Research Article Vol. 5, No. 1 / January 2018 / Optica 1

DiffuserCam: lensless single-exposure 3D imaging

Nick Antipa,† Grace Kuo,† Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller*

Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, California 94720, USA
**Corresponding author: lwaller@alum.mit.edu*

Received 4 October 2017; revised 22 November 2017; accepted 24 November 2017 (Doc. ID 308506); published 22 December 2017

We demonstrate a compact, easy-to-build computational camera for single-shot three-dimensional (3D) imaging. Our lensless system consists solely of a diffuser placed in front of an image sensor. Every point within the volumetric field-of-view projects a unique pseudorandom pattern of caustics on the sensor. By using a physical approximation and simple calibration scheme, we solve the large-scale inverse problem in a computationally efficient way. The caustic patterns enable compressed sensing, which exploits sparsity in the sample to solve for more 3D voxels than pixels on the 2D sensor. Our 3D reconstruction grid is chosen to match the experimentally measured two-point optical resolution, resulting in 100 million voxels being reconstructed from a single 1.3 megapixel image. However, the effective resolution varies significantly with scene content. Because this effect is common to a wide range of computational cameras, we provide a new theory for analyzing resolution in such systems.

OCIS codes: (110.6880) Three-dimensional image acquisition; (110.1758) Computational imaging.

https://doi.org/10.1364/OPTICA.5.000001

1. INTRODUCTION

Because optical sensors are two dimensional (2D), imaging 3D objects requires projection to 2D in such a way that the 3D information can be recovered. Scanning and multishot methods can achieve high spatial resolution 3D imaging, but sacrifices capture speed [1,2]. In contrast, single-shot 3D methods are fast, but may have low resolution or small field-of-view (FoV) [3,4]. Often, bulky hardware and complicated setups are required. Here, we introduce a compact, inexpensive single-shot lensless optical system for 3D imaging. We show how it can reconstruct a large number of voxels by leveraging compressed sensing.

Our lensless imager, DiffuserCam, encodes the 3D intensity of volumetric objects in a single 2D image. The diffuser, a thin phase mask, is placed a few millimeters in front of an image sensor. Each point source in the 3D space creates a unique pseudorandom caustic pattern that covers a large portion of the sensor. As result, compressed sensing algorithms can be used to reconstruct more voxels than pixels captured, provided that the 3D sample is sparse in some domain. We solve the inverse problem via a sparsity-constrained optimization procedure, using a physical model and simple calibration scheme to make the computation scalable. This approach allows us to reconstruct several orders of magnitude more voxels than related previous work [5,6].

We demonstrate a prototype DiffuserCam system built entirely from commodity hardware. It is efficient to calibrate, does not require precise alignment, and is light efficient (as compared to amplitude masks). We reconstruct 3D objects on a grid of 100 million voxels (nonuniformly spaced) from a single 1.3 megapixel image. Our reconstructions show true depth sectioning, allowing us to generate 3D renderings of the sample.

Our system, like many computational cameras, uses a nonlinear reconstruction algorithm, resulting in object-dependent performance. To quantify, we experimentally measure the resolution of our prototype with different objects. We show that the standard two-point resolution criterion is misleading and should be considered a best-case scenario. To better explain the variable resolving power of our system, we propose a new local condition number analysis that is consistent with our experiments.

DiffuserCam uses concepts from lensless camera technology and imaging through complex media, integrated together via computational imaging design principles. Our proposed architecture and algorithm could enable high-resolution, light-efficient lensless 3D imaging of large and dynamic 3D samples in an extremely compact package. We believe these cameras will open up new applications in remote diagnostics, mobile photography, and *in vivo* microscopy.

A. Previous Work

Lensless cameras for 2D photography have shown great promise because of their small form factors. Unlike traditional cameras, in which a point in the scene maps to a pixel on the sensor, lensless cameras map a point in the scene to many points on the sensor, requiring computational reconstruction. A typical lensless architecture replaces the lens with an encoding element placed directly in front of the sensor. These 2D lensless cameras have demonstrated passive incoherent imaging using amplitude

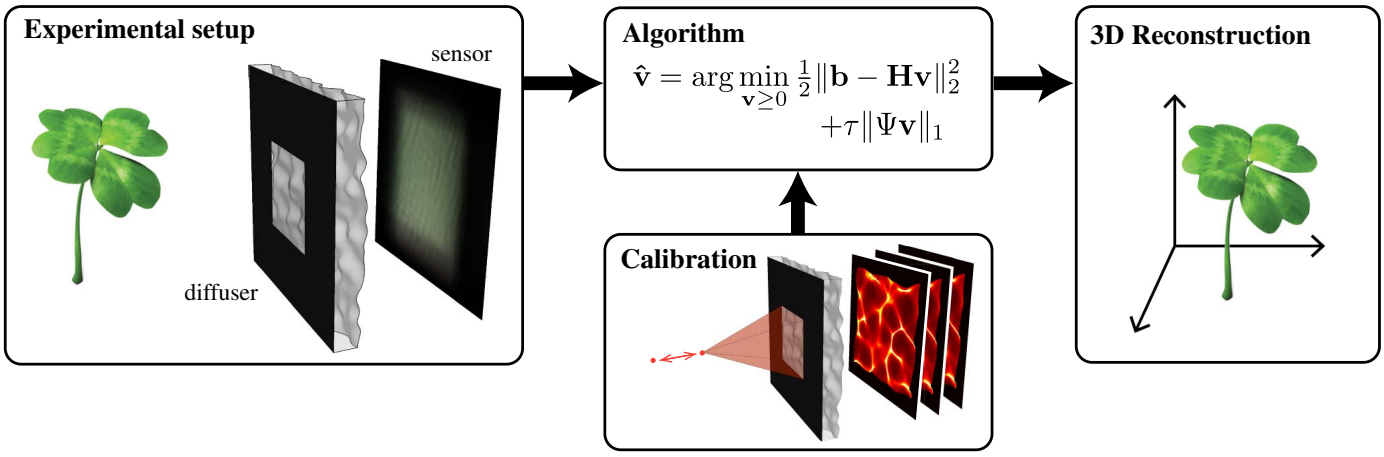

Fig. 1. DiffuserCam setup and reconstruction pipeline. Our lensless system consists of a diffuser placed in front of a sensor (bumps on the diffuser are exaggerated for illustration). The system encodes a 3D scene into a 2D image on the sensor. A one-time calibration consists of scanning a point source axially while capturing images. Images are reconstructed computationally by solving a nonlinear inverse problem with a sparsity prior. The result is a 3D image reconstructed from a single 2D measurement.

masks [7], diffractive masks [8,9], random reflective surfaces [10,11], and modified microlens arrays [12]. Our system uses a similar architecture with a diffuser as the encoding element, and also extends the design and image reconstruction to enable 3D capture.

Light field cameras (integral imagers) passively capture 4D space-angle information in a single-shot [13], which can be used for 3D reconstructions. This concept can be built into a thin form factor with microlens arrays [14] or Fresnel zone plates [15]. Lenslet array-based 3D capture schemes have also been used in microscopy [16], where wave-optical [3,17] or scattering [4,17] effects can be included. All of these systems, however, must trade resolution (or field-of-view) for single-shot capture, limiting the number of useful voxels. DiffuserCam improves upon this tradeoff, capturing large 3D volumes with high voxel counts in a single exposure.

Lensless imaging has also been demonstrated with coherent systems in both 2D [18–21] and 3D [22–26], but these methods require active (coherent) illumination, which limits applications. Further, many coherent methods do not generate unambiguous 3D reconstructions, but rather use digital refocusing to estimate depth. DiffuserCam, on the other hand, exhibits actual depth sectioning (in the absence of occlusions) for "true 3D."

Since methods for imaging through scattering often use diffusers as a proxy for general scattering media [27–29], our mathematical models will be similar. However, instead of trying to mitigate the effects of unwanted scattering, here we use the diffuser as an optical element in our system design. We choose a thin, optically smooth diffuser that refracts pseudorandomly, producing high-contrast patterns under incoherent illumination. Such diffusers have been used in light field imaging [30] and coherent holography [23,31]. Coherent multiple scattering has been demonstrated as an encoding mechanism for 2D compressed sensing [6], but necessitates a transmission matrix approach that does not scale well past a few thousand pixels. We achieve similar benefits without needing coherent illumination, and we reconstruct 3D objects, rather than 2D. Finally, an important benefit of our system over previous work is the simple calibration and efficient computation that allow for 3D reconstruction at megavoxel scales with superior image quality.

B. System Overview

DiffuserCam is part of the class of mask-based passive lensless imagers in which a phase or amplitude mask is placed a small distance in front of a sensor, with no main lens. Our mask (the diffuser) is a thin transparent phase object with smoothly varying thickness (see Fig. 1). When a temporally incoherent point source is placed in the scene, we observe a high-frequency pseudorandom caustic pattern at the sensor. The caustic patterns, termed point spread functions (PSFs), vary with the 3D position of the source, thereby encoding 3D information.

To illustrate how the caustics capture 3D information, Fig. 2 shows simulations of the PSFs for a point source at different locations in the object space. A lateral shift of the point source causes a lateral translation of the PSF [32]. An axial shift of the point source causes (approximately) a scaling of the PSF. Hence, each 3D position in the volume generates a unique caustic pattern. The structure and spatial frequencies present in the PSFs determine our reconstruction resolution. By using a phase mask (which concentrates light better than an amplitude mask) and designing the system to retain high spatial frequencies over a large range of depths, DiffuserCam attains good lateral resolution across the volumetric field-of-view.

By assuming that all points in the scene are incoherent with each other, the measurement can be modeled as a linear combination of PSFs from different 3D positions. We represent this as the matrix–vector multiplication,

$$\mathbf{b} = \mathbf{H}\mathbf{v}, \tag{1}$$

where $\mathbf{b}$ is a vector containing the 2D sensor measurement and $\mathbf{v}$ is a vector representing the intensity of the object at every point in the 3D FoV, sampled on a user-chosen grid. $\mathbf{H}$ is the forward model matrix whose columns consist of each of the caustic patterns created by the corresponding 3D points on the object grid. The number of entries in $\mathbf{b}$ and the number of rows of $\mathbf{H}$ are equal to the number of pixels on the image sensor, but the number of columns in $\mathbf{H}$ is set by the choice of reconstruction grid (discussed in Section 3). Note that this model does not account for partial occlusion of sources.

模型不考虑部分遮挡

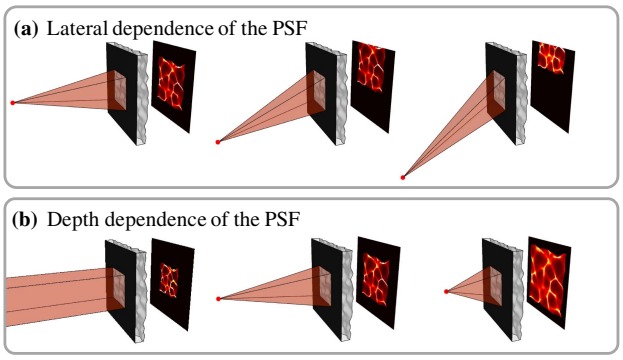

Fig. 2. Caustic pattern shifts with lateral shifts of a point source in the scene and scales with axial shifts. (a) Ray-traced renderings of caustics as a point source moves laterally. For large shifts, part of the pattern is clipped by the sensor. (b) The caustics magnify as the source is brought closer.

To reconstruct the 3D object, $\mathbf{v}$, from the measured 2D image, $\mathbf{b}$, we must solve Eq. (1) for $\mathbf{v}$. However, if we solve it on a 3D reconstruction grid that corresponds to the full optical resolution of our system (measured in Section 3.B), $\mathbf{v}$ will contain more voxels than there are sensor pixels. In this case, $\mathbf{H}$ has more columns than rows, so the problem is underdetermined and we cannot uniquely recover $\mathbf{v}$ simply by inverting Eq. (1). To remedy this issue, we rely on sparsity-based principles [33]. We exploit the fact that many 3D objects are sparse in some domain, meaning that the majority of coefficients are zero after a linear transformation. We enforce this sparsity as a prior and solve the ℓ_1 regularized nonnegativity-constrained inverse problem:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}\geq 0}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{b} - \mathbf{H}\mathbf{v}\|_2^2 + \tau\|\Psi\mathbf{v}\|_1. \tag{2}$$

Here, Ψ maps $\mathbf{v}$ into a domain in which it is sparse ($\Psi\mathbf{v}$ is mostly zeros), and τ is a tuning parameter that adjusts the degree of sparsity. For objects that are sparse in voxels, such as fluorescent particles in a volume, Ψ is the identity matrix. In our results, we show the reconstruction of objects that are not sparse in voxels but are sparse in the gradient domain. Hence, we choose Ψ to be the finite difference operator and $\|\Psi\mathbf{v}\|_1$ to be the 3D total variation (TV) semi-norm [34]. In general, any linear sparsity transformation may be used (e.g., wavelets), but we use only identity and gradient representations in this work.

有限差分算子

Equation (2) is the basis pursuit problem in compressed sensing [33]. For this optimization procedure to succeed, $\mathbf{H}$ must have distributed, uncorrelated columns. Since our diffuser creates high spatial frequency caustics that spread across many pixels in a pseudorandom fashion, any shift or magnification of the caustics leads to a new pattern that is uncorrelated with the original one (quantified in Supplement 1 Fig. S4). As discussed in Sections 2.B and 2.C, these properties allow us to reconstruct 3D images via compressed sensing.

2. METHODS

A. System Architecture

The hardware setup for our prototype DiffuserCam [Fig. 3(a)] consists of an off-the-shelf diffuser (Luminit 0.5°) placed at a fixed distance in front a sensor (PCO.edge 5.5 Color camera, 6.5 μm pixels). The diffuser has a flat input surface and an output surface described statistically as Gaussian low-pass-filtered white noise with an average spatial feature size of 140 μm and average slope magnitude of 0.7° (see Supplement 1 Fig. S1). The convex bumps on the diffuser surface can be thought of as randomly spaced microlenses that have statistically varying focal lengths and f-numbers. The average focal length determines the distance at which the caustics have highest contrast (the *caustic plane*), which is where we place the sensor [30]. This distance, measured experimentally, is 8 mm for our diffuser. However, the high average f-number of the bumps (8 mm/140 μm = 57) means that the caustics maintain high contrast over a large range of propagation distances. Therefore, the diffuser need not be placed precisely at the caustic plane (in our prototype, $d = 8.9$ mm). We also affix a 5.5×7.5 mm aperture on the textured side of the diffuser to limit the support of the caustics. 光圈值（F Number）=镜头焦距（mm）/光圈口径（mm）

Similar to a traditional camera, the sensor's pixel pitch should Nyquist sample the minimum features of the PSF. Since the f-number of the smallest bumps on the diffuser determines the minimum feature size of the caustics, it will also set the lateral optical resolution. In our case, the smallest features generated by the caustic patterns are roughly twice the pixel pitch of our sensor, so we perform 2×2 binning on the data, yielding 1.3 megapixel images, before applying our reconstruction algorithm.

B. Convolutional Forward Model

Recovering a 3D image requires knowing the system matrix, $\mathbf{H}$, which is extremely large. Measuring or storing the full $\mathbf{H}$ would be impractical, requiring millions of calibration images and operating on multi-terabyte matrices. Instead, we use the convolution model outlined below to drastically reduce the complexity of both the calibration and computation.

We describe the object, $\mathbf{v}$, as a set of point sources located at (x, y, z) on a non-Cartesian 3D grid. The relative radiant power collected by the aperture from each source is $\mathbf{v}(x, y, z)$. The caustic pattern at pixel (x', y') on the sensor due to a unit-powered point source at (x, y, z) is the PSF, $h(x', y'; x, y, z)$. Thus, $\mathbf{b}(x', y')$ is the sum of all 2D sensor measurements for each nonzero point in $\mathbf{v}$ after propagating through the diffuser and onto the sensor. This lets us explicitly write the matrix–vector multiplication $\mathbf{H}\mathbf{v}$ by summing over all voxels in the FoV, so

$$\mathbf{b}(x', y') = \sum_{(x,y,z)} \mathbf{v}(x, y, z) h(x', y'; x, y, z). \tag{3}$$

Our convolution model amounts to a shift invariance (or infinite memory effect [27,28]) assumption, which greatly simplifies the evaluation of Eq. (3). Consider the caustics created by point sources at a fixed distance, z, from the diffuser. Because the diffuser surface is slowly varying and smooth, the paraxial approximation holds. This implies that a lateral translation of the source by $(\Delta x, \Delta y)$ leads to a lateral shift of the caustics on the sensor by $(\Delta x', \Delta y') = (m\Delta x, m\Delta y)$, where m is the paraxial magnification. We validate this behavior in both simulations (see Fig. 2) and experiments (see Section 3.D). For notational convenience, we define the on-axis caustic pattern at depth z as $h(x', y'; z) \coloneqq h(x', y'; 0, 0, z)$. Thus, the off-axis caustic pattern is given by $h(x', y'; x, y, z) = h(x' + mx, y' + my; z)$. Plugging into Eq. (3), the sensor measurement is then given by

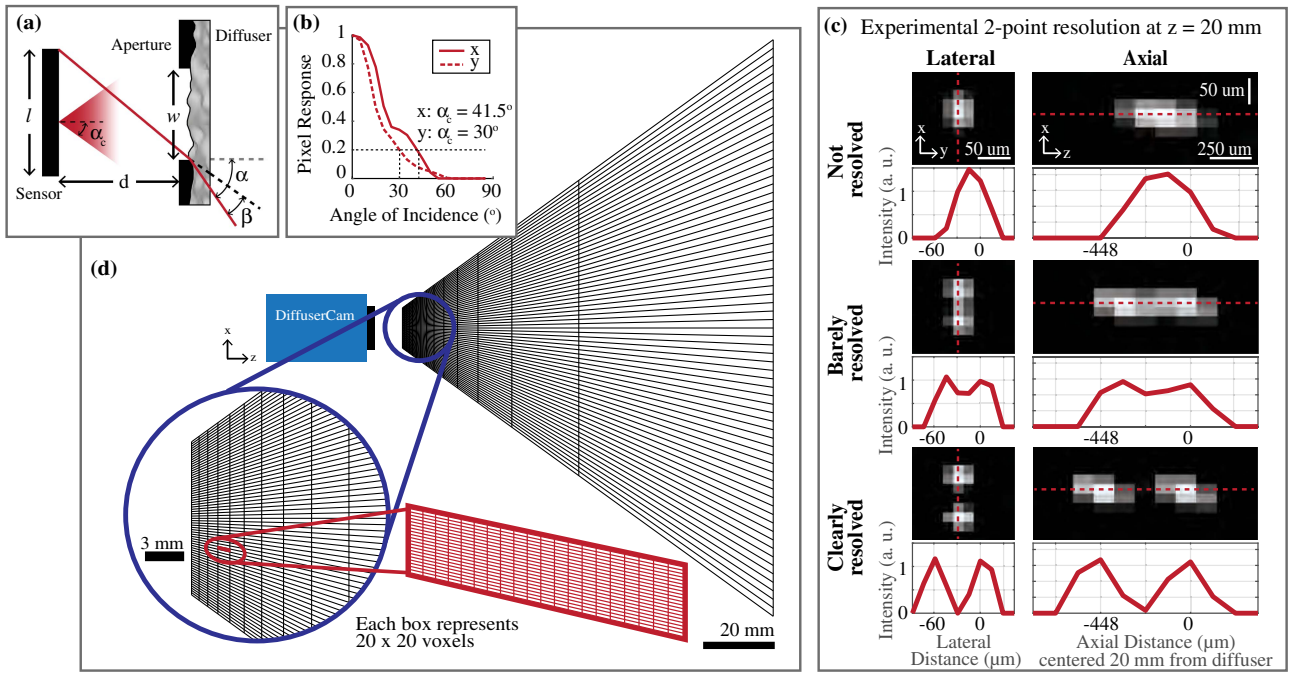

Fig. 3. Experimentally determined field-of-view (FoV) and resolution. (a) System architecture with design parameters. (b) Angular pixel response of our sensor. We define the angular cutoff (α_c) as the angle at which the response falls to 20%. (c) Reconstructed images of two points (captured separately) at varying separations laterally and axially, near the $z = 20$ mm depth plane. Points are considered resolved if they are separated by a dip of at least 20%. (d) To-scale nonuniform voxel grid for 3D reconstruction. The chosen voxel grid is based on the system geometry and Nyquist-sampled two-point resolution over the entire FoV. For visualization purposes, each box represents 20 × 20 voxels, as shown in red.

$$\mathbf{b}(x',y') = \sum_z \sum_{(x,y)} \mathbf{v}(x,y,z) h(x' + mx, y' + my; z)$$
$$= \mathbf{C} \sum_z \left[\mathbf{v}\left(\frac{-x'}{m}, \frac{-y'}{m}, z\right) * h(x',y';z) \right]. \quad (4)$$

Here, $*$ represents the 2D discrete convolution over (x', y'), which returns arrays that are larger than the originals. Hence, we crop to the original sensor size, denoted by the linear operator $\mathbf{C}$ (see Supplement 1 Fig. S5 for more details). For an object discretized into N_z depth slices, the number of columns of $\mathbf{H}$ is N_z times larger than the number of elements in $\mathbf{b}$ (i.e., the number of sensor pixels), so our system is underdetermined.

The cropped convolution model provides three benefits. First, it allows us to compute $\mathbf{Hv}$ as a linear operator in terms of N_z images, rather than instantiating $\mathbf{H}$ explicitly (which would require petabytes of memory to store). In practice, we evaluate the sum of 2D cropped convolutions using a single circular 3D convolution, implemented with 3D FFTs, which scale well to large arrays (see Supplement 1, Section 2.C). Second, it provides a theoretical justification of our system's capability for compressed sensing; derivations in [35] show that translated copies of a random pattern provide close-to-optimal performance.

The third benefit of our convolution model is that it enables simple calibration. Rather than measuring the system response for every voxel (hundreds of millions of images), we only need to capture a single calibration image of the caustic pattern from an on-axis point source. Though the scaling effect described in Section 1.B suggests that we could use only one image to calibrate the entire 3D space (by scaling it to predict PSFs at different depths), we obtain better results when we calibrate the PSF at each depth. A typical calibration thus consists of capturing images as a point source is moved axially. This takes minutes, but must only be performed once. The added aperture at the diffuser ensures that a point source at the minimum z distance generates caustics that just fill the sensor, so that the entire PSF is captured in each image (see Supplement 1 Fig. S2).

C. Inverse Algorithm

Our inverse problem is extremely large in scale, with millions of inputs and outputs. Even with the convolution model described above, using projected gradient techniques is extremely slow due to the time required to compute the proximal operator of 3D TV [36]. To alleviate this issue, we use the alternating direction method of multipliers (ADMM) [37] and derive a variable splitting that leverages the specific structure of our problem.

Our algorithm uses the fact that Ψ can be written as a circular convolution for both the 3D TV and native sparsity cases. Additionally, we factor the forward model in Eq. (4) into a diagonal component, $\mathbf{D}$, and a 3D convolution matrix, $\mathbf{M}$, such that $\mathbf{H} = \mathbf{DM}$ (details in Supplement 1). Thus, both the forward operator and the regularizer can be computed in 3D Fouier space. This enables us to use variable-splitting [38–40] to formulate the constrained counterpart of Eq. (2) as

$$\hat{\mathbf{v}} = \underset{w \geq 0, u, v}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{b} - \mathbf{D}v\|_2^2 + \tau \|u\|_1$$
$$\text{s.t. } v = \mathbf{Mv}, u = \Psi \mathbf{v}, w = \mathbf{v}, \quad (5)$$

where v, u, and w are auxiliary variables. We solve Eq. (5) by following the augmented Lagrangian arguments [41]. Using ADMM, this results in the following scheme at iteration k,

$$
\begin{aligned}
u^{k+1} &\leftarrow \mathcal{T}_{\frac{\tau}{\mu_2}}(\Psi \mathbf{v}^k + \eta^k/\mu_2)\\
v^{k+1} &\leftarrow (\mathbf{D}^T\mathbf{D} + \mu_1 I)^{-1}(\xi^k + \mu_1 \mathbf{M}\mathbf{v}^k + \mathbf{D}^T\mathbf{b})\\
w^{k+1} &\leftarrow \max(\rho^k/\mu_3 + \mathbf{v}^k, 0)\\
\mathbf{v}^{k+1} &\leftarrow (\mu_1 \mathbf{M}^T\mathbf{M} + \mu_2 \Psi^T\Psi + \mu_3 I)^{-1} r^k\\
\xi^{k+1} &\leftarrow \xi^k + \mu_1(\mathbf{M}\mathbf{v}^{k+1} - v^{k+1})\\
\eta^{k+1} &\leftarrow \eta^k + \mu_2(\Psi \mathbf{v}^{k+1} - u^{k+1})\\
\rho^{k+1} &\leftarrow \rho^k + \mu_3(\mathbf{v}^{k+1} - w^{k+1}),
\end{aligned}
$$

where

$$
r^k = (\mu_3 w^{k+1} - \rho^k) + \Psi^T(\mu_2 u^{k+1} - \eta^k) + \mathbf{M}^T(\mu_1 v^{k+1} - \xi^k).
$$

Note that $\mathcal{T}_\nu$ is a vectorial soft-thresholding operator with a threshold value of ν [42], and ξ, η, and ρ are the Lagrange multipliers associated with v, u, and w, respectively. The scalars μ_1, μ_2, and μ_3 are penalty parameters that we compute automatically using the tuning strategy in [37]. A MATLAB implementation of our algorithm is available at [43].

Although our algorithm involves two large-scale matrix inversions, both can be computed efficiently and in closed form. Since $\mathbf{D}$ is diagonal, $(\mathbf{D}^T\mathbf{D} + \mu_1 I)$ is itself diagonal, requiring complexity $\mathcal{O}(n)$ to invert using point-wise multiplication. Additionally, all three matrices in $(\mu_1 \mathbf{M}^T\mathbf{M} + \mu_2 \Psi^T\Psi + \mu_3 I)$ are diagonalized by the 3D discrete Fourier transform (DFT) matrix, so inversion of the entire term can be done using point-wise division in the 3D frequency space. Therefore, its inversion has good computational complexity, $\mathcal{O}(n^3 \log n)$, since it is dominated by two 3D FFTs being applied to n^3 total voxels. We parallelize our algorithm on the CPU using C++ and Halide [44], a high-performance programming language for image processing (see Supplement 1 Fig. S6 for runtime performance).

A typical reconstruction requires at least 200 iterations. Solving for $2048 \times 2048 \times 128 = 537$ million voxels takes 26 min (8 s per iteration) on a 144-core workstation and requires 85 gigabytes of RAM. A smaller reconstruction ($512 \times 512 \times 128 = 33.5$ million voxels) takes 3 min (1 s per iteration) on a four-core laptop with 16 gigabytes of RAM.

3. SYSTEM ANALYSIS

Unlike traditional cameras, the performance of computational cameras depends on properties of the scene being imaged (e.g., the number of sources). As a consequence, standard two-point resolution metrics may be misleading, as they do not predict resolving power for complex objects. To address this issue, we propose a new local condition number metric that we believe better predicts performance. We analyze resolution, FoV, and the validity of the convolution model, then combine these analyses to determine the appropriate sampling grid for our experiments.

A. Field-of-View

At every depth in the volume, the angular half-FoV is determined by the most extreme lateral position that contributes to the measurement. There are two possible limiting factors. The first is the geometric angular cutoff, α, set by the aperture size, w, the sensor size, l, and the distance from the diffuser to the sensor, d [see Fig. 3(a)]. Since the diffuser bends light, we also take into account the diffuser's maximum deflection angle, β. This gives a geometric angular half-FoV at every depth of $l + w = 2d \tan(\alpha - \beta)$. The second limiting factor is the angular response of the sensor pixels. Real-world sensor pixels may not accept light at the high angles of incidence that our lensless camera accepts, so the sensor angular response [shown in Fig. 3(b)] may limit the FoV. Defining the angular cutoff of the sensor, α_c, as the angle at which the camera response falls to 20% of its on-axis value, we can write the overall FoV equation as

$$
\text{FoV} = \beta + \min\left[\alpha_c, \tan^{-1}\left(\frac{l + w}{2d}\right)\right]. \tag{6}
$$

Since we image in 3D, we must also consider the axial FoV. In practice, the axial FoV is limited by the range of calibrated depths. However, the system geometry creates bounds on possible calibration locations. Point sources arbitrarily close to the sensor would produce caustic patterns that exceed the sensor size. To avoid this complication, we impose a minimum object distance at which an on-axis point source creates caustics that fill the sensor. Point sources arbitrarily far from the sensor theoretically can be captured, but axial resolution degrades with depth. The hyperfocal plane represents the axial distance beyond which no depth discrimination is available, establishing an upper bound. Objects beyond the hyperfocal focal plane can still be reconstructed to create 2D images for photographic applications [45], without any hardware modifications.

In our prototype, the axial FoV ranges from the minimum calibration distance (7.3 mm) to the hyperfocal plane (2.3 m). The angular FoV is limited by the pixel angular acceptance ($\alpha_c = 41.5°$ in x, $\alpha_c = 30°$ in y). Combined with our diffuser's maximum deflection angle ($\beta = 0.5°$), this yields an angular FoV of $\pm 42°$ in x and $\pm 30.5°$ in y. We validate the lateral FoV experimentally by capturing a scene at optical infinity and measuring the angular extent of the result (see Supplement 1 Fig. S3).

B. Resolution

Investigating optical resolution is critical for quantifying system performance and choosing our reconstruction grid. Although the raw data is collected on a fixed sensor grid, we can choose the nonuniform 3D reconstruction grid arbitrarily. This choice of reconstruction grid is important. When the grid is chosen with voxels that are too large, resolution is lost. When the voxels are too small, extra computation is performed without resolution gain. In this section we explain how to choose the grid of voxels for our reconstructions, with the aim of Nyquist sampling the two-point optical resolution limit.

1. Two-Point Resolution

A common metric for resolution analysis in traditional cameras is two-point distinguishablity. We measure our system's two-point resolution by imaging scenes containing two point sources at different separation distances, built by summing together images of a single point source (1 μm pinhole, wavelength 532 nm) at two different locations. We reconstruct the scene using our algorithm, with $\tau = 0$ to remove the influence of the regularizer. To ensure best-case resolution, we use the full 5 MP sensor data (no binning). The point sources are considered distinguishable if the reconstruction has a dip of at least 20% between the sources, as in the Rayleigh criterion. Figure 3(c) shows reconstructions with point sources separated both laterally and axially.

Our system has highly non-isotropic resolution [Fig. 3(d)], but we can use our model to predict the two-point distinguishability

over the entire volume from localized experiments. Due to the shift invariance assumption, the lateral resolution is constant within a single depth plane and the paraxial magnification causes the lateral resolution to vary linearly with depth. For axial resolution, the main difference between the two point sources is the size of their PSF supports. We find pairs of depths such that the difference in their support widths is constant,

$$c = \frac{1}{z_1} - \frac{1}{z_2}. \tag{7}$$

Here, z_1 and z_2 are neighboring depths and c is a constant determined experimentally.

Based on this model, we set the voxel spacing in our grid to Nyquist sample the 3D two-point resolution. Figure 3(d) shows a to-scale map of the resulting voxel grid. Axial resolution degrades with distance until it reaches the hyperfocal plane (~2.3 m from the camera), beyond which no depth information is recoverable. Due to the non-telecentric nature of the system, the voxel sizes are a function of depth, with the densest sampling occurring close to the camera. Objects within 5 cm of the camera can be reconstructed with somewhat isotropic resolution. In practice, this is where we place objects.

2. Multi-Point Resolution

In a traditional camera, resolution is a function of the system and is independent of the scene. In contrast, computational cameras that use nonlinear reconstruction algorithms may incur degradation of the effective resolution as the scene complexity increases. To demonstrate this in our system, we consider a more complex scene consisting of 16 point sources. Figure 4 shows experiments using 16 point sources arranged in a 4×4 grid in the (x, z) plane at two different spacings. The first spacing is set to match the measured two-point resolution limit ($\Delta x = 45$ μm, $\Delta z = 336$ μm). Despite being able to separate two points at this spacing, we cannot resolve all 16 sources. However, if we increase the source separation to ($\Delta x = 75$ μm, $\Delta z = 448$ μm), all 16 points are distinguishable [Fig. 4(d)]. In this example, the usable lateral resolution of the system degrades by approximately 1.7 × due to the increased scene complexity. As we show in Section 3.C, the resolution loss does not become arbitrarily worse as the scene complexity increases.

This experiment demonstrates that existing resolution metrics cannot be blindly used to determine the performance of computational cameras like ours. How can we then analyze resolution if it depends on object properties? In the next section, we introduce a general theoretical framework to assess resolution in computational cameras like ours.

C. Local Condition Number Theory

Our goal is to provide a new theory that describes how the effective reconstruction resolution of computational cameras changes with object complexity. To do so, we introduce a numerical analysis of how well our forward model can be inverted.

First, note that recovering the image $\mathbf{v}$ from the measurement $\mathbf{b} = \mathbf{H}\mathbf{v}$ entails simultaneous estimation of the locations of all nonzeros within our image reconstruction, $\mathbf{v}$, as well as the values at each nonzero location. To simplify the problem, suppose an oracle tells us the exact location of every source within the 3D scene. This corresponds to knowing *a priori* the support of $\mathbf{v}$, so we then need only determine the *values* of the nonzero elements in $\mathbf{v}$. This can be done by solving a least-squares problem using a sub-matrix consisting of only the columns of $\mathbf{H}$ that correspond to the indices of the nonzero voxels. If this problem fails, then the more difficult problem of simultaneously determining the nonzero locations *and* their values will certainly fail.

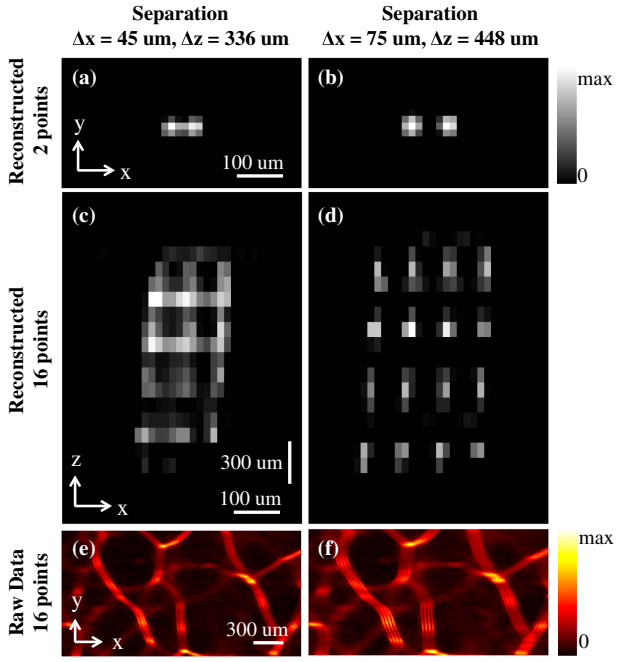

Fig. 4. Our computational camera has object-dependent performance, such that the resolution depends on the number of points. (a) To illustrate, we show here a situation with two points successfully resolved at the two-point resolution limit $(\Delta x, \Delta z) = (45\ \mu\text{m}, 336\ \mu\text{m})$ at a depth of approximately 20 mm. (c) When the object consists of more points (16 points in a 4×4 grid in the x–z plane) at the same spacing, however, the reconstruction fails. (b) and (d) Increasing the separation to $(\Delta x, \Delta z) = (75\ \mu\text{m}, 448\ \mu\text{m})$ gives successful reconstructions. (e) and (f) A close-up of the raw data shows noticeable splitting of the caustic lines for the 16-point case, making the points distinguishable. Heuristically, the 16-point resolution cutoff is a good indicator of resolution for real-world objects.

In practice, the measurement is corrupted by noise. The maximal effect this noise can have on the least-squares estimate of the nonzero values is determined by the condition number of the sub-matrix described above. We therefore say that the reconstruction problem is ill-posed if any sub-matrices of $\mathbf{H}$ are very ill-conditioned. In practice, ill-conditioned matrices result in increased noise sensitivity and longer reconstruction times, as more iterations are needed to converge to a solution.

In general, finding the worst-case sub-matrix is difficult. However, because our system measurements vary smoothly for inputs within a small neighborhood, the worst-case scenario is when multiple sources are in a contiguous block (i.e., nearby measurements are most similar, either by shift or scaling). Therefore, we compute the condition number of sub-matrices of $\mathbf{H}$ corresponding to a group of point sources with the separation varying by integer numbers of voxels. We repeat this calculation for different numbers of sources. The results are shown in Fig. 5. As expected, the conditioning is worse when sources are closer together. In this case, increased noise sensitivity means that

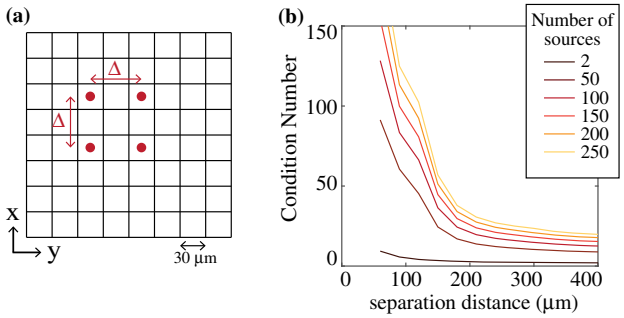

Fig. 5. Our local condition number theory shows how the resolution varies with the object complexity. (a) Virtual point sources are simulated on a fixed grid and moved by integer numbers of voxels to change the separation distance. (b) Local condition numbers are plotted for submatrices corresponding to grids of neighboring point sources with varying separation (at a depth 20 mm from the sensor). As the number of sources increases, the condition number approaches a limit, indicating that resolution for complex objects can be approximated by a limited number (but more than two) sources.

even small amounts of noise could prevent us from resolving the sources. This trend matches experiments in Figs. 3 and 4.

Figure 5 also shows that the local condition number increases with the number of sources in the scene, as expected. This means that the resolution will degrade as more and more sources are added. We see in Fig. 5, however, that as the number of sources increases, the conditioning approaches a limiting case. Hence, the resolution does not become arbitrarily worse with an increased number of sources. Therefore, we can estimate the system resolution for complex objects from distiguishability measurements with a limited number of point sources. This is experimentally validated in Section 4, where we find that the experimental 16-point resolution is a good predictor of the resolution for a USAF target.

Unlike the traditional two-point resolution metric, our new local condition number theory explains the resolution loss we observe experimentally. Since many optical systems are locally shift invariant, we believe that it is sufficiently general to be applicable to other computational cameras that use nonlinear algorithms, which likely exhibit similar performance loss.

D. Validity of the Convolution Model

In Section 2.B, we modeled the caustic pattern as shift invariant at every depth, leading to a simple calibration and efficient computation. Since our convolution model is an approximation, we should quantify its validity. Figures 6(a)–6(c) show registered close-ups of experimentally measured PSFs from plane waves incident at 0°, 15°, and 30°. The convolution model assumes that these are all exactly the same, although, they actually have subtle differences. To quantify the similarity across the FoV, we plot the inner product between each off-axis PSF and the on-axis PSF [see Fig. 6(d)]. The inner product is greater than 75% across the entire FoV and is particularly good within ±15° of the optical axis, indicating that the convolution model holds relatively well.

To investigate how the spatial variance of the PSF impacts the system performance, we use the peak width of the cross-correlation between the on-axis and off-axis PSFs to approximate the spot size off-axis. Figure 6(e) (solid) shows that we retain the on-axis resolution up to ±15°. Beyond that, the resolution

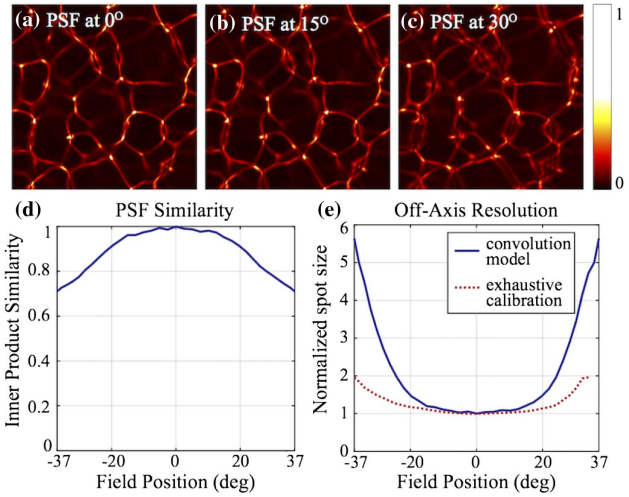

Fig. 6. Experimental validation of the convolution model. (a)–(c) Close-ups of registered experimental PSFs for sources at 0°, 15°, and 30°. The PSF at 15° is visually similar to that on-axis, while the PSF at 30° has subtle differences. (d) Inner product between the on-axis PSF and registered off-axis PSFs as a function of source position. (e) Resulting spot size (normalized by on-axis spot). The convolution model holds well up to ±15°, beyond which resolution degrades (solid). Exhaustive calibration would improve the resolution (dashed), at the expense of complexity in computation and calibration.

gradually degrades. To avoid model mismatch, one could replace the convolution model with exhaustive calibration over all positions in the FoV. This procedure would yield higher resolution at the edges of the FoV, as shown by the dashed line in Fig. 6(e). The gap between these lines is what we sacrifice in resolution by using the convolution model. However, in return, we gain simplified calibration and efficient computation, which makes the large-scale problem feasible.

4. EXPERIMENTAL RESULTS

Images of two objects are shown in Fig. 7, both illuminated using broadband white light and reconstructed with a 3D TV regularizer. We choose a reconstruction grid that approximately Nyquist samples the two-point resolution (by 2 × 2 binning the sensor pixels to yield a 1.3 megapixel measurement). Calibration images are taken at 128 different z-planes, ranging from $z = 10.86$ mm to $z = 36.26$ mm (from the diffuser), with spacing set according to conditions outlined in Section 3.B. The 3D images are reconstructed on a $2048 \times 2048 \times 128$ grid, but the angular FoV restricts the usable portion of this grid to the center 100 million voxels. Note that the resolvable feature size on this reconstruction grid varies based on the object complexity.

The first object is a negative USAF 1951 fluorescence test target, tilted 45° about the y-axis [Fig. 7(a)]. Slices of the reconstructed volume at different z planes are shown to highlight the system's depth sectioning capabilities. As described in Section 3.B, the spatial scale changes with depth. Analyzing the resolution in the vertical direction [Fig. 7(a) inset], we can easily resolve group 2/element 4 and barely resolve group 2/element 5 at $z = 24$ mm. This corresponds to resolving features 79 μm apart on the resolution target. This resolution is significantly worse than the two-point resolution at this depth (50 μm), but is similar to the 16-point resolution (75 μm). Hence, we reinforce our claim

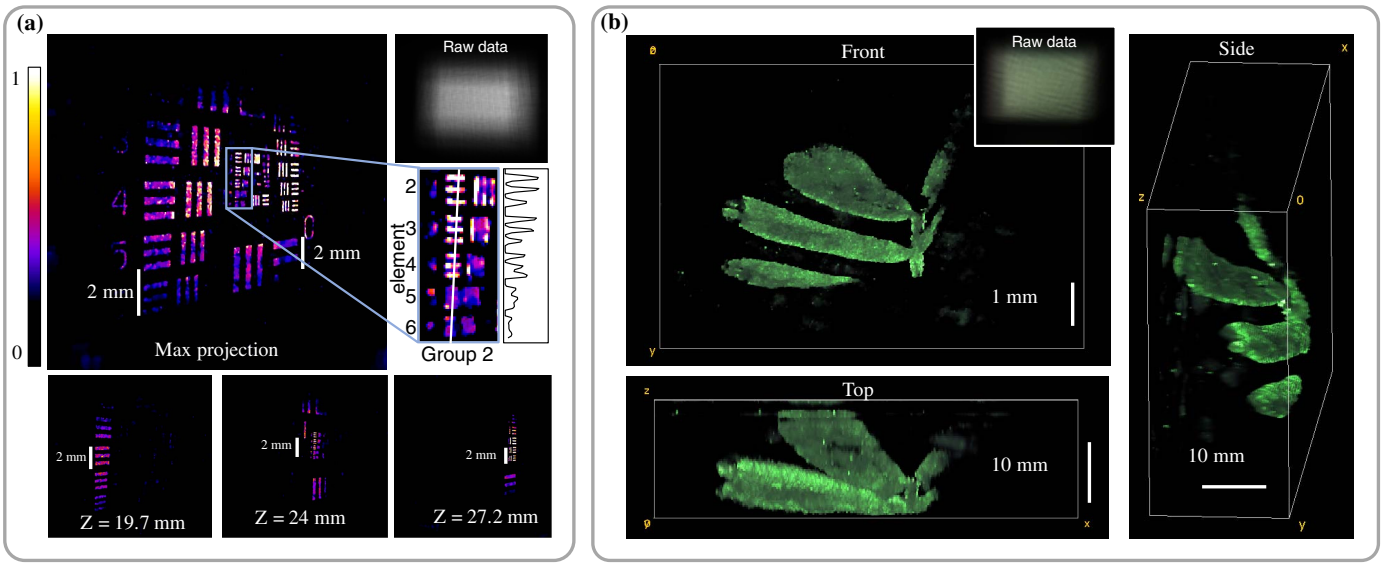

Fig. 7. Experimental 3D reconstructions. (a) Tilted resolution target, which was reconstructed on a 4.2 MP lateral grid with 128 z-planes and cropped to 640 × 640 × 50 voxels. The large panel shows the max projection over z. Note that the spatial scale is not isotropic. Inset is a magnification of group 2 with an intensity cutline, showing that we resolve element 5 at a distance of 24 mm, which corresponds to a feature size of 79 μm (approximately twice the lateral voxel size of 35 μm at this depth). The degraded resolution matches our 16-point distinguishability (75 μm at 20 mm depth). Lower panels show depth slices from the recovered volume. (b) Reconstruction of a small plant, cropped to 480 × 320 × 128 voxels, rendered from multiple angles.

that two-point resolution is a misleading metric for computational cameras, but multipoint distinguishability can be extended to more complex objects.

Finally, we demonstrate the ability of DiffuserCam to image natural objects by reconstructing a small plant [Fig. 7(b)]. Multiple perspectives of the 3D reconstruction are rendered to demonstrate the ability to capture the 3D structure of the leaves.

5. CONCLUSION

We demonstrated a simple optical system, with only a diffuser in front of a sensor, which is capable of single-shot 3D imaging. The diffuser encodes the 3D location of the point sources in caustic patterns, which allow us to apply compressed sensing to reconstruct more voxels than we have measurements. By using a convolution model that assumes that the caustic pattern is shift invariant at every depth, we developed an efficient ADMM algorithm for image recovery and simple calibration scheme. We characterized the FoV and two-point resolution of our system, and showed how resolution varies with object complexity. This motivated the introduction of a new condition number analysis, which we used to analyze how inverse problem conditioning changes with object complexity.

Funding. Hertz Foundation; U.S. Department of Defense (DOD); Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (SNF) (P2ELP2 172278, P2EZP2 159065); Defense Advanced Research Projects Agency (DARPA) (N66001-17-C-4015); Gordon and Betty Moore Foundation (GBMF4562); National Science Foundation (NSF) CAREER award.

Acknowledgment. Laura Waller is a Chan Zuckerberg Biohub Investigator. Ren Ng acknowledges support from the Alfred P. Sloan Foundation. Ben Mildenhall acknowledges funding from the Hertz Foundation and Grace Kuo is a National Defense Science and Engineering Graduate Fellow. Reinhard Heckel and Emrah Bostan acknowledge funding from the Swiss NSF. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. The authors thank Dr. Eric Jonas and the Rice FlatCam team for helpful discussions.

See Supplement 1 for supporting content.

†These authors contributed equally for this work.

REFERENCES

1. W. Denk, J. Strickler, and W. Webb, "Two-photon laser scanning fluorescence microscopy," Science **248**, 73–76 (1990).
2. T. F. Holekamp, D. Turaga, and T. E. Holy, "Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy," Neuron **57**, 661–672 (2008).
3. M. Broxton, L. Grosenick, S. Yang, N. Cohen, A. Andalman, K. Deisseroth, and M. Levoy, "Wave optics theory and 3-D deconvolution for the light field microscope," Opt. Express **21**, 25418–25439 (2013).
4. N. C. Pégard, H.-Y. Liu, N. Antipa, M. Gerlock, H. Adesnik, and L. Waller, "Compressive light-field microscopy for 3D neural activity recording," Optica **3**, 517–524 (2016).
5. M. F. Duarte, M. A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," IEEE Signal Process. Mag. **25**(2), 83–91 (2008).
6. A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron, "Imaging with nature: compressive imaging using a multiply scattering medium," Sci. Rep. **4**, 5552 (2014).
7. M. S. Asif, A. Ayremlou, A. Veeraraghavan, R. Baraniuk, and A. Sankaranarayanan, "Flatcam: replacing lenses with masks and computation," in *IEEE International Conference on Computer Vision Workshop (ICCVW)* (IEEE, 2015), pp. 663–666.
8. D. G. Stork and P. R. Gill, "Optical, mathematical, and computational foundations of lensless ultra-miniature diffractive imagers and sensors," Int. J. Adv. Syst. Meas. **7**, 201–208 (2014).

9. P. R. Gill, J. Tringali, A. Schneider, S. Kabir, D. G. Stork, E. Erickson, and M. Kellam, "Thermal escher sensors: pixel-efficient lensless imagers based on tiled optics," in *Computational Optical Sensing and Imaging* (Optical Society of America, 2017), paper CTu3B–3.
10. R. Fergus, A. Torralba, and W. T. Freeman, "Random lens imaging," Technical Report MIT-CSAIL-TR-2006-058 (Massachusetts Institute of Technology, 2006).
11. A. Stylianou and R. Pless, "Sparklegeometry: glitter imaging for 3D point tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2016), pp. 10–17.
12. J. Tanida, T. Kumagai, K. Yamada, S. Miyatake, K. Ishida, T. Morimoto, N. Kondou, D. Miyazaki, and Y. Ichioka, "Thin observation module by bound optics: concept and experimental verification," Appl. Opt. **40**, 1806–1813 (2001).
13. R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Computer Science Technical Report CSTR 2005-02, Stanford University, 2005), pp. 3418–3421.
14. R. Horisaki, S. Irie, Y. Ogura, and J. Tanida, "Three-dimensional information acquisition using a compound imaging system," Opt. Rev. **14**, 347–350 (2007).
15. K. Tajima, T. Shimano, Y. Nakamura, M. Sao, and T. Hoshizawa, "Lensless light-field imaging with multi-phased Fresnel zone aperture," in *IEEE International Conference on Computational Photography (ICCP)* (2017), pp. 76–82.
16. M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," in *ACM Trans. Graph. (Proc. SIGGRAPH)* (2006), Vol. **25**.
17. H.-Y. Liu, E. Jonas, L. Tian, J. Zhong, B. Recht, and L. Waller, "3D imaging in volumetric scattering media using phase-space measurements," Opt. Express **23**, 14461–14471 (2015).
18. W. Harm, C. Roider, A. Jesacher, S. Bernet, and M. Ritsch-Marte, "Lensless imaging through thin diffusive media," Opt. Express **22**, 22146–22156 (2014).
19. W. Chi and N. George, "Optical imaging with phase-coded aperture," Opt. Express **19**, 4294–4300 (2011).
20. A. Singh, G. Pedrini, M. Takeda, and W. Osten, "Scatter-plate microscope for lensless microscopy with diffraction limited resolution," Sci. Rep. **7**, 10687 (2017).
21. A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," Optica **4**, 1117–1125 (2017).
22. D. Brady, K. Choi, D. Marks, R. Horisaki, and S. Lim, "Compressive holography," Opt. Express **17**, 13040–13049 (2009).
23. K. Lee and Y. Park, "Exploiting the speckle-correlation scattering matrix for a compact reference-free holographic image sensor," Nat. Commun. **7**, 13359 (2016).
24. W. Bishara, T.-W. Su, A. F. Coskun, and A. Ozcan, "Lensfree on-chip microscopy over a wide field-of-view using pixel super-resolution," Opt. Express **18**, 11181–11191 (2010).
25. H. Faulkner and J. Rodenburg, "Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm," Phys. Rev. Lett. **93**, 023903 (2004).
26. A. Singh, D. Naik, G. Pedrini, M. Takeda, and W. Osten, "Exploiting scattering media for exploring 3D objects," Light Sci. Appl. **6**, e16219 (2017).
27. O. Katz, P. Heidmann, M. Fink, and S. Gigan, "Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations," Nat. Photonics **8**, 784–790 (2014).
28. E. Edrei and G. Scarcelli, "Memory-effect based deconvolution microscopy for super-resolution imaging through scattering media," Sci. Rep. **6**, 33558 (2016).
29. A. Singh, D. Naik, G. Pedrini, M. Takeda, and W. Osten, "Looking through a diffuser and around an opaque surface: a holographic approach," Opt. Express **22**, 7694–7701 (2014).
30. N. Antipa, S. Necula, R. Ng, and L. Waller, "Single-shot diffuser-encoded light field imaging," in *IEEE International Conference on Computational Photography (ICCP)* (2016), pp. 1–11.
31. Y. Kashter, A. Vijayakumar, and J. Rosen, "Resolving images by blurring: superresolution method with a scattering mask between the observed objects and the hologram recorder," Optica **4**, 932–939 (2017).
32. S. Feng, C. Kane, P. A. Lee, and A. D. Stone, "Correlations and fluctuations of coherent wave transmission through disordered media," Phys. Rev. Lett. **61**, 834–837 (1988).
33. E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," IEEE Signal Process. Mag. **25**(2), 21–30 (2008).
34. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D **60**, 259–268 (1992).
35. F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," Commun. Pur. Appl. Math. **67**, 1877–1904 (2014).
36. A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," IEEE Trans. Image Process. **18**, 2419–2434 (2009).
37. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Found. Trends Mach. Learn. **3**, 1–122 (2010).
38. M. S. C. Almeida and M. Figueiredo, "Deconvolving images with unknown boundaries using the alternating direction method of multipliers," IEEE Trans. Image Process. **22**, 3074–3086 (2013).
39. A. Matakos, S. Ramani, and J. A. Fessler, "Accelerated edge-preserving image restoration without boundary artifacts," IEEE Trans. Image Process. **22**, 2019–2029 (2013).
40. M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," IEEE Trans. Image Process. **19**, 2345–2356 (2010).
41. J. Nocedal and S. J. Wright, *Numerical Optimization* (Springer, 2006).
42. Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," SIAM J. Imag. Sci. **1**, 248–272 (2008).
43. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "DiffuserCam," http://www.laurawaller.com/research/diffusercam/ (2017). Accessed: 2017-11-17.
44. J. Ragan-Kelley, C. Barnes, A. Adams, S. Paris, F. Durand, and S. Amarasinghe, "Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines," in *ACM SIGPLAN Notices* (2013), Vol. **48**, pp. 519–530.
45. G. Kuo, N. Antipa, R. Ng, and L. Waller, "DiffuserCam: diffuser-based lensless cameras," in *Computational Optical Sensing and Imaging* (Optical Society of America, 2017), paper CTu3B–2.