Inverse Rendering: A Survey

Steven Gao
University of Illinois Urbana Champaign
hongyig3@illinois.edu

Asher Mai
University of Illinois Urbana Champaign
hanlinm2@illinois.edu

Abstract

Rendering is a process in computer graphics where artists define scene geometry, materials, lighting and a virtual camera's parameters which enable modern software to generate a photo-realistic image. The ability to reverse the process of rendering (i.e. from images to geometry, materials, lighting, etc.) is called Inverse Rendering. This technique gives creative professionals a tool to edit images in whole new ways, allowing them to perform object insertion, light editing, and materials editing with applications in augmented and virtual reality. Inverse Rendering is a large body of problems with algorithms tailored to various input representations and scene properties to estimate, some being optimization based and others utilizing learned priors. Neural Inverse Rendering has seen an explosion of growth in recent years with the advent of deep learning, radiance fields, and increased computing capabilities. In this survey, we dive into their variety of techniques and capabilities. Additionally, we evaluate current methods, trends and explore future research ideas.

1. Introduction

Inverse Rendering, sometimes also referred to as intrinsic image decomposition, aims to estimate the physical properties of a scene that is used to render a given image. These inverse rendering techniques typically take one or multiple images as input. Some methods also use scanned geometry, videos or user guidance to further improve and refine the quality of results. The estimated physical properties include materials, albedo, roughness, lighting, normal, geometry, etc.

Traditionally, this has been a highly ill-posed problem, requiring extensive assumptions and handcrafted models. Neural Inverse Rendering gained traction in recent years with increased computing resources and innovations on deep learning, which gave rise to techniques that benefit from the strong learned prior of deep learning models trained on large-scale real world datasets.

In this survey, we conduct extensive review of this topic.

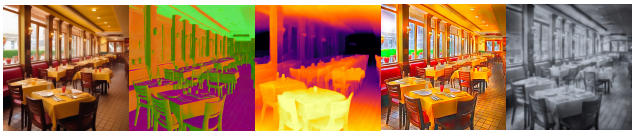

Figure 1. An example of inverse rendering and intrinsic image decomposition [5]. Given a single input image, geometric (normal, depth), materials (albedo) and lighting (shading) properties are estimated.

Inverse Rendering approaches can be broadly categorized by its conditional input, including: geometry, single view, multi-view, and 3D Geometry, and we give brief review of these methodologies. We also explore the capabilities including materials editing, light editing, novel view synthesis and object insertion. We summarize all papers reviewed in Tab. 1

1.1. Motivation of Inverse Rendering

The goal of inverse rendering is to decompose an image into its intrinsic physical properties that allow the image to be edited as desired by the user and then re-rendered into a new image that remains photo-realistic. Materials editing, re-lighting, and object insertion are among to most popular capabilities of recent inverse rendering techniques. Sec. 3 discusses some of these capabilities of inverse rendering techniques in detail. Applications include gaming, augmented reality (AR), virtual reality (VR), robotics simulation, and film production.

1.2. Background

In this section, we review preliminary knowledge on inverse rendering. Sec. 1.2.1 discusses lighting, materials and geometry that are used for rendering and are the outputs of most inverse rendering pipelines. Sec. 1.2.2 discusses various scene representations that inverse rendering techniques will use as their conditional input to their pipeline.

1.2.1 Scene Properties of Interest

We discuss the different scene properties that various inverse rendering techniques will output as part of pipeline and in many cases help to enable further editing of the scene.

Lighting. Illumination plays a crucial role in computer graphics and rendering. Modern graphics rendering engines are capable of utilizing a 360° image known as a High Dynamic Range Image (HDRI) to illuminate the rendered scene. Filmmakers who use visual effects to insert computer generated objects into a scene will typically use a mirror ball known as a light probe to capture the environmental lighting and generate an HDRI that matches the scene, so that the inserted objects look realistic with proper reflections and shadows. However, this method requires physical presence in same location with the same lighting condition as the captured scene, which is not always possible. Inverse lighting problems aim to estimate emissivities and positions of light sources, either jointly or given the other.

Materials. The physical properties of objects in a scene dictate the intensity, direction and color of light reflected from the objects. Representations of material properties include reflectance (albedo), shading (irradiance), roughness, and metallic. These can be stored as 2D images with per-pixel mappings. Roughness and metallic maps in particular can be stored as bidirectional reflectance distribution function (BRDF) maps.

Geometry. The shape of an object is crucial to determining the direction and intensity of light that enters the camera as well as the shape of the shadows and inter-reflections. In screen space, normal maps are used to represent per-pixel surface directions while depth maps show per-pixel distance away from the camera; 3D representations of geometry include triangular mesh, point cloud, and signed distance fields (SDF), etc.

1.2.2 Input Representations

Various techniques utilize different scene representations as conditional inputs to their pipeline.

Single Image. Estimating scene properties from a single image [5, 7, 9, 12, 14–16, 22, 26, 34, 36] is challenging and seem almost impossible due to the infinite number of illumination, geometry and materials that could produce the same single image. However, there are some explanations that are more likely than others. Many techniques use data-driver approach to produce the most likely explanations.

Multi-view images. As opposed to single image, inverse rendering using multi-view images [1, 4, 19, 20, 30, 32, 33, 35] provides a variety of information about the scene, which reduces the number of possible scene property explanations that produce the scene. One key challenge is the lack of large scale dataset available on multi-view HDR synthetic dataset. The handling of multi-view images in efficient pipeline designs is also less trivial and requires careful design of model architecture. However, they tend to produce better results than single images simply due to the amount of data available.

Scanned Geometry. Recent popularity on inexpensive mobile LiDAR and RGB-D sensors gave rise to advances in 3D geometry reconstruction. Conditioning on reconstructed 3D geometry [1, 19] allow differentiable renderers to jointly optimize on materials and lighting, providing a more accurate estimation of the scene properties.

2. Taxonomy

2.1. Datasets

Real Image Datasets It is very difficult to collect datasets of intrinsic images (albedo, normal, shading etc.) on real images due to the amount of information needed to accurately annotate each pixel. Therefore, existing datasets are often sparse rather than per-pixel, and often require crowdsourcing. Two notable datasets in this space include the Intrinsic Images in the Wild (IIW) and Shading Annotations in the Wild (SAW) datasets [3, 10], where reflectance and shading annotations are collected on real indoor images. However, they are hard to train dense networks due to their sparse annotations derived from only a few thousand samples.

Synthetic Datasets Computer generated synthetic datasets benefit from the ability to easily acquire per-pixel ground truth intrinsic image annotations at a much larger scale without crowd-sourced human annotations. SUNCG dataset [23] contains a large, diverse set of indoor scenes with complex geometry, containing 45,622 scenes with over 5M instances of 2644 unique objects in 84 object categories. Further modifications of this dataset has been seen in follow-up works such as the CGIntrinsics dataset [12] that provides ground truth reflectance and ground truth shading. Li et al. further improves the SUNCG dataset by photo-realistically maping microfacet materials to SUNCG geometries

2.2. Image-Space Estimation with Learned Priors

Recent innovations in deep neural networks and ever increasing GPU compute power enabled intrinsic image decomposition to benefit from strong learned prior of deep learning models trained on large-scale datasets. With data-driver approach, deep learning models are able to learn from a variety of different lighting conditions, materials, and geometry, allowing them to retain knowledge about what is the most likely solution. In this section, we discuss the various deep learning architectures and models that enable inverse rendering techniques to produce accurate estimation of scene properties.

2.2.1 Convolution Neural Network

Convolution Neural Network (CNN) has seen an explosion of popularity in recent years due to their success in solving tasks including image classification, object detection, and semantic segmentation. [22] and [24] use CNN to estimate HDR environmental light maps for object insertion tasks. Li et al. estimates lighting and materials using a CNN-based Cascade Network that progressively increases resolution and iteratively refines the predictions through global reasoning. Philip et al. uses a multi-scale convolutional network to output diffuse and specular image.

Many techniques also use the CNN-based U-Net architectures due to its ability to generalize to multiple scales and its image-to-image dense prediction capabilities. [12] and [13] use modified versions of U-Net to estimate albedo and shading given a single image. [14] and [26] use U-Net based architectures to produce depth and normal maps in addition to albedo and shading to achieve light editing and object insertion tasks respectively.

2.2.2 Diffusion Models

Diffusion Models have gained traction in recent years thanks to its capabilities to generate high-quality realistic images based on different conditioning such as texts, bounding boxes, images, etc. These diffusion models are trained on large scale real world data, often consisting of image and text pairs, and retain strong learned prior from a variety of different scenes. Kocsis et al. exploit the learned prior of recent diffusion models and transfer it to material estimation task. Du et al. use light weight Low-Rank Adaptors (LoRA) [6] to finetune diffusion models to produce normals, depth, albedo, and shading of a given input image.

2.2.3 Transformer Models

Vision Transformer (ViT) models benefit from its spacial attention layers that give them global reasoning capabilities as opposed to CNN whose receptive fields remain largely local throughout the consecutive layers, limiting the ability to capture long-range interactions that are crucial for light estimation. IRISformer [36] use 4 transformer layers with both encoders and decoders to estimate spacially-varying lighting. While not exactly using transformer models, MAIR [4] uses a multi-view attention network (MVANet) to aggregate multi-view images and estimate albedo, roughness, and normal that allows them to perform object insertion with high accuracy on reflected light of specular objects.

2.3. Differentiable Rendering and Radiance Fields

Scene representations are not limited to explicit, descriptive modeling of geometry, lighting, and reflectance, as they could also be approximately implicitly with neural methods as simple as multi-layer perceptrons (MLP). As long as the representation is differentiable, the primary optimization goal could be formulated in image-space to fit an implicit representation to multi-view images of a given scene. Shown in Figure 2, Mildenhall et al. proposed Neural Radiance Fields (NeRF) that models the scene with the plenoptic function [11], a continuous function over the product of 3D locations in the scene and viewing angle that outputs color and density. To render this representation, we accumulate this function and perform alpha blending at discrete intervals along a ray. In the original NeRF paper, the representation is parameterized by a MLP [18], but subsequent works have shown that voxel and point-based volumetric representations could also achieve similar or better results [8, 28, 29]. The radiance field is overfit for a specifc scene in training to minimize the previously described rendering loss. Many works in inverse rendering also capitalize on implicit and differentiable representations due to their flexibility in modeling, some inspired by the NeRF [18] and 3D Gaussian Splatting [8]. We broadly categorize these works utilizing differentiable rendering into two categories, optimization and learned approaches. Yet, these methods differ from the aforementioned image-space learned approaches in Section 2.2 as they aim to label scene properties in 3D space.

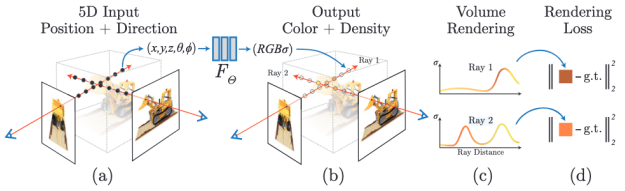

Figure 2. NeRF [18] models the scene with the plenotic function [11], which takes in position and viewing direction (x, y, z, θ, ϕ) and outputs color and density $(RGB\alpha)$. The loss is specified as the difference between the rendered and ground truth in image space (d). Radiance fields are the basis for many of the SOTA inverse rendering algorithms.

2.3.1 Optimization-Based Inverse Differentiable Rendering

Given coarse reconstructed geometry and object segmentation, Inverse Path Tracing estimates the parameters of a physically-based light transport function by performing path tracing on multi-view images, selecting different BRDFs for different objects [1]. Monte-Carlo sampling is used for the BRDFs to generate more rays in the path tracing process. Since this process is stochastic, the rendered image will be noisy; however, in optimization, this is equivalent to stochastic gradient descent when back propagating.

Due to the amount of computation involved, Inverse Path Tracing could only handle very simple geometry and relatively consistent materials properties for the same object.

To overcome these limitations, PhysSG models the scene geometry using signed distance functions (SDFs) approximated by MLPs, the reflectance using a monochrome isotropic BRDF, and the illumination using a environment map [30]. To generate a closed form solution and avoid Monte-Carlo sampling in the forward pass, the authors used a mixture of 128 spherical Gaussians (SG) [27] to approximate the environment map. The rendering process of PhysSG is differentiable, and all parameters are jointly optimized using multi-view images [30]. The model makes several assumption for simplicity in modeling and reduced computation, namely specular isotropic BRDFs and direct illumination.

Both Inverse Path Tracing [1] and PhysSG [30] model indirect illumination with path tracing or an environment map, which are either costly or oversimplified. Zhang et al. proposed capturing indirect illumination directly from the plenoptic function in a radiance field, first training a regular radiance field to capture geometry and lighting and optimizing a spatially varying BRDF (SVBRDF) on top of it [33]. The outgoing radiance is already encoded in the radiance field, resulting in reduced computation.

Modeling indirect illumination with the outgoing radiance is still an approximation, as it's not guaranteed that the ground-truth has enough guidance for rays that originate from inside a surface. I^2-SDF improves the physical accuracy by modeling neural SDF, radiance material, and emission fields separately and performing path tracing on top [35].

3D Gaussian Splatting [8] generally produces SOTA quality reconstructions with the same compute. However, it does not produce reliable normals or support for occlusion in indirect lighting, which GS-IR [17] addresses by concentrating depth gradients during optimization and baking-based approach to recover an occlusion cubemap for reflection. Along this direction, GIR [21] bakes the radiance field into a voxel grid for indirect lighting.

2.3.2 Radiance Fields with Pretrained Priors

As an under-contrained task, inverse rendering tasks would benefit from learned priors for works that build upon differentiable rendering and radiance fields as well. Li et al. propose using multiple encoder-decoder networks in a cascading fashion to sequentially predict scene properties and borrow the idea of differentiable rendering to back propagate through all of the networks. [16, 26] also use differentiable rerendering to further optimize the neural predictors. Zhu et al. use screen-space raytracing based on the G-buffer, which, along with SVBRDF and other properties, are generated by neural predictors. Similar to [33], NeRFactor jointly estimates geometry, SVBRDF, and indirect illumination using a radiance field but pretrains a BDRF encoder to provide better guidance for what BRDFs are empirically observed [32]. GIR [21] also produces an environment map using a learned component.

3. Current Capabilities and Evaluation

In this section, we discuss the capabilities of current techniques to make image edits after intrinsic image decomposition. These edits include material editing, light editing (re-lighting), and object insertion. We also provide discuss evaluations of these tasks through existing benchmarks.

3.1. Material Editing

Material editing allows creative professionals to change the color, roughness, and reflectance (albedo) of the scene. For example, an interior designer could change the color of the wall, floor and furniture of a given scene and decide what works best. In our literature review, we observe that the following papers are capable of material editing: [9, 15, 16, 16, 30, 32, 34, 36].

3.2. Light Editing

Light Editing, also known as re-lighting, is a way to change direction, intensity and light source while keeping the shadows and reflections in the scene consistent. Kocsis et al. demonstrated its ability to change the color of lamps in a bedroom while Nimier-David et al. showed accurate relighting of an indoor scene throughout different times of day, with the sun shining through the window with different angles and intensity. In addition, [14, 16, 20, 30, 32, 33, 35] are all capable of editing lights within a scene.

3.3. Object Insertion

Realistically insert objects into a scene has many applications, such as in augmented reality and robotic simulations. One of the earliest work on object insertion is by Karsch et al.. They showed realistic insertion of objects with small amounts of user interaction and showed its ability to confuse human perceptions in a user study. Srinivasan et al. use CNN based architecture to generate a spherical environment map and demonstrated its ability to insert highly specular virtual objects into real images with accurate reflections. Additionally, [4, 15, 22, 26, 34, 36] are all capable of object insertion tasks.

3.4. Metrics

Evaluation of capabilities such as material editing, light editing and object insertion are often subjective and therefore require conducting user studies. For quantitative analysis, many papers will use Peak Signal-to-Noise Ratio

(PSNR), Structural Similarity Index Measure (SSIM) [25], and Learned Perceptual Image Patch Similarity (LPIPS) [31] to evaluate image quality.

4. Future Directions

Current methods are only capable of inversely rendering of the scene that is in the view. However, there is a lack of literature that is able to estimate the geometry, lighting, and materials of objects that are out of view. With increased popularity of Diffusion Models that are capable of out-painting, future work can focus on expanding the scene in creative ways conditioned on text, and even generating 360 degree scene properties, enabling rendering for VR applications.

Fundamentally an under-constrained problem, inverse rendering would also benefit from more rigorous priors, with improvements either manifesting in improved quality or reduced compute. As foundational models grow larger by the power scaling law, they could provide stronger guidance for scene properties especially for single-image and few-shot cases where ray tracing or path tracing lack sufficient input data to generate SOTA results.

References

[1] Dejan Azinović, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse Path Tracing for Joint Material and Lighting Estimation, 2019. arXiv:1903.07145 [cs]. 2, 3, 4

[2] Jonathan T. Barron and Jitendra Malik. Shape, Illumination, and Reflectance from Shading, 2020. arXiv:2010.03592 [cs]. 2

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4), 2014. 2

[4] JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, and Junghyun Cho. MAIR: Multi-view Attention Inverse Rendering with 3D Spatially-Varying Lighting Estimation, 2023. 2, 3, 4

[5] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Intrinsic lora: A generalist approach for discovering knowledge in generative models. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024. 1, 2, 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[7] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering Synthetic Objects into Legacy Photographs, 2019. arXiv:1912.11565 [cs]. 2, 4

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 3, 4

[9] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic Image Diffusion for Indoor Single-view Material Estimation, 2024. arXiv:2312.12274 [cs]. 2, 3, 4

[10] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6998–7007, 2017. 2

[11] Michael Landy and J. Anthony Movshon. The Plenoptic Function and the Elements of Early Vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991. Conference Name: Computational Models of Visual Processing. 3

[12] Zhengqi Li and Noah Snavely. CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering, 2018. 2, 3

[13] Zhengqi Li and Noah Snavely. Learning Intrinsic Image Decomposition from Watching the World, 2018. arXiv:1804.00582 [cs]. 3, 2

[14] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans. Graph.*, 37(6):269:1–269:11, 2018. 2, 3, 4

[15] 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, 2019. 2, 3, 4

[16] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-Based Editing of Indoor Scene Lighting from a Single Image, 2022. arXiv:2205.09343 [cs]. 2, 4

[17] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. GS-IR: 3D Gaussian Splatting for Inverse Rendering, 2023. 4

[18] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, 2020. arXiv:2003.08934 [cs]. 3

[19] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, Anton Kaplanyan, Adrien Bousseau, and Morgan McGuire. Material and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering. In *EGSR (DL)*, pages 73–84, 2021. 2, 4

[20] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint Indoor Neural Relighting from Multi-view Stereo, 2021. arXiv:2106.13299 [cs]. 2, 3, 4

[21] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jian Zhang, Bin Zhou, Errui Ding, and Jingdong Wang. GIR: 3D Gaussian Inverse Rendering for Relightable Scene Factorization, 2023. 4

[22] Gowri Somanath and Daniel Kurz. HDR Environment Map Estimation for Real-Time Augmented Reality, 2021. arXiv:2011.10687 [cs]. 2, 3, 4

[23] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2

[24] Pratul P. Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T. Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting Lighting Volumes for Spatially-Coherent Illumination, 2020. arXiv:2003.08367 [cs]. 3, 4, 2

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

[26] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning Indoor Inverse Rendering with 3D Spatially-Varying Lighting, 2021. arXiv:2109.06061 [cs]. 2, 3, 4

[27] Ling-Qi Yan, Yahan Zhou, Kun Xu, and Rui Wang. Accurate Translucent Material Rendering under Spherical Gaussian Lights. *Comput. Graphics Forum*, 31(7):2267–2276, 2012. 4

[28] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. *arXiv*, 2021. 3

[29] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for Real-time Rendering of Neural Radiance Fields. *arXiv*, 2021. 3

[30] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting, 2021. arXiv:2104.00674 [cs]. 2, 4

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[32] Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. *ACM Transactions on Graphics*, 40(6):1–18, 2021. arXiv:2106.01970 [cs]. 2, 4

[33] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling Indirect Illumination for Inverse Rendering, 2022. 2, 4

[34] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Jiaxiang Zheng, Rui Tang, Hujun Bao, and Rui Wang. Learning-based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing, 2022. arXiv:2211.03017 [cs]. 2, 4

[35] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, and Rui Wang. I$ˆ2$-SDF: Intrinsic Indoor Scene Reconstruction and Editing via Raytracing in Neural SDFs, 2023. arXiv:2303.07634 [cs]. 2, 4

[36] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes, 2022. arXiv:2206.08423 [cs]. 2, 3, 4

Inverse Rendering: A Survey

Supplementary Material

A. Comparison of Methods

Paper	Input	Output	Method	Capability
Karsch et al.	Single Image, User guidance	Albedo, direct reflected light	Optimization	Object Insertion
Barron and Malik	Sngle Image	Shape, Normal, Lighting, Reflectance, Shading	Optimization	N/A
Kocsis et al.	Single Image	Albedo BRDF (Roughness, Metallic)	Diffusion Model	Material editing, Light Editing
Zhang et al.	Multi-View Images	Geometry (SDF), Meterials (BRDF), Environment Map (HDRI)	Differentiable Renderer	Material editing, Light Editing, Novel view synthesis
Zhang et al.	Multi-View Images and camera poses	Normals, Albedo, BRDF, Light visibility	NerF	Material Editing, Light Editing (with shadow)
Li et al.	Single Image	Albedo, Normal, Roughness	Neural Differential Renderer	Material Editing, Light Editing
Du et al.	Single Image	Normals, Depth, Albedo, Shading	Diffusion Model	N/A
Azinović et al.	Multi-View and 3D geometry	Albedo, Emission, Roughness	Differentiable inverse Monte Carlo renderer	N/A
Choi et al.	Multi-View Imags	Albedo, roughness and normal.	Multi-View Attention	Object Insertion
Philip et al.	Multi-View Images	Albedo	CNN	Light Editing, Novel view synthesis
Wang et al.	Single Image	albedo, normals, depth and lighting	Resnet and U-Net	Object Insertion
Zhu et al.	Multi-view images	shapes, incident radiance and materials	NerF, Differentiable Monte Carlo raytracing	Material editing, Light Editing
Zhu et al.	Single Image	albedo, normal, depth, metallic, and roughness	Differentiable Monte Carlo raytracing	Object Insertion, Material Editing
Zhu et al.	Single Image	depths, normals, albedo, roughness, lighting	Transformer	Object insertion and material editing
Li and Snavely	Single Image	albedo, shading	U-Net	N/A
Li and Snavely	Single Image	albedo, shading	U-Net	N/A
Li et al.	Single Image	depth, normals, albedo, roughness	U-Net	Light Editing
Li et al.	Single Image	albedo, normals, roughness, depth, lighting	CNN	Object insertion and material editing
Zhang et al.	Multi-view Images	albedo and roughness	Differentiable Renderer	Novel View Synthesis, Light Editing
Srinivasan et al.	stereo image	spherical environment (lighting) map	CNN	Object Insertion
Somanath and Kurz	Single Image (LDR env map)	HDR env map	CNN	Object Insertion
Nimier-David et al.	Multi-view Images and Scanned Geometry	Material, lighting	Differentiable Renderer	Light Editing

Table 1. Summary of input, output, method and capabilities of papers reviewed in this survey