Estimating Physically-Based Material Representations from Diffusion-Generated Lighting Variations

Felix Laarmann
TU Munich
felix.laarmann@tum.de

Lukas Haußmann
TU Munich
lukas.haussmann@tum.de

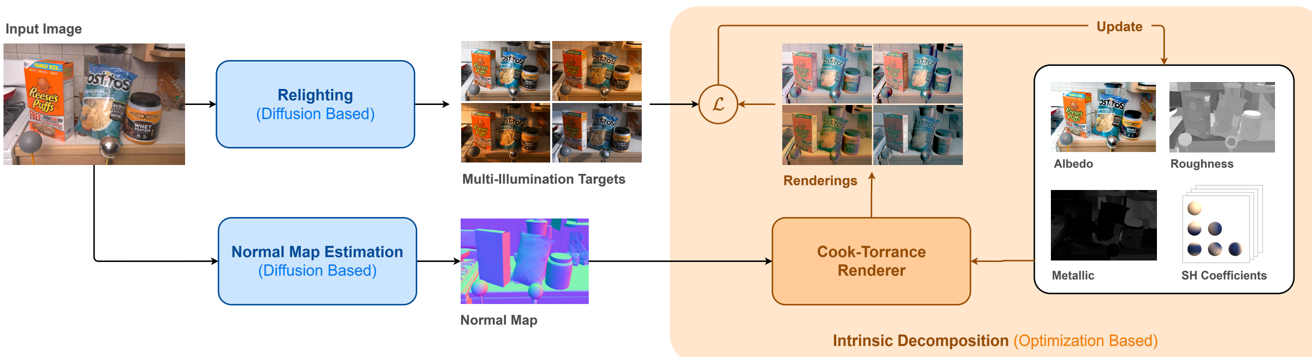

Figure 1. Our intrinsic decomposition pipeline: We generate a multi-illumination dataset from a single input image using a large diffusion-based image editing model. We optimize for an intrinsic decomposition by minimizing a regularized reconstruction loss against the target images through a differentiable Cook-Torrance renderer that is equipped with an estimated normal map.

Abstract

Recovering physically-based rendering (PBR) material parameters from a single image is ill-posed. Recent methods regress intrinsics directly with diffusion models that are fine-tuned on synthetic interiors and generalize poorly to photographs. We instead use the diffusion prior to generate observations. We recover per-pixel albedo, roughness and metallic maps jointly with per-image spherical-harmonic lighting through a differentiable Cook-Torrance renderer, and show that this yields more natural relighting of real photographs than direct diffusion baselines.

1. Introduction

Estimating PBR materials from RGB images is a active research problem in computer vision research. Obtaining the intrinsic components (albedo, roughness, metallic and normals) is required for various downstream tasks such as image relighting or material editing. The problem of decomposing a single image into its intrinsic components is ill-posed due to the inherent ambiguity between material and lighting. Classical methods rely on optimization-based decomposition using strong physical priors while recent methods leverage pretrained diffusion models to generate PBR material maps for a single input image. Such diffusion-based methods suffer from the non-existence of real-world ground-truth data and are therefore limited to synthetic datasets. As a result, those methods generalize poorly to out of distribution data. To overcome this challenge, we propose a method that leverages the implicit world knowledge of large diffusion-based image models to generate multi-illumination data from a single input image. We then use the generated image data to drive an optimization-based intrinsic decomposition strategy separating per-image lighting conditions from global per-pixel PBR materials.

2. Related Work

2.1. Classical Intrinsic Image Decomposition

Early work on intrinsic image decomposition originates from Retinex theory [2, 3], which separates an image into reflectance and illumination based on the assumption that illumination varies smoothly while reflectance changes abruptly at material boundaries. Later methods refine this

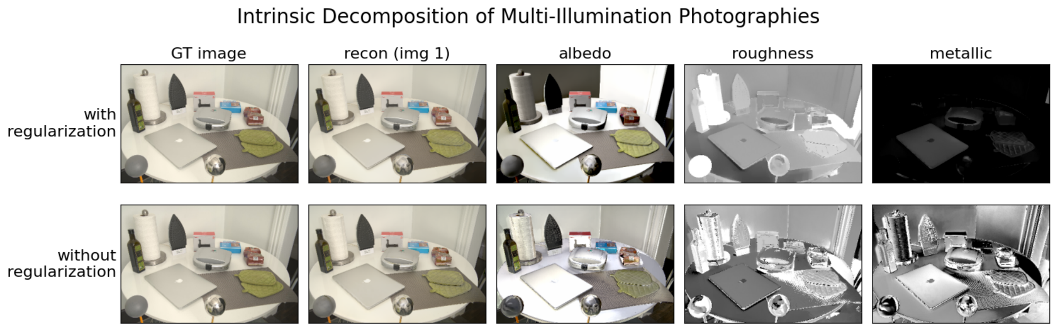

Figure 2. Exemplary decomposition of the `marlborough_kitchen_1` scene of the MIT dataset [1] with and without regularization. Regularization slightly increases the reconstruction error for the displayed image from 0.0397 to 0.0463 while producing significantly more plausible intrinsic maps.

idea with additional priors such as entropy minimization [4] and non-local texture cues [5].

2.2. Deep-Learning-Based Intrinsic Decomposition

With the rise of deep learning, later approaches train neural networks to directly regress intrinsic properties from an input image, typically supervised on synthetic datasets with known ground-truth material maps such as InteriorVerse [6] and Hypersim [7, 8]. More recent methods have leveraged the strong image priors of pretrained diffusion models for this task, for example, by fine-tuning a diffusion model to jointly predict multiple material channels [9, 10]. RGB→X [11] frames decomposition as image-to-image translation. Marigold [12] originally fine-tuned an image generation model for monocular depth and normal map estimation and expanded the approach to intrinsic decomposition [13].

3. Method

Our method consists of a multi-step pipeline shown in Fig. 1. Starting from a single RGB input image we use a large diffusion-based image editing model such as Flux.2 Klein [14] or GPT-Image-2 [15] and a list of prompts to generate a set of images of the same view with varying lighting conditions. In addition, we use a normal estimation model [12] to approximate normals from the input image. We then run an optimization-based intrinsic decomposition to separate into per-image lighting conditions and per-pixel albedo, roughness and metallic maps. The decomposition stage combines a differentiable renderer, a lighting model, a gradient-descent-based optimizer, and a set of regularization priors.

Rendering Model. We model surface reflectance with the Cook-Torrance microfacet BRDF [16], combining a GGX/Trowbridge-Reitz normal distribution [17, 18], Smith masking-shadowing with the Karis remapping [19, 20], and the Schlick Fresnel approximation [21]. Surfaces follow the metallic workflow [22], parameterized per-pixel by albedo $\rho \in [0,1]^3$, roughness $r \in [0,1]$, and metallic $m \in [0,1]$, which together determine the diffuse and specular BRDF terms. Outgoing radiance at a surface point with normal $\mathbf{n}$ and view direction $\mathbf{v}$ follows the rendering equation

$$L_o(\mathbf{v}) = \int_\Omega f_r(\mathbf{l},\mathbf{v})\, L_i(\mathbf{l})\, (\mathbf{n}{\cdot}\mathbf{l})\, d\mathbf{l}, \qquad f_r = f_d + f_s, \quad (1)$$

where $\mathbf{l}$ is the incident light direction integrated over the hemisphere Ω, and f_d and f_s are the diffuse and specular BRDF terms.

Lighting Model. Evaluating Equation 1 exactly requires integrating over the incident hemisphere for every pixel and optimization step, which is intractable at our scale. We instead represent the incident illumination L_i as a spherical-harmonic (SH) environment map truncated at $\ell \leq 3$ (16 real coefficients per color channel), fit independently per generated lighting image:

$$L_i(\boldsymbol{\omega}) = \sum_{\ell=0}^{3} \sum_{m=-\ell}^{\ell} L_{\ell m} Y_{\ell m}(\boldsymbol{\omega}). \quad (2)$$

This truncation admits closed-form solutions to both BRDF terms in Equation 1. The diffuse term reduces to the irradiance projection of Ramamoorthi and Hanrahan [23]; its band weights vanish beyond $\ell = 1$ under the clamped-cosine transfer function, so the $\ell = 3$ band contributes nothing to diffuse shading. The specular term exploits the GGX lobe's rotational symmetry about the mirror-reflection direction $\mathbf{r} = 2(\mathbf{n}\cdot\mathbf{v})\mathbf{n} - \mathbf{v}$. Convolving this zonal kernel with the environment reduces to a per-band scalar attenuation of its SH coefficients [24, 25], following the standard real-time approximation of evaluating the Fresnel and geometry terms at the view angle and treating the normal distribution function as zonal about $\mathbf{r}$ [20]. Unlike the diffuse case, the $\ell = 3$ band weight is generally non-zero for GGX, so it adds specular expressiveness beyond an $\ell \leq 2$ truncation. We derive these band weights analytically as closed-form functions of roughness and evaluate them in double

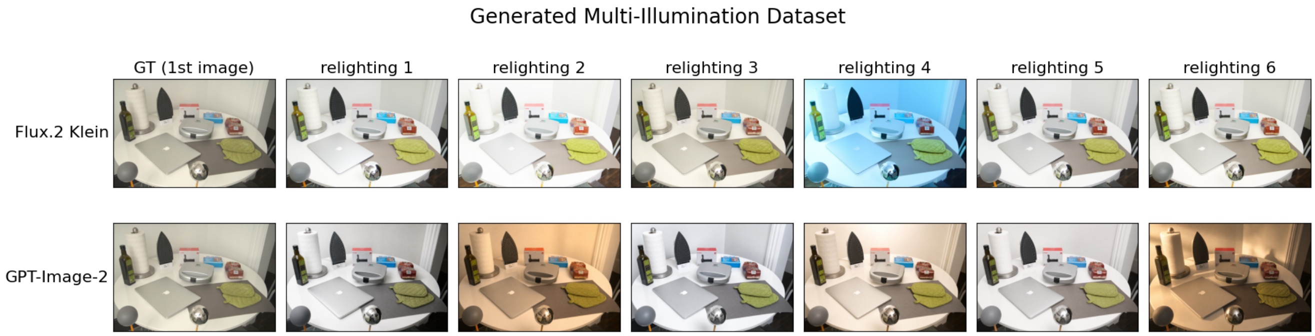

Figure 3. First six samples of multi-illumination datasets conditioned on a single image of the `marlborough_kitchen_1` scene
.

precision, avoiding cancellation errors that otherwise destabilize gradients at very low or very high roughness, rather than relying on numerical quadrature or a fixed-resolution lookup table. Together these closed forms give a fully differentiable renderer mapping per-pixel materials and per-image SH lighting directly to predicted radiance, with no numerical integration at render time.

Optimization. We jointly optimize all per-pixel materials and per-image SH lighting coefficients using L-BFGS for 500 steps, minimizing the L_2 rendering loss over all lighting variations. When using regularization, we combine the rendering loss with four priors, an L_1 sparsity prior on metallic favoring dielectric surfaces, a soft per-object cohesion prior pulling roughness and metallic toward their segment's mean for pixels classified by SAM2 [26], a spatial-smoothness prior over 4-connected neighbors that lets unclassified pixels blend with their neighbors, and a monochromatic light prior, that L_2 penalizes deviations of SH coefficients between the color channels.

4. Experiments

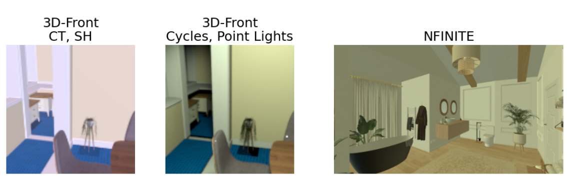

Figure 5. Samples from self-rendered databases

Dataset. To evaluate the decomposition, we require ground truth intrinsic information. For this we use the 3D-Front dataset [27] with CGAmbient PBR materials, filter for 67 setups that contain metallic materials and render dataset variants using our own renderer and SH2/SH3 lighting, Blender Cycles with an env map emulating SH lighting and point lights with and without rendering shadows to investigate different levels of model gap. We render 16 lighting variants for decomposition and another 16 variants to quantify relighting errors. We additionally render 10 scenes from the unpublished higher-detailed NFINITE dataset. Both contain interior scenes (Fig. 5). To investigate the performance on real images, we leverage the MIT Multi-Illumination dataset [1], that contains sets of photos captured with different light configurations from the same point of view.

	recon	albedo	rough.	metallic	relight
Base	0.001	0.065	0.232	0.186	0.166
SH2 decomp	0.002	0.064	0.242	0.147	0.154
Pred. normals	0.016	0.160	0.341	0.446	0.201
Cycles	0.040	0.386	0.408	0.630	–
NFINITE	0.043	0.222	0.334	0.435	0.210
Regularized	0.010	0.066	0.241	0.248	0.160
FOV=90	0.004	0.091	0.283	0.194	0.143

Table 1. RMSE of decomposition results. `Base`: SH3-decomposition of self-rendered 3D-Front dataset using no regularization, GT normals and GT FOV (60 deg). The `Cycles` configuration uses point light sources, so no relight error can be computed. `NFINITE` renders the NFINITE dataset emulating SH lighting with an env map.

Decomposition of Self-Rendered Sets. When decomposing 3D-Front scenes rendered by our CT renderer using SH lighting, no regularization and ground-truth normal maps, we achieve near-zero reconstruction and low relighting errors (Tab. 1). When decomposing scenes that were rendered with Blender's cycles renderer (including NFINITE) we observe higher errors due to the model gap. When using SH3 light for both dataset generation and decomposition we observe slightly lower recon RMSE in comparison to a SH2 setup. FOV deviations lead to minimal degradation.

Decomposition of Natural Multi-Illumination Images. To decompose photo sets of the MIT dataset we generate normal maps using Marigold [12]. On self-rendered 3D-Front data this raises the reconstruction RMSE well within

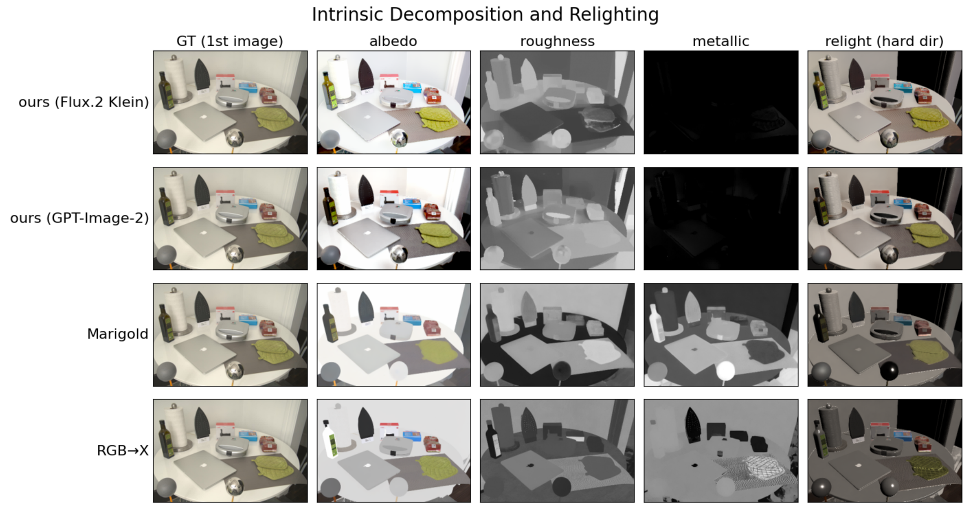

Figure 4. Exemplary decomposition of the `marlborough_kitchen_1` scene using generated datasets and comparing to reference methods. The relight rendering was produced using our CT renderer and an env map encoding hard directional light.

the margin already introduced by the model gap (Tab. 1). As no ground truth is available for the MIT dataset, we proceed with a qualitative review to determine an effective regularization strategy (Fig. 2).

Decomposition of Generated Multi-Illumination Data. We extend the pipeline to the full method by generating the set of multi-illumination images by conditioning a pretrained diffusion model on a single input image and a series of prompt instructions. We validate this approach using GPT-Image-2 as a large SOTA foundation model and Flux.2 Klein as a local alternative. For both models we use a preamble requesting consistency of geometry and camera perspective and a catalog of light conditions (Fig. 3). For the Flux.2 Klein variant, a high classifier-free guidance (cfg) value of 5 and the use of negative prompts significantly improves the dataset quality.

5. Evaluation

5.1. Results

In Fig. 4 results of the full pipeline are presented. Further visualizations are displayed in the supplementary material. Our method produces plausible intrinsic maps from datasets generated with both foundation models. When rendering the scene with novel light, our method produces more natural images in comparison to direct diffusion-based methods such as Marigold and RGB→X. We attribute this to the fact that datasets for purely diffusion-based intrinsic decomposition models are restricted to training on synthetic scenes, so real photos are out of distribution, while we can leverage prior knowledge on natural lighting encoded by large foundation models that trained on real world images as well.

5.2. Limitations and Discussion

Although our pipeline enables promising relighting results, it contains several approximations. We validated the estimation of normals and observed the decomposition to be robust against deviations from the assumed FOV and from the lighting model (e.g. by sharper light sources / finite light source distances, occlusions, mirrors), however, large model gaps and poorly conditioned parameter combinations will still result in poor local minima. Strong priors help to avoid baking shading information into the albedo map, but our segmentation prior again relies on the performance of the SAM model.

6. Conclusion

We demonstrated the feasibility of estimating image intrinsics by decomposing a set of multi-illumination variants generated by large image generation models conditioned on a single input image. An effective regularization strategy proved to be key as soon as the forward model doesn't match the data exactly (Fig. 2).

7. Future Work

In the absence of ground truth intrinsic data, we have to rely on qualitative feedback. To benchmark the performance of our method, a user study could be performed. Although our method does not predict the intrinsic maps in a single forward pass, our method relies on generated maps for normals, segments and relights. To automatically handle outliers, an agentic approach that reviews generated images and requests regenerations where necessary could be implemented. Finally the improvement of downstream applications such as relighting 3D Gaussian Splats can be explored.

References

[1] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A Dataset of Multi-Illumination Images in the Wild, October 2019. URL http://arxiv.org/abs/1910.08131. arXiv:1910.08131 [cs]. 2, 3

[2] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. 1

[3] Berthold KP Horn. Determining lightness from an image. *Computer graphics and image processing*, 3(4):277–299, 1974. 1

[4] Graham D Finlayson, Mark S Drew, and Cheng Lu. Intrinsic images by entropy minimization. In *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004, Proceedings, Part III*, pages 582–595. Springer, 2004. 2

[5] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. 2

[6] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, 2022. 2

[7] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding, August 2021. URL http://arxiv.org/abs/2011.02523. arXiv:2011.02523 [cs]. 2

[8] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2472–2481. Computer Vision Foundation / IEEE, 2020. 2

[9] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic Image Diffusion for Indoor Single-view Material Estimation, March 2024. URL http://arxiv.org/abs/2312.12274. arXiv:2312.12274 [cs]. 2

[10] Chris Careaga and Yağız Aksoy. Colorful Diffuse Intrinsic Image Decomposition in the Wild. *ACM Transactions on Graphics*, 43(6):1–12, December 2024. ISSN 0730-0301, 1557-7368. doi: 10.1145/3687984. URL https://dl.acm.org/doi/10.1145/3687984. 2

[11] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. RGBX: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, Denver CO USA, July 2024. ACM. ISBN 979-8-4007-0525-0. doi: 10.1145/3641519.3657445. URL https://dl.acm.org/doi/10.1145/3641519.3657445. 2

[12] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation, April 2024. URL http://arxiv.org/abs/2312.02145. arXiv:2312.02145 [cs]. 2, 3

[13] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis, May 2025. URL http://arxiv.org/abs/2505.09358. arXiv:2505.09358 [cs.CV]. 2

[14] Black Forest Labs. FLUX.2-klein-4B. https://huggingface.co/black-forest-labs/FLUX.2-klein-4B, 2026. 2

[15] OpenAI. GPT Image 2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026. 2

[16] Robert L. Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics*, 1(1):7–24, 1982. 2

[17] T. S. Trowbridge and K. P. Reitz. Average irregularity representation of a rough surface for ray reflection. *Journal of the Optical Society of America*, 65(5):531–536, 1975. 2

[18] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the Eurographics Symposium on Rendering (EGSR)*, pages 195–206, 2007. 2

[19] B. Smith. Geometrical shadowing of a random rough surface. *IEEE Transactions on Antennas and Propagation*, 15 (5):668–671, 1967. 2

[20] Brian Karis. Real shading in unreal engine 4. In *ACM SIGGRAPH 2013 Course Notes: Physically Based Shading in Theory and Practice*, 2013. 2

[21] Christophe Schlick. An inexpensive brdf model for physically-based rendering. *Computer Graphics Forum*, 13 (3):233–246, 1994. 2

[22] Brent Burley. Physically based shading at disney. In *ACM SIGGRAPH 2012 Course Notes: Practical Physically Based Shading in Film and Game Production*, 2012. 2

[23] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of ACM SIGGRAPH 2001*, pages 497–500, 2001. 2

[24] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of ACM SIGGRAPH 2001*, pages 117–128, 2001. 2

[25] Peter-Pike Sloan, Jan Kautz, and John Snyder. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 21(3):527–536, 2002. 2

[26] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[27] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang Cao Li, Zengqi Xun, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics, May 2021. URL http://arxiv.org/abs/2011.09127. arXiv:2011.09127 [cs.CV]. 3

Estimating Physically-Based Material Representations from Diffusion-Generated Lighting Variations

Supplementary Material

8. Additional Result Visualizations

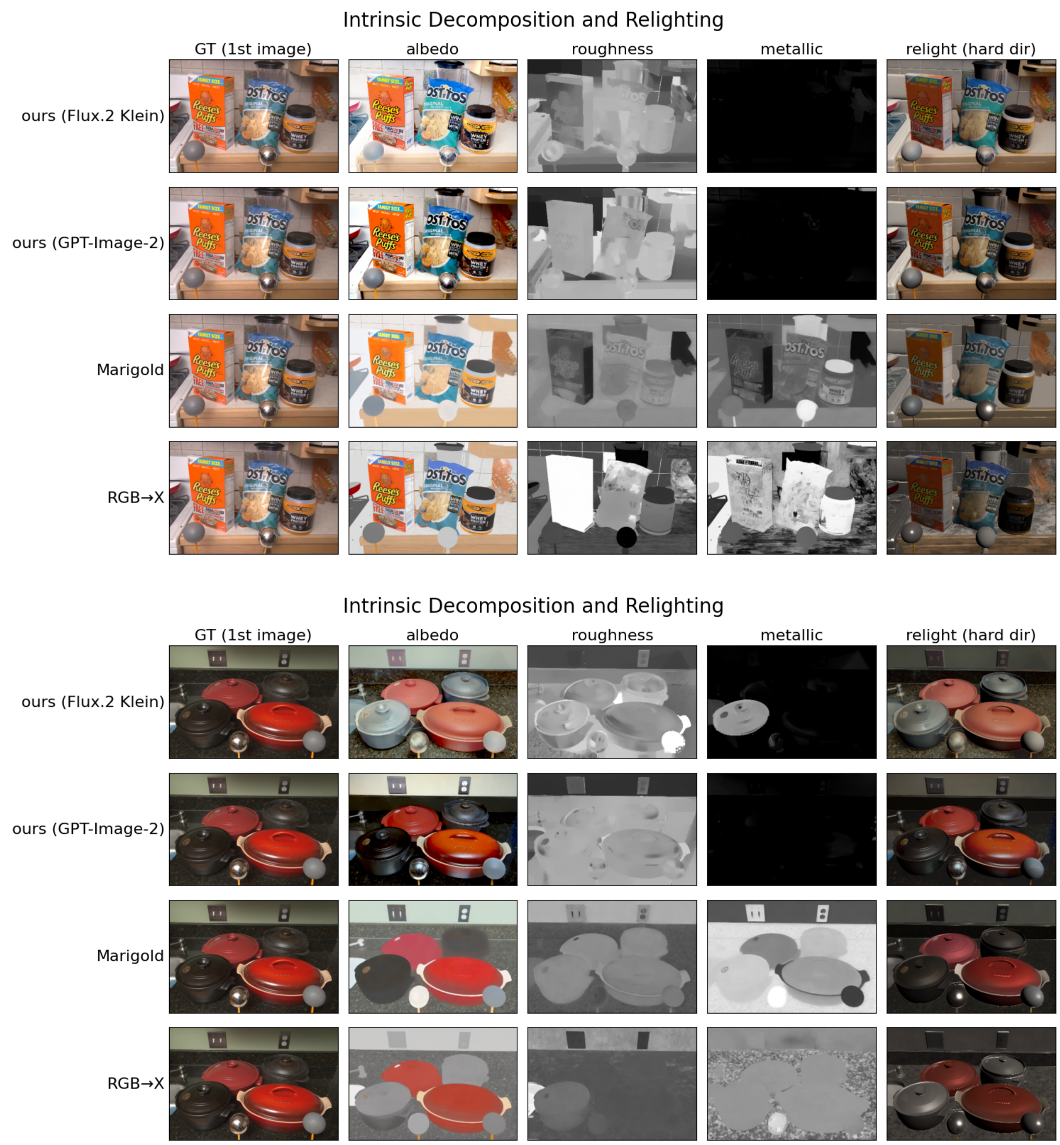

Figure 6. Exemplary decomposition of the `everett_kitchen_12` and `kingston_kitchen_18` scenes using generated datasets and comparing to reference methods. The relight renderings were produced using our CT renderer and an environment map encoding hard directional light.

9. Rendering and Lighting Model Details

9.1. BRDF terms

The diffuse and specular terms in Equation 1 are

$$f_d = \frac{(1-F)(1-m)\,\rho}{\pi}, \qquad f_s = \frac{D(\mathbf{h})\,G(\mathbf{n},\mathbf{v},\mathbf{l})\,F(\mathbf{v},\mathbf{h})}{4\,(\mathbf{n}\cdot\mathbf{v})(\mathbf{n}\cdot\mathbf{l})}, \tag{3}$$

where $\mathbf{h} = (\mathbf{v}+\mathbf{l})/\|\mathbf{v}+\mathbf{l}\|$.
The GGX normal distribution is

$$D(\mathbf{h}) = \frac{\alpha^2}{\pi\big((\mathbf{n}\cdot\mathbf{h})^2(\alpha^2-1)+1\big)^2}, \qquad \alpha = r^2, \tag{4}$$

the Smith masking term is $G = G_1(\mathbf{n}\cdot\mathbf{v})\,G_1(\mathbf{n}\cdot\mathbf{l})$ with $G_1(x) = x/(x(1-k)+k)$ and the Karis IBL remapping $k = \alpha^2/2$ [20], and the Schlick Fresnel term is

$$F(\mathbf{v},\mathbf{h}) = F_0 + (1-F_0)(1-\mathbf{v}\cdot\mathbf{h})^5, \qquad F_0 = (1-m)\,0.04 + m\,\rho, \tag{5}$$

so dielectric surfaces ($m = 0$) reflect a fixed 4% at normal incidence and metals ($m = 1$) tint the specular response by their own albedo.

9.2. Diffuse SH projection

Projecting f_d against the SH-truncated L_i gives

$$\int_\Omega f_d\, L_i(\mathbf{l})\,(\mathbf{n}\cdot\mathbf{l})\,d\mathbf{l} = \frac{(1-F)(1-m)\,\rho}{\pi}\sum_{\ell,m} A_\ell\, L_{\ell m}\, Y_{\ell m}(\mathbf{n}), \tag{6}$$

the classic irradiance projection of [23]. The transfer weights A_ℓ follow from convolving the clamped-cosine kernel with each band and vanish for all odd $\ell \geq 3$:

$$A_0 = \pi, \quad A_1 = \frac{2\pi}{3}, \quad A_2 = \frac{\pi}{4}, \quad A_3 = 0. \tag{7}$$

This is why, at our $\ell \leq 3$ truncation, the third band is a pure specular degree of freedom relative to the diffuse response.

9.3. Specular SH convolution

Writing $e = \alpha^2$ and $x = 1-e$, the zonal-harmonic band weights $h_\ell(r)$, $\ell = 0,\ldots,3$, of the GGX distribution admit the closed forms

$$h_0 = \frac{4}{1+e}, \tag{8}$$

$$h_1 = \frac{8e\ln\frac{2e}{1+e} + 4x}{x^2}, \tag{9}$$

$$h_2 = \frac{12ex + 4e(12-6x)\ln\frac{2e}{1+e}}{x^3} + \frac{4(6-6x+x^2)}{(1+e)x^2}, \tag{10}$$

$$h_3 = \frac{2(-17e^3 - 24e^2 + 39e + 2) + 48e(e^2+3e+1)\ln\frac{2e}{1+e}}{x^4}. \tag{11}$$

h_1, h_2, and h_3 each cancel catastrophically as $x \to 0$ (i.e. $r \to 1$); we switch to a Taylor expansion in x for $x < 10^{-3}$ for these three bands (e.g. $h_3 \approx -\frac{1}{4} + \frac{x^2}{8} + \frac{51x^3}{280}$), and evaluate both branches in float64. $h_0 = 4/(1+e)$ needs no such branch. Float32 is insufficient in either branch and destabilizes gradients at very low or very high roughness. With $\mathbf{r} = 2(\mathbf{n}\cdot\mathbf{v})\mathbf{n} - \mathbf{v}$, the specular integral is then

$$\int_\Omega f_s\, L_i(\mathbf{l})\,(\mathbf{n}\cdot\mathbf{l})\,d\mathbf{l} = \frac{F\,G_1(\mathbf{n}\cdot\mathbf{v})}{4}\sum_{\ell,m} h_\ell(r)\, L_{\ell m}\, Y_{\ell m}(\mathbf{r}), \tag{12}$$

following the standard real-time approximation of evaluating F and G_1 at the view angle and treating the GGX lobe as zonal about $\mathbf{r}$ [20, 24, 25], which is exact for light directions coplanar with $\mathbf{n}$ and $\mathbf{v}$ and a close approximation otherwise.