Locally Orderless Images for Optimization in Differentiable Rendering

Ishit Mehta Manmohan Chandraker Ravi Ramamoorthi

University of California San Diego

Abstract

Problems in differentiable rendering often involve optimizing scene parameters that cause motion in image space. The gradients for such parameters tend to be sparse, leading to poor convergence. While existing methods address this sparsity through proxy gradients such as topological derivatives or lagrangian derivatives, they make simplifying assumptions about rendering. Multi-resolution image pyramids offer an alternative approach but prove unreliable in practice. We introduce a method that uses locally orderless images — where each pixel maps to a histogram of intensities that preserves local variations in appearance. Using an inverse rendering objective that minimizes histogram distance, our method extends support for sparsely defined image gradients and recovers optimal parameters. We validate our method on various inverse problems using both synthetic and real data.

1. Introduction

Much of the recent work addresses inverse rendering with analysis-by-synthesis — start with an initial guess of scene parameters, render an image using a differentiable renderer [12, 16, 18, 21, 30], compare it with a given photograph, estimate gradients, and update the parameters iteratively. While the success of gradient-based optimization in machine learning validates this approach, significant challenges remain. Despite having algorithms that accurately estimate gradients for the physics of image formation, optimization hurdles such as local minima, noisy loss landscapes, and the search for good initialization and parameterization all still persist [23].

For many inverse problems, the ideal optimization trajectory requires long-range motion of pixels and image features. Consider the example shown in Fig. 1, where a shiny ball is lit by an unknown light source that we wish to recover from a given image. A primary effect of optimizing the light position here is on the motion of the specular highlight on the ball. Since image gradients are sparse (Fig. 1 ⑥) for such parameters, the optimization landscape is rife with local minima and plateau regions [7, 22]. Existing methods address the gradient-sparsity problem with proxy gradients such as topological derivatives [22], lagrangian derivatives [44], and variational derivatives [7] — all of which are either expensive to compute [7, 44], only work with implicit geometry [22], or are restricted to primary light-transport effects [22, 44]. We propose a complementary approach that is compatible with RGB gradients from a differentiable renderer [12, 16, 19] — which we demonstrate to be expressive for a variety of inverse problems that have multiple locally-optimal solutions (as in Fig. 1).

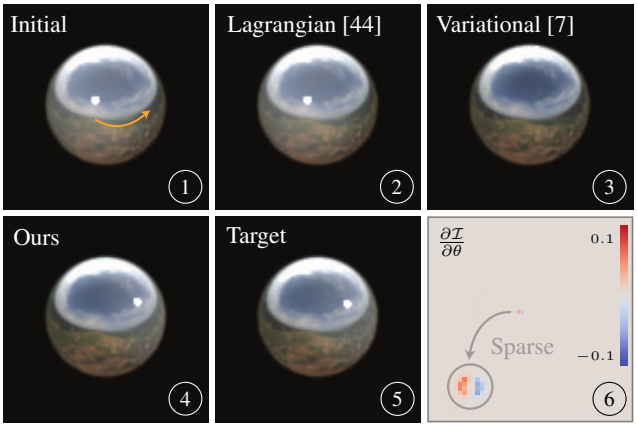


Figure 1. Image gradients are sparse with respect to optimization parameters that induce motion in the image space ⑥. We show an inverse problem with the goal of recovering the position (θ) of a distant light source from a synthetic image of a shiny ball, *i.e.* ① to ⑤. Existing methods compute proxy gradients such as: Lagrangian derivatives [44], which track only primary-ray intersections, or variational derivatives [7], which can be prone to local minima. Our method uses standard RGB gradients and uses an inverse rendering objective that matches locally orderless images.

Our method builds on scale-space matching [42], where signals are matched at multiple resolutions to measure similarity. In differentiable rendering, multi-scale matching has received limited attention, with a few works [7, 44] noting its unreliability. In this work, we observe that this unreliability

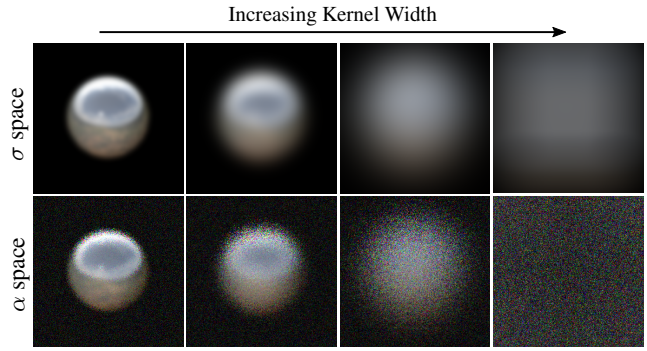


Figure 2. **Inner and Extent scale spaces.** The image representation $\mathcal{P}(\mathbf{x}, k, \alpha, \beta, \sigma)$, is composed of three distinct histogram-valued scale spaces. The σ-space (*top*) controls the effective resolution of the image and the α-space (*bottom*) defines the spatial extent of histogram integration. The rendered images shown here are intensities sampled as $\mathcal{I}(\mathbf{x}) = k, k \sim \mathcal{P}(\mathbf{x}, k, \theta, \alpha, \beta, \sigma)$ for given kernel parameters α and σ, and bin width β. We recover inverse rendering parameters θ by matching these locally orderless structures for rendered and reference images.

stems from the standard approach of using multi-resolution images or Gaussian Pyramids [1], which average out local appearance and geometric details. Instead, we repurpose the inverse rendering objective to match local histograms that preserve the *full* distribution of intensity values within a neighborhood. Originally proposed by Griffin [11] to address imprecision in radiance measurements, and later extended by Koenderink and van Doorn [15] for reasoning about image topology — we postulate that the idea of viewing images as a family of histograms might be relevant in the seemingly unrelated field of inverse rendering.

Formally, we render images in three distinct scale spaces [15]: 1) the inner (σ) scale, describing the effective resolution of the image (Fig. 2 *top*), 2) the tonal (β) scale, which quantifies the imprecision in measured radiance at a location ($\mathbf{x}$), and 3) the extent (α) scale, which relates to the spatial region over which a histogram is computed (Fig. 2 *bottom*). The resulting representation, $\mathcal{P}(\mathbf{x}, \alpha, \beta, \sigma)$, is considered *locally orderless* [15] as the histograms eliminate any spatial variations in intensity, preserving only their distribution. We find that matching images within this locally orderless structure extends gradient support (Fig. 3) and is effective in solving inverse rendering problems that require long-range motion of image features (Fig. 7). Our method is straightforward to implement, requires no modifications to a differentiable renderer, is robust under noise, and works with arbitrary geometry representations and complex light-transport effects.

Overall, our method achieves good local minima for diverse optimization problems arising in differentiable rendering (§ 4). Results on inverse vectorization (§ 4.1), path tracing (§ 4.2) and rasterization (§ 4.4) validate that our method can reliably use standard RGB gradients, while methods that require computing additional gradients may fail. Lastly, we show that our method is also compatible with variational optimization on scenes with complex-light transport effects (§ 4.3) and real data (Fig. 9).

2. Related Work

The goal of differentiable rendering is to compute derivatives of the rendering integral [13] with respect to scene parameters. Estimating the rendering derivative has a long history in vision and graphics [2, 29, 40], and proves useful for accurate shape estimation [24], inverse scattering [9], material acquisition [5, 14], and various vision tasks [3]. While early works focus on deriving gradients for specific applications, more recent works by Loper and Black [21], Li *et al*. [18], Jakob *et al*. [12], Laine *et al*. [16], and Ravi *et al*. [30] propose general-purpose differentiable renderers. These renderers come with different tradeoffs, however we focus specifically on maximizing the utility of the estimated gradients for solving inverse problems.

In contrast to neural networks, where parameter gradients can have global support [10], rendering gradients are more local, especially for parameters that define visibility in the scene. Optimization using gradient descent is difficult in this case, as shown in [7, 22, 44]. Existing methods extend gradient support either by using alternate geometry gradients [22, 44] or through the notion of *differentiating through blurring* — either by blurring discontinuities [8, 20, 39, 45] or blurring the parameter space [7] as a form of stochastic approximation [6] of the gradient. Our approach operates entirely in image space and uses scale-space matching [42]. While this makes our method conceptually orthogonal to other methods, we show that they are complementary (§ 4.3).

Scale-space matching remains relatively underexplored in differentiable rendering. While Li *et al*. [18] use Gaussian Pyramids to avoid local minima when optimizing geometry primitives, and Vicini *et al*. [37] apply them for recovering signed distance functions, Xing *et al*. [44] show that using multi-resolution images is unreliable. In § 3, we illustrate why using image pyramids may fail. Our approach uses three distinct scale spaces implemented as Locally Orderless Images (LOIs), as proposed by Koenderink and van Doorn [15]. Unlike Gaussian pyramids that only match mean intensities within neighborhoods, our method matches entire distributions, leading to better recovery. LOIs have proven effective for various vision tasks [35], including image retrieval [17], object tracking [25], and non-linear filtering [33, 34].

3. Method

Consider an image $\mathcal{I}(\mathbf{x}; \theta) : \mathbb{R}^2 \mapsto \mathbb{R}$ as a set of direct or indirect observations of parameters θ that model the geometry and the appearance of a scene. Starting from an unknown set

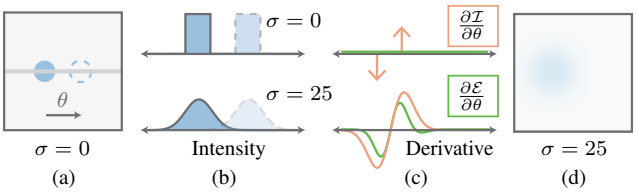


Figure 3. **Scale-space matching extends gradient support.** Given an image (a) of a disk we recover its position θ on the horizontal axis. At stationary resolution ($\sigma = 0$), the initial and target (dotted) disks do not overlap, as shown in the corresponding 1D signals in (b). The image gradient $\frac{\partial \mathcal{I}}{\partial \theta}$ is sparse (orange) and is non-zero only at the boundaries of the disk (c-*top*). The error gradient $\frac{\partial \mathcal{E}}{\partial \theta}$ is zero everywhere (green) and the optimization is stuck in a local minimum. When matching at coarser scales (d), the gradients are no longer sparse (c-*bottom*), leading to optimal recovery.

of parameters θ, the goal is to recover an optimal set such that the rendered image matches identically to a given reference. We use a differentiable renderer to compute derivatives $\frac{\partial \mathcal{I}}{\partial \theta}$ and minimize an error function $\mathcal{E}$ through gradient descent. As shown in the example in Fig. 1, image gradients can be sparse and ineffective for inverse problems [7, 18, 22, 44].

As an illustration, consider the 1D toy example shown in Fig. 3 A reference image shows a disk for which we wish to recover its position θ along the horizontal axis. For any initialization, the image gradients are sparsely localized at the disk's silhouette. In the very likely case of no overlap between the initial and the target disks, $\frac{\partial \mathcal{E}}{\partial \theta} = 0$, providing no signal to update the initial choice of θ. Even with some overlap between the two disks, $\frac{\partial \mathcal{E}}{\partial \theta}$ is non-zero but still sparse, making the optimization sensitive to noise.

Inner Scale To increase gradient support, we render images at different resolutions [42] and match them in scale space [41]. We apply a linear filter with a progressively increasing kernel width $\sigma \in \mathbb{R}_+$ to images rendered at their stationary resolution ($\sigma = 0$):

$$\mathcal{I}(\mathbf{x}; \theta, \sigma) = (G * I)(\mathbf{x}; \sigma), \tag{1}$$

where $*$ is the convolution operator and G is the Gaussian aperture function that defines the *inner scale*:

$$G(\mathbf{x}; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mathbf{x} \cdot \mathbf{x}}{2\sigma^2}\right), \quad \sigma > 0. \tag{2}$$

Let us revisit the example in Fig. 3. At coarser scales ($\sigma > 0$), we find that the image gradients have an extended support with a strong enough signal for convergence as the disks turn into larger overlapping blobs. This example reveals a key insight: while gradients are inherently local, blurring images with a scale-space kernel increases their

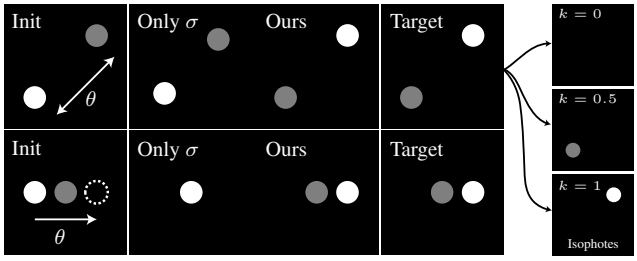


Figure 4. **Tonal Separation.** Shown are two (a-*top* and a-*bottom*) 1D inverse problems where we recover disk positions (θ) from images (*left*). Image matching within σ-space measures only the errors in the mean of the intensity distributions at each scale. In inverse settings that involve multiple objects with different appearances, this approach is likely to get stuck in a local minimum (a-*center-left*). The α-space integration kernels are intensity-aware and treat images as sets of distinct equal-intensity isophotes (b). When images are matched in all three scale spaces, the optimization is less prone to getting stuck in local minima (a-*center-right*).

support and simplifies the problem of alignment in image space. Note that this idea of *gradient diffusion* is similar to variational optimization [7, 31, 32], where a blur kernel is alternatively used in the *parameter space* to achieve a similar effect of extending gradient support and enabling long-range matching in image space.

Scale-space matching performs well for simple, controlled problems, but it remains unreliable in more realistic scenarios. As shown in recent work [7, 44], for inverse problems with complex geometry and light-transport effects, blurring and multi-scale strategies are not robust. We identify two primary reasons for this. *First*, blurring has an averaging effect that changes the local appearance — this makes matching image features more difficult. *Second*, blurring suppresses high-frequency details. With minor tweaks to the previous toy example, such as including image noise or multiple objects (Fig. 4 and 5), we find that inner-scale matching still suffers from local minima problems. Our method uses two additional scale parameters to achieve optimal recovery.

Tonal Scale We use scale-imprecision space [11] and relax the assumption of using images as real-valued functions. Instead, at each pixel location, we model a radiance distribution that captures the uncertainty in the measurements. Even in ideal settings, this is reasonable, as pixels integrate light over a non-zero area and measurements have finite precision [28]. In the context of inverse rendering, images can be noisy due to Monte Carlo integration [13] or external factors such as sensors and lenses. Explicitly modeling uncertainty in estimated parameters can lead to better recovery, as also observed in [47]. Our method models the uncertainty in ra-

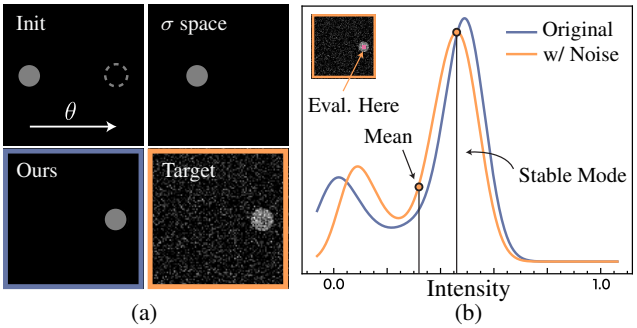


Figure 5. **Histogram matching is less sensitive to noise.** To recover the position (θ) of a circular disk from a noisy reference image (a-*bottom-right*), methods that match images only at their stationary resolution or in σ-scale space fail — as they overlook imprecision and uncertainty in radiance measurements. Our method uses a tonal parameter (β) to account for intensity uncertainty and an extent scale-space to preserve the distribution modes at coarser scales (b), leading to optimal recovery of θ.

diance estimates using 1D kernels with bandwidth $\beta \in \mathbb{R}_+$. The result is a family of spatially-varying probability distributions $\mathcal{P}$ defined on the image plane:

$$\mathcal{P}(\mathbf{x}, k; \theta, \sigma, \beta) = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left(-\frac{(k - \mathcal{I}(\mathbf{x}; \theta, \sigma))^2}{2\beta^2}\right), \quad (3)$$

where β is the scale in intensity domain and controls the tonal resolution, and $k \in [0, 1]$ is intensity. For a given k, $\mathcal{P}(\mathbf{x}, k)$ is a soft *isophote* that measures the probability of intensity being k at $\mathbf{x}$. This approach provides the benefit of *tonal separation*, where each isophote is treated as a distinct image, leading to improved separation both in visual features and image gradients. The inverse problem in Fig. 4 illustrates the benefit of using a tonal scale. In practice, we use β as the width of the bins used for discretizing $\mathcal{P}$ as histograms.

Extent Scale To capture spatial relationships between the local distributions $\mathcal{P}$, we can extend this representation further using the idea of locally orderless images [15]. Similar to integrating image intensity over a spatial extent as in σ scale space, we use an aperture function A that integrates the locally-defined histograms as follows:

$$\mathcal{H}(\mathbf{x}, k; \theta, \sigma, \beta, \alpha) = \int_{\mathbb{R}^2} A(\mathbf{x} - \mathbf{y}; \alpha) \mathcal{P}(\mathbf{y}, k; \theta, \sigma, \beta)\, d\mathbf{y}, \quad (4)$$

where A is similar to the kernel defined in Eq. 2. Unlike conventional scale-space blurring which operates directly on radiance values, Eq. 4 blurs histogram contributions. This distinction enables the method to preserve the effective *modes* of the radiance distribution across scales — thereby retaining both the appearance and geometric characteristics of the image even at coarser resolutions. Consequently, our method

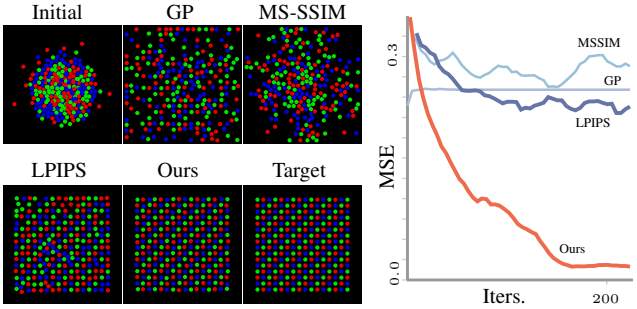


Figure 6. **Comparisons with multi-scale methods.** Given an image we optimize positions of 256 disks starting from random intialization. We use diffvg [19] to compute RGB gradients. Our method recovers the closest optimal arrangement. The error plot on the right shows mean-squared error (MSE) between rendered and target images *vs*. iterations.

is well-suited to handle perturbations in the reference images as the modes of the intensity distribution remain stable under noise [34]. See Fig. 5 for an illustrative example.

Histogram Matching Equipped with three separate scale spaces, defined using the parameters α, β and σ, we can now pose the inverse rendering objective as minimizing the difference between the distributions of rendered ($\mathcal{H}'$) and reference ($\mathcal{H}^{\text{gt}}$) images. Matching histograms and computing related distance metrics is a well-studied task in graphics and optimal transport literature [26, 27]. We use the Wasserstein distance which has a closed form for 1D distributions. At given scales α, β and σ, the distance between two histograms is estimated as the summation of point-wise errors between their cumulative distribution functions (cdf):

$$\mathcal{E}(\theta, \alpha, \beta, \sigma) = \int_{\mathbb{R}^2} \int_0^1 \left[\text{cdf}_{\mathcal{H}'}(\mathbf{x}, k; \theta, \alpha, \beta, \sigma) - \text{cdf}_{\mathcal{H}^{\text{gt}}}(\mathbf{x}, k; \alpha, \beta, \sigma)\right]^{1/p} dk d\mathbf{x}, \quad (5)$$

$$\text{where,} \quad \text{cdf}_{\mathcal{H}}(\mathbf{x}, k; \cdot) = \int_0^k \mathcal{H}(\mathbf{x}, j; \cdot)\, dj$$

We compute the total error as the summation over all the scales, *i.e.* $\mathcal{E}_{\text{total}}(\theta) = \sum_{\alpha,\beta,\sigma} \mathcal{E}(\theta, \alpha, \beta, \sigma)$. We use $p = 1$ in our experiments.

4. Results

We evaluate our method using three differentiable renderers: diffvg [19] for vector graphics, a path tracer [12], and a rasterizer [16]. Our experiments are designed for three goals: First, we compare our method against other multi-scale approaches like Gaussian Pyramids [1] and MS-SSIM [38] (§ 4.1). Second, we evaluate how parameter-space blurring (PRDPT [7]) performs compared to scale-space matching (§ 4.2). Third, we test the reliability of RGB gradients versus proxy gradients like RGBXY [44]. We also show how our

Table 1. **Quantitative comparisons for 2D.** We recover positions of $n = [4, 16, 32, 64, 256]$ disks with *known* color from a reference 128×128 image. Image gradients are computed using diffvg [19] and sparsely defined along the silhouettes of the disks [22]. Compared to other scale-space approaches, we find our method to be best suited for this task.

n Disks	4		16		32		64		256	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
GP [1]	25.05	0.97	19.83	0.91	16.83	0.83	15.34	0.76	9.60	0.34
MS-SSIM [38]	27.44	0.98	21.41	0.94	17.93	0.87	15.42	0.78	9.83	0.35
LPIPS [46]	27.79	0.98	20.94	0.93	17.88	0.86	17.04	0.82	15.2	0.76
Ours	30.40	0.99	33.55	0.99	27.43	0.97	28.23	0.96	21.57	0.90

method complements parameter-space blurring through two example scenes (§ 4.3), including one with real data.

4.1. Differentiable Vectorization

Multi-scale feature matching is widely used in vision and graphics for various tasks. Three prominent techniques include: (a) the Gaussian pyramid [1], which approximates the σ-scale space and matches the mean of intensity distributions across scales; (b) multi-scale structural similarity (MS-SSIM) [38], which matches both the mean and covariance of the distributions; and (c) LPIPS [46], which uses a hierarchy of features from a deep neural network to measure perceptual similarity. Our approach, using three distinct scale spaces, is similar to these methods but is specifically designed for inverse rendering problems.

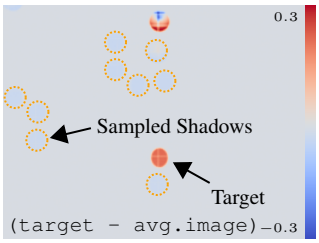
We design a test benchmark to evaluate multi-scale methods. Similar to previous experiments, the objective is to recover the positions of 2D disks with known colors from a given image. This is a representative task for inverse rendering problems that require long-range feature matching with sparse gradient information. The benchmark includes five problem sets, with difficulty determined by the number of disks in each image. Note that this a much more difficult task than optimizing color, since appearance gradients are not as sparse as geometry gradients [22]. Optimization parameters are randomly initialized, while target positions form a uniform grid. Differentiable vectorization [19] is used to compute image gradients. For other multi-scale approaches, we measure the L2 error between features at different scales, while our method uses Wasserstein distance (Eq. 5) to compare histograms. The reference images are 128×128 at stationary resolution and we choose $\alpha = [1, 5, 15], \sigma = [1, 5, 15, 45]$ and $\beta = 0.125$ as the scale space parameters. For consistency, we perform 10 runs, each with a different initialization, and report mean PSNR and SSIM in Table 1. The optimized images for $n = 256$ are shown in Fig. 6 along with the corresponding error plot.

4.2. Differentiable Path Tracing

Following the evaluation framework in [7], we test our method on five scenes. The goal here is to evaluate methods that use the notion of scale either in image space or parameter space (*i.e.* variational optimization). We use Mitsuba 3 [12] as the default renderer with the `prb_reparam` integrator [36]. Quantitative results evaluating visual appearance (PSNR) and parameter recovery (MAE) are in Table 2. We use the same kernel parameters, $\alpha = [1, 5, 15, 45], \beta = 0.125$, and $\sigma = [1, 5]$ for all the scenes. All images are rendered at 256×192 resolution. We use Adam with a learning rate of 10^{-2} for all methods.

Shadow We use the `Shadow` scene from [7]. Given an image, we recover the position of a sphere that is not visible in the view frustum but projects a shadow onto a visible plane. This is a representative example for when RGB gradients perform poorly [7]. Using gradients from Mitsuba 3 [12], our method recovers the position of the sphere. Qualitative results for our method and variational optimization are shown in Fig. 7. While variational optimization is able to recover the target shadow, other multi-scale methods get stuck in local minima.

Shadow Mini We modify the previous scene with two changes: (a) an initialization that is further away from the target, and (b) changing the scale of the sphere and hence the area of the shadow. Variational optimization struggles in such scenes. It achieves blurring in parameter space through multiple samples that correspond to images with perturbed positions of the sphere. Ideally, for this approach to work, the shadows in the sampled images require some overlap with the target in the image space to update the parameters reliably. Since that is unlikely to happen (see inset) with the small shadow footprint in the reference, the optimization converges to the sphere being pushed out of the scene. The inset figure plots the difference between the target image and the variational estimate (avg. image) using 10 samples. As mentioned, shadows from the variational samples do not overlap with the target shadow. We test with different kernel parameters and sampling rates (τ in [7]), all of which perform similarly on this scene. Since our method operates directly in image space, it reliably recovers the reference independent of the initial position of the shadows. Other multi-scale methods converge at different local minima.



Why does PRDPT [7] fail?

Caustics and Lights This scene requires optimization in two separate domains: (a) recovering the positions of glass spheres, and (b) adjusting lights that are not directly visible in the view frustum. We normalize the range of both domains within $[-5, 5]$. This task is challenging because, at initialization, the caustic patterns do not overlap with the

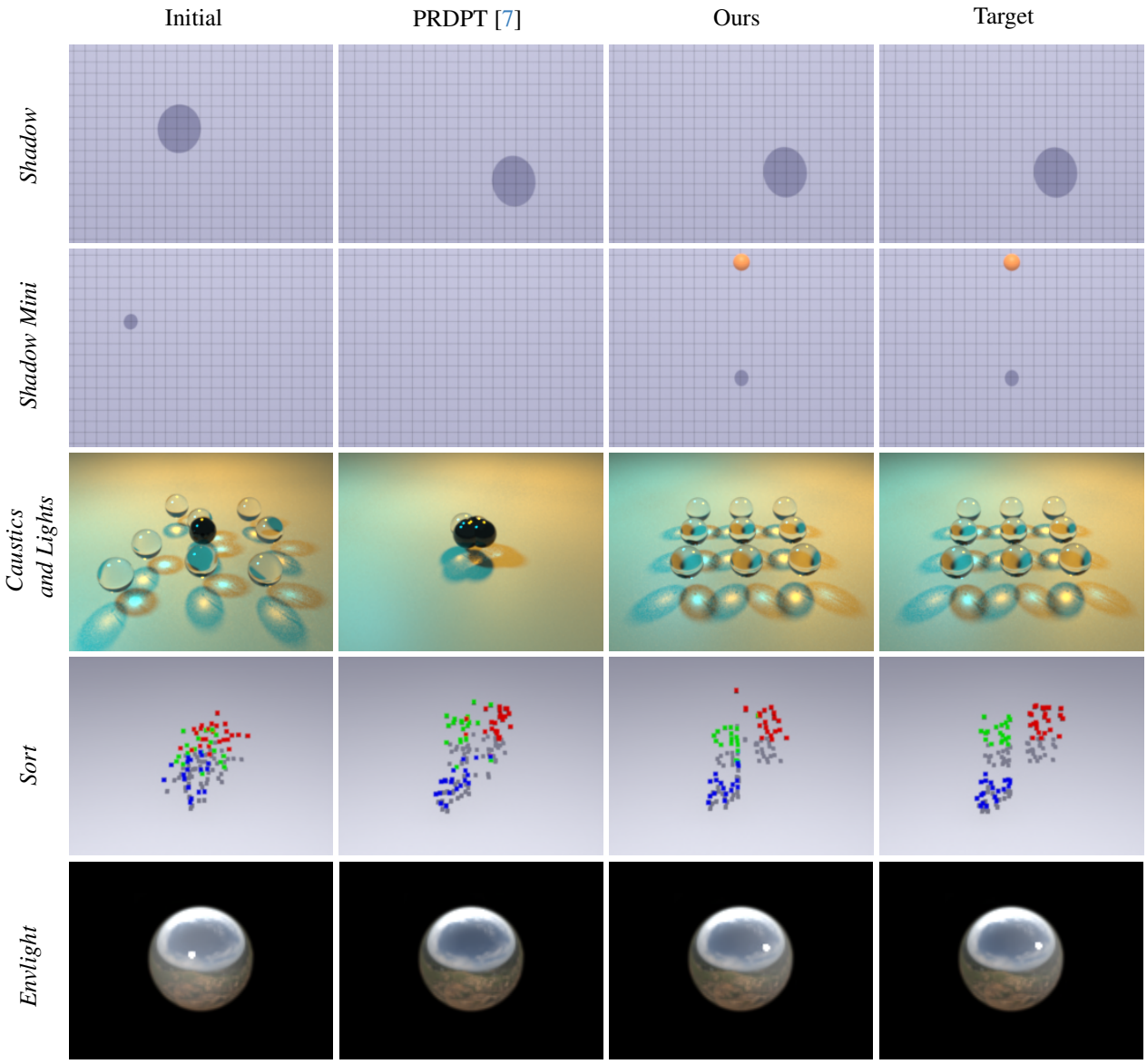


Figure 7. **Quantitative comparisons with parameter-space blurring.** We show five separate scenes (one in each row) with varying degrees of difficulty. Starting from a random initialization (left-most column), the goal is to recover the positions of different primitives so that the rendered image matches the reference (right-most column). Compared to parameter-space blurring (PRDPT [7]), we find our method that uses image-space local histogram matching is more suitable for these tasks. For quantitative comparisons, refer to Table 2.

reference in either appearance or position. We observe that matching histograms is particularly suited for this task due to the presence of Monte Carlo noise in forward rendering and high-frequency caustics patterns.

Sort In the `Sort` scene [7], the objective is to recover the positions of 65 differently colored primitives from an image. This involves optimizing the x and z translation coordinates of these primitives, resulting in a 130-dimensional optimization problem. Since optimization in this scene is highly sensitive to the choice of initialization, we perform 20 different runs with varying random seeds and report the average results in Table 2. Comparisons for one of the runs are shown in Fig. 7. Since stochastic gradients have high variance for high dimensional problems [4], variational optimization is not ideal for such scenes.

Envlight In this scene, we recover the position of a distant light source from an image of a reflective sphere rendered within a high-frequency environment. The light source is not directly visible, casting only a small specular highlight

Table 2. We evaluate our method and PRDPT [7], which employs parameter-space blurring, across five synthetic scenes of varying difficulty. We also present results using other multi-scale baselines. Our method recovers the optimal parameter configuration for all scenes, despite relying on sparsely defined RGB gradients.

	Shadow		*Shadow Mini*		*Caustics and Lights*		*Sort*		*Envlight*	
	PSNR	MAE	PSNR	MAE	PSNR	MAE	PSNR	MAE	PSNR	MAE
PRDPT [7]	31.77	0.0035	27.74	6.0147	17.09	4.8801	22.25	0.9201	31.87	0.7543
Ours	54.33	0.0034	54.73	0.0033	32.58	0.3475	25.23	0.4855	36.17	0.0775
Mitsuba [12] (GP)	24.44	1.1349	24.06	2.3491	21.02	0.7317	23.55	0.5189	28.73	0.8072
Mitsuba [12] (MSSSIM)	24.35	0.9320	24.34	0.9061	10.30	86.2286	18.91	0.7742	30.26	0.6703

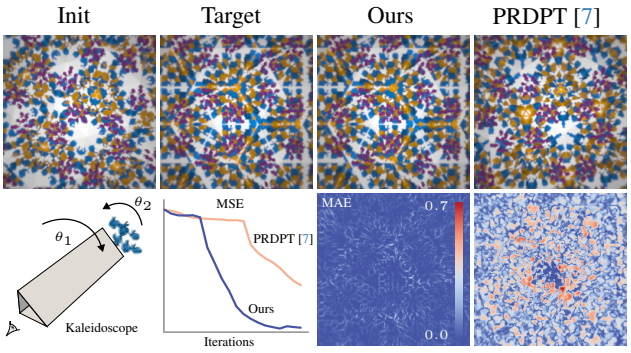


Figure 8. **Variational Optimization with Locally Orderless Images.** Our method works well with other optimization techniques like variational optimization. Given an image of a pattern inside a kaleidoscope, we recover the configuration of the kaleidoscope (θ_1) and the particles (θ_2) reflecting light through it. Using variational derivatives and matching images at their stationary resolution leads to sub-optimal recovery. By matching local image histograms (Ours), we successfully recover the target configuration.

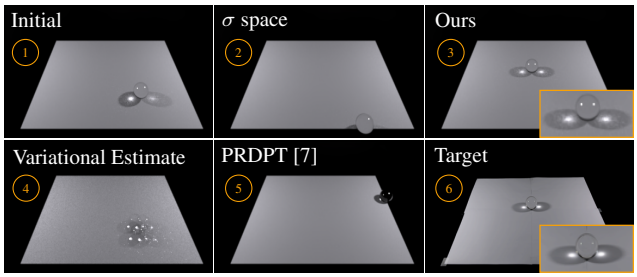


Figure 9. **Using a real photograph.** We show a preliminary experiment with real data that validates our method's robustness against sensor noise and calibration errors. Starting with known material, camera, and lighting, and a randomly positioned glass orb ①, we recover the optimal translation from a real photograph ⑥ using variational optimization. The variational estimate with 8 samples is shown in ④. Matching at stationary resolution ⑤ or only in the σ space ② fails in this task. Our method achieves optimal recovery starting from any initial position on the surface.

on the sphere, creating an optimization landscape with multiple local minima. Similar to the `Shadow Mini` scene, PRDPT [7] is stuck in a local minimum, as one of the sampled images need the highlight to be exactly aligned with the target. Our method, despite using sparse RGB gradients, successfully recovers the optimal configuration. Using image pyramids fails at this task as the high-frequencies and modes of intensity distribution are lost at coarser scales.

4.3. Variational Optimization

In the following two scenes, we demonstrate cases where RGB gradients alone are not reliable, and show how combining our method with variational derivatives [7] can achieve optimal recovery.

Kaleidoscope As shown in [43], RGB gradients are unstable and inefficient for scenes involving complex light transport. Consider the virtual kaleidoscope scene shown in Fig. 8. Our goal is to recover both the kaleidoscope configuration and the positions of light-reflecting particles from a single image. The complex light paths within the kaleidoscope result in high-variance RGB gradients that cannot be reliably used for gradient-based optimization. Moreover, small parameter perturbations can dramatically change the kaleidoscope patterns, leading to high variance even in variational derivatives and preventing optimal recovery. We find that our approach of combining multi-scale histogram matching with parameter-space blurring is ideal for this task and successfully recovers the optimal configuration. We use the same scale-space parameters as in § 4.2 and render images at 200×200 resolution.

Real Scene Inverse rendering with real photographs presents additional challenges, such as noise from sensors and lenses, and calibration errors. We test our method on a real scene, where we aim to recover the position of a glass orb on a known planar surface from a given photograph. The scene is initialized with calibrated light sources, camera parameters, and known material (borosilicate glass), but the position of the orb is unknown (Fig. 9 ①). Similar to the previous scene, we use variational optimization. We observe that using matching stationary images ($\sigma = 0, \alpha = 0$) is sen-

Table 3. **Quantitative results with differentiable rasterization.** We evaluate our method using gradients (∇) from Nvdiffrast along with other multi-scale methods on test scenes provided in [44]. We show that our method using RGB gradients can reliably recover optimal solutions for low-dimensional problems (such as Translation and Rotation) and is ideal for higher dimensional problems (such as Camera Pose or Material recovery). Each image is rendered at 128×128. We use the scale parameters $\alpha = [1, 5, 15, 45], \sigma = [0, 5]$ and $\beta = 0.03125$.

	Parameters	Translation		Rotation		Shape		Trans. + Rot.		Camera Pose		Material		Env. Map	
	Num. views	6		6		1		4		1		6		6	
Method	∇ type	PSNR	MAE	PSNR	MAE	PSNR	MAE	PSNR	MAE	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Xing *et al*. [44]	RGBXY	63.62	0.00	26.66	0.45	30.75	0.30	30.23	0.68	23.50	0.92	33.01	0.96	32.43	0.91
Ours	RGB	51.77	0.00	30.73	0.33	28.22	0.54	36.23	0.35	27.22	0.96	49.07	1.00	55.22	1.00
Nvdiffrast [16] (GP [1])	RGB	31.16	0.18	22.40	0.49	25.55	0.59	19.53	2.11	14.82	0.73	42.48	0.99	47.79	0.99
Nvdiffrast [16] (MS-SSIM [38])	RGB	23.51	0.60	17.61	0.88	11.20	0.83	17.74	2.37	09.01	0.33	11.19	0.72	23.30	0.73

sitive to noise and fails in this task (Fig. 9 ⑤). We attribute this failure to the same alignment issue as in `Shadown Mini`, where one of the samples in the variational estimate (Fig. 9 ④ shows the mean of all samples) needs to overlap with the target position of the orb. Our local-histogram matching approach does not require perfect alignment between the variational estimate and the reference images, allowing it to reliably recover the target position of the orb from any initialization on the planar surface (Fig. 9 ③). As an ablation on the α and β scale spaces, we also test matching only in the σ space with the same kernel parameters, which fails in this task (Fig. 9 ②).

4.4. Differentiable Rasterization

We evaluate our method on scenes provided by Xing *et al*. [44]. The tested scenes vary in complexity, requiring optimization for both geometric and appearance parameters. Quantitative results are in Table 3, using metrics to assess both visual appearance (PSNR) and parameter recovery accuracy (MAE). For low-dimensional parameter spaces (e.g., `Translation` and `Rotation`), our method using *only* RGB derivatives performs comparably to Xing *et al*.'s [44] approach, which utilizes both RGB and Lagrangian derivatives. For high dimensional parameter spaces, such as optimizing the full camera-pose matrix or texture map, RGB gradients combined with local histogram matching performs better. We use the same scale parameters as in § 4.1.

5. Discussion

Figure 10. Our method is not designed to recover parameters that do not influence the image, such as the rotation of the mug [7].

Ablations Our method relies on two critical components: the α and β scale spaces. Using only the σ space reduces our method to standard Gaussian Pyramids (GPs), as shown in our quantitative evaluations (Tables 1, 2, and 3). While using only the β space preserves tonal variations, it fails to extend the spatial influence of gradients, leading to poor performance across our experiments . Since the α space is dependent on β (Eq. 4), it cannot be used in isolation.

Limitations As our method operates exclusively in image space, parameters that do not influence primary or secondary effects still yield zero gradients in the LOI representation. For example, in the `mug` scene (Fig. 10) from [7], optimizing the mug's rotation when its handle is not visible in the initial view is infeasible. In such cases, our method is most effective when combined with variational optimization approaches, where parameter-space blurring can reveal occluded features like the handle through sampling.

Conclusion In this work, we revisit the idea of locally orderless images (LOIs) in the context of differentiable rendering. One of the key contributions of this work is bridging these previously disconnected areas in computer vision and graphics. Through various experiments, we show that our method yields better recovery in optimization problems where existing methods are likely to fail. Our method is straightforward to implement and integrates seamlessly with existing differentiable renderers without requiring any modifications to their core functionality.

Acknowledgments This work was supported in part by NSF grant 2402583. We also acknowledge NSF grant 2110409, a Qualcomm Innovation Fellowship, gifts from Adobe, Google, Qualcomm and Rembrand, the Ronald L. Graham Chair and the UC San Diego Center for Visual Computing.

References

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984. 2, 4, 5, 8

[2] James Arvo. The irradiance jacobian for partially occluded polyhedral sources. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 343–350, 1994. 2

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 2

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 6

[5] Adrien Bousseau, Emmanuelle Chapoulie, Ravi Ramamoorthi, and Maneesh Agrawala. Optimizing environment maps for material depiction. In *Computer graphics forum*, pages 1171–1180. Wiley Online Library, 2011. 2

[6] Thomas Deliot, Eric Heitz, and Laurent Belcour. Transforming a non-differentiable rasterizer into a differentiable one with stochastic gradient estimation. *arXiv preprint arXiv:2404.09758*, 2024. 2

[7] Michael Fischer and Tobias Ritschel. Plateau-reduced differentiable path tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4285–4294, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[8] Michael Fischer and Tobias Ritschel. Zerograds: Learning local surrogates for non-differentiable graphics. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 2

[9] Ioannis Gkioulekas, Shuang Zhao, Kavita Bala, Todd Zickler, and Anat Levin. Inverse volume rendering with material dictionaries. *ACM Transactions on Graphics (TOG)*, 32(6): 1–13, 2013. 2

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`. 2

[11] Lewis D Griffin. Scale-imprecision space. *Image and Vision Computing*, 15(5):369–398, 1997. 2, 3

[12] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 1, 2, 4, 5, 7

[13] James T Kajiya. The rendering equation. 20(4), 1986. 2, 3

[14] Pramook Khungurn, Daniel Schroeder, Shuang Zhao, Kavita Bala, and Steve Marschner. Matching real fabrics with micro-appearance models. *ACM Trans. Graph.*, 35(1):1–1, 2015. 2

[15] Jan J Koenderink and Andrea J Van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31:159–168, 1999. 2, 4

[16] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 1, 2, 4, 8

[17] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce, et al. Spatial pyramid matching. *Object Categorization: Computer and Human Vision Perspectives*, 3(4), 2009. 2

[18] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37 (6):222:1–222:11, 2018. 1, 2, 3

[19] Tzu-Mao Li, Michal Lukáč, Gharbi Michaël, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 39(6):193:1–193:15, 2020. 1, 4, 5

[20] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7708–7717, 2019. 2

[21] Matthew M Loper and Michael J Black. Opendr: An approximate differentiable renderer. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 154–169. Springer, 2014. 1, 2

[22] Ishit Mehta, Manmohan Chandraker, and Ravi Ramamoorthi. A theory of topological derivatives for inverse rendering of geometry. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 419–429, 2023. 1, 2, 3, 5

[23] Luke Metz, C Daniel Freeman, Samuel S Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021. 1

[24] Shree K Nayar, Katsushi Ikeuchi, and Takeo Kanade. Shape from interreflections. *International Journal of Computer Vision*, 6:173–195, 1991. 2

[25] Shaul Oron, Aharon Bar-Hillel, Dan Levi, and Shai Avidan. Locally orderless tracking. *International Journal of Computer Vision*, 111:213–228, 2015. 2

[26] Ofir Pele and Michael Werman. The quadratic-chi histogram distance family. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*, pages 749–762. Springer, 2010. 4

[27] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 4

[28] Henri Poincaré. Science and hypothesis., 1905. 3

[29] Ravi Ramamoorthi, Dhruv Mahajan, and Peter Belhumeur. A first-order analysis of lighting, shading, and shadows. *ACM Transactions on Graphics (TOG)*, 26(1):2–es, 2007. 2

[30] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1, 2

[31] Joe Staines and David Barber. Variational optimization. *arXiv preprint arXiv:1212.4507*, 2012. 3

[32] Hyung Ju Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do differentiable simulators give better policy gradients? In *International Conference on Machine Learning*, pages 20668–20696. PMLR, 2022. 3

[33] Joost van de Weijer and Theo Gevers. Color mode filtering. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, pages 125–128. IEEE, 2001. 2

[34] Joost Van de Weijer and Rein Van den Boomgaard. Local mode filtering. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages II–II. IEEE, 2001. 2, 4

[35] Bram van Ginneken and Bart M ter Haar Romeny. Applications of locally orderless images. *Journal of Visual Communication and Image Representation*, 11(2):196–208, 2000. 2

[36] Delio Vicini, Sébastien Speierer, and Wenzel Jakob. Path replay backpropagation: Differentiating light paths using constant memory and linear time. *Transactions on Graphics (Proceedings of SIGGRAPH)*, 40(4), 2021. 5

[37] Delio Vicini, Sébastien Speierer, and Wenzel Jakob. Differentiable signed distance function rendering. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022. 2

[38] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 4, 5, 8

[39] Zichen Wang, Xi Deng, Ziyi Zhang, Wenzel Jakob, and Steve Marschner. A simple approach to differentiable rendering of sdfs. *arXiv preprint arXiv:2405.08733*, 2024. 2

[40] Greg Ward and Paul S Heckbert. Irradiance gradients. 1992. 2

[41] Andrew Witkin, Demetri Terzopoulos, and Michael Kass. Signal matching through scale space. *International Journal of Computer Vision*, 1(2):133–144, 1987. 3

[42] Andrew P Witkin. Scale-space filtering. In *Readings in computer vision*, pages 329–332. Elsevier, 1987. 1, 2, 3

[43] Liwen Wu, Rui Zhu, Mustafa B Yaldiz, Yinhao Zhu, Hong Cai, Janarbek Matai, Fatih Porikli, Tzu-Mao Li, Manmohan Chandraker, and Ravi Ramamoorthi. Factorized inverse path tracing for efficient and accurate material-lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3848–3858, 2023. 7

[44] Jiankai Xing, Fujun Luan, Ling-Qi Yan, Xuejun Hu, Houde Qian, and Kun Xu. Differentiable Rendering Using RGBXY Derivatives and Optimal Transport. *ACM Transactions on Graphics*, 41(6):1–13, 2022. 1, 2, 3, 4, 8

[45] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, 2021. 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[47] Zhiqian Zhou, Cheng Zhang, Zhao Dong, Carl Marshall, and Shuang Zhao. Estimating uncertainty in appearance acquisition. 2024. 3