Vision, Modeling, and Visualization (2023)
T. Grosch and M. Guthe (Eds.)

N-SfC: Robust and Fast Shape Estimation from Caustic Images

Marc Kassubeck[1], Moritz Kappel[1], Susana Castillo[1], and Marcus Magnor[1]

[1] Institut für Computergraphik, TU Braunschweig, Germany
{kassubeck, kappel, castillo, magnor}@cg.cs.tu-bs.de

Abstract

This paper handles the highly challenging problem of reconstructing the shape of a refracting object from a single image of its resulting caustic. Due to the ubiquity of transparent refracting objects in everyday life, reconstruction of their shape entails a multitude of practical applications. While we focus our attention on inline shape reconstruction in glass fabrication processes, our methodology could be adapted to scenarios where the limiting factor is a lack of input measurements to constrain the reconstruction problem completely. The recent Shape from Caustics (SfC) method casts this problem as the inverse of a light propagation simulation for synthesis of the caustic image, that can be solved by a differentiable renderer. However, the inherent complexity of light transport through refracting surfaces currently limits the practical application due to reconstruction speed and robustness. Thus, we introduce Neural-Shape from Caustics (N-SfC), a learning-based extension incorporating two components into the reconstruction pipeline: a denoising module, which both alleviates the light transport simulation cost, and also helps finding a better minimum; and an optimization process based on learned gradient descent, which enables better convergence using fewer iterations. Extensive experiments demonstrate that we significantly outperform the current state-of-the-art in both computational speed and final surface error.

CCS Concepts
• ***Computing methodologies*** → ***Image-based rendering; Shape modeling;*** *Machine learning;*

1. Introduction

Recent advances in physics-based differentiable rendering have enabled to incorporate ever more complex light transport effects such as caustics into inverse vision problems. Previously, these effects had to be modeled by hand in a time consuming fashion, while readily available gradients from differentiable rendering frameworks allow application-specific reconstructions to focus more on efficient regularization and general optimization schemes. One of these application-specific examples is the Shape from Caustics (SfC) problem as formulated by Kassubeck *et al.* [KBC*21], which deals with the under-constrained problem of reconstructing the shape of a refracting object from its resulting caustic image, *i.e.* the brightness distribution as seen on a screen surface under illumination from a known light source. Tackling this problem is enabler for many applications in inline quality control of optical components, especially in the emerging field of integrated optical manufacturing, which is fuelled by novel production processes such as laser glass deposition (LGD). The requirements on the feedback and quality control loop necessitate taking the measurement *in-situ*, such as after each printed layer, and being fast enough to take the reconstruction result into account when the next layer is ready to be deposited. Only in this way small deviations in previous layers can be adaptively compensated, without unduly slowing down the production process. Thus, such an inline measurement is not meant to replace established high accuracy methods, but complement them in the most time-sensitive cases. As a solution, SfC [KBC*21] proposed an optical measurement setup with no moving parts, which requires only a single caustic image as the input to their classical constrained optimization based reconstruction. However, when trying to adapt the SfC method to the real production process, the reconstruction algorithm revealed deficits in terms of reconstruction speed and robustness. Furthermore, with the trend in fabrication to produce smaller and smaller batch sizes down to batch sizes of one in *individualized manufacturing*, the control system needs to robustly assess a large variety of possible shapes, which are not known a priori.

We address these problems by introducing two learned components into the reconstruction loop: a denoiser and a learned gradient descent scheme. Since the image formation simulation is based on Monte-Carlo integration and unless a large amount of samples are taken into account, the resulting caustics are invariably noisy, hindering the forward simulation and backpropagation. Our denoiser alleviates this problem by allowing us to fix the number of forward simulation samples, keeping the runtime cost in check. Secondly, our learned gradient descent scheme allows to leave out a significant amount of gradient descent steps, while still producing results with lower total shape error. We show empirically that this combination allows to avoid spurious local minima, enabling more robust and faithful reconstructions. Source code, including all dataset generation scripts, is available at https://graphics.tu-bs.de/publications/kassubeck2023n-sfc.

DOI: 10.2312/vmv.20231224

2. Related Work

As our work relates to differentiable and inverse rendering as well as to computational caustic design, denoising, learned gradient descent, and refractive reconstruction methods, we will group the related works in the following into these categories:

Differentiable rendering describes a subset of the field of inverse rendering, *i.e.* estimating physically valid parameters from imaging data, by means of a differentiable renderer. Those renderers formulate the forward image formation process in a differentiable manner and allow to efficiently compute gradients with respect to relevant free parameters, which can in turn be used in gradient-based local parameter search strategies. Redner [LADL18], Mitsuba 2, Path replay backpropagation [NDVZJ19, VSJ21]and lastly Mitsuba 3 [JSRV22] are increasingly efficient and fully featured renderers but in their current iteration mainly use unidirectional path tracing for image formation and are thus not applicable to our caustics-based optimization problem without further adaptations. Several other authors tackle the difficult problem of providing gradients with respect to scene geometry, which is in the mesh case not trivially differentiable due to visibility discontinuities. Of these methods, Zhang *et al.* [ZMY*20, ZYZ21] explicitly handle bidirectional methods and even participating media, but a source code release was missing at the start of this project. Unbiased gradient estimators [BLD20, ZMY*20] have been shown to be beneficial in terms of convergence behavior and final parameter estimation results [LZBD21], but for our problem even perfect gradients would only achieve convergence to an improper local minimum in many cases.Thus we focus our efforts on a learned gradient descent system to circumvent those spurious local minima an achieve significantly improve reconstruction for the 3D printing process at hand. We follow the simulation approach from SfC [KBC*21] for image formation and gradient generation, as it in turn builds upon the work of Frisvad *et al.* [FSES14], which allows for sharper caustic edges with fewer samples, further reducing the computational burden of the simulation.

Caustic design is stated as the problem of finding a refractor or reflector, which produces a desired caustic image on a given screen surface. Schwartzburg *et al.* [STTP14] define caustic design as a two step process by first solving an *optimal transport problem* and then calculating a heightfield achieving said transport. Meyron *et al.* [MMT18] expand upon the theory of optimal transport and provide a general algorithm for different computational design tasks. Another widely researched approach is modeling the freeform optics as the solution of a PDE [RM02, FFL16, WXL*13]. However, regarding physical constraints on the shape, those methods mainly enforce smoothness and reduced curvature to ensure physical realizability, whereas we are concerned with reconstructing true shapes given data of achievable geometries. Thus our objective more closely aligns with true general *refractive reconstructions* as described below, while having a less restrictive setup.

Denoising plays an important role in many physics-based image formation simulations [HY21], as the Monte-Carlo nature of path-tracing-based methods quickly leads to correct but highly noisy results, which clear up at the rate of the square root of the number of samples. Thus, filtering methods to alleviate the strain on computational resources are a component of even commercial rendering systems [Aut21]. With the rise of neural-network-based image processing, state-of-the-art denoising methods often integrate learned components into the processing pipeline. One can categorize these methods, based on whether they operate on the final output image [KBS15, VRM*18]or act deeper in the path tracing process to predict global illumination effects [NAM*17], like calculating high-resolution radiance maps from low resolution samples [JK21]. Our method directly operates in image-space, however we note that noise statistics of photon mapping methods differs from regular unidirectional path-tracing due to the bias in the estimator [ZWW*20, ZXJ*20]. In this case regular pre-trained denoiser is expected to perform sub-par, thus we opt to create our own dataset and network. By design, our denoiser performs a similar task to learned density estimation [ZXJ*20] as part of photon mapping based image generation pipeline. However, we further incorporate this denoiser into an inverse parameter estimation problem, which has not been presented before to the best of our knowledge.

Learned gradient descent as a subset of *meta-learning* or *learning to learn* replaces the classical hand-designed optimizer with a learned component [Sch93, YHC01, HYC01], improving ill-posed inverse problems by including a prior on reachable solutions by modification of gradients. Additionally, larger parameter-specific steps can be taken, allowing for faster convergence. We take inspiration from recent work [ADG*16, FBD*19], which cast the optimization trajectory as steps in a recurrent neural network, conditioned on (approximate) gradients and other problem specific inputs. In contrast to these methods, we do not train our recurrent network with truncated backpropagation through time, as this would necessitate computation of second order derivatives of the simulator. Instead, we use a simple point-wise training scheme, which we evaluate by applying our network recurrently at test time and demonstrating, that it has nevertheless learned to smoothly and efficiently minimize the relevant error metric.

Several works tackle the **reconstruction of refractive shapes** by proposing the inclusion of other sources of information.For example, a recent work [LWL*20] reports estimation of shape of general glass objects with a differentiable rendering system utilizing multiple views and Gray coded structured light. This approach achieves high quality results of free form shapes, but the physical setup complexity is beyond integration in existing manufacturing machines and the scope of this paper. We restrict ourselves to a single view and an arbitrary light source, but create a large dataset to build a prior over achievable shapes. In the preparation of this dataset we are related to but more specialized than Mousavi & Estrada [ME21], who provide a general dataset for computer vision tasks for scenes with transparent refracting objects.

3. Method

Following the approach of SfC [KBC*21] we represent the solution space of our shape reconstruction as a heightfield $h \in \mathbf{R}^{n \times n}$ over the flat base substrate of known thickness d. To simulate the resulting caustic image we place a light source above the substrate and calculate the wavelength-dependent irradiance $E \in \mathbf{R}_{+}^{n_w \times m \times m}$ per pixel of a sensor surface below the substrate. The simulation (cf. [FSES14]) and subsequent gradient calculation through backpropagation mainly depends on three hyperparameters: The number

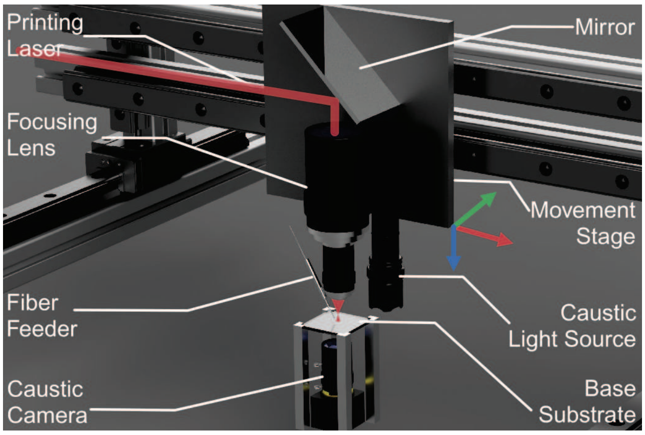

Figure 1: *Schematic sketch of* ***in-situ*** *feedback during production.*

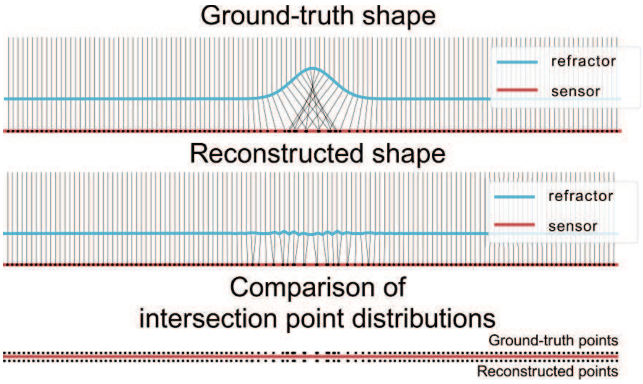

Figure 2: ***Underdeterminism.****The reconstructed shape, albeit near perfect in the metric considered, is still far from the desired solution.*

of samples n_l, *i.e.* the number of paths outgoing from the light; the number of distinct wavelengths associated with each ray n_w; and a smoothing parameter s, which controls the size of an elliptical footprint over which the energy of each transported photon is distributed. The smoothing parameter represents a trade-off between the minimum feature size, which can occur in the caustic image and the sampling noise of the simulation. An overview of further parameters and typical values can be found in our supplemental.

To illustrate a potential real-world application of this setup we provide a sketch in Fig. 1, where this setup is integrated into a manufacturing process, which concerns glass 3D printing by selectively melting glass fibers onto a base substrate. The functional goal in such a setup is validating the printed shape as specified by the part designer and possibly correct for deviations in subsequent steps. This is under special consideration, as functional tests of waveguide structures require additional steps, such as cleaving of the beginning and end of the fiber and precise placement of e.g. diodes and sensors at each end of respective fibers, which are difficult to achieve in-situ. Thus we propose a method, which does not directly measure such transverse functional specifications, but rather use the longitudinal caustic as a measure by which to infer adherence to design specifications, for which such functionality is known to hold. Note that the only moving parts are the ones which were already present in the production process, we simply added the light source above the substrate and a camera with appropriate field of view and focus distance below. While having the advantage of being easily integrable into even existing setups, this also leads to added complexity for the reconstruction process. We note that this production example informs physical measurements and dataset distribution, but our methodology is independent of this specific application case. We see this methodology as a framework for a class of problems, where camera and light placement, as well as the compute budget, restrict the amount of data that is needed for reconstruction.

3.1. Underdeterminism of the Problem

We show that reconstruction of a refracting shape from a single caustic image is severely under-constrained by considering a simple 2D toy example. Fig. 2 shows a ground-truth shape and resulting refracted light paths using a fixed refractive index as well as a reconstruction using the Hausdorff metric between the point sets of ground truth and estimated intersection. This can be thought of as roughly equivalent to the irradiance-based metric in Sec. 3.2, when each light path transports the same amount of energy. The result is obtained by optimizing the height of the refracting shape with the (biased) gradients of SfC [KBC*21] and with the Adam optimizer [KB15] of PyTorch [PGM*19]. The initial guess was set to a flat surface. When comparing the intersection distributions it becomes clear that for this intersection plane the estimation has reached near perfect parity, even with only very minor and local shape changes. Note that this problem would also persist when using unbiased gradients, which is in contrast to other recent work [LZBD21], which reported convergence to better minima when using unbiased estimators. This is because even the biased estimator finds a local minimum that is close to a true global minimum in the given loss function; *i.e.* , the main problem lies in the ambiguity of the loss landscape due to lack of sufficient data, not in the quality of the gradients. Even though the problem is exacerbated by the regular and fixed sampling scheme employed in this example, the principle also holds when randomly sampling new light paths in each step. Gradient-based schemes can only rely on very localized information to locally adapt the refractive surface patch responsible for the , like distance between corresponding intersection points here or discrepancy between overlapping radiance patches during photon mapping in the full problem considered in this paper. Thus every optimization progress, with stable convergence, would converge to a surface, which is globally much closer to the initial flat guess, because there is no indication in any loss that operates on caustics, that rays have ’crossed over each other’ as seen in the elevated focal point in the top of Fig. 2, invalidating such solutions. Classical approaches to deal with this problem are to include more data about the true light paths into the problem. This could be achieved by obtaining the intersection points (or equivalently caustic images) at different depths, coding the light paths spatially by projecting different colors [WRHR11], or temporally by projecting varying patterns [MK05], and capturing multiple frames. However, as we do not wish to slow down the underlying production process, for the rest of the paper, we consider the case where only a single image under a given light source can be obtained.

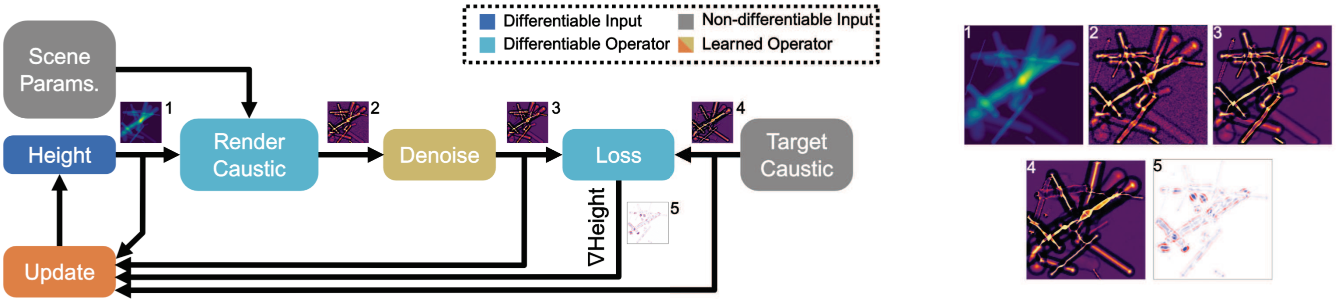

Figure 3: ***Our processing pipeline*** *includes differentiable building blocks and trained components.* `Height`, `Scene` *and* `Target Caustic` *are inputs into the method and the updated* `Height` *is the final output. 1-5 display representative outputs of each step. 1 and 5 are scalar valued and 2-4 are vector valued images with n_w channels. This case displays the L_2-norm over the channel dimension.*

Given those constraints, the modelling has to include a prior on the solution space to disambiguate solutions. The prior can be handcrafted [KBC*21], or learned from data. We opt for the latter as the generating process we consider (LGD) constrains the feature size and makes very small rapidly changing structures like in the reconstruction of Fig. 2 impossible. Furthermore, the integration into such a manufacturing process allows continuous improvement of the prior by adding newly manufactured and measured samples into the dataset, making it an ideal candidate for continuous learning.

3.2. Pipeline Overview

Fig. 3 shows an overview of the processing steps of the proposed method. Starting from an initial guess of the refractor heightfield and other non-differentiable scene parameters, we compute the caustic with a differentiable rendering module [KBC*21] and pass the result through a learned denoising module. This caustic image is then compared to the desired target caustic. After backpropagating the gradient to the initial heightfield through the differentiable renderer, this gradient is passed into an update module along with the initial guess, the denoised caustic image, and the target caustic image, to compute the final adjustment of the heightfield. The whole process is then potentially looped until a pre-defined convergence criterion is met. To be more specific, our non-differentiable scene parameters define the rest of the scene setup, which interact with the light paths. Summarizing these parameters as a vector $\theta \in \mathbb{R}^{n_\theta}$, we define our rendering function as $R : \mathbb{R}^{n\times n} \times \mathbb{R}^{n_\theta} \to \mathbb{R}_+^{n_w \times m \times m}$, where $n \times n$ and $m \times m$ indicate the pixel resolution of the height map and the caustic image respectively, and n_w denotes the number of wavelengths in the simulation. Its output is the wavelength-dependent irradiance E at the sensor plane. Note that, unlike many other image processing tasks, we cannot impose an a priori upper limit on this quantity, since it is largely dependent on the intensity of the light source and focus due to the estimated geometry. Consequently, all subsequent steps need to handle these full dynamic range images. The first of these is the denoising network D, which is an *endofunction*, $D : \mathbb{R}_+^{n_w \times m \times m} \to \mathbb{R}_+^{n_w \times m \times m}$. The intuition behind this component in the context of inverse problems is that usually computing and memory budget is limited when trying to design a fast feedback algorithm. This directly leads to limiting the number of samples n_l in image and gradient computation. However, the target caustic is of a different distribution, because it is either obtained by direct measurement or is a high quality simulation that is largely devoid of noise. Recent work [ZGB21] suggests that unbiased noise in gradient descent optimization is not the limiting factor, when combined with appropriate optimizers, but note that in this case we have additional bias in the simulation, due to using a photon mapping variant [FSES14]. Thus, we make an effort to transform the estimated caustic image into the same distribution as the target caustics under consideration with the denoising component. An important observation is that the adjoint of the denoising operator degrades the gradient signal coming from the loss function and subsequent update steps. Therefore, we exclude the denoiser from the backward pass and replace it with an identity function.

4. Network Architecture

Subsequent processing of caustic images is dependent on two neural-network components, which belong to the same architectural family (see our supplemental). Both networks share a structure similar to UNet [RFB15] as the main component, and differ in a few blocks with respect to the input and output. The denoiser includes a single `Conv + nonlin` block, which expands the number of channels to $c_{init} \in [1,32]$ channels for the UNet part of the network. An equivalent block contracts those channels after the UNet part of the denoising network.

The update network directly starts with the first block of the UNet part of the network, albeit with a fixed number of channels and a different small output network that potentially contracts the channels over multiple steps and integrates information over a larger receptive field. The general update scheme is motivated by the success of learned gradient descent methods [ADG*16, FBD*19] in solving ill-posed parameter estimation problems. The main idea is to replace the classical gradient-based local update rule $x_{i+1} = \mathbf{S}(x_i - \alpha_i \nabla x_i)$, with $\alpha_i > 0$ being the step size for step i and $\mathbf{S}$ being projection operators and further heuristics arising in constrained optimization [KBC*21] with a fully learned update rule: $x_{i+1} = x_i - U(x_i, \nabla x_i)$. This has the advantage that the update network can learn an appropriate prior distribution, thus avoiding unwanted local minima and taking larger adaptive steps. In our specific case the updater is conditioned on the current heightfield as well as its computed gradient and the simulated as well as the target caustic image, scaled to the same spatial resolution ($m = n$). Independent processing of the channels as in the denoiser is not applied here, since these channels are mutually dependent of each other. Thus, we define the updater as: $U : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n_w\times m\times m} \times \mathbb{R}^{n_w\times m\times m} \to \mathbb{R}^{n\times n}$.

Overall we consider a family of networks for both the denoising and the update parts, which are parameterized via the hyperparameters provided in our supplemental. At training time we search for the best network architecture over the parameter space defined therein. We additionally provide PyTorch Lightning [Fe19] code, which includes all implementation details of our models.

The loss function we employ in Fig. 3 is the mean squared error of irradiances and total variation of the estimated heightfield, so the full objective function is

$$\mathbf{L}(h,\hat{E}) = \frac{1}{n_w m^2}\sum_{i=0}^{n_w m^2} (D(R(h,\theta)) - \hat{E})_i^2 + \lambda_{TV}\sum_{i=0}^{n^2}\|\nabla h_i\|_2^2, \qquad (1)$$

where h is the heightfield θ non-differentiable scene parameters, $\hat{E}$ the given irradiance of the target caustic image and λ_{TV} the weighting term for the heightfield regularization. Note that in contrast to our evaluation of the resulting height field error, we do not use a relative error here, since the normalization by the L_2 norm of the ground truth caustic is nothing more than a fixed scaling of the first part of the loss and does not change the optimization, aside from having to re-balance both loss terms by adapting λ_{TV}.

4.1. Datasets

We provide two synthetic datasets in addition to the *test dataset*: the *denoising* and the *updater* datasets. They both are rendered by drawing samples from distributions carefully chosen to closely match the real data distribution, by considering the LGD printing process and the achievable fiber diameters. Instead of real fiber networks, which engineers might design, we sample a large number of height fields to cover a wide range of possible – albeit mostly nonsensical designs – to further accurately simulate the effect of printing a fiber on top of pre-existing ones, which might be the case, when correcting for a printing mistake in a previous layer. Consistently, the physical values for the non-differentiable scene parameters were selected to match usual process values with current technology. All details for the generation of the datasets as well as explored ranges are provided in our supplemental.

For the **denoising dataset** we draw 50000 samples from the heightfield distribution and render 2 caustic images with different quality levels: one with n_l = 10^6 light samples and one with n_l = $1.6\cdot 10^7$ light samples. The first four entries of the dataset are shown in Fig. 4. It is clearly visible that the filament lines create complex caustic patterns when crossing over each other, giving the denoising network many image patches to learn representative and varied caustic patterns.

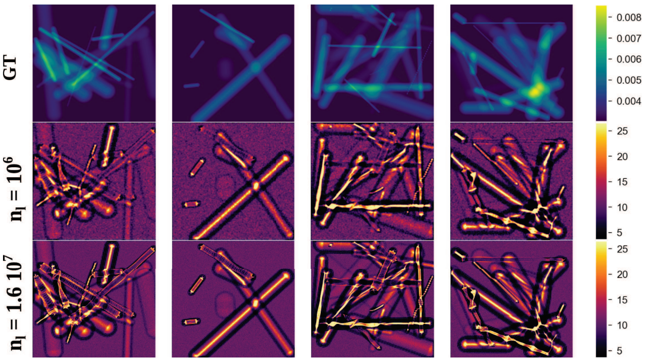

Figure 4: ***Denoising dataset****: we generate training data for our denoiser by sampling random ground-truth heightfields from line distribution (top), which are then used to render low quality (middle) and high quality (bottom) caustics images. During training, we use the low quality renderings as network inputs and the high quality versions for supervision.*

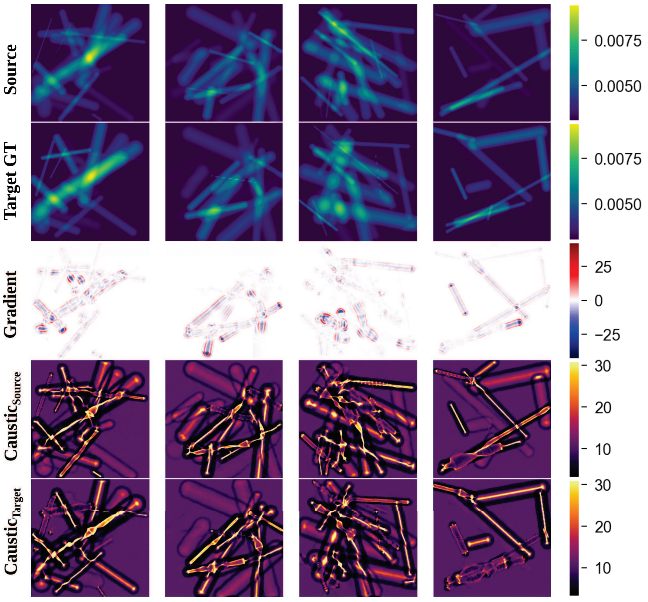

Figure 5: ***Updater dataset****: similar to our denoising dataset, we sample pairs of source and target heightfields to train our updater.*

For the **updater dataset** we sample two heightfields from two distributions. One is used to represent the initial heightfield in the updater network, and the other (an offset of the current estimate) is used as the target heightfield of the updater. We then render caustic images for the current estimated heightfield and the target heightfield and use the previously trained denoiser to produce respective caustic images. We further compute the MSE loss between the two to generate the gradient with respect to the currently estimated heightfield. The current heightfield and its caustic, the target heightfield and its caustic and the heightfield gradient are then saved for this dataset. We generated 100000 samples using this procedure, from which the first four can be seen in Fig. 5.

Lastly, we provide a **test dataset** with 10 samples. One with hand-picked lines in the same distribution as the training data , but not present in any training data set. This sample serves as direct comparison with SfC [KBC*21], where it acted as the main test sample. Our test set also includes six further samples with varying

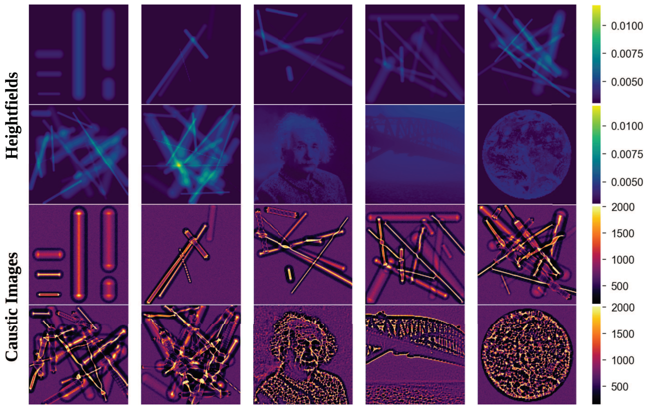

Figure 6: ***Test set**: we evaluate our framework on a dedicated test set containing* 10 *heightfields of varying complexity.*

complexity, *i.e.* with 5 to 30 random lines from the same distribution as the training data, but not present in any training data set. Finally, we include 3 creative commons gray-scale images converted into heightfields and scaled to the same value range as the training data as out-of distribution samples. The samples and resulting high-quality ($1.6 \cdot 10^7$ light paths) caustic images are shown in Fig. 6.

4.2. Evaluation and Metrics

We compare our method against SfC [KBC*21] and (HCCCD) [STTP14]. All reconstructions are initialized with a constant heightfield of $d = 3$mm. When comparing the accuracy of different reconstructions, we report the *relative* heightfield L_2 error: $\mathbf{L_{rel}}(h,\hat{h}) = \frac{\|h-\hat{h}\|_2}{\|\hat{h}\|_2}$, where h is the reconstruction result and $\hat{h}$ is the ground-truth heightfield. We chose this metric because it is independent of the absolute height of the ground-truth samples and thus allows for a fair comparison of different methods. Furthermore, we chose to not report perceptual metrics like SSIM [WBSS04], or PSNR because they are not suited for comparing heightfields in the context of production feedback.

As far as parameters are concerned, N-SfC is parameter-free at test-time, whereas SfC depends on several hyperparameters, namely α_p, τ_p and γ for the variant with extended thresholded nonlinear Landweber scheme and volume heuristic (called M2V1 in [KBC*21]). Analogous to the network parameter search we optimize these parameters on the first test sample and leave them constant for the remaining samples. We do the same for the reconstruction parameters for HCCCD.

With respect to evaluating our method for iterated application, the reconstruction quality depends on the number of update steps and thus the choice of the stopping criterion. We compare four choices for the stopping criterion of SfC and N-SfC, while for HCCCD only the fully converged solution is available: L_{rel} after one update step *(iter. 1)*, for cases, where the reconstruction is severely time-constrained; L_{rel} after 10 update steps *(iter. 10)*, for a balance between reconstruction quality and time. L_{rel} after full convergence of the optimization *(conv.)*; *i.e.* when time is no constraint. L_{rel} after full convergence of the system, but reported at the time, when the optimization has reached the minimal caustic image error, wrt. the MSE *(crit.)*. The names in brackets refer to the abbreviation used in Tabs. 1 and 2. The last criterion *(crit.)* is of particular interest, because it highlights the difference between caustic design methods, such as HCCCD and surface reconstruction methods as well as the feasibility of using the caustic image as a sole supervision signal.

4.3. Results and Comparisons

We present renderings of reconstruction results in Fig. 7, where we increased the distance between screen and substrate for better visibility. A full overview of numerical errors on the test dataset is given in Tab. 1. From those results it is clear that SfC [KBC*21] fails to converge to a good reconstruction and is outperformed by N-SfC under all considered criteria, only managing to slightly improve upon the initial guess after with the *iter.1* and *crit.* criteria. We hypothesize that this is due to reduced smoothing of and subsequently sharper and higher-dynamic range caustic images of our setup, when compared to the original SfC paper. In contrast N-SfC can smoothly decrease the shape error by repeated execution of update steps as further illustrated in Fig. 8, which shows the convergence averages and standard deviations of $\mathbf{L_{rel}}$ over 50 iterations. Despite not being trained in a recurrent manner, the updater seems to have learned a good representation of gradient descent dynamics, since for the more complex samples the error continues to decrease as displayed in the *conv.* and *crit.* columns in Tab. 1. However, this comes at the cost of significantly increased runtime and lower performance in for simpler samples, such that those criteria can not be recommended for practical use. With respect to HCCCD [STTP14] we can see that it generally outperforms SfC in most cases, the exception being SfC after 1 iteration in $\mathbf{L_{rel}}$ metric. However, on average and in most samples it is still outperformed by our full N-SfC method. Lastly we wish to address runtimes of our method compared to SfC and HCCCD. This is not a trivial comparison, since the former failed to converge on our test cases and the latter is only present as a CPU implementation. However, the authors of SfC reported note a runtime of 2.95 min for 72 iterations on a heightfield, which is structurally very similar to our first test set sample, albeit with different scene parameters, such as increased smoothing. All samples, when reconstructed with HCCCD took more than 15 minutes with several hundred iterations each, which would not benefit much

Table 1: ***Comparisons on the relative heightfield reconstruction errors*** *of our method with several stopping criteria against SfC [KBC*21] and HCCCD [STTP14], according to* $\mathbf{L_{rel}}$. *Green highlights absolute per-row minima, while red highlights errors, which exceed the the error of the initial (flat heightfield) guess.*

Test Img.	Initial Error	HCCCD	SfC				N-SfC			
			iter. 1	iter. 10	conv.	crit.	iter. 1	iter. 10	conv.	crit.
1	0.1954	0.1549	0.1918	0.4273	0.4271	0.1917	0.1652	0.1104	0.1448	0.1447
2	0.1102	0.0890	0.0991	0.4844	0.4869	0.0991	0.1003	0.1118	0.1536	0.1344
3	0.1477	0.1244	0.1333	0.4599	0.4604	0.1333	0.1225	0.1182	0.1592	0.1589
4	0.1538	0.1280	0.1379	0.4195	0.4201	0.1379	0.1285	0.1260	0.1710	0.1710
5	0.2770	0.2588	0.2613	0.4055	0.4068	0.2613	0.2419	0.1477	0.1576	0.1609
6	0.3061	0.2835	0.2887	0.3835	0.3855	0.2887	0.2640	0.1497	0.1660	0.1510
7	0.3984	0.3845	0.3832	0.3955	0.3939	0.3832	0.3578	0.2103	0.1870	0.3578
8	0.2346	0.2269	0.2133	0.3467	0.3455	0.2133	0.2143	0.1281	0.1104	0.1151
9	0.2823	0.2791	0.2700	0.2955	0.2940	0.2700	0.2701	0.1972	0.1711	0.1711
10	0.2372	0.2219	0.2081	0.3697	0.3680	0.2081	0.2109	0.1278	0.1304	0.1302
Avg.	0.2343	0.2151	0.2187	0.3989	0.3988	0.2187	0.2076	0.1427	0.1551	0.1695

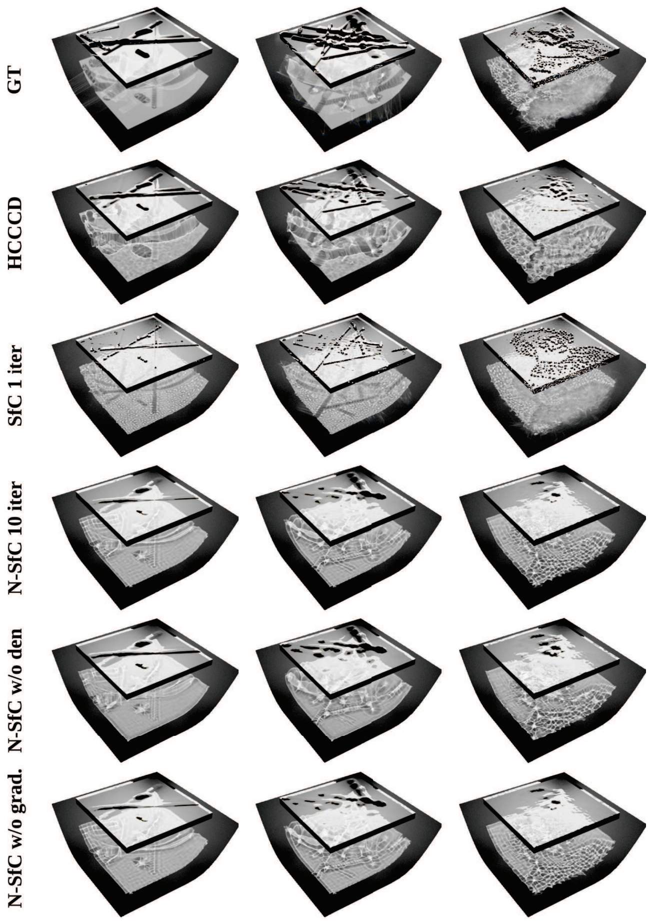

Figure 7: ***Qualitative Reconstruction results.*** *Please note, that while SfC may at times seem to more faithful caustics, the main point of comparison is the height field, as this is the design variable, which is supposed to be measured in a production environment (Sec. 3).*

from GPU acceleration due to its sequential nature. We however created all of our reconstructions in only one to ten iterations with an average iteration time of 1.6 sec per iteration, which is an improvement factor of 11 to 111 in speed compared to SfC and a factor of 56 to 560 compared to HCCCD for the iter. 1 and iter. 10 criteria. This advantage diminishes for the *conv.* and *crit.* criteria, since the runtime is dominated by the up to thousands of iterations required for full caustic image convergence.

4.4. Ablation Study

In the following, we ablate parts of our processing to illustrate the influence of single components. We trained a variant of our best performing updater network with the direct output from our render module, leaving out the denoiser. We denote this variant as *N-SfC w/o den.* in Tab. 2. While, at first glance, leaving out denoising does not significantly impair the reconstruction, and even outperforms our full model on a quantitative basis in the best performing *iter. 10* criterion, we find that this model converges slower with a higher overall error and variance, as evident by the errors *conv.* criterion and Fig. 8. In contrast, using our denoiser allows for faster, smoother and more robust convergence over the entire test set under more convergence criteria settings.

Table 2: ***Ablations*** *on the relative heightfield reconstruction errors of our method with several stopping criteria, according to* $\mathbf{L_{rel}}$. *Green highlights improvement over and red highlights decreases over the full method with the same stopping criterion as Tab. 1.*

Test	N-SfC w/o den.				N-SfC w/o grad.			
Img.	iter. 1	iter. 10	conv.	crit.	iter. 1	iter. 10	conv.	crit.
1	0.1618	0.0696	0.2021	0.0833	0.1633	0.1154	0.1456	0.1456
2	0.1025	0.0904	0.0921	0.0923	0.0996	0.1213	0.1559	0.1558
3	0.1229	0.0874	0.0989	0.0989	0.1211	0.1262	0.1611	0.1611
4	0.1318	0.1078	0.2811	0.1328	0.1266	0.1315	0.1706	0.1709
5	0.2407	0.1303	0.2510	0.1663	0.2400	0.1495	0.1576	0.1668
6	0.2612	0.1382	0.2682	0.1487	0.2618	0.1505	0.1630	0.1575
7	0.3545	0.1917	0.2344	0.3545	0.3558	0.2101	0.1892	0.3558
8	0.2198	0.1527	0.1622	0.1342	0.2121	0.1241	0.1095	0.1130
9	0.2781	0.2586	0.2448	0.2448	0.2679	0.1888	0.1696	0.1696
10	0.2144	0.1191	0.2570	0.1114	0.2087	0.1272	0.1305	0.1304
Avg.	0.2088	0.1346	0.2092	0.1567	0.2057	0.1445	0.1553	0.1727

Another ablation we performed is training an updater without the guidance of our differentiable renderer to assess the impact of local gradient information on the final update steps predictions. We denote this variant as *N-SfC w/o grad*. The quantitative evaluation in Tab. 2 yields that withholding this information leads to less accurate reconstructions in the majority of test samples with the notable exception of the *iter. 1* criterion, however, the increase in error is not major and in some cases even leads to better results.

From the above experiments we observe the following: Firstly, our method can be trained and executed without the presence of a differentiable renderer, only the forward caustic image simulation is strictly necessary. Secondly, to get the best possible reconstruction results, gradients from a differentiable renderer help in most cases, though they only contribute very local information (see Sec. 3.1). Thirdly, using a dedicated denoiser in the loop increases robustness by decreasing the convergence variance caused by noise in Monte Carlo simulation. Lastly, training the learned gradient descent updater in a recurrent manner is not necessary for a well-behaved and converging updater, which finds better global minima than purely local gradient-based search.

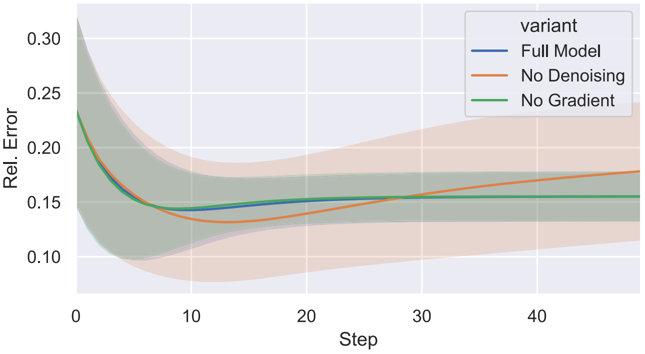

Figure 8: ***Mean Convergence*** *plot of* $\mathbf{L_{rel}}$ *and standard deviation for iterated application of model variants over the test set.*

4.5. Real-World

We further performed an experiment to validate our method on a real-world glass sample in a prototypical setup. The capture system is composed of simple torch as light source as well as a Fresnel lens to collimate the beam, the sample and finally the camera as depicted in Fig. 9 (top left). Since this setup is sensitive to slight misalignments we employ an optical diffusor (a sheet of paper) after the test specimen to isolate the caustic from the background. The resulting caustic measurement is shown in Fig. 9 (bottom left) and depicts a branching glass fiber of about 2.75 mm width and 0.85 mm height over the base substrate of 6 mm thickness. A macro photograph of the glass surface itself is depicted on Fig. 9 (top right), where a small fiber branching off is better visible. As can be seen in Fig. 9 (bottom right), the main fiber structure is evident, however the secondary branch is not fully formed, likely due to the reduced signal in the caustic image from flatness deviations in the optical diffusor. The reconstructed fiber width of 2.83 mm of the main fiber closely matches the true fiber width, however, the maximal fiber height is overestimated with a center height of 1.5 to 2.55 mm, which is likely due to the sample having a base substrate thickness that is double those in our training set. We estimate that this can be overcome by training our model on a dataset of thicker base substrates. Despite the challenges this out-of-distribution sample poses, it demonstrates that our method is a good indicator for such feedback applications as detecting a malfunctioning fiber feeder.

5. Discussion and Limitations

The above experiments show that our method outperforms the current state-of-the art by a significant margin. At the same time, our current implementation and datasets make several assumptions on the concrete area of application, thus implying some implicit limitations. As previously mentioned, our fully trained model is completely free of any additional parameters, which avoids the need for manual hyperparameter tuning, including the choice of step sizes. However the applied training data and learning process are adjusted towards the physical scene configuration (such as the light and screen positions) of our *in-situ* manufacturing setup. Generating more diverse training data and providing the networks with explicit knowledge about certain scene parameters could help to further improve the generalizability of our model.

In the ablation study, we found that passing the gradient from a differentiable rendering module only marginally improves reconstruction quality, which is in line with the literature (see the work of Morris & Kutulakos [MK05] and Fig. 2), stating that local information can be counterproductive for settings like ours. The denoiser however was a critical part of reducing variance in the reconstruction error and thus improving robustness of the model.

Lastly, our reconstructions usually reach minimal shape error before converging to a solution with slightly higher overall error as depicted in Fig. 8, but these points strongly depends on the currently considered sample. Thus another supervision mechanism, which limits the number of steps to achieve the best trade-off between reconstruction quality and number of necessary steps would be desirable in the context of time-limited feedback for production loops.

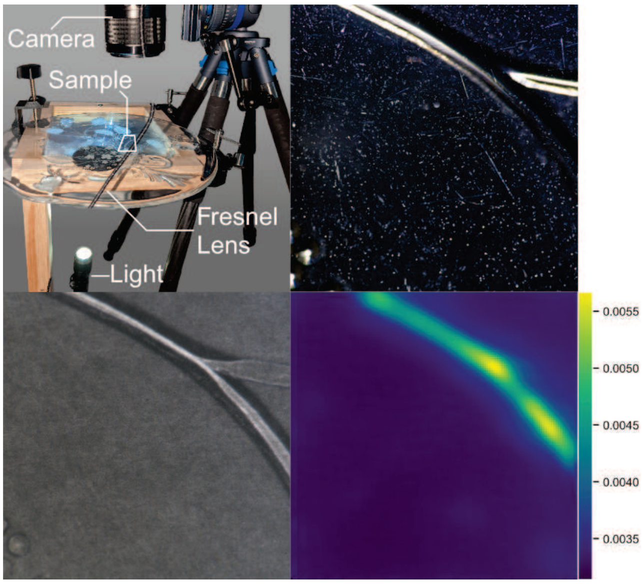

Figure 9: ***Real-World*** *setup, macro photo of sample area, captured caustic image, and resulting reconstructed heightfield.*

6. Conclusion

We presented Neural-Shape from Caustics (N-SfC), a fast and flexible method for reconstructing the shape of translucent objects from a single caustic image. We combine recent work on differentiable light transport simulation with our novel neural denoising components and learned gradient descent optimizer to significantly improve both the stability and quality of the iterative reconstruction process. Our quantitative and qualitative analysis showed that our neural approach outperforms current state-of-the-art approaches in terms of final reconstruction error and compute requirements. Furthermore, we found that our learned gradient-based update scheme enables better generalization and overall flexibility, making the approach adaptable to practical applications such as integrated quality feedback control for glass manufacturing processes.

Acknowledgments

The authors gratefully acknowledge funding by the DFG under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122, Project ID 390833453).

References

[ADG*16] ANDRYCHOWICZ M., DENIL M., GOMEZ S., HOFFMAN M. W., PFAU D., SCHAUL T., SHILLINGFORD B., DE FREITAS N.: Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems* (2016), pp. 3981–3989. 2, 4

[Aut21] AUTODESK: Denoiser, 2021. Online; Last accessed 15-06-2023; https://docs.arnoldrenderer.com/display/A5AF3DSUG/Denoiser#Denoiser-OptiXDenoiser. 2

[BLD20] BANGARU S., LI T.-M., DURAND F.: Unbiased warped-area sampling for differentiable rendering. *ACM Transactions on Graphics 39*, 6 (2020), 245:1–245:18. 2

[FBD*19] Flynn J., Broxton M., Debevec P., DuVall M., Fyffe G., Overbeck R., Snavely N., Tucker R.: Deepview: View synthesis with learned gradient descent. In *IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 2367–2376. 2, 4

[Fe19] Falcon W., *et al.*: Pytorch lightning, 2019. Software available from pytorchlightning.ai. 5

[FFL16] Feng Z., Froese B. D., Liang R.: Freeform illumination optics construction following an optimal transport map. *Applied Optics 55*, 16 (2016), 4301–4306. 2

[FSES14] Frisvad J., Schjøth L., Erleben K., Sporring J.: Photon differential splatting for rendering caustics. *Computer Graphics Forum 33*, 6 (2014), 252–263. 2, 4

[HY21] Huo Y., Yoon S.: A survey on deep learning-based Monte Carlo denoising. *Computational Visual Media 7*, 2 (2021), 169–185. 2

[HYC01] Hochreiter S., Younger A. S., Conwell P. R.: Learning to learn using gradient descent. In *Artificial Neural Networks* (2001), Springer, pp. 87–94. 2

[JK21] Jiang G., Kainz B.: Deep radiance caching: Convolutional autoencoders deeper in ray tracing. *Computers & Graphics 94* (2021), 22–31. 2

[JSRV22] Jakob W., Speierer S., Roussel N., Vicini D.: DR.JIT: A just-in-time compiler for differentiable rendering. *ACM Transactions on Graphics 41*, 4 (2022). 2

[KB15] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. In *International Conference on Learning Representations* (2015). 3

[KBC*21] Kassubeck M., Bürgel F., Castillo S., Stiller S., Magnor M.: Shape from caustics: Reconstruction of 3D-printed glass from simulated caustic images. In *EEE/CVF Winter Conference on Applications of Computer Vision* (2021), pp. 2877–2886. 1, 2, 3, 4, 5, 6

[KBS15] Kalantari N. K., Bako S., Sen P.: A machine learning approach for filtering Monte Carlo noise. *ACM Transactions on Graphics 34*, 4 (2015), 122–1. 2

[LADL18] Li T.-M., Aittala M., Durand F., Lehtinen J.: Differentiable Monte Carlo ray tracing through edge sampling. *ACM Transactions on Graphics 37*, 6 (2018). 2

[LWL*20] Lyu J., Wu B., Lischinski D., Cohen-Or D., Huang H.: Differentiable refraction-tracing for mesh reconstruction of transparent objects. *ACM Transactions on Graphics 39*, 6 (2020). 2

[LZBD21] Luan F., Zhao S., Bala K., Dong Z.: Unified shape and svbrdf recovery using differentiable Monte Carlo rendering. *Computer Graphics Forum 40*, 4 (2021), 101–113. 2, 3

[ME21] Mousavi M., Estrada R.: SuperCaustics: Real-time, open-source simulation of transparent objects for deep learning applications. In *International Conference on Machine Learning and Applications* (2021), IEEE, pp. 649–655. 2

[MK05] Morris N., Kutulakos K.: Dynamic refraction stereo. In *International Conference on Computer Vision* (2005), vol. 2, pp. 1573–1580. 3, 8

[MMT18] Meyron J., Mérigot Q., Thibert B.: Light in power: a general and parameter-free algorithm for caustic design. *ACM Transactions on Graphics 37*, 6 (2018), 1–13. 2

[NAM*17] Nalbach O., Arabadzhiyska E., Mehta D., Seidel H.-P., Ritschel T.: Deep shading: Convolutional neural networks for screen space shading. *Computer Graphics Forum 36*, 4 (2017), 65–78. 2

[NDVZJ19] Nimier-David M., Vicini D., Zeltner T., Jakob W.: Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics 38*, 6 (2019). 2

[PGM*19] Paszke A., Gross S., Massa F., Lerer A., *et al.*: Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (2019). 3

[RFB15] Ronneberger O., Fischer P., Brox T.: U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (2015), Springer, pp. 234–241. 4

[RM02] Ries H., Muschaweck J.: Tailored freeform optical surfaces. *Journal of the Optical Society of America 19*, 3 (2002), 590–595. 2

[Sch93] Schmidhuber J.: A neural network that embeds its own meta-levels. In *IEEE International Conference on Neural Networks* (1993), vol. 1, pp. 407–412. 2

[STTP14] Schwartzburg Y., Testuz R., Tagliasacchi A., Pauly M.: High-contrast computational caustic design. *ACM Transactions on Graphics 33*, 4 (2014). 2, 6

[VRM*18] Vogels T., Rousselle F., McWilliams B., Röthlin G., Harvill A., Adler D., Meyer M., Novák J.: Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics 37*, 4 (2018), 1–15. 2

[VSJ21] Vicini D., Speierer S., Jakob W.: Path replay backpropagation: Differentiating light paths using constant memory and linear time. *ACM Transactions on Graphics 40*, 4 (2021). 2

[WBSS04] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612. 6

[WRHR11] Wetzstein G., Roodnick D., Heidrich W., Raskar R.: Refractive shape from light field distortion. In *International Conference on Computer Vision* (2011), pp. 1180–1186. 3

[WXL*13] Wu R., Xu L., Liu P., Zhang Y., Zheng Z., Li H., Liu X.: Freeform illumination design: a nonlinear boundary problem for the elliptic Monge–Ampére equation. *Optics Letters 38*, 2 (2013), 229–231. 2

[YHC01] Younger A., Hochreiter S., Conwell P.: Meta-learning with backpropagation. In *International Joint Conference on Neural Networks* (2001), vol. 3, pp. 2001–2006. 2

[ZGB21] Zhao S., Gkioulekas I., Bangaru S.: CVPR tutorial on physics-based differentiable rendering, 2021. 4

[ZMY*20] Zhang C., Miller B., Yan K., Gkioulekas I., Zhao S.: Path-space differentiable rendering. *ACM Transactions on Graphics 39*, 4 (2020). 2

[ZWW*20] Zeng Z., Wang L., Wang B.-B., Kang C.-M., Xu Y.-N.: Denoising stochastic progressive photon mapping renderings using a multi-residual network. *Journal of Computer Science and Technology 35* (2020), 506–521. 2

[ZXJ*20] Zhu S., Xu Z., Jensen H. W., Su H., Ramamoorthi R.: Deep kernel density estimation for photon mapping. *Computer Graphics Forum 39*, 4 (2020), 35–45. 2

[ZYZ21] Zhang C., Yu Z., Zhao S.: Path-space differentiable rendering of participating media. *ACM Transactions on Graphics 40*, 4 (2021). 2

Proceedings published by Eurographics - The European Association for Computer Graphics.