This high-dynamic range property of our diffuse shading enables image enhancement applications as shown in Figure 5. As a result of this property, our estimated residual has both negative and positive values. The positive part of the residual contains non-diffuse illumination effects such as specularities and visible light sources, while the negative residual shows over-exposed regions in the input image as shown in Figures 5 and 10.

3.3.1 Training dataset. High-resolution synthetic datasets are scarce for diffuse shading and lack diversity, while real-world datasets are non-existent. This is the main reason why prior methods that focus on the residual model limit their application scenario to specific object classes. In our pipeline, however, our diffuse shading network readily gets the albedo and S_c as input, which eases the contextual nature of its task. We train our diffuse network solely on the synthetic indoor Hypersim dataset [Roberts et al. 2021]. However, as our qualitative results demonstrate, our method generalizes to a wide range of in-the-wild images. This shows that by simplifying the task of each network, we are able to utilize the generalizability of our albedo estimation pipeline to achieve in-the-wild non-diffuse intrinsic decomposition.

3.4 Network Structure and Training

We utilize the same encoder-decoder architecture from [Ranftl et al. 2020] that has been shown to be useful for various mid-level vision tasks for all of our networks. We use a sigmoid activation to output quantities strictly in the $[0-1]$ range. We train all the networks using the Adam optimizer with a learning rate of 10^{-5}. Since intrinsic decomposition is an inherently scale-invariant task, typical formulations utilize scale-invariant losses when predicting intrinsic components. Due to the instability of these losses, we adopt the methodology of [Careaga and Aksoy 2023] and set the arbitrary scale of ground-truth according to the input decomposition of each network. In doing these, regular loss functions can be used, training the networks to rely on the scale of the input to make their predictions. We provide further details in the supplementary.

4 EXPERIMENTS

We present quantitative evaluations of our method on common benchmarks, as well as qualitative comparisons to recent work. We extend our qualitative comparisons and show all the different components we estimate in a large set of in-the-wild images in the supplementary material.

Fig. 8. We train a large single network baseline on the same datasets as our full method (bottom row). Our pipeline achieves superior albedo estimation (middle row), especially on out-of-distribution high-resolution imagery. We attribute our performance to our multi-stage approach wherein each network accomplishes its simpler sub-task, learning generalizable behavior. Images from Unsplash by Alli Stefanova, Shalev Cohen, and Judith Girard-Marczak.

In order to show the effectiveness of our multi-stage pipeline, we compare our method to a single large model trained on the same datasets as a baseline. Specifically, we compare the albedo estimation of our approach (grayscale shading → shading chromaticity → albedo) to a single network that takes the input image and estimates albedo directly. The single network has 485 million parameters compared to our 4 networks which have 337 million parameters, cumulatively. We train the network for 1 million iterations with a batch size of 2 as that is the maximum size allowed by a 40GB GPU. We refer to this network as the "single-network baseline". It should be noted that such a single-network baseline is not practical for diffuse shading estimation due to the lack of real-world training datasets.

4.1 Quantitative Evaluation

Due to the lack of ground-truth benchmarks on diffuse shading, we report our quantitative analysis on the common test sets in the literature that focus on albedo estimation.

4.1.1 MAW Dataset. Measured Albedo in the Wild (MAW) Dataset [Wu et al. 2023] has recently been introduced to measure real-world albedo accuracy in terms of intensity and color. The dataset consists of ~850 indoor images and measured albedo within specific masked regions in the image. The albedo is measured using a known gray card placed on areas of homogeneous albedo. We focus our evaluation on two metrics that measure the accuracy of albedo intensity and chromaticity, respectively. The results are reported in Table 1. As shown by the discrepancy between the intensity and chromaticity scores of the work by Careaga and Aksoy [2023], the grayscale shading assumption results in large inaccuracies in the color of the estimated albedo. Our initial shading chroma estimation is already able to compensate for these color shifts and scores the second-best in all metrics. Our final refined albedo estimation further improves the results, outperforming all prior methods in terms of both intensity and chromaticity. When compared to the single-network baseline, our method achieves significantly higher performance, especially in terms of accurate albedo chromaticity. We attribute this improvement to our multi-stage approach which allows our method to generalize to the real-world images in MAW.

Table 1. Numerical results on the Measured Albedo in the Wild (MAW) Dataset [Wu et al. 2023]. We achieve state-of-the-art performance on albedo estimation in terms of both intensity and chromaticity. Methods with an asterisk use the grayscale shading assumption and therefore have a fixed chromaticity score. For the first 7 methods we use the results computed by the authors of the MAW dataset.

Method	Intensity (× 100)↓	Chromaticity↓
Bell et al. [2014]	3.11	6.61
Li and Snavely [2018b]	2.71	5.15
Li and Snavely [2018a]	1.72	6.56*
Sengupta et al. [2019]	2.17	6.39
Liu et al. [2020]	2.62	6.00
Li et al. [2020]	1.41	5.64
Luo et al. [2020]	1.24	4.73
Lettry et al. [2018]	2.77	8.05
Zhu et al. [2022b]	1.44	4.94
Kocsis et al. [2024]	1.13	5.35
Chen et al. [2024]	0.98	4.12
Careaga and Aksoy [2023]	0.57	6.56*
Single-Network Baseline	0.69	4.15
Ours ($\hat{A}_c$)	0.56	3.50
Ours	**0.54**	**3.37**

Table 2. Zero-shot albedo evaluation on the synthetic ARAP Dataset [Bonneel et al. 2017]. Our proposed method estimates the most accurate albedo across all zero-shot methods, even out-performing a non-zero-shot method marked with an asterisk in terms of SSIM.

Method	LMSE↓	RMSE↓	SSIM↑
Chromaticity	0.038	0.193	0.710
Constant Shading	0.047	0.264	0.693
Luo et al. [2020]*	0.023	0.129	0.788
Lettry et al. [2018]	0.050	0.193	0.732
Kocsis et al. [2024]	0.030	0.160	0.738
Zhu et al. [2022b]	0.029	0.184	0.729
Chen et al. [2024]	0.038	0.171	0.692
Careaga and Aksoy [2023]	0.035	0.162	0.751
Single-Network Baseline	0.022	0.150	**0.796**
Ours ($\hat{A}_c$)	0.025	0.156	0.752
Ours	**0.021**	**0.149**	**0.796**

4.1.2 ARAP Dataset. In order to quantify the generalization abilities of each method to out-of-distribution scenes, we evaluate albedo estimation on the As Real as Possible (ARAP) Dataset [Bonneel et al. 2017]. The dataset consists of about 50 rendered scenes, from various sources. We augmented the dataset with three scenes from the MIST Dataset [Hao and Funt 2020] and also removed duplicated images

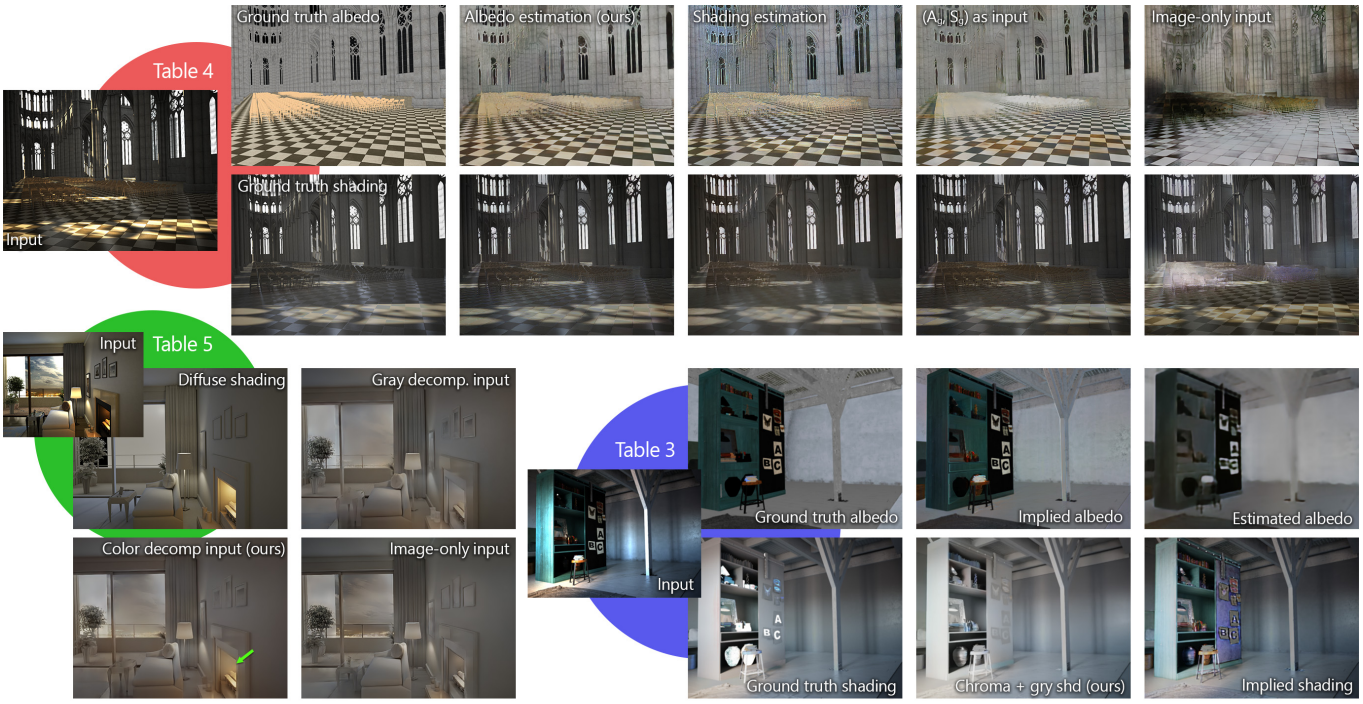

Fig. 9. Qualitative examples from the Hypersim dataset for each of the three ablations discussed in Section 4. In Table 3 (bottom right) we show that estimating the chromaticity is a much easier task than estimating albedo directly, due to it's low-frequency nature. In Table 4 (top) we show that estimating albedo, given an initial colorful decomposition, is the best way to remove residual shading effects due to the inherent smoothness of the albedo component. In Table 5 we show that our colorful decomposition is a vital input when estimating diffuse shading, and unlocks in-the-wild diffuse shading estimation.

and made sure each scene is equally represented in the dataset. We follow the same experimental setup as Careaga and Aksoy [2023] for computing evaluation metrics on the albedo. The results are reported in Table 2, with similar conclusions to Table 1. We observe that in the rendered data setting of the ARAP dataset, the single-network baseline performs similarly to our network quantitatively despite its relatively poor performance on the MAW dataset. This shows that although the single network is able to accurately model the distribution of the synthetic training data, our multi-stage approach unlocks superior generalization to in-the-wild images.

4.2 Qualitative Evaluation

Figure 7 show the results by Careaga and Aksoy [2023], Luo et al. [2020], Das et al. [2022] and Liu et al. [2020] which all adopt the grayscale intrinsic diffuse model, as well as the work by Lettry et al. [2018] that adopts the RGB intrinsic model in their unsupervised formulation. Additionally, we compare against Zhu et al. [2022b] and our single-network baseline. Since these two methods only provide albedo estimations, we compute shading by dividing the input image and estimated albedo. When the grayscale model is enforced on the albedo-shading pair, the color of secondary illuminations creates color shifts in the albedo, as the results by Careaga and Aksoy [2023] and Das et al. [2022] show. Although the method of Zhu et al. [2022b] estimates unconstrained albedo, their estimation still exhibits residual colors in the cast shadows. This color cast is removed in the result by Luo et al. [2020]. However, since they still work within the grayscale model, their intrinsic components fail to faithfully reconstruct the image. The single-network baseline is able to remove the color cast from some shadows, but the network misses many shading effects (e.g. on the leaves), and generally shifts the colors of the input image (e.g. on the gloves). We see a strong color cast and residual albedo colors in the shading by Lettry et al. [2018], while our adoption of the intrinsic residual model allows us to estimate a clean albedo with colors of the secondary illuminations represented in our colorful shading.

We show in-the-wild comparisons against recent diffusion-based methods by Zeng et al. [2024], Kocsis et al. [2024], and Chen et al. [2024] in Figure 6. The work by Chen et al. [2024] suffers from low resolution in in-the-wild scenes, while in indoor scenes they are sometimes susceptible to a string tiling effect due to their high-resolution refinement. The work by Kocsis et al. [2024], on the other hand, struggles in out-of-distribution scenes and generates a low-resolution result due to the averaging of their results. The refinement and averaging strategies adopted by these methods result in $>$ 10 seconds run times, while our full pipeline takes around a second on average to generate a high-resolution result. The method by Zeng et al. [2024] can generate sharp results but suffers from typical diffusion-based generation artifacts around text and may cause cartoonization of human faces. Our analytical modeling of the problem remains faithful to the input image and is able to generalize to out-of-distribution images effectively.

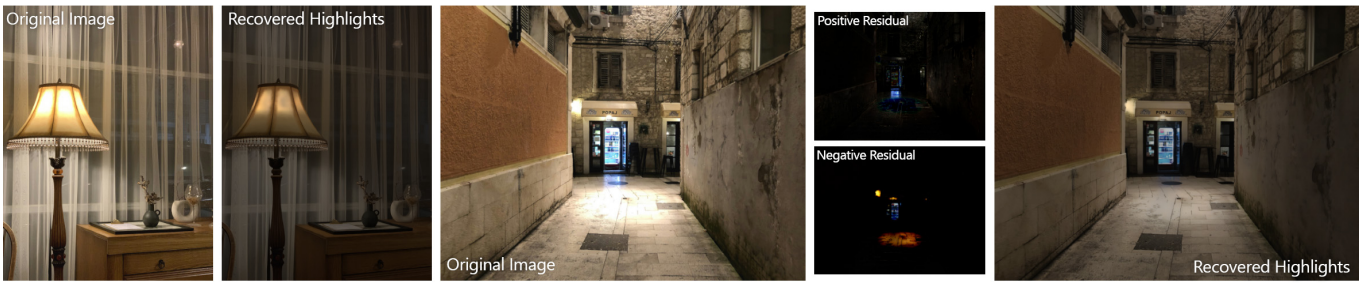

Fig. 10. Since our estimated diffuse shading can represent unbounded light intensities, our method is able to recover information that was originally clipped in the input image. This clipped information results in negative values in our residual layer. Image from Unsplash by Jiwoo Park

Table 3. Ablation experiment of our low-resolution chroma network. Our proposed method of estimating two-channel color information achieves much better results than directly estimating the albedo layer.

Method	Shading			Albedo		
	LMSE↓	RMSE↓	SSIM↑	LMSE↓	RMSE↓	SSIM↑
direct albedo estimation	0.233	3.024	0.554	0.027	0.096	0.710
ours - w/ chroma network	**0.187**	**2.866**	**0.689**	**0.022**	**0.083**	**0.815**

Table 4. Ablation of albedo network formulation. By estimating the albedo given $\hat{S}_c$ our method yields better results than when directly estimating the shading at high resolution. This shows that the network can exploit the sparse nature of the albedo to make accurate predictions. Using any other inputs other than $\hat{S}_c$ and $\hat{A}_c$ results in decreased performance.

Method	Shading			Albedo		
	LMSE↓	RMSE↓	SSIM↑	LMSE↓	RMSE↓	SSIM↑
image-only input	0.140	1.635	0.551	0.026	0.174	0.652
(S_g, A_g) as input	0.125	1.298	0.648	0.015	0.092	0.752
shading estimation	**0.097**	**1.187**	0.624	0.019	0.097	0.751
ours - albedo estimation	0.116	**1.188**	**0.657**	**0.012**	**0.090**	**0.757**

4.3 Ablations

In order to measure the performance impact of each individual design choice in our pipeline, we carry out multiple controlled experiments using the Hypersim dataset. For all ablations, we create a random scene split of the Hypersim dataset, with 66,000 images for training and 6,000 for evaluation. Each model variant is trained with a batch size of 8 for 25,000 iterations which is enough to give reasonable convergence given the similarity in the training and testing distributions.

4.3.1 Chromaticity Estimation. Table 3 shows the result of removing our chromaticity estimation formulation. The first row provides the scores for our proposed approach, estimating two color components and using the decomposition of Careaga and Aksoy [2023] as the luminance. The second row shows the result of directly estimating the albedo instead of using color components. Both networks receive the same input consisting of the input image and the grayscale decomposition from Careaga and Aksoy [2023]. The networks are evaluated at the receptive field size of the network (384px) in order to measure the global accuracy of each variant. We can see that by estimating the color components, the network learns the task much more effectively than when trying to directly reason about the albedo layer. This shows that this is an effective first step to estimating accurate albedo from the grayscale decomposition.

Table 5. Diffuse shading ablation experiment. When providing our diffuse shading network with diffuse albedo A_d and the corresponding shading S_c our network yields the best results. Any other input configuration results in worse performance, highlighting the effectiveness of our multi-step pipeline.

Method	LMSE↓	RMSE↓	SSIM↑
image input	0.045	0.352	0.696
grayscale decomp. input	0.043	0.340	0.723
ours - diffuse estimation	**0.040**	**0.329**	**0.728**

4.3.2 Albedo Estimation. Table 4 shows the result of alternatives to our second network that estimates high-resolution albedo. The first row shows our proposed approach of estimating the albedo given the decomposition with low-resolution chromaticity predicted by our first network. The second row shows that if we were to instead estimate high-resolution shading, our performance would drop in both albedo and shading estimation, showing that albedo estimation is a conceptually easier task for the network to model. The third and fourth rows show our albedo estimation network with different input configurations. Omitting the low-resolution chromaticity in the input shading results is a large performance decrease across all metrics, showing the value of estimating the albedo with our two-step approach. When completely omitting the intrinsic inputs and only providing the input image, the performance drops even more drastically, further showing the effectiveness of our multi-step approach.

4.3.3 Diffuse Shading Estimation. Table 5 shows the impact of different input configurations on the diffuse shading network. The first row shows our proposed approach of using our estimated albedo and RGB shading layer as input. The second row shows that only providing the network with a grayscale decomposition, makes the task more difficult as the network has to reason about the shading color, resulting in lower scores on all metrics. Finally, the last row shows that our multi-step approach is essential for being able to estimate diffuse shading as trying to estimate it directly from the input image yields poor results. We present qualitative comparisons accompanying Tables 3–5 in Figure 9.

Fig. 11. By removing the specular residual produced by our method, we are able to achieve specularity removal. Even though our diffuse shading network is solely trained on indoor images, it can still generate accurate estimations for diverse images. This capability is enabled by our multi-step formulation.
Images from Unsplash by Kostiantyn Li (teapot) and Israel Albornoz (coffee).

Fig. 12. As our pipeline starts with a grayscale intrinsic decomposition as input, it may not be able to fix strong mistakes made by the initial model, such as the hard shadows of the balconies incorrectly included in the initial albedo on the left. However, we show that these mistakes can be fixed when the input albedo is corrected, in this case using Photoshop's content-aware inpainting tool.
Image from Unsplash by Jon Flobrant.

5 APPLICATIONS

The intrinsic residual model allows for several computational photography applications by estimating a color component for the shading and separating diffuse and non-diffuse illumination effects. As demonstrated in Figures 2 and 11, specularities in an image can be removed by computing the diffuse image $A_d * S_d$. Estimating the shading in color allows for per-pixel multi-illuminant white balancing, as shown in Figure 2. Our unbounded estimation of the diffuse shading allows us to recover details that are lost to clipping in the input image, as demonstrated in Figure 10.

6 LIMITATIONS

Our method builds the intrinsic residual components by starting from the estimation of an existing method. While our networks are trained to account for erroneous initial estimations, they may also propagate some of the challenging mistakes as shown in Figure 12. In some simple cases, such mistakes can roughly be edited out in the input using commercial software, which in turn allows our pipeline to correct its estimation.

7 CONCLUSION AND FUTURE WORK

In this work, we present an intrinsic decomposition method that can successfully separate diffuse and non-diffuse lighting effects in the wild and at high resolutions. Our high-resolution performance and generalization ability come from our modeling of this highly under-constrained problem in physically-motivated sub-tasks. We demonstrate through quantitative analysis as well as qualitative examples that despite training our final diffuse network only on a synthetic indoor dataset, we are able to generalize to a wide variety of scenes including human faces and outdoor landscapes. We demonstrate new illumination-aware image editing applications that are made possible by adopting the intrinsic residual model.

We believe our method opens up multiple avenues for future work in this area. Our intrinsic residual model has the potential to improve intrinsics-based computational photography applications, some of which have been explored but could be improved by our approach, such as relighting [Careaga et al. 2023], flash photography [Maralan et al. 2023], and HDR reconstruction [Dille et al. 2024]. Our method represents a large step towards developing physically accurate inverse rendering methods that generalize to in-the-wild images, and our components have the potential to be further decomposed into explicit lighting, BRDF parameters, and single- vs. multi-bounce contributions using more complex image formation models.

ACKNOWLEDGMENTS

We would like to thank Zheng Zeng for promptly providing results for their method. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RGPIN-2020-05375].

REFERENCES

Jonathan T. Barron. 2015. Convolutional Color Constancy. In *Proc. ICCV*.

Jonathan T. Barron and Jitendra Malik. 2015. Shape, Illumination, and Reflectance from Shading. *IEEE Trans. Pattern Anal. Mach. Intell.* (2015).

Anil S. Baslamisli, Thomas T. Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. 2018a. Joint Learning of Intrinsic Images and Semantic Segmentation. In *Proc. ECCV*.

Anil S. Baslamisli, Hoang-An Le, and Theo Gevers. 2018b. CNN based learning using reflection and retinex models for intrinsic image decomposition. In *Proc. CVPR*.

Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic Images in the Wild. *ACM Trans. Graph.* 33, 4 (jul 2014), 1–12.

Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. 2017. Intrinsic Decompositions for Image Editing. *Comput. Graph. Forum* 36, 2 (2017).

Chris Careaga and Yağız Aksoy. 2023. Intrinsic Image Decomposition via Ordinal Shading. *ACM Trans. Graph.* 43, 1, Article 12 (2023), 24 pages.

Chris Careaga, S. Mahdi H. Miangoleh, and Yağız Aksoy. 2023. Intrinsic Harmonization for Illumination-Aware Compositing. In *Proc. SIGGRAPH Asia*.

Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. 2024. IntrinsicAnything: Learning Diffusion Priors for Inverse Rendering Under Unknown Illumination. In *Proc. ECCV*.

Partha Das, Sezer Karaoglu, and Theo Gevers. 2022. PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition. In *Proc. CVPR*.

Sebastian Dille, Chris Careaga, and Yağız Aksoy. 2024. Intrinsic Single-Image HDR Reconstruction. In *Proc. ECCV*.

Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. 2012. Intrinsic Images by Clustering. *Comput. Graph. Forum* 31, 4 (2012), 1415–1424.

Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. 2022. A Survey on Intrinsic Images: Delving Deep Into Lambert and Beyond. *Int. J. Comput. Vision* (2022).

Xiangpeng Hao and Brian Funt. 2020. A multi-illuminant synthetic image test set. *Color Research & Application* (2020).

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

Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. 2014. Automatic scene inference for 3d object compositing. *ACM Trans. Graph.* (2014).

Dongyoung Kim, Jinwoo Kim, Seonghyeon Nam, Dongwoo Lee, Yeonkyung Lee, Nahyup Kang, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han, and Seon Joo Kim. 2021. Large Scale Multi-Illuminant (LSMI) Dataset for Developing White Balance Algorithm Under Mixed Illumination. In *Proc. ICCV*.

Hyeongwoo Kim, Hailin Jin, Sunil Hadap, and Inso Kweon. 2013. Specular reflection separation using dark channel prior. In *Proc. CVPR*.

Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. 2024. Intrinsic Image Diffusion for Single-view Material Estimation. *Proc. CVPR*.

Philipp Krahenbuhl. 2018. Free Supervision from Video Games. In *Proc. CVPR*.

Hoang-An Le, Partha Das, Thomas Mensink, Sezer Karaoglu, and Theo Gevers. 2021. EDEN: Multimodal Synthetic Dataset of Enclosed Garden Scenes. In *Proc. WACV*.

Louis Lettry, Kenneth Vanhoey, and Luc van Gool. 2018. Unsupervised Deep Single-Image Intrinsic Decomposition using Illumination-Varying Image Sequences. *Comput. Graph. Forum* 37, 7 (2018), 409–419.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. 2023. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proc. ICCV*.

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proc. CVPR*.

Zhengqi Li and Noah Snavely. 2018a. CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering. In *Proc. ECCV*.

Zhengqi Li and Noah Snavely. 2018b. Learning Intrinsic Image Decomposition from Watching the World. In *Proc. CVPR*.

Zhengqi Li and Noah Snavely. 2018c. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Proc. CVPR*.

Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi Yang. 2022. PhyIR: Physics-based Inverse Rendering for Panoramic Indoor Images. In *Proc. CVPR*.

Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. 2020. Unsupervised Learning for Intrinsic Image Decomposition from a Single Image. In *Proc. CVPR*.

Jundan Luo, Zhaoyang Huang, Yijin Li, Xiaowei Zhou, Guofeng Zhang, and Hujun Bao. 2020. NIID-Net: Adapting Surface Normal Knowledge for Intrinsic Image Decomposition in Indoor Scenes. *IEEE Trans. Vis. Comp. Graph.* (2020).

Jundan Luo, Nanxuan Zhao, Wenbin Li, and Christian Richardt. 2023. CRefNet: Learning Consistent Reflectance Estimation With a Decoder-Sharing Transformer. *IEEE Trans. Vis. Comp. Graph.* (2023).

Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. 2018. Single Image Intrinsic Decomposition Without a Single Intrinsic Image. In *Proc. ECCV*.

Sepideh Sarajian Maralan, Chris Careaga, and Yağız Aksoy. 2023. Computational Flash Photography through Intrinsics. *Proc. CVPR*.

Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. 2018. LIME: Live Intrinsic Material Estimation. In *Proc. CVPR*.

Abhimitra Meka, Mohammad Shafiei, Michael Zollhoefer, Christian Richardt, and Christian Theobalt. 2021. Real-time Global Illumination Decomposition of Videos. *ACM Trans. Graph.* (2021).

S. Mahdi H. Miangoleh, Mahesh Reddy, and Yağız Aksoy. 2024. Scale-Invariant Monocular Depth Estimation via SSI Depth. In *Proc. SIGGRAPH*.

Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019. A Multi-Illumination Dataset of Indoor Object Appearance. In *Proc. ICCV*.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *Proc. ICCV*.

Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. 2019. Neural Inverse Rendering of an Indoor Scene From a Single Image. In *Proc. ICCV*.

Viraj Shah, Svetlana Lazebnik, and Julien Philip. 2023. JoIN: Joint GANs Inversion for Intrinsic Image Decomposition. *arXiv preprint 2305.11321 [cs.CV]* (2023).

Jianbing Shen, Xiaoshan Yang, Yunde Jia, and Xuelong Li. 2011. Intrinsic images using optimization. In *Proc. CVPR*.

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

Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. 2017. Learning Non-Lambertian Object Intrinsics Across ShapeNet Categories. In *Proc. CVPR*.

Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. 2005. Recovering intrinsic images from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 9 (2005), 1459–1472.

Yujie Wang, Qingnan Fan, Kun Li, Dongdong Chen, Jingyu Yang, Jianzhi Lu, Dani Lischinski, and Baoquan Chen. 2022. High quality rendered dataset and non-local graph convolutional network for intrinsic image decomposition. *Journal of Image and Graphics* (2022).

Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. 2021. Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting. In *Proc. ICCV*.

Jiaye Wu, Sanjoy Chowdhury, Hariharmano Shanmugaraja, David Jacobs, and Soumyadip Sengupta. 2023. Measured Albedo in the Wild: Filling the Gap in Intrinsics Evaluation. In *Proc. ICCP*.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Trans. Graph.* (2022).

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yiwei Hu Yannick Hold-Geoffroy, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. RGB↔X: Image Decomposition and Synthesis Using Material- and Lighting-aware Diffusion Models. In *Proc. SIGGRAPH*.

Qian Zhang, Vikas Thamizharasan, and James Tompkin. 2022. Learning Physically-based Material and Lighting Decompositions for Face Editing. *Computational Visual Media* (2022).

Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. 2012. A Closed-Form Solution to Retinex with Nonlocal Texture Constraints. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 7 (2012), 1437–1444.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. In *Proc. ECCV*.

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. 2022b. Learning-Based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing. In *Proc. SIGGRAPH Asia*.

Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. 2022a. IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes. In *Proc. CVPR*.