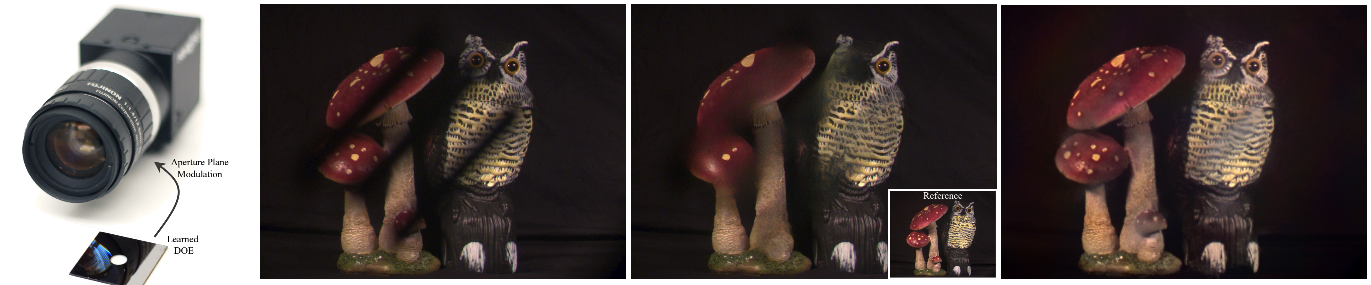
Seeing Through Obstructions with Diffractive Cloaking

ZHENG SHI, YUVAL BAHAT, SEUNG-HWAN BAEK, Princeton University
QIANG FU, HADI AMATA, King Abdullah University of Science and Technology
XIAO LI, PRANEETH CHAKRAVARTHULA, Princeton University
WOLFGANG HEIDRICH, King Abdullah University of Science and Technology
FELIX HEIDE, Princeton University



Occluded Conventional Capture — Obstruction Inpainting (LaMa [2022]) — Proposed Optical Diffraction Cloaking

Fig. 1. We propose a computational monocular camera that optically cloaks unwanted obstructions, such as raindrops or dirt stains on lens cover glass, or a fence near the camera. Instead of inpainting the obstructed information post-capture, we learn a custom diffractive optical element that, when placed in front of the camera lens, acts as a depth-dependent scatterer. The learned optical element sits on the aperture plane of an existing camera setup, and produces a point spread function with large spatial support for close objects, scattering light away from what would otherwise be a focal spot. All while preserving spatial resolution at long distances without increasing the camera footprint. In conjunction with the optical element, we jointly optimize a feature-based deep learning reconstruction network to recover the unobstructed image.

Unwanted camera obstruction can severely degrade captured images, including both scene occluders near the camera and partial occlusions of the camera cover glass. Such occlusions can cause catastrophic failures for various scene understanding tasks such as semantic segmentation, object detection, and depth estimation. Existing camera arrays capture multiple redundant views of a scene to see around thin occlusions. Such multi-camera systems effectively form a large synthetic aperture, which can suppress nearby occluders with a large defocus blur, but significantly increase the overall form factor of the imaging setup. In this work, we propose a *monocular single-shot* imaging approach that *optically cloaks* obstructions by emulating a large array. Instead of relying on different camera views, we learn a diffractive optical element (DOE) that performs depth-dependent optical encoding, scattering nearby occlusions while allowing paraxial wavefronts to be focused. We computationally reconstruct unobstructed images from these superposed measurements with a neural network that is trained jointly with the optical layer of the proposed imaging system. We assess the proposed method in simulation and with an experimental prototype, validating that the proposed computational camera is capable of recovering occluded scene information in the presence of severe camera obstruction.

Authors' addresses: Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Princeton University; Qiang Fu, Hadi Amata, King Abdullah University of Science and Technology; Xiao Li, Praneeth Chakravarthula, Princeton University; Wolfgang Heidrich, King Abdullah University of Science and Technology; Felix Heide, Princeton University.

ACM Reference Format:
Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2021. Seeing Through Obstructions with Diffractive Cloaking . *ACM Trans. Graph.* 39, 4, Article 1 (July 2021), 15 pages. https://doi.org/http://dx.doi.org/10.1145/8888888.7777777

1 INTRODUCTION

Cameras can often be subjected to non-ideal environmental conditions, resulting in obstructions that are detrimental to photography and computer vision tasks. For example, in the case of automotive vehicles and robotics, dirt from road debris, insects, raindrops or even spray from close-by vehicles can accumulate on the windshield or camera cover glass. Beyond reducing photographic quality for human viewers, this can result in catastrophic failure of downstream scene understanding tasks [Gaylard et al. 2017; Hnewa and Radha 2020]. Active cleaning approaches [Ficosa 2017; Monrad 2017; Orlaco 2013; S3 2016] which use nozzles, wipers, or spinning cover glasses to remove obstructions are often inadequate, especially in the case of strong obstructions (e.g., insects stuck to the windshield) or when obstructions are part of the scene itself (e.g., a wiper).

Existing approaches to this challenge rely on increasing the aperture of the imaging system. This enables the imaging system to see through unwanted obstructions by capturing a larger portion of the incident light field that is unobstructed. The obstructions can be either explicitly detected and suppressed for a downstream task [Uricar et al. 2021], or a light field can be reconstructed from all available viewing angles, making it possible to recover an unobstructed refocused image [Wilburn et al. 2005]. Using a camera with a very large physical aperture is limited to specific domains, such as telescopic imaging and self-driving cars, where sensor stacks can span the entire roof [Caesar et al. 2020; Waymo 2019]. In its extreme form, this is known as a synthetic aperture array [Vaish et al. 2006; Wilburn et al. 2005]. While camera arrays are effective at handling partial occlusions, they come at the cost of a large form factor, prohibiting applications that rely on a single camera.

Restricting camera systems to smaller form factors, researchers have investigated machine learning for obstruction removal using captures from single-camera hand-held devices, robotics, or automotive imaging systems. This often requires multiple spatially or temporally varying captures [Li et al. 2021; Liu et al. 2020; Xue et al. 2015], mandating static scenes. Also, such approaches often fail on test inputs deviating from the training input distributions. Single-image inpainting approaches that aim to recover the latent obstructed image regions [Farid et al. 2016; Gupta et al. 2021; Qian et al. 2018; Yang et al. 2021; Yu et al. 2019] do not constitute a valid alternative, as they often produce fictitious hallucinations for large occluded regions.

In this work, we introduce a monocular single-shot obstruction-free imaging method using a single camera combined with a diffractive optical element (DOE) and a computational reconstruction process. Instead of capturing diverse view angles in an array configuration, we introduce a diffractive optical element acting as a depth-dependent diffuser to increase the angle of the incident light cone per-pixel. Conventional refractive lens stacks restrict this light cone by their smooth surface profile. Instead, we design a DOE that produces strong depth-dependent aberrations, allowing us to generate PSFs that can optically disambiguate foreground obstruction from background scenes. Specifically, we introduce a differentiable occlusion-aware forward model and learn an optical assembly such that paraxial wavefronts from the background scenes can arrive at the imaging sensor, whereas the light from foreground obstructions is diffused. The proposed method does not increase the form factor, and the single-shot capability allows dynamic scene applications. We recover unoccluded images with a reconstruction network learned in an end-to-end fashion. Our physics-aware reconstruction network utilizes the PSFs of the DOE to better generalize to unseen test cases. We assess the proposed approach both in simulation and with an experimental prototype, validating that our method can see through obstructions and recover background content where *all* compared methods cannot.

Specifically, we make the following contributions:

- We introduce a monocular single-shot imaging method that recovers an unobstructed scene with a learned DOE, enabling imaging through occlusions without the need for inpainting.
- We learn the proposed DOE using end-to-end optimization, relying on a differentiable depth- and obstruction-aware image formation model and a physics-based reconstruction network.
- We validate the proposed method in simulation and with an experimental prototype, confirming that the method is capable of recovering image details lost when using conventional methods.

Optimized lens designs, network checkpoints, fabrication details, and all code needed to reproduce the results presented in the manuscript are available under https://light.princeton.edu/seeing-through-obstructions.

Limitations. We design the proposed DOE assuming narrow band RGB input illumination. Although finer wavelength sampling may further improve the generalization capability to in-the-wild scenes, the computational efficiency and memory requirements of multi-spectral DOE training would exceed the resources available to us. Simulated multispectral datasets and training infrastructure with an order of magnitude larger memory may lift this limitation in the future. Existing inference hardware allows us to achieve real-time performance and low latencies, albeit requiring high-power GPUs for high sensor resolutions. These are impractical for low-power consumer applications. In the future, efficient implementation of the proposed reconstruction method on FPGAs or custom ASICs may allow for fast inference on edge devices.

2 RELATED WORK

Obstruction-free Light-field Imaging. Light field camera arrays have been investigated to a great extent by researchers and practioners [Isaksen et al. 2000; Pei et al. 2013; Vaish et al. 2005, 2006, 2004; Wang et al. 2020; Wilburn et al. 2005; Xiao et al. 2017; Yang et al. 2014]. With synthetic aperture refocusing [Isaksen et al. 2000; Vaish et al. 2005], it is possible to see through foreground obstructions and reconstruct occluded background objects. The light field sub-aperture images facilitate reliable obstruction segmentation and background estimation, using techniques such as entropy cost [Vaish et al. 2006, 2004], stereo matching [Pei et al. 2013], handcrafted feature extraction [Xiao et al. 2017; Yang et al. 2014], or deep neural networks [Wang et al. 2020]. While light field refocusing can effectively suppress obstructions, it requires a large form factor, which limits its applicability.

Multi-frame Obstruction Removal. Using small devices such as smart-phone cameras and single-camera automotive imaging systems, multiple spatial or temporal varying captures are typically needed to robustly identify obstructions and reconstruct background. Existing depth-based approaches identify foreground obstructions by estimating scene depth using multi-view stereo [Liu et al. 2020; Xue et al. 2015] and depth from focus [Yamashita et al. 2010]. These methods separate foreground obstructions from the background scene based on parallax, and the latent background can be recovered since pixels occluded in one viewpoint are likely to be visible in other viewpoints. Yamashita et al. [2010] perform fence detection using three multi-focus images. While avoiding multiple cameras, such approaches still require multiple captures, which prohibits the use in highly dynamic environments. Handcrafted features [You et al. 2013] as well as convolutional neural networks [Jonna et al. 2015, 2016] have been proposed to detect common obstructions such as raindrops and fences from the video data. Li et al. [2021] additionally utilize a recurrent structure to better leverage temporal cues. While video-based obstruction removal fundamentally requires at least a handful of captures to recover a single frame, the proposed method is a single-shot approach that relies on a custom optical stack.

Image Inpainting of Obstructions. With only one capture, single image inpainting methods fill the occluded area to restore a visually pleasing image by relying on nearby visible image content, and global image priors. Assuming an obstruction mask is given, this problem becomes a generic image inpainting problem. Yan et al. [2018] propose a copy-and-paste approach to fill in the gaps using similar patterns from the same image or from a large image database. The quality of such conventional image inpainting methods heavily depends on the handcrafted distance metrics for the search algorithm, which often fails to generalize to out-of-distribution test cases. To perform automatic obstruction detection, earlier methods assume regular or near-regular obstructions [Farid et al. 2016; Liu et al. 2008]. More recently, deep learning methods have been proposed for more robust detection [Gupta et al. 2021; Hao et al. 2019; Qian et al. 2018]. Such methods usually focus on a particular type of obstruction to achieve robust obstruction detection, and struggle with complex scene structure. All single image inpainting approaches, including state-of-the-art generic inpainting methods [Suvorov et al. 2021], are inherently ill-posed due to the lack of information from the occluded background. As a result, they often produce unrealistic hallucinations, especially in the presence of a domain gap or large occlusions.

Optical Cloaking Devices. A large body of work has investigated optical devices that aim to optically cloak portions of a scene. Cloaking devices typically rely on optical assemblies in the scene that redirect light around the region one aims to "hide" from an observer at a specific viewing position [Howell et al. 2014]. While perfect optical cloaking in broad-band from all viewing positions and with all polarization states is an open problem, existing methods [Cai et al. 2007; Chen et al. 2013, 2011; Ergin et al. 2010; Jiang et al. 2020] have demonstrated cloaking at diverse wavelengths, hiding objects under a surface [Ergin et al. 2010], cloaking large objects [Chen et al. 2011], cloaking in incoherent natural light [Chen et al. 2013] and 3D cloaking for all polarization and azimuthal angles of incident waves [Jiang et al. 2020]. Departing from existing works on optical cloaking, which only hide objects within the cloaking device, we propose to learn an optical element as part of the camera lens, which hides objects in front of the camera, in conjunction with a computational reconstruction method.

Differentiable Optics Design. Conventional imaging systems are typically designed in a sequential approach, where lens and sensors are hand-engineered based on specific metrics, such as PSF spot size or dynamic range, independent of the downstream camera task. Departing from this conventional design approach, a large body of work in computational imaging has explored jointly optimizing the optics and reconstruction algorithms, with successful applications in color image restoration [Chakrabarti 2016; Peng et al. 2019], microscopy [Horstmeyer et al. 2017; Kellman et al. 2019; Nehme et al. 2020; Shechtman et al. 2016], monocular depth imaging [Chang and Wetzstein 2019; Haim et al. 2018; He et al. 2018; Wu et al. 2019], super-resolution and extended depth of field [Sitzmann et al. 2018; Sun et al. 2021], time-of-flight imaging [Chugunov et al. 2021; Marco et al. 2017; Su et al. 2018], high-dynamic range imaging [Metzler et al. 2019; Sun et al. 2020], active-stereo imaging [Baek and Heide 2021], hyperspectral imaging [Baek et al. 2021], and other computer vision tasks [Tseng et al. 2021b].

We propose an end-to-end optimization method for imaging through nearby occluders. Drawing inspiration from depth-encoding PSF designs in microscopy [Gustavsson et al. 2018; Pavani and Piestun 2008], we optimize a strongly depth-dependent diffractive optic element that, instead of preserving 3D information, destroys information by behaving as a scattering layer at selected distances where occluders may be present. The ample design space of DOEs allows for rich optical encodings but has the unintended consequence of being challenging to optimize for spatially large DOEs, as quadratic phase profile simulation imposes additional requirements on sampling rate [Cottrell et al. 1990]. To handle the resulting computational load and lift the sampling rate constraints, we rely on an effective far-field approximation to estimate the joint effect of the quadratic phase profile of a refractive lens and the near-field Fresnel propagation. This allows us to design large DOE patterns with depth-dependent PSFs together with a PSF-aware feature-based deconvolution method.

3 LEARNING TO SEE THROUGH OBSTRUCTIONS

To learn a computational camera that is capable of seeing through nearby obstructions, we propose a differentiable obstruction-aware image formation model, which is illustrated in Fig. 2. This image formation model comprises a DOE suitable for seeing through obstruction, followed by a computational reconstruction procedure in the form of a PSF-aware deep neural network. We describe both components in the following sections.

3.1 Obstruction-aware Camera Image Formation

We model the obstructions and the latent target scene as located in two depth layers, with the target scene at optical infinity (i.e., ≥ 5m) and the obstruction lying on a plane situated closer to the aperture, i.e. with distance smaller than optical infinity. With obstructions and target scene in foreground and background, respectively, simulating the image formation requires computing depth-dependent light transport. To this end, we rely on a depth-dependent PSF model and an occlusion-aware image convolution method for realistic observations in the presence of occlusions. We next describe the depth-dependent PSF design of the proposed system and the resulting image formation model.

Depth-dependent PSF. The proposed optical system consists of a DOE and a refractive lens, followed by an intensity sensor. The point spread function (PSF) of such an imaging system can be obtained efficiently using wave optics with a set of approximations. Specifically, we consider the light wave emitted from a scene point (x, y, z) located at depth z with respect to the DOE plane. The propagating

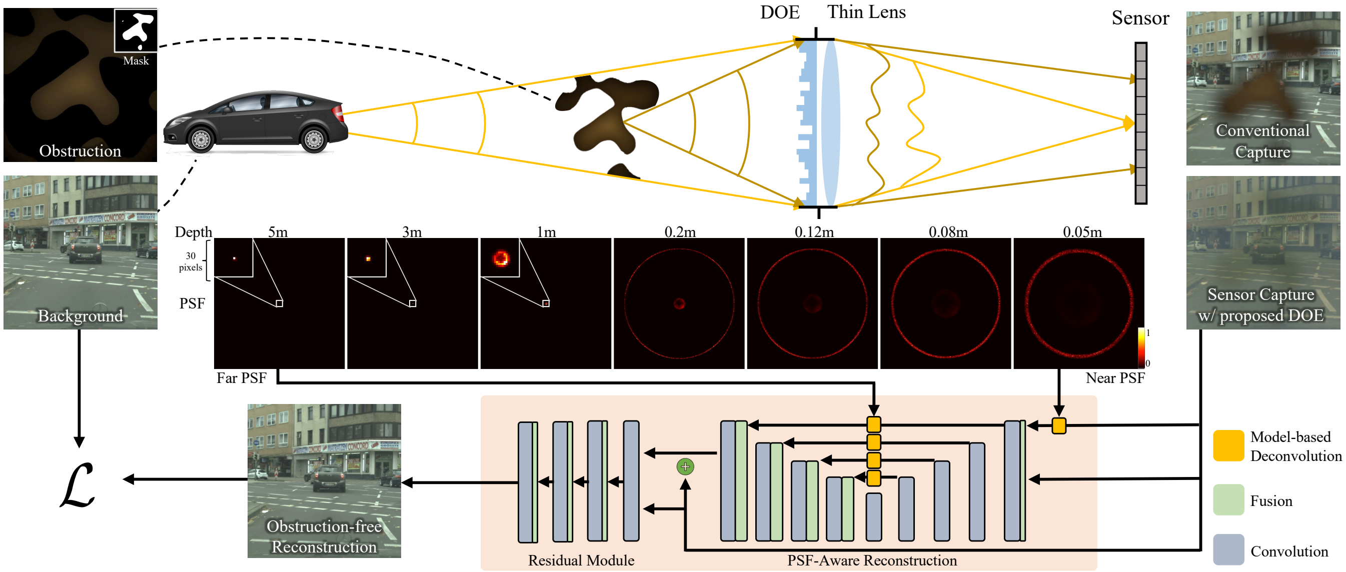


Fig. 2. **Learning Obstruction-aware Diffractive Optical Elements.** We learn a DOE placed in front of a camera lens jointly with an optics-aware reconstruction network to see through thin occluders close to the camera. To this end, we introduce a differentiable obstruction-aware image formation model that simulates the sensor capture of the proposed optical system. The proposed camera (top left) modulates the latent scene at a long distance and the obstruction close to the observer, e.g., dirt on the wind shield, with the depth-dependent point-spread function of the proposed optical system. The resulting sensor capture (top right) is fed into a feature deconvolution network (bottom) along with the depth-dependent point spread function (center). The entire computational imaging system, including optical and compute layers, is jointly learned (bottom left), resulting in a DOE that learns to scatter light for close objects, with a ring-shaped PSF, while preserving angular resolution for objects at optical infinity.

light arrives at the DOE with a spherical phase profile

$$\phi_s = \frac{2\pi}{\lambda}\sqrt{x^2 + y^2 + z^2}, \tag{1}$$

where λ is the wavelength. Our DOE then modulates the phase of the incident light based on its height profile h and the wavelength-dependent refractive index μ_λ of the DOE material as

$$\phi_{\text{DOE}} = \frac{2\pi(\mu_\lambda - 1)}{\lambda} h. \tag{2}$$

Right after the DOE, the refractive lens provides focusing power with a corresponding quadratic phase profile, which is of the form

$$\phi_{\text{focus}} = \frac{-2\pi}{\lambda} \frac{(x^2 + y^2)}{2f}, \tag{3}$$

where f is the focal length of the lens, which corresponds to the distance from the lens to the sensor (i.e., the camera is focused at infinity). Note that we assume here that there is no gap between the DOE and the lens. In summary, the resulting light after being modulated by the refractive lens has the accumulated phase

$$\phi_l = \phi_s + \phi_{\text{DOE}} + \phi_{\text{focus}}. \tag{4}$$

This light then propagates to the sensor, resulting in the PSF

$$p = |\mathcal{F}^{-1}\{\mathcal{F}\{u_l\} \cdot \mathcal{H}\}|^2, \tag{5}$$

where u_l is the complex-valued light wave before the propagation, $\mathcal{H} = \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2\pi}(x^2+y^2)}$ is the Fresnel propagation kernel, $k = \frac{2\pi}{\lambda}$ is the wave number and $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse, respectively.

However, the simulation of the quadratic phase profile of the refractive lens, as well as the following Fourier-propagation kernel, require sampling the light field at a minimum resolution prescribed by Nyquist limit [Cottrell et al. 1990; Moreno et al. 2020]. Specifically, for an $N \times N$ light field,

$$F_N = \frac{N\delta^2}{\lambda} \leq f \tag{6}$$

is required to avoid aliasing, where F_N is the Nyquist limit focal length, N is the dimension in pixels and δ is the wavefront sampling resolution. For instance, assuming a wavelength of 500 nm and an 8mm focal length camera, simulating a 4 mm lens requires sampling at $\delta \leq 1\ \mu m$ (with corresponding $N \geq 4000$), suggesting that learning a full 2D phase profile jointly with a reconstruction network is computationally restrictive as it would require over 50 GB of GPU memory.

To lift this sampling constraint, we simulate the refractive lens and Fourier propagation jointly using a Fraunhofer far field approximation, since with a spherical light source the quadratic phase term in the Fresnel propagation is canceled out by the quadratic phase profile of the refractive lens. We implement this propagation by computing the Fourier transform of the complex-valued light wave u_l' of phase $\phi_l' = \phi_s + \phi_{\text{DOE}}$. This results in an estimated PSF p' without having to sample the wave in the near field, that is

$$p' = |\mathcal{F}(u_l')|^2, \tag{7}$$

where u_l' is the complex-valued light wave after passing through the DOE and before being modulated by the refractive lens. Note that the corresponding wave sampling resolution required for accurately representing p' becomes

$$\delta' = \frac{\lambda f}{\delta N}, \tag{8}$$

where f is the focal length (propagation distance), δ is the incoming wavefront sampling resolution, and N is the corresponding wavefront dimension [Goodman 2005]. As a result, a scale factor $\frac{N}{\lambda f}$ appears, providing us with the sensor PSF

$$p = \text{resize}(p', \frac{N}{\lambda f}), \tag{9}$$

where resize$(\cdot)$ is a nearest neighbor resizing operator.

Lens Obstructions. We formulate the image formation in the presence of obstruction with an alpha composition to combine an obstruction-free image I_{far} with an obstruction image I_{obs}, while taking into account their corresponding depth PSFs, p_{far} and p_{obs}, respectively.

$$I_s = \alpha \odot (p_{\text{obs}} * I_{\text{obs}}) + (1 - \alpha) \odot (p_{\text{far}} * I_{\text{far}}), \tag{10}$$

We use alpha-mask $\alpha = p_{\text{obs}} * M$ to handle the invalid regions in I_{obs} (where the obstruction does not exist), with M being a binary mask corresponding to the obstruction. $*$ and $\odot$ denote convolution and element-wise product, respectively. To accurately model the sensor measurement, we adopt the noise model from [Tseng et al. 2021b] in this work.

3.2 Depth-dependent Obstruction Simulation

We simulated three types of obstruction corresponding to different imaging scenarios, see Fig. 3: (i) thin fence-like occluders, which are common in point-and-shoot photographs, (ii) dirt obstructions, and (iii) raindrops on a windshield or a lens cover, which are common in robotics and automotive imaging scenarios.

Depth-dependent Obstruction Simulation. To realistically simulate obstructions at different depths, we rescale the obstruction such that

$$s_p = \frac{s_m f}{z_{\text{obs}} \Delta_p}, \tag{11}$$

where s_p is the size of simulated obstructions in pixels, and the other parameters are given in meters: s_m is the physical width of the obstruction (e.g, the fence wire diameter), f is the focal distance, z_{obs} is the depth of the obstruction with respect to the sensor and Δ_p is the pixel pitch size. We describe each obstruction type in the following.

Thin Fence Occluders. We model thin fence occluders (see Fig. 3, right) using the fence dataset from [Du et al. 2018], containing 545 fence training images. As this dataset contains a variety of different fence types captured in the wild, we use the fence masks in the dataset to extract and scale all fence types to a normalized occluder width in screen space. Assuming a typical fence wire diameter of 4 mm, we simulate fence occlusion at randomly sampled depth in [40cm, 80cm]. We augment the fence patterns with random rotations in [0°, 90°], random horizontal flips and random contrast adjustments and color jitter.

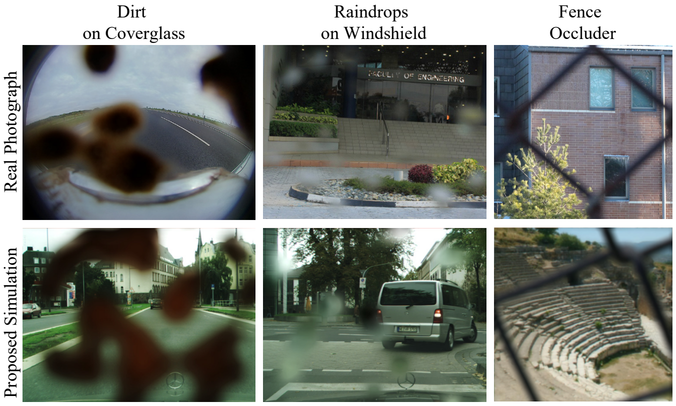


Fig. 3. **Obstruction Model.** We model three types of obstructions that are common for robotic imaging and photography. Specifically, we simulate dirt and mud from road debris on the coverglass (left column); raindrops on a windshield as seen by a front-facing camera (center column), common in both self-driving vehicles and driver-assistance systems; and thin meshed fence occluders for point-and-shoot photography (right column). Real example captures (top row) validate that the proposed obstruction models (bottom row) reflect realistic scenarios, see text for more details.

Raindrops. We simulate raindrops of diverse sizes (see Fig. 3, center) with shapes varying from a disk to an oval, to simulate the effect of wind pressure applied to the drop, which would otherwise be perfectly circular [Iseringhausen et al. 2017]. To this end, we combine a randomly generated oval with a half of a disk. Assuming forward-moving vehicles, we set the orientation of the oval drop to align with the airflow direction. Due to the shape of the raindrops and their placement on a diagonally oriented windshield, they act as fish-eye lenses, where light-rays transported through them are tilted upwards. This means that the image reflected from a raindrop corresponds to a larger and higher region in the background scene [You et al. 2013]. The resulting image is then blurred by the PSF corresponding to the raindrop's depth, which is randomly sampled between [5cm, 12cm] to account for different spacing to the windshield. The blurred image is then blended with the background using the generated oval-shaped drop mask, following Eq. (10). We randomly sample drop position and number of drops during training.

Dirt Obstructions. We consider dirt and soil that is deposited on the lens cover glass or camera enclosure glass (see Fig. 3, left). Soil and road debris on the camera can be sprayed by preceding vehicles on wet roads, or by exposure to the elements during off-road driving. Real-world examples are available in the Woodscape surround view dataset [2019]. Unfortunately, the soil annotations in the dataset are manually labeled polygons, that are not per-pixel accurate, containing background mixed into the soil. As such, we chose to simulate realistic soil obstruction on the camera cover glass by simulating

Perlin Noise [Perlin 1985], which is then low-pass filtered with a Gaussian blur and thresholded with a random threshold to retrieve diverse connected soil-like patterns. We simulate additive color jitter, soil intensity augmentation and per-sample random soil patterns for a distance sampled randomly in [5cm, 12cm].

3.3 End-to-End Obstruction Free Imaging

Once we have captured measurements in hand, we recover the latent background using a learned reconstruction network that is trained jointly with the optical stack. We describe this network and the end-to-end training procedure in the following.

Optics-aware Reconstruction Network. We devise a reconstruction network that performs model-based deconvolution, along with feature-based dehazing: our model-based deconvolution method takes the PSFs (either simulated or calibrated post-fabrication) of the proposed optics system as input to ensure generalization, while suppressing unwanted residual aberrations from nearby occlusions (that manifest as "haze").

Specifically, the proposed network takes in the sensor capture, as well as the PSF measurements corresponding to the designed optics system at optical infinity and in the near distance, where the occluder is expected to appear. By using these PSFs to perform deconvolution of the captured image, the network gains access to information about the captured scene of interest, as well as valuable information concerning the occluder itself, which can be exploited to further refine the image. To incorporate the PSFs into our model, we rely on differentiable inverse filtering blocks inspired by [Tseng et al. 2021a].

The proposed architecture is illustrated in Fig. 2 and comprises of two sequential components: a reconstruction block which performs multi-scale feature extraction, feature propagation through PSF-aware inverse filters and feature decoding; and a refinement block which further refines the reconstruction output of the first stage. The residual-learning block follows a U-Net architecture design with four downsampling stages. At each downsampling stage, features extracted by block f_{FE} are filtered using the inverse of the far PSF in $f_{\mathrm{Z}\to\mathrm{W}}$. In other words, $f_{\mathrm{Z}\to\mathrm{W}}$ deconvolves these features (i.e., it propagates features Z to their deconvolved spatial positions W). Both the extracted and deconvolved features are fed into the upsampling and decoding stage f_{DE}, that combines the propagated features into the resulting image output of the first stage. The first stage becomes

$$\mathbf{O} = \underset{\substack{\uparrow\\ \text{Decoder}}}{f_{\mathrm{DE}}}\left(\overset{\substack{\text{Inverse Filter}\\ \downarrow}}{f_{\mathrm{Z}\to\mathrm{W}}}\left(\underset{\substack{\uparrow\\ \text{Feature Extraction}}}{f_{\mathrm{FE}}}\left(\overset{\substack{\text{Concat.}\\ \downarrow}}{C}\left(\mathbf{I}, f_{\mathrm{Z}\to\mathrm{W}}\left(\mathbf{I}, p_{\mathrm{near}} \right) \right) \right), p_{\mathrm{far}} \right) \right), \quad (12)$$

where $\mathbf{I}$ is the sensor measurement and $\mathbf{O}$ is the output of the first stage. In addition, we concatenate the inverse-filtered output of $f_{\mathrm{Z}\to\mathrm{W}}(\mathbf{I}, p_{\mathrm{near}})$ as a separate channel with the concatenation operator C. While deconvolving with the near PSF may appear uninformative to a human observer, partly deconvolved obstructions help the network identify and remove the obstruction residual. Although not limited to a specific inverse filter block, we use a differentiable implementation of the Wiener filter for $f_{\mathrm{Z}\to\mathrm{W}}$. Unlike traditional deconvolution approaches, our network benefits from the robustness of feature learning, and therefore does not require a pixel-level accurate PSF, which is typically unavailable in practice. The second stage provides further refinement to the output of the first stage. This module is entirely learned and consists of 3 residual blocks. We provide not only the output of the first stage to the refinement module but also the original sensor capture, to allow it to further improve the reconstruction quality. Please refer to the Supplemental Document for network architecture details.

End-to-End Loss. We train our framework by minimizing the loss function

$$\mathcal{L} = \mathcal{L}_{\ell_1} + \mathcal{L}_{\mathrm{perc}} + \mathcal{L}_{\mathrm{obs}}, \quad (13)$$

where $\mathcal{L}_{\ell_1}$ is a per-pixel ℓ_1 loss between I_{recon} and I_{far}, and $\mathcal{L}_{\mathrm{perc}}$ is an LPIPS based perceptual loss. Term $\mathcal{L}_{\mathrm{obs}} = M \odot |I_{\mathrm{recon}} - I_{\mathrm{far}}|$ is an obstruction-focus loss that encourages the reconstruction model to pay extra attention to areas degraded by the obstructions, with M being the binary mask corresponding to the obstruction and $\odot$ is the Hadamard product. As illustrated in Fig. 2, the proposed loss function is directly applied to the reconstructed network output, to supervise both optics optimization and image reconstruction, in an end-to-end fashion.

4 ANALYSIS

Before assessing the proposed method on experimental measurements, we first analyze the method using simulations on synthetic data and compare to existing baseline methods. To this end, we consider the following two representative imaging tasks.

The first task involves a front-facing automotive imaging camera mounted behind a windshield. We assume a typical automotive sensor with a 1.85μm pixel pitch, 4000 × 3000 sensor resolution and an 8 mm focal length. We assume the camera is mounted in a position near the rear-view mirror, which places the windshield 5 to 10 cm away from the lens. We simulate raindrops and dirt obstructions as both can be commonly found on vehicle windshields. We use captured scenes from the Cityscapes dataset [Cordts et al. 2016] as the obstructed background at a distance of more than 5 m from the camera.

We consider a point-and-shoot photography task as the second representative imaging scenario. Specifically, we assume a DSLR camera with large 4.3μm pixel pitch at a resolution of 4000×4000 and a 50 mm focal length. We simulate fence-like thin obstructions at depths between 0.4 to 0.8 m, and we again assume that the obstructed scene objects are more than 5 m away from the camera. We use the Places365-Standard dataset [Zhou et al. 2017] for representative background scenes. This dataset contains 1.8 million training images sampled from 365 scene types.

4.1 Depth-dependent PSF Analysis

Next, we compare the proposed learned optical system to a conventional camera lens (modeled as a thin lens without DOE) and a Fresnel refocusing DOE learned alongside the thin lens. We additionally compare to a random diffuser placed in front of the thin lens, using narrow-band RGB wavelengths (656nm, 589nm and 486nm) as

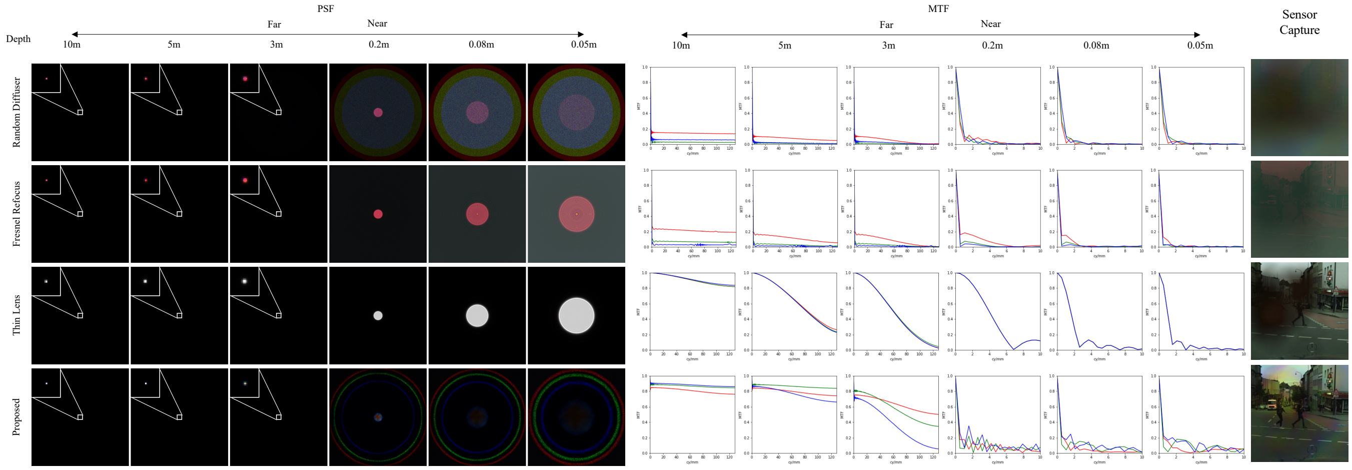


Fig. 4. **PSF and MTF Analysis.** We report the near-scene and far-scene PSFs and MTFs corresponding to the proposed learned optical design, and to other heuristic baseline optical designs that do not increase the aperture. Specifically, we compare the proposed approach (Proposed) against a random diffuser DOE profile paired with a thin lens (Random), a learned Fresnel DOE profile paired with a thin lens (Fresnel) and a thin-lens-only design (Thin Lens). The thin lens is a fixed 8 mm lens in all cases, and we assume a machine vision sensor with 1.85μm pixel pitch. All baseline designs are heuristic choices, except for the Fresnel design which we optimize in the same end-to-end approach as the proposed lens, but only learning a free focal length parameter. Sensor simulations corresponding to each of the designs are shown on the right, with dirt obstruction 5 cm from the entrance pupil. The proposed learned optical system suppresses the obstruction, while preserving high MTF for long distances.

a proxy for broadband illumination. Fig. 4 reports the corresponding simulated PSFs for the automotive imaging task.

The conventional camera lens is designed to focus at far distances, whereas close objects are blurred with a blur kernel determined by the aperture shape. As such, the corresponding sensor image exhibits full occlusions of background content, even for nearby, small occluders. Combining a conventional lens with a DOE allows us to modulate the phase at micron-scale resolution, in contrast to grinded optics that allows only smooth phase modulation. Taking the DOE design to the extreme, one could implement a random phase mask that acts as a diffuser in front of the refractive focusing lens. In this scenario, however, the wavefronts would be scrambled across the entire depth range, resulting in low MTF values for far distances and an overall blurry sensor image, that does not preserve the distant background scene. Optimizing a conventional Fresnel lens design DOE in the proposed end-to-end optimization fashion does not constitute a valid alternative. Specifically, learning a Fresnel profile jointly with a fixed refractive lens instead results in a focus change from 8mm (sensor distance) to 2mm, a compromise between the refractive-only configuration and scrambling the wavefronts across all depths. However, this design comes at the cost of severe chromatic aberration in all depths, while still obtaining low MTF values at long distances.

In contrast to the optical systems discussed above, the proposed design achieves high MTF values at long distances corresponding to background scene depths, and disperses the radiance from nearby objects over the entire sensor. Specifically, the PSF at optical infinity resembles that of a conventional lens with a small focal spot, while the PSF for nearby distances resembles a large ring, with a radius prescribed by the maximum diffraction angle supported by the phase plate.

The proposed lens configuration effectively suppresses foreground objects by distributing their energy over the entire sensor as a low-frequency signal, while preserving the background image as an additive component in the lens design.

Note that this PSF is obtained using our *end-to-end learning approach* without using any design heuristics. To further validate our end-to-end optimization approach, we hand-crafted a loss on the PSF such that the far PSF approaches a dirac delta and the near-scene PSF spreads light evenly across the sensor plane, followed by separate optimization for the reconstruction network given the fixed DOE. As we show quantitatively in Tab. 2, hand-crafting an intermediate loss directly on PSFs results in sub-par performance, see Supplemental Document for details. This demonstrates the benefits of our end-to-end approach.

4.2 Reconstruction Network Ablation Study

We validate our network architecture choices with an ablation study. As we show in Fig. 5 and quantitatively report in Tab. 1, when removing the inverse filtering block which incorporates the PSFs into our model, the proposed model becomes an image-to-image encoder-decoder mapping network. Similar to vanilla image-to-image mapping architectures (e.g. the popular UNet [Ronneberger et al. 2015]), aberrations resulting from the foreground obstructions make it difficult for the network to recover the true color of the background, resulting in "hazy" outputs.

Nonetheless, this ablated design performs better than a standard UNet, emphasizing the important role of other architecture choices in our network, such as the use of the refinement block, which allows the reconstruction block to focus on removing the occluder residuals. The results in Fig. 5 and Tab. 1 also suggest that removing the feature extraction block hinders our network's ability to

Table 1. **Quantitative Reconstruction Network Analysis.** We evaluate different reconstruction network architectures in simulation, including removing components from the proposed network (rows 2-5), existing image-to-image networks (rows 6-7), and conventional deconvolution methods (rows 8-10).

	SSIM	PSNR,	LPIPS
Proposed	**0.93 (0.94)**	**23.09 (23.96)**	**0.97 (0.94)**
Proposed w/o inverse filter	0.92 (**0.94**)	22.92 (23.73)	0.96 (0.93)
Proposed w/o residual block	0.91 (0.93)	21.95 (22.55)	0.96 (0.93)
Proposed w/o feature extraction	0.90 (0.92)	20.78 (21.02)	0.94 (0.89)
Sequential optimization	0.92 (0.93)	21.68 (22.07)	0.90 (0.80)
DeblurGAN-V2 [Kupyn et al. 2019]	0.84 (0.85)	20.14 (19.21)	0.93 (0.85)
UNet [Ronneberger et al. 2015]	0.91 (0.92)	22.19 (22.49)	0.95 (0.90)
Richardson-Lucy Deconvolution (far kernel)	0.89 (0.91)	20.48 (20.78)	0.93 (0.84)
Wiener Deconvolution (far kernel)	0.91 (0.92)	20.77 (20.95)	0.94 (0.87)
Wiener Deconvolution (close kernel)	0.16 (0.15)	13.00 (12.70)	0.61 (0.24)
Sensor Capture	0.79 (0.81)	17.45 (17.87)	0.92 (0.85)

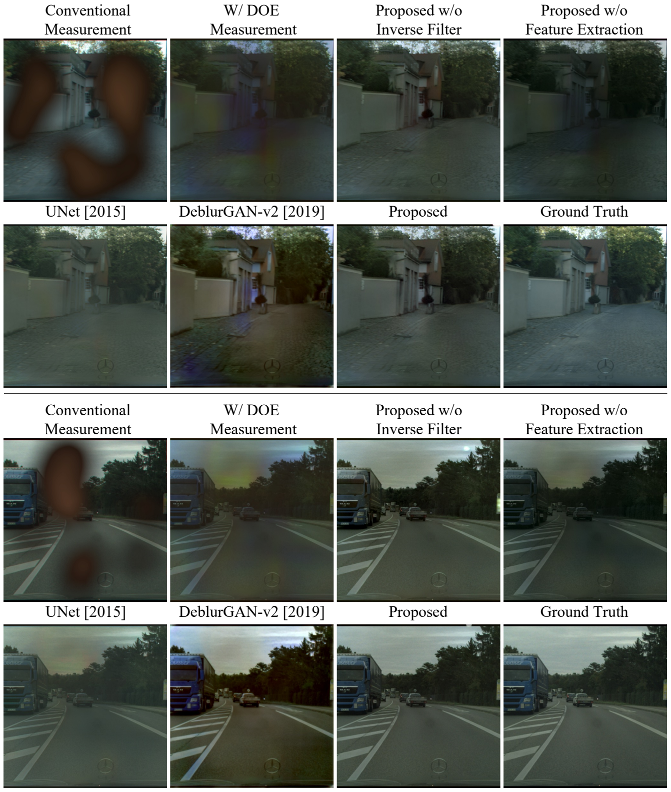


Fig. 5. **Network Ablation Experiments.** We report the effect of eliminating core network elements from the proposed method, specifically, the inverse-filtering and feature extraction blocks. Next we compare the proposed approach to recent image deblurring (DeblurGAN-v2 [Kupyn et al. 2019]), conventional image-to-image mapping (UNet [Ronneberger et al. 2015]), and the corresponding ground truth obstruction-free images. Please refer to the Supplemental Document for additional qualitative ablation study results.

mitigate color abberation and dark region artifacts, similar to the performance obtained by applying Wiener filtering or Richardson-Lucy deconvolution. Finally, we also compare to the DeblurGAN-v2 method [Kupyn et al. 2019], aimed at the closely related image deblurring task, which produces inferior results.

4.3 Synthetic Assessment

Next, we compare the proposed method to existing *single-image* obstruction removal methods, that are designed for conventional monocular cameras. To this end, we consider three types of baseline methods: (i) Generic single-image inpainting methods, which can be applied post-capture to predict user-defined occluded areas, represented by the recently published texture-based CTSDG inpaiting approach [2021] and the large-kernel inpainting method LaMa [Suvorov et al. 2021]. (ii) Methods for obstruction removal, specialized to a particular type of obstruction, represented by DefenceNet [2019] for visually removing fences, and AttentiveGAN [2018] for post-capture raindrop artifacts removal. (iii) Optics-only methods allowing no post-processing computations, which we implement by learning a DOE and refractive lens pair, without using a reconstruction network.

Qualitative and quantitative comparisons are reported in Fig. 6 and Tab. 2, respectively, while additional comparisons are presented in the Supplemental Document. For these comparisons, we simulate conventional sensor capture assuming a perfect thin lens, following the forward model from Sec. 3.2. We give the competing baseline methods an advantage over ours, by providing *all* of them with the *ground-truth* obstruction masks M from Eq. 10.

Our synthetic test set (withheld from training) contains 1500 pairs of obstruction and latent background pairs for each type of obstruction. We fine-tune the pre-trained models provided by the authors using our training set whenever the code provided by the authors allows it. All baseline methods use the same model for both dirt and raindrops removal, as is the case for our proposed method.

DefenceNet [2019] employs a two-phase reconstruction approach, where fence regions are first detected by a UNet-based network (DetectNet), and then inpainted using a different ResNet-based network (RemoveNet). While the use of conventional high-pass filters in the fence detection is effective for all-in-focus images, a fence closer than optical infinity will feature blurry edges for larger apertures in our case. This makes it difficult for the edge-filter-based DetectNet to produce adequate mask outputs, and such errors are carried over to the next inpainting stage. Moreover, even when provided with ground truth masks, nearby fences induce large areas to be inptained, in which case the Gaussian inpainting step fails to provide a good starting point for the ResNet refinement process.

AttentiveGAN [Qian et al. 2018] uses an attentive generative network for raindrop removal, where visual attention is used in both generative and discriminative networks, such that the inpainting is focused on the raindrop regions and the nearby surrounding structures*. While the combination of the attention map and adversarial training generally yields high-quality outputs, it also biases the method to prefer raindrops of similar size and shape, resulting in

*Replacing the estimated raindrop masks with ground truth ones yielded similar quality outputs.

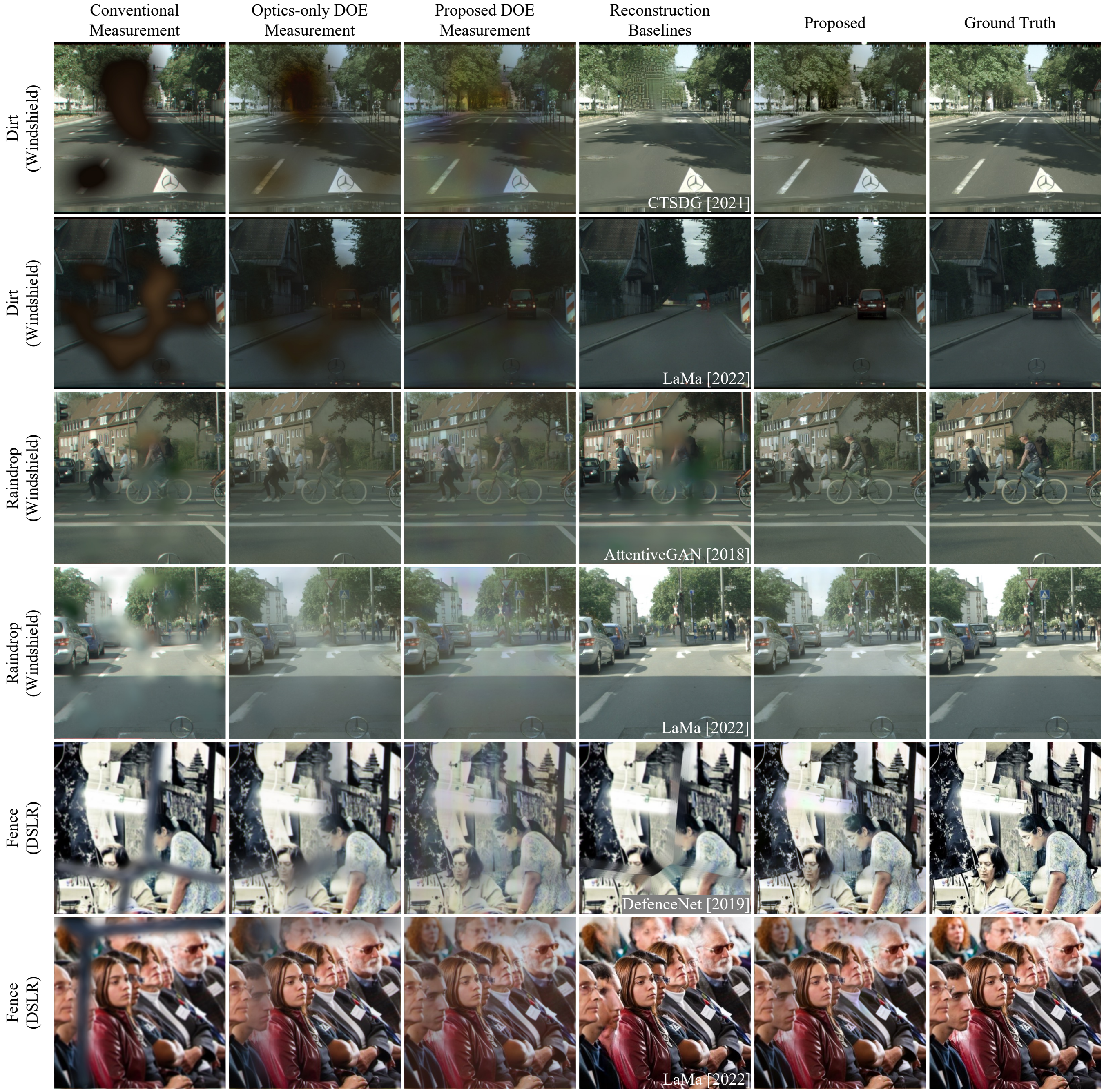


Fig. 6. **Analysis for Different Obstruction Types.** We analyze the proposed method for different types of obstructions, including dirt (top two rows) and raindrops (middle two rows) in a driving scenario, and fence obstructions, which are often encountered in point-and-shoot photography (bottom two rows). Our DOE (optimized to operate with or without a subsequent neural network) allows us to see through occlusions. Feeding the captured image into our neural network results in an almost obstruction-free image. Existing inpainting methods instead hallucinate the occluded background regions.

Table 2. **Quantitative Reconstruction Quality.** We evaluate reconstruction quality using SSIM, PSNR and 1- LPIPS (higher is better) for dirt, raindrops and fence obstructions. We compare our method against conventional inpainting methods (CTSDG [Guo et al. 2021], LaMa [Suvorov et al. 2021]), a method tailored to raindrop degradation (AttentiveGAN [Qian et al. 2018]) and a method specialized to fence inpainting (DefenceNet [Matsui and Ikehara 2020]). We further compare to optics-only DOE designs and using the proposed DOE while ablating the neural network.

	Reconstruction Performance of the Occluded Region (and Full Image)								
		Fence - DSLR			Raindrop - Windshield			Dirt - Windshield	
	SSIM	PSNR	1-LPIPS	SSIM	PSNR	1-LPIPS	SSIM	PSNR	1-LPIPS
Proposed	**0.88** (**0.93**)	**23.22** (28.14)	**0.97** (0.91)	**0.97** (**0.98**)	**29.17** (**29.89**)	**0.98** (**0.96**)	**0.93** (**0.94**)	**23.09** (23.96)	**0.97** (**0.93**)
LaMa[2021]**	0.73 (**0.93**)	22.31 (**28.60**)	0.95 (**0.94**)	0.83 (0.92)	24.25 (27.65)	0.92 (0.90)	0.78 (0.87)	21.68 (**24.39**)	0.88 (0.86)
CTSDG [2021]**	0.68 (0.92)	21.08 (27.37)	0.93 (0.91)	0.81 (0.91)	23.88 (27.28)	0.89 (0.86)	0.74 (0.85)	21.27 (23.98)	0.83 (0.80)
DefenceNet [2020]**	0.55 (0.77)	17.86 (22.78)	0.89 (0.71)	-	-	-	-	-	-
AttentiveGAN [2018]	-	-	-	0.82 (0.88)	21.56 (23.00)	0.89 (0.75)	-	-	-
Optics-only DOE Capture	0.72 (0.84)	19.33 (23.54)	0.94 (0.82)	0.92 (0.94)	24.82 (26.36)	0.96 (0.91)	0.84 (0.88)	17.81 (18.40)	0.91 (0.86)
Conventional Camera Capture	0.55 (0.77)	15.76 (20.90)	0.90 (0.71)	0.82 (0.89)	22.81 (25.59)	0.89 (0.80)	0.64 (0.73)	14.22 (14.66)	0.81 (0.70)
Proposed DOE Capture	0.84 (0.89)	22.01 (25.71)	**0.97** (0.89)	0.91 (0.93)	23.83 (24.38)	0.95 (0.87)	0.79 (0.81)	17.45 (17.87)	0.92 (0.85)

failures when raindrop size varies, e.g., due to different raindrop depths.

CTSDG [2021] is a generic two-stream inpainting method, jointly optimizing texture synthesis and texture-guided structure reconstruction. LaMa [Suvorov et al. 2021] is a recent inpainting method using fast Fourier convolutions, which facilitates a large network receptive field, performing well even for large occluded regions. As generic image inpainting methods, they are both able to handle diverse obstruction types. We provide the ground truth mask to these methods, giving it some advantage over our proposed method. We further make it easier for these inpainting methods by applying them on the ground truth images rather than on the simulated refractive lens captures. Nonetheless, the missing occluded background information makes the task ill-posed, such that single-image inpainting approaches can only attempt to hallucinated the missing image content.

The optics-only baseline learns to scatter away some of the near scene (obstruction) while preserving information from the background scene. However, without a reconstruction network present, this optical setup minimizes chromatic aberration at the cost of occlusions. Hence, much of the background remains occluded, and the overall image quality is drastically reduced.

The proposed approach achieves a 1 to 5 dB PSNR margin over existing approaches when reconstructing the occluded regions, by optically cloaking the obstruction while relying on complementary computational reconstruction to remove image aberrations, thus effectively allowing the occluded background to be “seen” by the recovery module.

Qualitative comparisons in Fig. 6 validate that the method is able to remove diverse obstructions, and recover “hidden” details, which can merely be hallucinated by existing methods.

5 EXPERIMENTAL ASSESSMENT

In this section, we evaluate the proposed method with experimental captures. To this end, we fabricate the learned diffractive optical element for the automotive machine vision scenario described in Sec 4. We first describe the experimental setup and validate that the measured PSFs feature a depth-dependent ring structure, before presenting experimental reconstructions using a prototype camera system.

5.1 Experimental Prototype

We implement the proposed method experimentally with the prototype system shown in Fig 7. To this end, we fabricate the DOE designed for the front-facing automotive imaging task in a 16-level photolithography process on a fused silica wafer. The diameter of the fabricated DOE is 4.4 mm, and we use a chrome layer as an optical baffle. We use a FLIR Blackfly S USB3 camera with an 8mm lens (Edmund Optics 33-307). While this compound lens ensures low distortion imaging with high MTF across all fields, this comes at the cost of a complex optical assembly with half a dozen elements. To place the DOE on the aperture plane, instead of cutting an existing lens barrel, we employ a 4F relay system, as shown and illustrated in Fig 7. This allows us to swap different DOEs (or remove them) in a reproducible fashion, without changing the target objective or the camera alignment. We use two large-aperture lenses (Pentax SMC FA 75mm f/2.8) for our 4F-configuration with the aperture plane and DOE plane placed at 75 mm distance from both lenses. Note that, although the physical size of the proposed prototype system is larger than the objective lens, the DOE plane is relayed to the aperture plane, so that the *proposed aperture-plane DOE does not effectively increase the size of the camera objective.* To reproduce obstruction positions, we place the representative foreground obstructions on an AR-coated glass holder (Edmund Optics $\lambda/4$ N-BK7 75 mm × 75 mm Window) in front of the optical system.

5.2 DOE Fabrication and PSF Analysis

The prototype DOEs are fabricated by a combination of photolithography and reactive-ion etching techniques. Since it is challenging to fabricate continuous height profiles, we first quantize the optimized DOEs into $2^4 = 16$ levels (75 nm per level), which allows us to faithfully approximate the continuous phase function. Though the theoretical diffraction efficiency with 16 levels can reach above 95%, we observe that manufacturing deviations result in a significant

**Methods have access to ground truth obstruction masks. Note that in a practical application, this mask would have to be accurately estimated for the respective baseline approach to perform comparably.

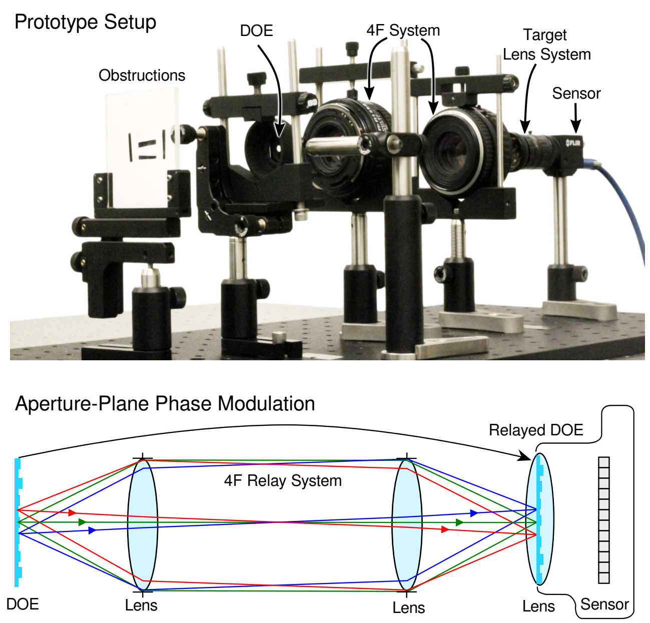


Fig. 7. **Experimental Prototype System.** We evaluate the proposed method experimentally with the prototype system shown on top. We fabricate a phase-only DOE on a fused silica wafer using a 16-level photolithography process. This phase plate is designed to be *in the aperture plane* of the target camera configuration. Rather than cutting open an off-the-shelf compound lens or building our own compound lens, we relay the phase DOE to the aperture plane using a 4F system (bottom). This facilitates experimentation and allow us to acquire images with or without a DOE, without replacing the objective lens. To make obstructions reproducible, we place them on an AR-coated glass holder at different depths.

0^{th}-order component, which we discuss below. We repeat 4 iterations of the basic photolithography with different masks, and then etch the same substrate with doubled etching time sequentially. Finally, a Chromium aperture is deposited to block the light outside of the clear aperture. See the Supplementary document for additional details on the fabrication procedure.

To validate the fabrication, we measure the depth-dependent PSFs using the prototype system from above with and without the proposed DOE being present in the optical path. To acquire and compare the optical aberrations in a reproducible fashion, we use the fiber tip coupled with a 520 nm laser diode as source, see Fig. 8. This source is a spherical emitter at close distances, allowing us to capture the near PSF. To acquire the PSF at optical infinity, we add a collimation lens (planoconvex 100 mm lens), resulting in a plane wave illumination. The measured PSFs with and without the DOE are shown in Fig. 8. We note that all PSFs are measured with optical systems that have the same physical aperture size. The measurements validate that the proposed phase plate matches the shape of the simulated PSF, also shown in the figure for the corresponding distance. The measured PSF exhibits the desired ring shape at close (obstruction) distance, while contracting to a small spot at long distances, even slightly outperforming the conventional system, as reported also in Fig. 4. However, the measurement also reveals that the PSF has a substantially stronger 0^{th}-order component resulting from manufacturing deviations. The fabrication process available to us (described in the Supplementary Document) involves manual operations, and therefore deviations from industry standard processes are not uncommon. As a result, the fabricated DOE features a central component that is significantly stronger than in simulation.

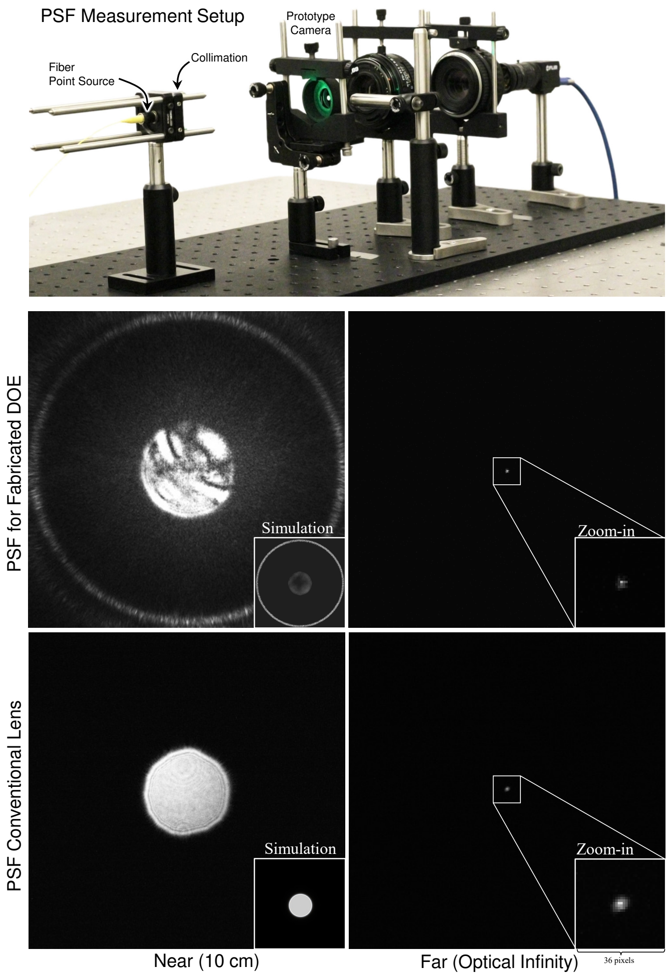


Fig. 8. **PSF Measurements.** After fabricating the proposed diffractive phase plate correspond to the machine vision design, see Sec. 4, we measure the PSF of the entire optical system from Fig. 7 with (center row) and without (bottom row) the DOE. We measure the PSF at different depths using a fiber optic source (top left). The fiber tip makes for a spherical source, accurately representing close objects. To measure the PSF at optical infinity, we collimate the source with an additional collimation lens, resulting in a plane wave illumination. The measurements validate that the designed phase plate produces a ring-shaped PSF for close occluders, while resulting in a small spot size for far scene content. Manufacturing inaccuracies result in a 0^{th}-order component (center left) at close distances, for which we compensate by providing the measured PSF to the reconstruction network. We measure the light throughput of the DOE as 86%, and the relative intensity between the ring and the disk in the 1^{st} order to be 1:1, compared to 3:1 in simulation.

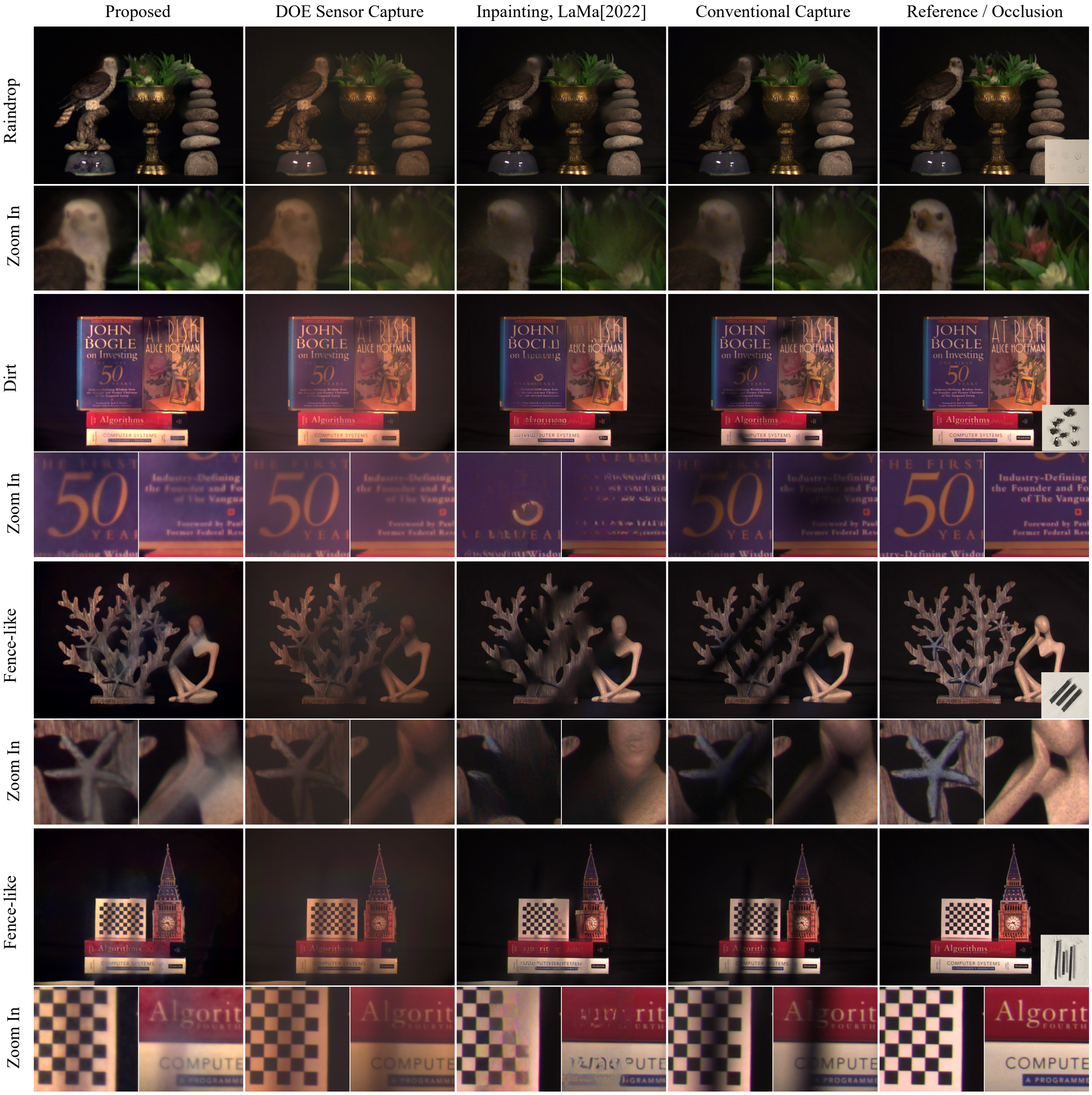


Fig. 9. **Experimental Assessment.** We experimentally validate the proposed method for three different occluder types (alternating rows), after fabricating the learned DOE and fine-tuning the complementary reconstruction network. Our method (left column) is able to restore image regions that would otherwise be occluded by the obstruction, see the conventional camera column. Merely using the DOE without subsequent processing (second column) already provides more visual information compared to using a conventional camera. With help of the reconstruction network, the proposed computational camera significantly outperforms the recent image inpainting method LaMa [Suvorov et al. 2021]. Post-capture inpainting fails to recover scene details (e.g., text) and sometimes entire regions due to lack of information.

The proposed method accounts for such deviations by feeding the fabricated PSF (measured for RGB illumination) into the proposed reconstruction network.

5.3 Assessment

To compensate for fabrication inaccuracies of the optimized DOE, we finetune our reconstruction network using a real-world dataset captured by our prototype, as we explain in the Supplemental Document. We mimic three types of occluders in a way that allows reproducing results. We use black mascara for dirt, black tape for thin occluders and clear nail-polish to mimic raindrops. Each scene is first captured without the occluder, to serve as reference for qualitative evaluation. We then insert the occluder and capture the scene with our fabricated DOE, and again using a conventional camera, to be used as a baseline. We then feed the DOE-captured image into our reconstruction network, to obtain the reconstructed background scene. Results of the three different occluder types are presented in Fig. 1 and Fig. 9. For each scene, we show the reconstruction by our method, as well as the raw DOE sensor capture, the image captured using a conventional camera, the occluder, and the reference image captured without the occluder. We additionally compare to the SOTA LaMa image inpainting method [Suvorov et al. 2021], which is applied on the conventional camera capture. For LaMa, which additionally requires a mask indicating the missing regions (like any inpainting method), we manually mark the occluded regions. This involves the subtle task of determining the optimal inpainting mask for each image; an overly conservative mask might result in the output being too similar to the occluded input, while an aggressive mask might discard potentially useful information, yielding a visually smoother output. We therefore repeat the marking process several times for each scene, and present the best result. Please refer to the Supplemental Document to view LaMa outputs corresponding to different-sized masks. While our method is capable of reconstructing fine details in occluded regions (e.g., text), results by LaMa often fail and hallucinate the wrong content, and in some cases even produce results visually worse than the corresponding conventional camera capture input. This is to be expected, since as an inpainting method LaMa does not exploit partially occluded regions, which are discarded by the (binary) input mask. The proposed method reveals hidden details that are lost in all competing approaches, further validating the designed DOE and reconstruction network experimentally.

6 CONCLUSION

We propose a method for monocular single-shot imaging in the presence of image obstructions. To see through near obstructions, we encode the incoming wavefront from the scene with a learned diffractive optical element, which together with a physics-aware reconstruction method allows us to suppress light scattered from close obstructing objects, and recover the portion of the wavefront that stems from the latent scene. Placed in the aperture of the camera, this learned phase plate does not increase the footprint of the optical system. The learned optics act as a depth-dependent diffuser, such that paraxial wavefronts from the background scenes arrive unperturbed, whereas the off-axis light from foreground obstructions is diffused over the entire sensor. We validate the proposed design through extensive simulations, and using experimental prototype captures. While the experiments validate the approach and the fabricated devices in a lab setting, the synthetic experiments confirm that the designed optics and reconstruction network are effective in diverse application scenarios. In the future, learning diffractive modulation layers in a fabrication-in-the-loop approach may allow for prototyping domain-specific optical systems like ours, with higher accuracy than existing low volume experimental fab processes, potentially reducing the barrier to high-quality fabrication methods. Learning hybrid camera systems that exploit array optics and diffractive encoding may make more complex vision tasks beyond obstruction removal possible – a stepping stone towards practical optical compute layers for imaging and vision.

REFERENCES

Seung-Hwan Baek and Felix Heide. 2021. Polka Lines: Learning Structured Illumination and Reconstruction for Active Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5757–5767.

Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. 2021. Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2651–2660.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

Wenshan Cai, Uday K Chettiar, Alexander V Kildishev, and Vladimir M Shalaev. 2007. Optical cloaking with metamaterials. *Nature photonics* 1, 4 (2007), 224–227.

Ayan Chakrabarti. 2016. Learning sensor multiplexing design through back-propagation. *Advances in Neural Information Processing Systems* 29 (2016).

Julie Chang and Gordon Wetzstein. 2019. Deep Optics for Monocular Depth Estimation and 3D Object Detection. *ArXiv* abs/1904.08601 (2019).

Hongsheng Chen, Bin Zheng, Lian Shen, Huaping Wang, Xianmin Zhang, Nikolay I Zheludev, and Baile Zhang. 2013. Ray-optics cloaking devices for large objects in incoherent natural light. *Nature communications* 4, 1 (2013), 1–6.

Xianzhong Chen, Yu Luo, Jingjing Zhang, Kyle Jiang, John B Pendry, and Shuang Zhang. 2011. Macroscopic invisibility cloaking of visible light. *Nature Communications* 2, 1 (2011), 1–6.

Ilya Chugunov, Seung-Hwan Baek, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2021. Mask-ToF: Learning Microlens Masks for Flying Pixel Correction in Time-of-Flight Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9116–9126.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Don M Cottrell, Jeffrey A Davis, Theodore R Hedman, and Rodger A Lilly. 1990. Multiple imaging phase-encoded optical elements written as programmable spatial light modulators. *Applied optics* 29, 17 (1990), 2505–2509.

Chen Du, Byeongkeun Kang, Zheng Xu, Ji Dai, and Truong Nguyen. 2018. Accurate and efficient video de-fencing using convolutional neural networks and temporal information. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

Tolga Ergin, Nicolas Stenger, Patrice Brenner, John B Pendry, and Martin Wegener. 2010. Three-dimensional invisibility cloak at optical wavelengths. *science* 328, 5976 (2010), 337–339.

Muhammad Shahid Farid, Arif Mahmood, and Marco Grangetto. 2016. Image de-fencing framework with hybrid inpainting algorithm. *Signal, Image and Video Processing* 10, 7 (2016), 1193–1201.

Ficosa. 2017. Ficosa Sensor and Camera Cleaning System. Accessed Oct 18, 2021. (2017). https://www.ficosa.com/products/underhood/sensor-and-camera-cleaning/

Adrian P Gaylard, Kerry Kirwan, and Duncan A Lockerby. 2017. Surface contamination of cars: A review. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 231, 9 (2017), 1160–1176.

J.W. Goodman. 2005. *Introduction to Fourier Optics*. W. H. Freeman. https://books.google.com/books?id=ow5xs_Rtt9AC

Xiefan Guo, Hongyu Yang, and Di Huang. 2021. Image Inpainting via Conditional Texture and Structure Dual Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14134–14143.

Divyanshu Gupta, Shorya Jain, Utkarsh Tripathi, Pratik Chattopadhyay, and Lipo Wang. 2021. A robust and efficient image de-fencing approach using conditional generative adversarial networks. *Signal, Image and Video Processing* 15, 2 (2021), 297–305.

Anna-Karin Gustavsson, Petar N Petrov, Maurice Y Lee, Yoav Shechtman, and WE Moerner. 2018. 3D single-molecule super-resolution microscopy with a tilted light sheet. *Nature communications* 9, 1 (2018), 1–8.

Harel Haim, Shay Elmalem, Raja Giryes, Alex Bronstein, and Emanuel Marom. 2018. Depth Estimation From a Single Image Using Deep Learned Phase Coded Mask. *IEEE Transactions on Computational Imaging* 4 (2018), 298–310.

Zhixiang Hao, Shaodi You, Yu Li, Kunming Li, and Feng Lu. 2019. Learning from synthetic photorealistic raindrop for single image raindrop removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

Lei He, Guanghui Wang, and Zhanyi Hu. 2018. Learning Depth From Single Images With Deep Neural Network Embedding Focal Length. *IEEE Transactions on Image Processing* 27 (2018), 4676–4689.

Mazin Hnewa and Hayder Radha. 2020. Object Detection Under Rainy Conditions for Autonomous Vehicles: A Review of State-of-the-Art and Emerging Techniques. *IEEE Signal Processing Magazine* 38, 1 (2020), 53–67.

Roarke Horstmeyer, Richard Y. Chen, Barbara Kappes, and Benjamin Judkewitz. 2017. Convolutional neural networks that teach microscopes how to image. *ArXiv* abs/1709.07223 (2017).

John C. Howell, J. Benjamin Howell, and Joseph S. Choi. 2014. Amplitude-only, passive, broadband, optical spatial cloaking of very large objects. *Applied Optics* 53, 9 (Mar 2014), 1958. https://doi.org/10.1364/ao.53.001958

Aaron Isaksen, Leonard McMillan, and Steven J Gortler. 2000. Dynamically reparameterized light fields. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 297–306.

Julian Iseringhausen, Bastian Goldlücke, Nina Pesheva, Stanimir Iliev, Alexander Wender, Martin Fuchs, and Matthias B Hullin. 2017. 4D imaging through spray-on optics. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.

Zijie Jiang, Qingxuan Liang, Zhaohui Li, Tianning Chen, Dichen Li, and Yang Hao. 2020. A 3D Carpet Cloak with Non-Euclidean Metasurfaces. *Advanced Optical Materials* 8, 19 (2020), 2000827.

Sankaraganesh Jonna, Krishna K Nakka, and Rajiv R Sahay. 2015. My camera can see through fences: A deep learning approach for image de-fencing. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 261–265.

Sankaraganesh Jonna, Krishna K Nakka, and Rajiv R Sahay. 2016. Deep learning based fence segmentation and removal from an image using a video sequence. In *European Conference on Computer Vision*. Springer, 836–851.

Michael Kellman, Emrah Bostan, Michael Chen, and Laura Waller. 2019. Data-Driven Design for Fourier Ptychographic Microscopy. In *2019 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–8.

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. 2019. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V. Sander. 2021. Let's See Clearly: Contaminant Artifact Removal for Moving Cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2011–2020.

Yanxi Liu, Tamara Belkina, James Hays, and Roberto Lublinerman. 2008. Image defencing. *2008 IEEE Conference on Computer Vision and Pattern Recognition* (2008), 1–8.

Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2020. Learning to see through obstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14215–14224.

Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrián Jarabo, Min H. Kim, Xin Tong, and Diego Gutierrez. 2017. DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graph.* 36 (2017), 219:1–219:12.

Takuro Matsui and Masaaki Ikehara. 2019. Single-image fence removal using deep convolutional neural network. *IEEE Access* 8 (2019), 38846–38854.

T. Matsui and M. Ikehara. 2020. Single-Image Fence Removal Using Deep Convolutional Neural Network. *IEEE Access* 8 (2020), 38846–38854.

Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. 2019. Deep Optics for Single-shot High-dynamic-range Imaging. *arXiv preprint arXiv:1908.00620* (2019).

Rolf Monrad. 2017. Mostad Mekaniske - Apparatus for Cleaning Object Surface. Accessed Oct 18, 2021. (2017). https://uspto.report/patent/app/20200188965

Ignacio Moreno, Don M Cottrell, Jeffrey A Davis, María M Sánchez-López, and Benjamin K Gutierrez. 2020. In-phase sub-Nyquist lenslet arrays encoded onto spatial light modulators. *JOSA A* 37, 9 (2020), 1417–1422.

Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. 2020. DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning. *Nature methods* 17, 7 (2020), 734–740.

Orlaco. 2013. Orlaco All Time Vision Camera. Accessed Oct 18, 2021. (2013). https://rmtequip.com/en/products/orlaco-all-time-vision-camera

Sri Rama Prasanna Pavani and Rafael Piestun. 2008. Three dimensional tracking of fluorescent microparticles using a photon-limited double-helix response system. *Optics express* 16, 26 (2008), 22048–22057.

Zhao Pei, Yanning Zhang, Xida Chen, and Yee-Hong Yang. 2013. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognition* 46, 1 (2013), 174–187.

Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph. (Proc. Siggraph Asia)* 38, 6 (2019), 219–1.

Ken Perlin. 1985. An image synthesizer. *ACM Siggraph Computer Graphics* 19, 3 (1985), 287–296.

Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. 2018. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2482–2491.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Rotoclear S3. 2016. Rotoclear S3 Spin Window. Accessed Oct 18, 2021. (2016). https://cromar.co.uk/products/accessories/roto-clear-spin-window/

Yoav Shechtman, Lucien E Weiss, Adam S. Backer, Maurice Y. Lee, and W E Moerner. 2016. Multicolour localization microscopy by point-spread-function engineering. *Nature photonics* 10 (2016), 590–594.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 114.

Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. 2018. Deep End-to-End Time-of-Flight Imaging. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 6383–6392.

Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2020. Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qilin Sun, Congli Wang, Fu Qiang, Dun Xiong, and Heidrich Wolfgang. 2021. End-to-end complex lens design with differentiable ray tracing. *ACM Transactions on Graphics (Proc. Siggraph)* 40, 4 (2021), 1–13.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. *arXiv preprint arXiv:2109.07161* (2021).

Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. 2021a. Neural Nano-Optics for High-quality Thin Lens Imaging. *Nature Communications* 12, 1 (2021), 6493.

Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. 2021b. Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design. *ACM Transactions on Graphics (TOG)* 40, 2, Article 18 (2021).

Michal Uricar, Ganesh Sistu, Hazem Rashed, Antonin Vobecky, Varun Ravi Kumar, Pavel Krizek, Fabian Burger, and Senthil Yogamani. 2021. Let's Get Dirty: GAN Based Data Augmentation for Camera Lens Soiling Detection in Autonomous Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 766–775.

Vaibhav Vaish, Gaurav Garg, Eino-Ville Talvala, Emilio Antunez, Bennett Wilburn, Mark Horowitz, and Marc Levoy. 2005. Synthetic aperture focusing using a shear-warp factorization of the viewing transform. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 129–129.

Vaibhav Vaish, Marc Levoy, Richard Szeliski, C Lawrence Zitnick, and Sing Bing Kang. 2006. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 2331–2338.

Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. 2004. Using plane+ parallax for calibrating dense camera arrays. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Vol. 1. IEEE, I–I.

Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. 2020. DeOccNet: Learning to see through foreground occlusions in light fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 118–127.

Waymo. 2019. Waymo Open Dataset: An autonomous driving dataset. (2019).

Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. 2005. High Performance Imaging Using Large Camera Arrays. *ACM Trans. Graph.* 24, 3 (Jul 2005), 765–776. https://doi.org/10.1145/1073204.1073259

Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. 2019. PhaseCam3D — Learning Phase Masks for Passive Single View Depth Estimation. *2019 IEEE International Conference on Computational Photography (ICCP)* (2019), 1–12.

Zhaolin Xiao, Lipeng Si, and Guoqing Zhou. 2017. Seeing beyond foreground occlusion: a joint framework for sap-based scene depth and appearance reconstruction. *IEEE Journal of Selected Topics in Signal Processing* 11, 7 (2017), 979–991.

Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. 2015. A Computational Approach for Obstruction-Free Photography. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 34, 4 (2015).

Atsushi Yamashita, Akiyoshi Matsui, and Toru Kaneko. 2010. Fence removal from multi-focus images. In *2010 20th International Conference on Pattern Recognition*.

IEEE, 4532–4535.

Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*. 1–17.

Tao Yang, Yanning Zhang, Jingyi Yu, Jing Li, Wenguang Ma, Xiaomin Tong, Rui Yu, and Lingyan Ran. 2014. All-in-focus synthetic aperture imaging. In *European Conference on Computer Vision*. Springer, 1–15.

Wenxia Yang, Xin Li, and Liang Zhang. 2021. Toward semantic image inpainting: where global context meets local geometry. *Journal of Electronic Imaging* 30, 2 (2021), 023028.

Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. 2019. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9308–9318.

Shaodi You, Robby T Tan, Rei Kawakami, and Katsushi Ikeuchi. 2013. Adherent raindrop detection and removal in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1035–1042.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4471–4480.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).