## Fullwave design of cm-scale cylindrical metasurfaces via fast direct solvers

Wenjin Xue,<sup>1,2,\*</sup> Hanwen Zhang,<sup>2,3,4,\*</sup> Abinand Gopal,<sup>4</sup> Vladimir Rokhlin,<sup>4</sup> and Owen D. Miller<sup>2,3</sup>

<sup>1</sup>Department of Electrical Engineering, Yale University, New Haven, Connecticut 06511, USA

<sup>2</sup>Energy Sciences Institute, Yale University, New Haven, Connecticut 06511, USA

<sup>3</sup>Department of Applied Physics, Yale University, New Haven, Connecticut 06511, USA

<sup>4</sup>Department of Mathematics, Yale University, New Haven, Connecticut 06511, USA

(Dated: August 21, 2023)

Large-scale metasurfaces promise nanophotonic performance improvements to macroscopic optics functionality, for applications from imaging to analog computing. Yet the size scale mismatch of centimeter-scale chips versus micron-scale wavelengths prohibits use of conventional full-wave simulation techniques, and has necessitated dramatic approximations. Here, we show that tailoring "fast direct" integral-equation simulation techniques to the form factor of metasurfaces offers the possibility for accurate and efficient full-wave, large-scale metasurface simulations. For cylindrical (two-dimensional) metasurfaces, we demonstrate accurate simulations whose solution time scales *linearly* with the metasurface diameter. Moreover, the solver stores compressed information about the simulation domain that is reusable over many design iterations. We demonstrate the capabilities of our solver through two designs: first, a high-efficiency, high-numerical-aperture metalens that is 20,000 wavelengths in diameter. Second, a high-efficiency, large-beam-width grating coupler. The latter corresponds to millimeter-scale beam design at standard telecommunications wavelengths, while the former, at a visible wavelength of 500 nm, corresponds to a design diameter of 1 cm, created through full simulations of Maxwell's equations.

### I. INTRODUCTION

Simulating Maxwell's equations underpins the exploration of wave phenomena across nanophotonics, from complex interference effects [1-3] to exotic band topology [4-6]. Yet for the emerging field of metasurfaces [7-9, where diameters are readily centimeter-scale (cmscale) or larger, standard tools such as finite difference [10, 11] and finite element [10, 12] methods cannot provide accurate solutions in a reasonable time. One typically must employ dramatic approximations, such as the unit-cell approximation [13–15], for proof-of-principle demonstrations. In this paper, we develop a "fast direct" solver, based on well-conditioned integral formulations of electromagnetic scattering, to achieve fast and accurate fullwave simulation and design of large-scale metasurfaces. Fast-direct solver techniques have been proposed and developed in the applied mathematics community over the past decade [16–22], descendent from fast multipole methods [23-27]; we show that a modified fastdirect implementation tailored to the unique geometry of metasurfaces can lead to accurate simulations with computation times scaling *linearly* (technically, quasilinearly) with the diameter of the metasurface. We implement the fast-direct technique for cylindrical metasurfaces (translation-invariant in one direction) of both TE and TM polarizations, and demonstrate superior scaling compared with a finite-difference approach. The speed and accuracy of this approach enables a striking demonstration: the fullwave design (by adjoint optimization) of a high-NA metalens that is twenty thousand wavelengths in diameter. For a 500 nm visible wavelength,

this corresponds to a metasurface diameter of 1 cm. Our simulation technique readily accepts any source field, which we illustrate in a second demonstration of a highefficiency, large-area, aperiodic grating coupler. Our results showcase the promise of fast-direct fullwave solvers for large-area metasurfaces, enabling the full complexity of nanophotonic wave physics to be harnessed at macroscopic scales.

The workhorse simulation tools in nanophotonics are finite-difference and finite-element methods [10-12]. There are a few advantages of these methods that have made them so popular. First, matrix discretizations of differential equations are relatively easy to implement. Second, the matrices are sparse, thanks to the local nature of electromagnetic interactions, such that solutions of the linear equations can easily be accelerated [28]. Finally, for a subset of these approaches—including the finite-difference time-domain (FDTD) methodparallelization can be efficient and relatively straightforward to implement. However, these collective advantages also come with drawbacks. Differential operators are unbounded (typically exhibited as eigenvalues going to infinity), leading to ill-conditioned matrix representations [16, 29]. Finite-resolution sampling creates wavespeed inaccuracies that cause increasingly large phaseaccumulation errors at increasing scattering-region sizes. Hence, even if the amount of time to run an FDTD simulation nominally increases linearly with the size of a domain (for a fixed resolution), to keep the error fixed one has to increase the resolution, quickly leading to impractically large computation times for large scatterers.

Integral equations are a well-known alternative to their differential counterparts, arising through consideration of the current/polarization response fields within a scattering body (or at its interfaces) and enforcing consis-

<sup>\*</sup> equal contribution

tency between these currents and the fields they excite. The central operator in an integral equation is a convolutional Green's-function operator that represents the field excited from a point source in the *absence* of the scattering body. Green's functions encode the correct speed of light, preventing the phase error plaguing differential approaches. Moreoever, integral equations can be numerically well-conditioned by the nature of integral operators [16]. These advantages, however, come with their own drawbacks. First, the singular nature of Green's functions (e.g., divergent fields at a point source) requires careful discretization. Second, the resulting matrices are dense, which can quickly lead to impossible long simulation times without clever acceleration techniques. A common approach for acceleration is to exploit the convolutional nature of Green's functions in an iterative linear-equation solver, which relies on fast matrix-vector multiplication.

Two leading integral-equation techniques are the discrete dipole approximation (DDA) [30-35] and the method of moments (MoM) [10, 36]. Their approaches highlight some of the tensions in overcoming the drawbacks discussed above. Both DDA and MoM techniques are typically implemented with iterative solvers that rely on fast matrix-vector multiplication. In DDA this is particularly simple, as the integral equation is discretized on a uniform grid whose translational symmetry enables use of fast Fourier transforms. However, it has traditionally been difficult to accurately discrete the singular Green's functions on a uniform grid, and DDA solutions tend to offer a low accuracy that further degrades with modest refractive indices or scatterer sizes [34]. Conversely, MoM methods typically use a reasonable numerical integration (quadrature) scheme, but incur significant complexity in generating triangular/tetrahedral meshes and the corresponding matrix elements, and then in developing algorithms for fast matrix-vector multiplication without the translational symmetry of the uniform grid. Both DDA and MoM solvers can achieve quasilinear computation times per iteration [36], O(N) for N degrees of freedom ("order N," up to logarithmic corrections), but the number of iterations needed for convergence increases with the scatterer size, ultimately stalling these approaches before they reach the largest possible sizes.

"Fast direct solvers" [16] represent a promising new approach to solving integral equations. A noteworthy recent example is given in Ref. [22], which starts with a simple uniform grid, but uses the high-order quadrature scheme developed in Ref. [29] to accurately resolve the Green's-function singularity. Then, instead of using an iterative solver that suffers at large scatterer sizes, Ref. [22] use a direct solver build from a multilevel compression of the Green's function matrix, exploiting the information-sparsity of long-range interactions [16]. This "fast-direct" solver successfully reduces the computation time of two-dimensional simulations from cubic scaling with N simulation degrees of freedom, i.e.,  $O(N^3)$  (for standard dense linear algebra solvers) to  $O(N^{3/2})$ . Ideally, however, one would like a simulation approach that scales linearly with the problem size, in order to enable simulations at the largest size scales.

Complementary to such "full-wave" solvers are approximation techniques, exploiting physical simplifications to reduce computational complexity. For metasurfaces, for example, unit-cell [13–15] and overlapping-domain [37] approximation methods are popular schemes for building large-area, single-layer devices from small-region constituents. Similarly, for waveguide and grating-coupler devices, adiabaticity is used to simplify in-plane propagation and out-of-plane coupling [38–40]. However, in each case the set of devices that satisfy the requisite approximations is strongly constrained, and much of the multiple-scattering physics of electromagnetic waves must be discarded.

In this work, we improve the fast-direct, integralequation-based simulation technique of Ref. [16] to efficiently simulate metasurface geometries at extraordinarily large scales (tens of thousands of wavelengths in diameter). We start by reviewing fast-direct solvers and describing our proposed improvements (Sec. II). Our fast solver exploits the large aspect ratio of metasurface geometries and has a simulation time that scales linearly with metasurface diameter. We demonstrate its capabilities in Sec. III, where we test its simulation prowess against a standard FDTD simulation tool, and then use it design metasurfaces that are thousands of wavelengths in diameter. Finally, we conclude with a brief summary and discussion of new avenues to pursue (Sec. IV).

#### II. A FAST-DIRECT SOLVER FOR METASURFACES

For a scattering domain V with material susceptibility distribution  $\chi(\mathbf{r})$ , within which an incident field  $\mathbf{E}_{inc}$ induces a polarization-field response  $\mathbf{P}(\mathbf{r})$ , the volume integral (Lippmann–Schwinger) equation is

$$\mathbf{P}(\mathbf{r}) = \chi(\mathbf{r}) \left( \mathbf{E}_{\text{inc}}(\mathbf{r}) + \int_{V} \mathbf{G}(\mathbf{r}, \mathbf{r'}) \mathbf{P}(\mathbf{r'}) d\mathbf{r'} \right), \quad (1)$$

where **G** is the background Green's function, whose convolution with **P** determines the scattered field. Throughout, we will assume a background of free space. Integralequation methods can be adapted to non-vacuum backgrounds, though care is needed to ensure that the simulations remain fast and efficient. Equation (1) applies to any nonmagnetic scatterer in three dimensions, if  $\chi$  is interpreted as a tensor field and **G** as a dyadic Green's function. For the remainder of this article, for simplicity of notation we will assume an isotropic scalar susceptibility in two dimensions. Generalizations to tensor susceptibilities are simple and straightforward. Generalizations to three-dimensional Green's functions are possible and, as discussed in the Conclusions section, represent an important future step. In this section, we describe a fast-direct solver for Eq. (1), tailored to metasurface geometries, i.e., scatterers with large aspect ratios. Starting from a square grid discretization, which is easy to set up but low accuracy, local correction terms (by "Duan–Rokhlin quadrature") can increase accuracy with only a few near-diagonal corrections (Sec. II A). This Green's function can then be compressed by interpolative decomposition, using a binary tree structure that further aids in computing a complete or approximate inverse of the integral-equation matrix (Sec. II B). If the approximate inverse is built, then a few iterations will quickly converge to the final solution.

#### A. Duan-Rokhlin quadrature

Square grids are powerfully simple for discretization. They are especially common in finite difference methods [10, 11], where they yield relatively simple secondorder-accurate approximations of differential operators. For integral equations, however, a difficulty arises for a square-grid discretization (or any "Nyström" discretization, in which the unknown variables are the field values at specific positions, instead of basis-function coefficients) of Eq. (1): the diagonal elements of the matrix representation of the Green's function operator would be proportional to the self-term  $\mathbf{G}(\mathbf{r},\mathbf{r})$ , which diverges. The simplest approach, which yields the DDA method, is to ignore the self term in the summation. In a twodimensional scattering domain with discrete points  $\mathbf{r}_i$  on the grid, and grid spacing h, the DDA approximation of Eq. (1) is:

$$\mathbf{P}(\mathbf{r}_i) = \chi(\mathbf{r}_i) \left( \mathbf{E}_{\text{inc}}(\mathbf{r}_i) + \sum_{j \neq i} \mathbf{G}(\mathbf{r}_i, \mathbf{r}_j) \mathbf{P}(\mathbf{r}_j) h^2 \right), \quad (2)$$

which can be understood mathematically as the application of the "punctured trapezoidal rule" to Eq. (1). As resolution increases and h decreases, the error of the solution of Eq. (1) decays at best as  $h^2$ . Various versions of DDA approximate the self-term through physical polarizability arguments [30–35], but none of these approaches change the underlying error scaling as a function of resolution. The sizeable errors in the DDA approximation, and their relatively slow scaling with h, ultimately limit the size of the scattering body (e.g., metasurface) that can be simulated quickly and efficiently.

There is a simple-to-implement correction to Eq. (2) that can dramatically increase its accuracy, without sacrificing the simplicity of the square-grid discretization. The idea was developed in Ref. [29], based on the following observation. Since a polarization field  $\mathbf{P}$  is compactly supported and smooth for a compactly supported smooth scatterer (as discussed above), any integration of  $\mathbf{P}$  by standard trapezoidal rule will converge superalgebraically, i.e., it will incur an error smaller than  $h^m$  for any m > 0 (cf. Ref. [41]). Hence, the errors in Eq. (2) as an approximation of Eq. (1) come entirely from the singular elements of **G**, along the diagonal. In Ref. [29], such errors are systematically corrected by modifying a few weights near the diagonal. The first correction that one can make is to the singularity (diagonal) coefficient itself; for 2D scalar waves, correcting this term leads to  $\sim h^4$  error scaling. Correcting successive neighbors by moment matching leads to higher-order accuracies. For a set  $N_i$  that includes the point  $\mathbf{r}_i$  and its nearest neighbors (and possibly next nearest neighbors, etc.), one arrives at a modified version of Eq. (2):

$$\mathbf{P}(\mathbf{r}_{i}) = \chi(\mathbf{r}_{i}) \left( \mathbf{E}_{\text{inc}}(\mathbf{r}_{i}) + \sum_{j \neq i} \mathbf{G}(\mathbf{r}_{i}, \mathbf{r}_{j}) \mathbf{P}(\mathbf{r}_{j}) h^{2} + \sum_{\mathbf{r}_{j} \in N_{i}} \tau(\mathbf{r}_{i}, \mathbf{r}_{j}) \mathbf{P}(\mathbf{r}_{j}) h^{2} \right)$$
(3)

Equation (3) successfully achieves both simple, squaregrid discretization, as well as high accuracy. Ref. [29] demonstrated up to *tenth*-order accuracy for TE (scalar) waves with a 25-point stencil for  $\tau(\mathbf{r}_i, \mathbf{r}_i)$ ; we find below that even fourth-order accuracy leads to significant improvements over conventional DDA, and requires only a single correction, along the diagonal, for each source point. The straightforward corrections of Eq. (3) highlight the more general idea that it is easier to achieve stable, high-order discretizations for integral equations than for their differential counterparts because numerical integration is free of numerical cancellations that plague numerical differentiation. The order of accuracy of a discretization of Eq. (1) is solely determined by the order of the numerical quadrature for convolving the Green's function against the polarization field [42].

The simplicity of the above high-order quadrature scheme comes from the assumption that the scatterer is *smooth*, i.e., without discontinuous boundaries. This might seem useless for the typical scenario in nanophotonics, with discrete materials and sharp boundaries. However, any discontinuous material interface can be well-approximated by a narrow but smooth transition. One can smooth susceptibility distributions using Planck-taper, sigmoid, and other filter functions, tailored to the discretization resolution to ensure that the boundary is sufficiently sampled, to obtain results close enough to exact solutions after applying quadrature corrections.

By combining the quadrature scheme and boundary smoothing described above, we arrive at a matrix form of Eq. (1), the volume integral equation:

$$(\mathbf{I} - \mathbf{B}\mathbf{G})\boldsymbol{\psi} = \mathbf{f} \tag{4}$$

where **I** is an identity matrix, **B** is a diagonal matrix with elements  $\mathbf{B}_{i,i} = \chi(\mathbf{r}_i)h^2$ , **G** is the free-space Green's function matrix, including correction terms  $\tau$ , and  $\psi$  and **f** are the vector representations of **P** and  $\chi \mathbf{E}_{inc}$ , respectively. The next question, then, is how to efficiently *solve* the matrix equation of Eq. (4). It is well-conditioned (due to its integral-equation origins), but the Green's-function matrix **G** is dense. A fast solver is needed to compute  $\psi = (\mathbf{I} - \mathbf{BG})^{-1}\mathbf{f}$ .

#### B. Fast direct solver techniques

Fast direct solvers compute the solution  $\psi = (\mathbf{I} - \mathbf{I})^{-1}$  $\mathbf{BG}$ )<sup>-1</sup>**f** directly and efficiently based on compressed representations of G [22]. A low-rank matrix is easy to compress: one can simply store the singular vectors corresponding to nontrivial singular values. Unfortunately, **G** is neither low rank nor approximately low rank. However, it is hierarchically off-diagonal low rank (HODLR), which means that the off-diagonal blocks of  $\mathbf{G}$  are low rank, at every scale. For example, writing **G** as a block  $2 \times 2$  matrix, the two off-diagonal blocks are low rank. Moreover, each of the diagonal blocks can be decomposed into smaller  $2 \times 2$  blocks, each of which have off-diagonal blocks which themselves are low rank. And so on, recursively. This property is depicted in Fig. 1(a), which shows both the magnitude of the elements of a prototype Green's function, and the recursive decomposition of a Green's-function matrix into diagonal (shaded) and low-rank (white) blocks. Physically, the HODLR property arises because even though wave amplitudes decay slowly in space, the information contained in those amplitudes decays quickly [16]. (For example, accurate modeling of the radiation into a fixed volume V can require significantly fewer multipole moments as the separation increases.)

A HODLR representation can yield simple recursive expressions for the needed matrix inverse, but recursive computation can be inefficient in practice, and especially difficult to efficiently implement in a parallel computing environment. A non-recursive approach is to store each constituent component of the matrix in a binary tree structure, as in Fig. 1(b). Physically, this corresponds to physical subdivisions of a scatterer, as depicted in Fig. 1(b). A compressed representation of  $\mathbf{G}$ can consist of the information of each block (of varying sizes) in Fig. 1(a). The diagonal (shaded) blocks require dense storage, but can all be relatively small. The off-diagonal blocks can be stored with any decomposition (singular value, interpolative, etc.) that exploits the low-rank nature of those blocks. The total storage scales as  $O(kN \log N)$ , where k is the upper limit of the off-diagonal rank of **G** [16]. In two dimensions, k is proportional to  $N^{1/2}$ , while in three dimensions,  $k \sim O(N^{2/3})$ . Then matrix multiplication and matrix inversion can operate directly on these stored representations, in times that scale as  $O(kN \log N)$ . For our simulations, we compress **G** into hierarchically block separable (HBS) form [22], and we use interpolative decomposition (ID) for the low-rank approximation [16, 22], as it saves compression time and memory.

Further compression savings can typically be made by exploiting the surface equivalence principle: the fields emanating from a volume of currents can be exactly reproduced by appropriate surface currents. Hence, the low-rank approximation of the interaction **G** between any two blocks in Fig. 1(c) can replace the source volume with discrete surface points (on that volume), which is referred to as a "proxy surface" [16, 22]. By utilizing a number of points proportional to the surface instead of the volume, the compression complexity can be reduced, with a compression time that scales os  $O(N^{3/2})$  instead of  $O(N^2)$  [22].

However, for metasurface geometries that are thin in the direction perpendicular to their diameter, with enormous aspect ratios, the surface area to volume ratio is much larger than for a box domain, rendering such proxy surfaces ineffective in reducing computation time. We introduce a new twist on the proxy-surface approach, specifically tailored to metasurfaces. In a metasurface, at the first few tree levels (the larger domains in Fig. 1(b)), the dominant field interaction is typically horizontal wave propagation, such that discretizing the surfaces with equal number of points along top/bottom sidewalls as their left/right counterparts is wasteful (and leads to artificial inflation of the apparent rank of the interaction). Instead, we propose using an *open* proxy surface, vertical cuts extending above and below the metasurface, as depicted in Fig. 1(d), for these large-domain, first tree levels, which unlocks faster compression via proxy surfaces in metasurface geometries.

Given the HBS compression of  $\mathbf{G}$ , exploiting our vertical-cut proxy surfaces, the next step is to enable matrix-vector multiplication with the "solution" matrix  $(I-BG)^{-1}$ , without actually computing and storing this large, dense matrix. The key, again, is to exploit the rank structure of the matrices involved. The hierarchical block matrix form of  $\mathbf{I} - \mathbf{B}\mathbf{G}$  has the same HODLR structure as G. After separating the solution matrix (TBD: OR ITS INVERSE?) into a product of the diagonal and offdiagonal parts of the matrix (for example, see Sec. 5.6 of Ref. [16]), the inverse of the product becomes the inverse of a block-diagonal matrix multiplied by the inverse of an identity-plus-low-rank matrix; the latter itself of identityplus-low-rank form. Hence one can compute the inverses of the small blocks at the base of the hierarchy, then traverse the tree upward to build the matrices comprising the solution matrix,  $(\mathbf{I} - \mathbf{B}\mathbf{G})^{-1}$ .

Since we have not formed the large, dense solution matrix itself, we then need an algorithm to compute the product of the solution matrix with a source (incident) field **f**. We can do this by reversing the order of operations in building the inverse matrices, and doing a downward traversal of the tree, from root to leaves, sequentially adding the products of the relevant matrices with **f**. Given the compression and proxy-surface techniques outlined above, instead of the usual  $O(N^{3/2})$  complexity for general geometries, the total simulation complexity for metasurfaces is O(N).

The collective procedure for a fast and accurate integral-equation-based metasurface simulation solver is given in Fig. 2. Starting from a square-grid discretization of a smoothed geometry (step 1), one adds quadraturecorrection terms to the Green's-function matrix to enable higher-order accuracy (step 2). A binary tree data structure is then created for the hierchical compression



FIG. 1. (a) Left panel: the magnitude of Green's function matrix  $\mathbf{G}$ , a dense matrix with slowly decaying off-diagonal components. Right panel: the HODLR rank structure of  $\mathbf{G}$ . The hierarchical off-diagonal components (white) are low-rank, leaving only a small number of small, full-rank diagonal blocks (gray). (b) A binary tree structure for non-recursive storage of  $\mathbf{G}$ . (c) Hierarchical partition of a metasurface simulation region. Each partitioned box corresponds to a node in the binary tree. (d) A tailored vertical-cut proxy surface to most efficiently compress the interactions of one volume of the scatterer with another. Usage of this proxy surface in tandem with compression and acceleration techniques leads, up to logarithmic corrections, to a computation time scaling linearly with diameter.

and fast solution algorithms (step 3). As part of the compression, vertical-cut proxy surfaces are used, while random-matrix algorithms can be used for fast interpolative decomposition (ID) in steps 4 and 5. These matrices can then be used to represent the solution matrix, either accurately or with crude accuracy, where the latter is useful for preconditioning (steps 6,7). Finally, either the accurate inverse is multiplied by the excitation vector, or an iterative algorithm is used with the preconditioned system matrix (step 8). The orange boxes highlight the high-order quadrature ideas first proposed in Ref. [29], the green boxes highlight fast-direct solver ideas from Ref. [22], while the blue box highlights the metasurface-adapted proxy surface adaptations that we propose towards simulating, and designing, the largest possible metasurfaces. These steps are nearly independent of the dimensionality and polarization of the problem (2D TE, 2D TM, 3D), except that in the vector cases some reordering of the indices is helpful, and that the Green's function itself requires unique quadratures for each case. For the largest diameters, we use all of steps 1 through 8 for maximum computational efficiency, but one can use only a subset of the steps for smaller problems, to reduce implementation complexity.

#### **III. COMPUTATIONAL EXAMPLES**

In this section, we demonstrate the capability of this fast-direct, linear-in-diameter algorithm to simulate large- and very-large-scale metasurfaces. We start with a prototypical problem of small to modest sizes, where we can compare the simulation times of the fast-direct integral solver with a finite-difference time-domain solver. The key result is that the FDTD simulation error increases steadily as the domain size increases, whereas the fast-direct solver error does not. Next, we turn our efforts to the *design* of very-large-scale metasurfaces. First we design a metalens that is  $20,000\lambda$  in diameter, with a high efficiency even at a high numerical aperture. Finally, we show that this simulation approach works for any incident field, including in-plane waveguide-mode excitations, and design a large-area ( $1000\lambda$ -diameter) non-periodic grating coupler, for large-area beam generation.

#### A. Test case: fast-direct integral solver vs. FDTD

As a first example, we compare the performance of the integral solver proposed above with Meep, a state-of-theart FDTD solver [11]. For benchmarking purposes, we consider 2D TE simulations (electric field out of plane) of the geometries in Fig. 3(a). These geometries constitute non-periodic, multilayered metasurfaces, generated by starting with a non-symmetric center region comprising two blazed-grating regions (with randomly situated slanted pillars) with opposite blazing directions, then mimicking the left/right-pattern outwards, consistently doubling the size. We use this structure to induce strong nonlocality (waves that propagate left-to-right in addi-



FIG. 2. Procedure for fast and accurate integral-equation-based metasurface simulations. The orange boxes outline the highorder quadrature method developed in Ref. [29], while the green boxes highlight fast-direct solver techniques from Ref. [16]. The blue box highlights the metasurface-adapted proxy surface and randomized ID adaptations that we propose, enabling linear-in-diameter scaling of the simulation time. Smaller structures can be simulated with iterative methods (steps 1,2,8), but the largest structures require compression and a direct solver (steps 1–6,8) or an approximate direct solve with a few "polish" iterative solves at the end (steps 1–8).

tion to upward/downward propagation). Metasurfaces in which the interaction is almost entirely local (predominantly up/down propagation) can already be simulated by techniques exploiting the locality [15, 37, 40, 43–45], but such metasurfaces have limitations in performance if they cannot take advantage of nonlocality [46, 47]. A primary goal in the metasurface community is in taking steps to nonlocal [48–52] and multilayer [53–59] metasurfaces, which Fig. 3(a) effects. For generic wavelength  $\lambda$ , we consider 7 multilayer designs: the first has diameter 50 $\lambda$ , the second 100 $\lambda$ , and so forth (doubling each time), up to  $25 \times 2^6 \lambda = 1600 \lambda$  diameter. The grating thickness is taken to be  $1\lambda$ , the substrate  $0.45\lambda$ , and the air gaps (mimicking spacer layers [55, 56, 58, 59])  $0.2\lambda$ . The scatterer material is taken to have refractive index of 2, typical of transparent materials at visible frequencies (e.g. SiN [60]). For each diameter we create 5 sample metasurfaces to simulate (with different random pillar positions), to smooth the error rates. The resolution of the FDTD solver is taken to be 80 points per free-space wavelength (chosen to ensure no greater than 10% error at the smallest diameter), while the fast-direct integral solver uses 40 points per wavelength. To compute error rates, we find the "ground truth" solution at each diameter by using the fast-direct solver at high enough resolution to ensure converge to  $\lesssim 0.1\%$  accuracy. We also verify for the smallest diameter structure that the FDTD solutions indeed converge (albeit slowly) to the ground-truth solution.

Figure 3(b-d) depict data from the fast-direct integral-

equation solver and the FDTD solver for various scatterer diameters. Fig. 3(b) shows the total electric field (from our fast-direct solver) in the smallest-diameter case considered, with  $50\lambda$  diameter. The image shows the complex interference patterns common to nonlocal metasurfaces. Fig. 3(c) shows the simulation times for both solvers as a function of the diameter of the scatterer. The number of grid points per wavelength is kept fixed for each simulation, and the FDTD termination criterion measures the field energy in the scattering body. The absolute value of the simulation time is not crucial: the fastdirect solver is about twice as fast, but these numbers can be scaled based on CPU/GPU implementations, number of parallel cores, etc. (The integral-equation solver is run on a high-memory, 8-node, 3.3 GHz compute node, while the FDTD solver is run on a 24-cpu, 2.9 GHz compute node.) The total simulation time for a given application will also highly depend on whether a multi-frequency simulation is required (where FDTD can be very fast), multifrequency design is required (where FDTD is less fast due to slow adjoint computations [61]), multi-angle simulations or designs (where the integral-equation solver can be very fast), multi-iteration designs (where the integralequation solver has additional speedups), and so forth. In the quest for pushing simulations to the largest size scales, the key question is the amount of time required for accurate simulation, as a function of scatterer size.

As seen in Fig. 3(c), both solvers have simulation times that scale linearly with the size of the solver. This is almost a trivial fact for the FDTD solver, since we



FIG. 3. Comparison between the fast-direct integral-equation method and FDTD for nonlocal metasurfaces. (a) Geomety of the nonlocal metasurfaces: three layers, each generated from a non-symmetric center region comprising two blazed-grating regions (with randomly located slanted pillars) with opposite blazing directions, then adding the same the left/right-pattern outwards, consistently doubling the size. (b) Total electric field in our  $50\lambda$ -diameter metasurface. (c) Average simulation time over 5 metasurfaces at each diameter, with the simulation resolution fixed, which automatically leads to linear-in-time scaling for FDTD. The integral solver showing linear scaling is a demonstration of the predicted scaling from our compression and proxy-surface techniques. (d) Average simulation error for each technique. Whereas the FDTD solver average error increases from 10% to 18% as the diameter increases (with the largest single-simulation error exceeding 20%), the integral solver maintains less than 1% error.

keep the resolution fixed, and the number of time steps hardly changes. For the fast-direct integral-equation solver, however, the linear scaling is a significant achievement. This shows that our implementation of densematrix compression as well as hierarchical simulation algorithms have indeed correctly led to linear-in-time, integral-equation-based simulation.

The payoff for the linear-in-time integral-equation solver is shown in Fig. 3(d). The FDTD solver error increases as the size of the simulation region increases, as expected for a finite-difference discretization of the poorly conditioned differential form of Maxwell's equations. One can see that the possibly tolerable 10% errors at the smaller diameters are approaching 20% at the  $1600\lambda$ -diameter structure. At the extraordinarily large sizes that we can consider in the next section, the errors are too large to be useful, and increasing the resolution to the point of bringing the errors back to say 5% would incur enormous increases in simulation time. By contrast, the matrices created by discretizing the integral equations are well-conditioned, implying the possibility for high accuracy at very large sizes, which is born out in the integral-solver data in Fig. 3(d). The fast-direct solver can maintain accuracies better than 1%, which is important not only for simulation purposes, but additionally for iterative design algorithms, where errors in consecutive geometries can hamper the convergence speed of an optimization algorithm itself.

# B. Fullwave inverse design of a centimeter-scale metalens

Next we use the fast-direct integral-equation solver to design metasurfaces at extreme size scales. We consider a cylindrical metalens with a refractive index of 2 (as for SiN in the visible [60], under TE-polarized illumination (scalar electric field), with a diameter that is 20.000 wavelengths in size. This size corresponds to a 1 cm diameter for 500 nm wavelength. To the best of our knowledge, this is significantly larger than the largest fullwave metalens designs to date [59], which had diameters of  $129\lambda$  (for single-wavelength thickness) or  $100\lambda$  (for  $10\lambda$  thickness). We design for a numerical aperture of 0.9, a high NA that requires strong light focusing, and where unit-cell and related approximations break down. We consider a metasurface of "pillars"  $1\lambda$  in height, with arbitrary widths and separations, which together comprise about 200,000 design variables (the total number is not fixed as pillars can be created or destroyed during the optimization process). We use adjoint-based inverse design [62-64] to compute the gradients with respect to all degrees of freedom at every iteration, and we use gradient ascent for the optimization algorithm. We also impose a minimumfeature-size constraint by forbidding widths smaller than  $\lambda/50$  (10 nm for 500 nm wavelength). To balance the tradeoff between speed and accuracy, we use a simulation grid spacing of  $\lambda/36$ , which corresponds to field errors at a reasonable 2-3%. The initial compression of **G** re-



FIG. 4. Optimized 2D metalens with a 20,000 $\lambda$  diameter. (a) Evolution of the figure of merit over 50 optimization iterations, requiring less than three days to complete on a 16-CPU cluster. The final Strehl ratio reaches 56%. (b) Electric field intensity distribution on the focal plane. Inset: field intensity map near the focal plane (white dashed line), in a region  $10\lambda \times 20\lambda$ . (c) Electric field from the metalens through the focal plane. The absolute field value is plotted in log scale, with the colormap floor restricted to  $10^{-0.5}$  to de-emphasize the fields within the scatterer itself. The location of the  $1.45\lambda \times 20,000\lambda$  metalens is represented by the black line. (d) Detailed geometric structure of the metalens. Due to the large aspect ratio, the metalens is divided into 64 vertically stacked segments, each of which is  $312.5\lambda$  wide. Shaded portions in (c) are mapped to their corresponding rows in (d) to illustrate the segmentation.

quires 230 seconds, but can be reused for every iteration. Within every iteration, building  $(\mathbf{I} - \mathbf{BG})^{-1}$  takes 2330 seconds, and the solution via BICGSTAB takes 1150 sec for both direct and adjoint simulations, for a total time per iteration of 4630 seconds.

Figure 4(a) shows the evolution of the focal-point intensity of the metalens over the course of 50 optimization steps. By the end of the optimization, the Strehl ratio (defined as the focal-point intensity as a fraction of an ideal Airy-beam focusing intensity) reaches 56%, quite a high efficiency consider the numerical aperture of 0.9. This Strehl ratio is defined relative to *all power* incident upon the metasurface, not only the power transmitted through the surface. Figure 4(b,c) show the fields of the optimal design: one can clearly see nearly ideal performance in the focal plane (Fig. 4(b)), and the excellent focusing performance (Fig. 4(c)). Figure 4(d) shows the full metalens design. Given the difficulty of visualizing a structure with an approximately 20,000:1.5 aspect ratio, we partition the metasurface into 64 segments, each  $312.5\lambda$  wide, and vertically stack the neighboring segments. The shaded circles in Fig. 4(c) map different segments within the metasurface to the corresponding design for that segment in Fig. 4(d). The entire fullwave optimization process for the 20,000 $\lambda$  diameter metasurface requires only 64 hours to complete on a single 16-CPU compute node cluster. Further speedups should be possible with a more extensive parallelization effort.

Inverse design offers little guidance as to why an optimized structure performs well. In Fig. 5 we present a



FIG. 5. (a) Phase performance of the ideal lens (black), optimized metalens (gray), and a binary Fresnel lens (red). (b) The optimized metalens geometry shows an average-height profile similar to a Fresnel lens (averaged over one-micron regions), albeit in a lithographic form factor and with significant corrections. (c) Intensity enhancements of the binary Fresnel lens versus the optimized metalens.

partial mechanism for understanding the optimized design. First, Fig. 5(a) shows that the optimized lens effectively creates the correct outgoing-field phase profile across the surface of the metalens, even at the extreme edges where the phase varies more rapidly. Then, in Fig. 5(b) we see that a locally averaged (over every one micron) height profile of the optimized metalens resembles that of a Fresnel lens, albeit with some significant modifications near the height discontinuities, presumably to smooth and enhance the performance. A binary Fresnel lens, depicted in (red) in Fig. 5(b), does not create an effective phase profile in Fig. 5(a), which leads to subpar intensity focusing as depicted in Fig. 5(c). The optimized metalens offers a 40X relative intensity enhancement. Hence, we can interpret the complex structural patterns of the optimized metalens as a device that creates the ideal outgoing-field phase profile, while accounting for all multiple-scattering effects present in such strong-fieldbending scenarios.

### C. Fullwave inverse design of a millimeter-scale waveguide coupler

Large-area, nonperiodic metasurfaces can also be excited from the side, scattering waves into free space, thereby acting as grating couplers. Such side-coupled excitations experience very long propagation lengths, rendering typical unit-cell approximations [54] ineffective or of modest efficiency [13–15, 65]. The fast-direct solver approach is effective for any excitation. Hence the same algorithms and solver described above for conventional metasurface applications can seamlessly be applied to grating couplers and related beam-coupling devices. In this section, we design a 1000-wavelength-long grating coupler, well beyond the limits of conventional finite-difference and finite-element solvers, for coupling a waveguide mode to a large-area Gaussian beam.

For a telecommunications wavelength of 1550 nm, the coupler length of 1000 wavelengths corresponds to a physical length of 1.55 mm. We consider a material with permittivity 2.2, close to that of  $SiO_2$  [66], and TE-polarized propagation. The grating is taken to comprise pillars of  $0.4\lambda$  height on a  $0.45\lambda$ -thick substrate. We use the same shape-optimization algorithm described above, utilizing adjoint gradients at each iteration of the optimization, to maximize the out-coupling efficiency to a target Gaussian beam with field pattern  $E = E_0 \exp\left(-x^2/2w^2\right)$ , with parameter w set to  $341\lambda$ , creating a full-width at halfmaximum of about  $800\lambda$ , corresponding to a  $1.24 \,\mathrm{mm}$ beam size for  $\lambda = 1550$  nm. The amplitude  $E_0$  is chosen to correspond to all incoming waveguide power being scattered into the Gaussian mode. The waveguide mode is excited by an electric dipole source located multiple wavelengths from the start of the design region, to prevent back-reflection of the source near field.



FIG. 6.  $1,000\lambda$ -long waveguide coupler inverse designed via 750 iterations, each with two full simulations by our fast-direct solver. (a) Target (red) and optimal (blue) scattered field on a horizontal plane  $90\lambda$  above the coupler, excited from the left by the primary TE mode of the waveguide. The coupling efficiency is 53%, defined as the power flow into the target mode. (b) Field intensity, showing strong coupling to a vertically propagating Gaussian mode. The primary loss mechanism is downward scattering, which is likely unavoidable given the refractive indices and design constraints (height, lithography, etc.). (c) Depiction of the optimal metalens, partitioned into 10 segments of  $100\lambda$  widths.

To balance the tradeoff between simulation speed and accuracy, we use resolution  $\lambda/40$ , for which we find errors of apprximately 5%. The compression stage takes 30 seconds, which can be reused for all simulations during the optimization. Then simulating the forward and adjoint excitations takes a little more than two minutes per iteration, so that a 750-iteration optimization on an 8-core, 2.9 GHz compute node can be completed in about 33 hours. Starting from a random initial pattern, the design process iterates to a grating coupler with 53% coupling efficiency to a large-area Gaussian beam.

The final design and its corresponding field patterns are shown in Fig. 6. Figure 6(a) shows the target field (red) alongside the simulated field of the optimal design (blue), along a plane  $90\lambda$  above the grating-coupler surface. The full spatial distribution is shown in Fig. 6(b), while the optimal design is shown in Fig. 6(c). (For visualization purposes, the intensity is shown in log scale, with small-scale oscillations filtered out by Gaussian convolution. Figure 6(a) clearly depicts the oscillations. The colormap ceiling is saturated to de-emphasize the fields within the scatterer.) To accommodate the high aspect ratio of the device we split it into ten segments and arrange them vertically. From Fig. 6(a,b), one can see that the optimal field closely approaches the target. The key factor inhibiting even higher efficiency is power that is scattered downwards, due to the symmetric air regions above and below the device (pattern and thin substrate). For example, in the unpatterned structure, a dipole source placed directly above the waveguide radiates nearly symmetrically in the up and down directions: approximately 50–55% upwards, and 45–50% downwards. An examination of the polarization fields induced in the optimal design shows little vertical phase variation within the grating-coupler pillars, suggesting that the pillars cannot effectively cancel the downwards radiation. Hence, the optimal design is closely approaching the upper limit of what is possible for a grating coupler in this architecture.

#### IV. CONCLUSION AND OUTLOOK

In this work, we propose a fast-direct integral-equation solver for large-scale two-dimensional metasurface simulations. The solver takes advantage of high-order quadrature schemes to attain high simulation accuracy for modest resolution, as well as tailored algorithmic improvements within the emerging fast-direct solver framework [16, 22, 29, 67], to ultimately achieve linear-in-time scaling of the computations as a function of metasurface diameter. Our simulations show significant scalability advantages relative to finite-difference simulations, offering the possibility to reach extreme sizes and scales without sacrificing any aspects of wave scattering. We realize these possibilities with two optimal designs; first, for a high-NA metalens with 20,000 $\lambda$  diameter (corresponding to a cm-scale design for 500 nm wavelength), then, for a high-efficiency grating coupler with a 1000 $\lambda$  design-region length, targeting large-area beam generation applications.

The key next steps are the translation of these design principles to three dimensions. Two-dimensional simulations with rotational symmetries would already be of interest for many three-dimensional applications with cylindrical symmetry (e.g. Ref. [59]), after which the three-dimensional case follows. Each of these cases requires significant numerics and code development, as the Green's functions are different in each scenario. One interesting potential tradeoff in three dimensions would be to consider a surface-integral-equation approach, which has more complex quadrature implementation requirements, but which may yield computational speed advantages. As these simulations tools are developed, one can expect that fast-direct integral-equation solvers will unlock the possibility to do nanophotonic design at the scale of conventional optics.

#### V. ACKNOWLEDGEMENTS

W. Xue, H. Zhang and O.D. Miller were partially supported by the Air Force Office of Scientific Research Grant No. FA9550-22-1-0393. H. Zhang, A. Gopal, and V. Rokhlin were supported by ONR grant N00014-18-1-2353 and by NSF grant DMS-195275.

- J. Sol, A. Alhulaymi, A. D. Stone, and P. Del Hougne, Reflectionless programmable signal routers, Science Advances 9, eadf0323 (2023).
- [2] C. Sauvan, J.-P. Hugonin, I. S. Maksymov, and P. Lalanne, Theory of the spontaneous optical emission of nanosize photonic and plasmon resonators, Physical Review Letters **110**, 237401 (2013).
- [3] V. Ganapati, O. D. Miller, and E. Yablonovitch, Light trapping textures designed by electromagnetic optimization for subwavelength thick solar cells, IEEE Journal of Photovoltaics 4, 175 (2013).
- [4] H. Zhou, C. Peng, Y. Yoon, C. W. Hsu, K. A. Nelson, L. Fu, J. D. Joannopoulos, M. Soljacic, and B. Zhen, Observation of bulk fermi arc and polarization half charge from paired exceptional points, Science **359**, 1009 (2018).
- [5] A. Cerjan, C. W. Hsu, and M. C. Rechtsman, Bound states in the continuum through environmental design, Physical review letters 123, 023902 (2019).
- [6] L. Yuan, A. Dutt, and S. Fan, Synthetic frequency dimensions in dynamically modulated ring resonators, APL Photonics 6, 10.1063/5.0056359 (2021).
- [7] N. Yu and F. Capasso, Flat optics with designer metasurfaces, Nature materials **13**, 139 (2014).
- [8] S. M. Kamali, E. Arbabi, A. Arbabi, and A. Faraon, A review of dielectric optical metasurfaces for wavefront control, Nanophotonics 7, 1041 (2018).
- [9] P. Lalanne and P. Chavel, On the prehistory of optical metasurfaces, Photoniques, 41 (2023).
- [10] J.-M. Jin, Theory and computation of electromagnetic fields (John Wiley & Sons, 2015).
- [11] A. F. Oskooi, D. Roundy, M. Ibanescu, P. Bermel, J. D. Joannopoulos, and S. G. Johnson, Meep: A flexible free-software package for electromagnetic simulations by the FDTD method, Computer Physics Communications 181, 687 (2010).
- [12] J. N. Reddy, Introduction to the finite element method (McGraw-Hill Education, 2019).
- [13] M. Khorasaninejad, F. Aieta, P. Kanhaiya, M. A. Kats, P. Genevet, D. Rousso, and F. Capasso, Achromatic metasurface lens at telecommunication wavelengths, Nano letters 15, 5358 (2015).

- [14] W. T. Chen, A. Y. Zhu, V. Sanjeev, M. Khorasaninejad, Z. Shi, E. Lee, and F. Capasso, A broadband achromatic metalens for focusing and imaging in the visible, Nature nanotechnology 13, 220 (2018).
- [15] S. Wang, P. C. Wu, V.-C. Su, Y.-C. Lai, M.-K. Chen, H. Y. Kuo, B. H. Chen, Y. H. Chen, T.-T. Huang, J.-H. Wang, *et al.*, A broadband achromatic metalens in the visible, Nature nanotechnology **13**, 227 (2018).
- [16] P.-G. Martinsson, Fast direct solvers for elliptic PDEs (SIAM, 2020).
- [17] Y. Chen, A fast, direct algorithm for the Lippmann– Schwinger integral equation in two dimensions, Advances in Computational Mathematics 16, 175 (2002).
- [18] A. Gillman, Fast direct solvers for elliptic partial differential equations, Ph.D. thesis, University of Colorado at Boulder (2011).
- [19] A. Gillman and A. Barnett, A fast direct solver for quasiperiodic scattering problems, Journal of Computational Physics 248, 309 (2013).
- [20] E. Corona, P.-G. Martinsson, and D. Zorin, An O(N) direct solver for integral equations on the plane, Applied and Computational Harmonic Analysis 38, 284 (2015).
- [21] S. Ambikasaran, C. Borges, L.-M. Imbert-Gerard, and L. Greengard, Fast, adaptive, high-order accurate discretization of the Lippmann–Schwinger equation in two dimensions, SIAM Journal on Scientific Computing 38, A1770 (2016).
- [22] A. Gopal and P.-G. Martinsson, An accelerated, highorder accurate direct solver for the Lippmann–Schwinger equation for acoustic scattering in the plane, Advances in Computational Mathematics 48, 42 (2022).
- [23] L. Greengard and V. Rokhlin, A fast algorithm for particle simulations, Journal of computational physics 73, 325 (1987).
- [24] L. Greengard, The rapid evaluation of potential fields in particle systems (MIT press, 1988).
- [25] J. Carrier, L. Greengard, and V. Rokhlin, A fast adaptive multipole algorithm for particle simulations, SIAM journal on scientific and statistical computing 9, 669 (1988).
- [26] L. Greengard and V. Rokhlin, A new version of the fast multipole method for the Laplace equation in three di-

mensions, Acta numerica 6, 229 (1997).

- [27] H. Cheng, W. Y. Crutchfield, Z. Gimbutas, L. F. Greengard, J. F. Ethridge, J. Huang, V. Rokhlin, N. Yarvin, and J. Zhao, A wideband fast multipole method for the helmholtz equation in three dimensions, Journal of Computational Physics **216**, 300 (2006).
- [28] T. A. Davis, *Direct Methods for Sparse Linear Systems* (Society for Industrial and Applied Mathematics, 2006).
- [29] R. Duan and V. Rokhlin, High-order quadratures for the solution of scattering problems in two dimensions, Journal of Computational Physics 228, 2152 (2009).
- [30] E. M. Purcell and C. R. Pennypacker, Scattering and absorption of light by nonspherical dielectric grains, Astrophysical Journal, Vol. 186, pp. 705-714 (1973) 186, 705 (1973).
- [31] B. T. Draine, The discrete-dipole approximation and its application to interstellar graphite grains, Astrophysical Journal 333, 848 (1988).
- [32] B. T. Draine and J. Goodman, Beyond clausius-mossottiwave propagation on a polarizable point lattice and the discrete dipole approximation, Astrophysical Journal 405, 685 (1993).
- [33] B. T. Draine and P. J. Flatau, Discrete-dipole approximation for scattering calculations, JOSA A 11, 1491 (1994).
- [34] M. A. Yurkin and A. G. Hoekstra, The discrete dipole approximation: an overview and recent developments, Journal of Quantitative Spectroscopy and Radiative Transfer 106, 558 (2007).
- [35] M. A. Yurkin and A. G. Hoekstra, The discrete-dipoleapproximation code adda: capabilities and known limitations, Journal of Quantitative Spectroscopy and Radiative Transfer 112, 2234 (2011).
- [36] W. C. Chew, E. Michielssen, J. Song, and J.-M. Jin, Fast and efficient algorithms in computational electromagnetics (Artech House, Inc., 2001).
- [37] Z. Lin and S. G. Johnson, Overlapping domains for topology optimization of large-area metasurfaces, Optics express 27, 32445 (2019).
- [38] S. G. Johnson, P. Bienstman, M. Skorobogatiy, M. Ibanescu, E. Lidorikis, and J. Joannopoulos, Adiabatic theorem and continuous coupled-mode theory for efficient taper transitions in photonic crystals, Physical review E 66, 066608 (2002).
- [39] C. Pérez-Arancibia, R. Pestourie, and S. G. Johnson, Sideways adiabaticity: beyond ray optics for slowly varying metasurfaces, Optics express 26, 30202 (2018).
- [40] R. Pestourie, C. Pérez-Arancibia, Z. Lin, W. Shin, F. Capasso, and S. G. Johnson, Inverse design of large-area metasurfaces, Optics express 26, 33732 (2018).
- [41] L. N. Trefethen and D. Bau, Numerical linear algebra, Vol. 181 (SIAM, 2022).
- [42] D. Colton and R. Kress, Integral Equation Methods in Scattering Theory (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013) https://epubs.siam.org/doi/pdf/10.1137/1.9781611973167.
- [43] F. Aieta, M. A. Kats, P. Genevet, and F. Capasso, Multiwavelength achromatic metasurfaces by dispersive phase compensation, Science 347, 1342 (2015).
- [44] A. Arbabi, Y. Horie, M. Bagheri, and A. Faraon, Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission, Nature nanotechnology 10, 937 (2015).
- [45] M. Khorasaninejad, W. T. Chen, R. C. Devlin, J. Oh, A. Y. Zhu, and F. Capasso, Metalenses at visible wave-

lengths: Diffraction-limited focusing and subwavelength resolution imaging, Science **352**, 1190 (2016).

- [46] K. Shastri and F. Monticone, Nonlocal flat optics, Nature Photonics 17, 36 (2023).
- [47] A. Chen and F. Monticone, Dielectric nonlocal metasurfaces for fully solid-state ultrathin optical systems, ACS Photonics 8, 1439 (2021).
- [48] A. Overvig and A. Alù, Diffractive nonlocal metasurfaces, Laser & Photonics Reviews 16, 2100633 (2022).
- [49] H. Goh and A. Alù, Nonlocal scatterer for compact wavebased analog computing, Physical Review Letters 128, 073201 (2022).
- [50] A. C. Overvig, S. A. Mann, and A. Alù, Thermal metasurfaces: complete emission control by combining local and nonlocal light-matter interactions, Physical Review X 11, 021050 (2021).
- [51] H. Kwon, A. Cordaro, D. Sounas, A. Polman, and A. Alu, Dual-polarization analog 2d image processing with nonlocal metasurfaces, ACS Photonics 7, 1799 (2020).
- [52] H. Kwon, D. Sounas, A. Cordaro, A. Polman, and A. Alù, Nonlocal metasurfaces for optical signal processing, Physical review letters **121**, 173004 (2018).
- [53] A. Arbabi, E. Arbabi, Y. Horie, S. M. Kamali, and A. Faraon, Planar metasurface retroreflector, Nature Photonics 11, 415 (2017).
- [54] M. Mansouree, H. Kwon, E. Arbabi, A. McClung, A. Faraon, and A. Arbabi, Multifunctional 2.5 d metastructures enabled by adjoint optimization, Optica 7, 77 (2020).
- [55] E. Arbabi, A. Arbabi, S. M. Kamali, Y. Horie, M. Faraji-Dana, and A. Faraon, Mems-tunable dielectric metasurface lens, Nature communications 9, 812 (2018).
- [56] H. Kwon, E. Arbabi, S. M. Kamali, M. Faraji-Dana, and A. Faraon, Single-shot quantitative phase gradient microscopy using a system of multifunctional metasurfaces, Nature Photonics 14, 109 (2020).
- [57] A. Arbabi, E. Arbabi, S. M. Kamali, Y. Horie, S. Han, and A. Faraon, Miniature optical planar camera based on a wide-angle metasurface doublet corrected for monochromatic aberrations, Nature communications 7, 13682 (2016).
- [58] Y. Zhou, I. I. Kravchenko, H. Wang, J. R. Nolen, G. Gu, and J. Valentine, Multilayer noninteracting dielectric metasurfaces for multiwavelength metaoptics, Nano letters 18, 7529 (2018).
- [59] R. E. Christiansen, Z. Lin, C. Roques-Carmes, Y. Salamin, S. E. Kooi, J. D. Joannopoulos, M. Soljačić, and S. G. Johnson, Fullwave maxwell inverse design of axisymmetric, tunable, and multi-scale multi-wavelength metalenses, Optics Express 28, 33854 (2020).
- [60] L. Y. Beliaev, E. Shkondin, A. V. Lavrinenko, and O. Takayama, Optical, structural and composition properties of silicon nitride films deposited by reactive radiofrequency sputtering, low pressure and plasma-enhanced chemical vapor deposition, Thin Solid Films **763**, 139568 (2022).
- [61] A. M. Hammond, A. Oskooi, M. Chen, Z. Lin, S. G. Johnson, and S. E. Ralph, High-performance hybrid time/frequency-domain topology optimization for large-scale photonics inverse design, Optics Express 30, 4467 (2022).
- [62] J. S. Jensen and O. Sigmund, Topology optimization for nano-photonics, Laser Photon. Rev. 5, 308 (2011).

- [63] O. D. Miller, Photonic Design: From Fundamental Solar Cell Physics to Computational Inverse Design, Ph.D. thesis, University of California, Berkeley (2012).
- [64] A. Y. Piggott, J. Lu, K. G. Lagoudakis, J. Petykiewicz, T. M. Babinec, and J. Vučković, Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer, Nature Photonics 9, 374 (2015).
- [65] H. Chung and O. D. Miller, High-NA achromatic metalenses by inverse design, Optics Express 28, 6945 (2020).
- [66] Y. Arosa and R. de la Fuente, Refractive index spectroscopy and material dispersion in fused silica glass, Optics Letters 45, 4268 (2020).
- [67] P.-G. Martinsson and V. Rokhlin, A fast direct solver for scattering problems involving elongated structures, Journal of Computational Physics **221**, 288 (2007).