LASER & PHOTONICS REVIEWS

www.lpr-journal.org

Numerical Optimization Methods for Metasurfaces

Mahmoud M. R. Elsawy, Stéphane Lanteri, Régis Duvigneau, Jonathan A. Fan, and Patrice Genevet**

In recent years, metasurfaces have emerged as revolutionary tools to manipulate the behavior of light at the nanoscale. These devices consist of nanostructures defined within a single layer of metal or dielectric materials, and they offer unprecedented control over the optical properties of light, leading to previously unattainable applications in flat lenses, holographic imaging, polarimetry, and emission control, amongst others. The operation principles of metaoptics include complex light–matter interactions, often involving insidious near-field coupling effects that are far from being described by classical ray optics calculations, making advanced numerical modeling a requirement in the design process. In this contribution, recent optimization techniques used in the inverse design of high performance metasurfaces are reviewed. These methods rely on the iterative optimization of a Figure of Merit to produce a final device, leading to freeform layouts featuring complex and non-intuitive properties. The concepts in numerical inverse designs discussed herein will push this exciting field toward realistic and practical applications, ranging from laser wavefront engineering to innovative facial recognition and motion detection devices, including augmented reality retro-reflectors and related complex light field engineering.

1. Introduction

During the last decade, metasurfaces have received lots of attention due to their ability to precisely control the phase, amplitude, and wavefront of light. These light–matter interactions are mediated by ensembles of subwavelength meta-atoms, made of plasmonic or high dielectric refractive index materials, which have thicknesses within the range of the operating wavelength. The design of metasurfaces is generally achieved using the following two approaches. In the first approach, which we will refer to as the direct design approach, a rigorous full wave electromagnetic solver is used to study different classes of meta-atoms. Large sets of key meta-atom parameters are parametrically swept to create meta-atom libraries, and ensembles of meta-atoms from these libraries are assembled together to create metasurfaces with given desired optical responses. This method works well for certain applications. However, it does not incorporate near-field electromagnetic coupling effects between neighboring meta-atoms and does not generalize to large area, freeform devices.

The second approach is inverse design. With this approach, the desired optical response is defined as an objective cost function (for example, the deflection or focusing efficiencies) and the inverse problem solves for the shape and dimensions of the metasurface in a manner that maximizes the cost function value. In principle, the inverse design approach can account for near-field interactions by optimizing relatively large metasurface regions at a time. However, it requires rigorous and computationally efficient optimization techniques that can deal with large parameter spaces and multi-objective cost functions. There are two main classes of optimization methods that have currently been used in inverse design of metasurfaces: gradient-based algorithms and gradient-free approaches.

Gradient-based methods depend strongly on the initial guess of the solution and are efficient in finding local optima. Gradient-based method require the knowledge of the derivatives of the cost function with respect to design parameters, which can be evaluated analytically sometimes, or approximated numerically in most cases.[1] These methods have a lengthy history in metasurface design and were used early in the field[2,3] to create devices that maximized light diffraction at visible wavelengths. Today, several descent methods, ranging from steepest-descent to quasi-Newton methods for both constrained and unconstrained problems, have been applied to metasurfaces. In this short review, we focus on the most common methods used recently in the literature, including the objective-first[1,4] and topology

Dr. M. M. R. Elsawy, Prof. S. Lanteri, Dr. R. Duvigneau
Université Côte d'Azur
Inria, CNRS, LJAD
06902 Sophia Antipolis Cedex, France
E-mail: stephane.lanteri@inria.fr

Prof. J. A. Fan
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Dr. P. Genevet
CNRS, CRHEA
Université Côte d'Azur
06560 Sophia Antipolis Cedex, France
E-mail: patrice.genevet@crhea.cnrs.fr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/lpor.201900445

DOI: 10.1002/lpor.201900445

optimization algorithms, which have proven to be efficient and rigorous design methodologies.[4–8]

Gradient-free approaches are capable of capturing global optima,[9] albeit for a limited number of parameters, thereby overcoming the local minimum trapping issue of gradient-based algorithms. In other words, with gradient-free global optimization techniques, the final optimized results are not influenced by the initialization of the optimizer. In addition, some of these algorithms can deal with discrete optimization parameters and nondifferentiable objective functions, which are conditions that are generally not handled by gradient-based algorithms. Without gradients to help guide the optimization, convergence with global optimization algorithms is often considerably slower than convergence with gradient-based algorithms. To date, several global optimization techniques have been proposed in the context of metasurface design. To properly deal with the extensive (often discrete) parameter space and the existence of several local optima, the majority of inverse design methods of interest for metasurface design are stochastic and include genetic algorithms and evolutionary strategies.[10–18]

In addition to the methods discussed above, emergent approaches, including artificial neural networks and Bayesian optimization, have the potential to uncover surprising new metasurface designs. We will highlight in this review the key ideas behind these techniques and illustrate their versatility and advantages for the optimization of practical metasurfaces.[19,20]

We note there have been a number of reviews about inverse design for nanophotonics that have been published recently.[21–23] In Ref. [21], the authors focus on inverse design in the framework of deep learning. In Ref. [22], the authors provide general guidance for different optimization techniques, indicating their performances with respect to a range of nanophotonics problems. In Ref. [23], the authors present a collection inverse design techniques in nanophotonics and their specific application to nonlinear optics, integrated optics, and topological photonics. What differentiates our review from the previous one is that we we focus here our on numerical methods used for designing metasurfaces, providing a simple and clear illustration of these techniques and demonstrating their advantages and drawbacks. Indeed, we present an overview (with a discussion on the numerical implementation) of the main and most common optimization methods for metasurface designs. We do not intend to present an exhaustive literature review on nanophotonic structures optimized using inverse design techniques. This topic has already been covered in recently published works.[21–25]

The article is organized as follows. In Section 2, we discuss the main gradient-based optimization techniques used in the metasurface literature: the objective-first and topology optimization approaches. In Section 3, we introduce the most popular gradient-free techniques, that is, genetic algorithms, particle swarm optimization, and covariance matrix adaptation evolution strategy. In Section 4, we give a simple and practical introduction to artificial neural networks and how they can be applied to the inverse design of metasurface devices. In Section 5, we introduce the concept of Bayesian optimization and discuss its application to inverse design for nanophotonics. Finally, in Section 6, we discuss methods to incorporate robustness into the optimization process.

2. Gradient-Based Optimization Techniques

2.1. Objective-First Algorithm

The objective-first algorithm is a widely used optimization technique in nanophotonics.[1,9] As described in Chapter 6 in Ref. [1] and in refs. [6, 26], the algorithm begins by defining quantities to optimize. After specifying an objective or target function describing these quantities, the algorithm searches for the spatial distribution of dielectric material that maximizes this target while satisfying Maxwell's equations as accurately as possible.[6] For most of the applications discussed in this manuscript, we consider nonmagnetic materials, such that the algorithm searches for permittivity distributions within a design window.

Following the notations in Chapter 6 of Ref. [1] and assuming a time-harmonic dependency of the electromagnetic field, we write the general optimization problem as

$$\begin{aligned} &\underset{x,p}{\text{minimize}} \quad f(x) \\ &\text{subject to} \quad A(p(x))x - b(p(x)) = 0, \end{aligned} \tag{1}$$

A change of variables is applied to match this problem with electromagnetics: $H \to x$, $\epsilon^{-1} \to p$ such that $A(p) = \nabla \times \epsilon^{-1} \nabla \times -\mu_0 \omega^2$, and $b(p) = \nabla \times \epsilon^{-1} J$. J is the current density vector, μ_0 is the vacuum permeability, and ω is the angular frequency. Equation (1) indicates that the minimization of the target function, achieved by varying H and ϵ^{-1} simultaneously, is performed while satisfying the wave equation.[1] Note that Equation (1) is *a priori* a non-convex problem, since this requires solving for $p(x)$ and x simultaneously, which is in general a difficult problem.

The objective-first algorithm splits the optimization problem into two convex subproblems. One of these subproblems deals with the fields: given the permittivity, it solves for the usual wave equation and determines the fields that minimize the residual. The second subproblem solves for the permittivity distribution given constant electromagnetic fields. After this minimization, both subproblems are then merged together[1]:

Loop:

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad \| A(p)x - b(p) \|^2 \\ &\text{subject to} \quad f(x) = f_{ideal}, \\ &\underset{p}{\text{minimize}} \quad \| B(x)p - d(x) \|^2 \\ &\text{subject to} \quad p_{min} \le p \le p_{max}, \end{aligned} \tag{2}$$

f_{ideal} is the ideal performance for function $f(x)$. We see that the first subproblem tries to converge to an ideal solution that satisfies the wave equation up to some residual. The second subproblem, with $B(x) = \nabla \times (\nabla \times H) - \nabla \times J$ and $d(x) = \mu_0 \omega^2 H$, seeks to solve for the permittivity with the condition that it takes continuous values. In practice, the optimization domain that corresponds to the physical space containing the nanostructures is decomposed into equally spaced pixels, each specified by a given local permittivity value. The optimization process is iteratively

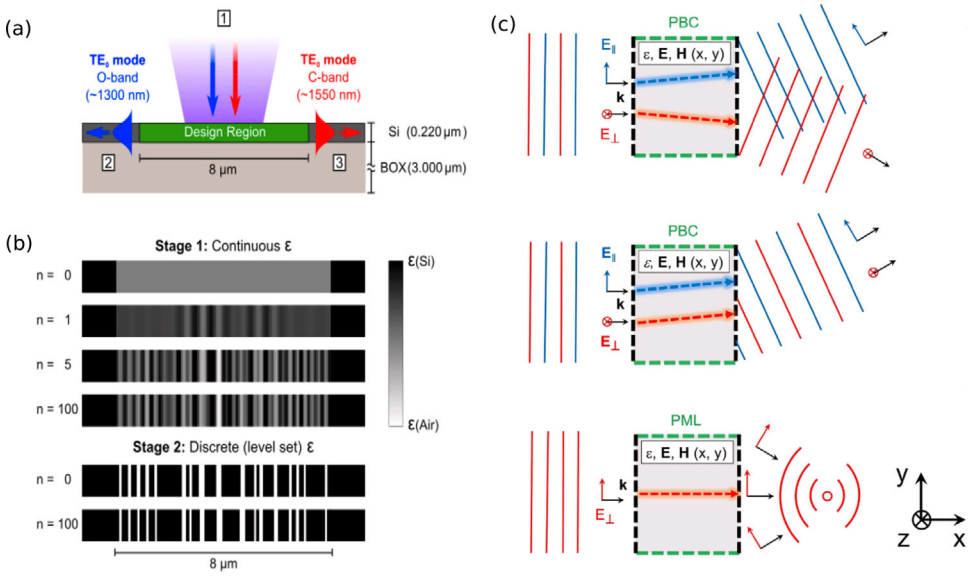


Figure 1. Optimized devices using the objective-first algorithm. a) Schematic representation of a wavelength demultiplexing grating coupler device. This device is illuminated with a normally incident light from above, and splits the light (according to the wavelength). b) The optimization process for directional modal coupling. The optimization process supposes, during the first evaluation round, that the permittivity varies in a continuous way (stage 1). In the second evaluation round, the permittivity is converted to a binary-set representation between Si and air (stage 2). The right column figure is taken from Ref. [5], in which several metadevices have been optimized using the objective-first algorithm. Here, the desired optical functionality is defined by a set of input and output conditions at the boundaries of the design space (black dots). The top figure represents a polarization splitter, the middle one is a bending device, and the bottom one, a device to convert an incident plane wave into lens to focus light at a single focal point. (a,b) Reproduced with permission.[7] Copyright 2015, Springer Nature. (c) Reproduced with permission.[5] Copyright 2018, Springer Nature.

employed until convergence, and until it reaches a final structure that reasonably satisfies the objective.[1,6,7,27,28]

To summarize, and as it is mentioned in Ref. [28], the objective first algorithm ingeniously splits Equation (1) into two convex sub-problems as shown in Equation (2), and uses local optimization approaches based on convex optimization[29] to effectively explore the huge parameter space. With respect to classical gradient-based optimization methods, which stick to physically realistic solution able to satisfy the wave equation, the objective first method treats each sub-problem sequentially, employing the alternating directions method, solving for p and x[30] iteratively. The resolution of the first sub-problem optimizes the performance but does not check if the solution satisfies the wave equation. This "violation," as discussed in refs. [27, 28], is regularized through the minimization of the second sub-problem. The first experimental demonstration of the objective first algorithm has been reported in Ref. [28], where, the objective first algorithm is followed by an adjoint-based gradient method to finely tune the structure[28] by implementing classical steepest-descent algorithm. It has been shown that the final device performs better than those optimized using only adjoint optimization method. For more details about the mathematical implementation of this method, we refer to Chapter 7 and Appendix. C in Ref. [30] together with Refs. [27, 28].

It is important to distinguish this objective first optimization method from other topology optimization (TO) and genetic algorithm (GA) methods discussed in the next paragraph. Objective first solutions are constrained to satisfy Maxwell equations using convergence process of initially nonphysical solutions while other methods, including TO and GA, optimize the solutions satisfying Maxwell equation all the way along the optimization procedure.

We note that since the algorithm relies on a gradient-based technique, it does not directly apply to structures comprising discrete representations of the permittivity. As such, a subsequent stage of discrete optimization, based on a binary representation of the structure, is required and achieved using a separate optimization method (see refs. [5, 6, 26] for more details). The latter step is of critical importance when considering practical experimental device realizations.

In **Figure 1**, we present several examples of nanophotonic devices optimized using this objective-first algorithm. In the first column, a demultiplexing grating waveguide that splits an incident free-space Gaussian beam into left-going O-band (1300 nm) and right-going C-band (1550 nm) waveguided modes has been optimized.[7] The fabricated device has a measured splitting ratio of 17 dB at 1310 nm and 12 dB at 1540 nm, whereas the designed values at these wavelengths were 19.6 and 22.2 dB. In the second column of **Figure 1**, the objective-first algorithm was used to optimize all-dielectric devices with different functionalities[5] including polarization splitting (top figure), light deflection (middle figure), and light focusing. Several other nanophotonic devices have been optimized using this technique, see for example Ref. [1] for optical cloaks, Ref. [6] for the optimization of a broadband optical diodes and Ref. [26] for the optimization of a 1D grating coupler.

It is worth mentioning that in most of the reported cases, the agreement in efficiency between optimized devices and the fabricated ones is quite low. Discrepancies are generally related to the difficulties in properly addressing the change from continuous to the binary representation. To correctly binarize designs in

a precise way, subsequent efficient optimization techniques that can mimic this clear-cut simplification must be considered.[26]

2.2. Topology Optimization

In this section, we will review a second class of optimization methods, namely topology optimization (TO), used in the inverse design of nanophotonic devices.[31] TO historically has been applied to a broad range of physical systems such as mechanical structures, MEMS, and materials design Ref. [32], and it was introduced to nanophotonics in the early 2000s in the context of photonic crystal-based technologies.[33,34] More recently, it has been applied to the design of linear[35] and nonlinear[36] metasurfaces. TO is based on a gradient-based algorithm that is able to produce freeform geometric configurations. This method works as a local optimizer, starting from an initial guess for the configuration of the device and then undergoing an iterative process to achieve a locally optimal configuration.[35]

With TO, the device is subdivided into an ensemble of pixels, for which each pixel is associated to a design variable such as the dielectric permittivity. The dimensions of these pixels can be as small as a few square nanometers for visible light devices, such that the final devices can comprise smooth, curvilinear shapes. Often, the goal is to produce a final device that consists of dielectric materials with discrete permittivity values, such as an isotropic semiconducting or insulating material with air voids. However, as TO is a gradient-based algorithm, the optimization is required to be performed on a continuum of permittivity values. Consider as an example a final device that comprises two dielectric materials with permittivities, ϵ_1 and ϵ_2. These permittivities can be normalized to take values of "0" and "1," respectively, and the actual parameter undergoing optimization will be a continuous parameter $0 \leq s \leq 1$.[31,32] This continuous representation leads to optimized regions with intermediate dielectric values between the two desired discrete permittivities. To ensure that TO produces fully discrete devices, terms can be added to the objective function that penalize the presence of grayscale dielectric values.[31,32]

We present below a general and quick overview of the optimization problem in the framework of TO (see also refs. [31, 36]). The optimization problem reads[4]:

$$\begin{aligned} &\underset{\tilde{\epsilon}}{\text{minimize}} \quad F(\mathbf{E}, \tilde{\epsilon}) \\ &\text{subject to} \quad G(\mathbf{E}, \tilde{\epsilon}) \leq 0, \quad 0 \leq \tilde{\epsilon} \leq 1. \end{aligned} \tag{3}$$

Here, $\tilde{\epsilon}$ is the normalized dielectric permittivity associated to each pixel in a specific volume. The values of $\tilde{\epsilon}$ are related to the position-dependent dielectric profile via some linear interpolation function.[36] Note that both the objective function (F) and the constraints (G) are a function of the permittivity and the electric field **E**, which is a solution to Maxwell's equations. Equation (3) can be solved using mathematical techniques such as the Method of Moving Asymptotes.[31] However, in the framework of TO, the derivative of the objective function and the constraints with respect to $\tilde{\epsilon}$ have to be computed at each pixel. This can be treated using methods such as adjoint variable method.[31,35,36]

Metagrating devices that maximize deflection efficiency at high deflection angles and near-infrared wavelengths have been optimized using TO and demonstrated experimentally in Ref. [37] The optimization algorithm starts with a random and continuous distribution of permittivity values between the values of Si and air, and these permittivity are iteratively optimized in a manner that optimizes the cost function value. As can be seen from **Figure 2**a, the gray scale values of the dielectric constant are pushed toward air or Si as the optimization proceeds, which is driven by the use of penalty terms in the cost function. As a result, final devices have a binary layout of Si in air. These devices are able to achieve high efficiencies due to non-trivial multiple scattering effects mediated by the presence of high order optical modes in high contrast dielectric structures.[38,39] TO has been readily extended to other variants of periodic diffractive optical structures, including those exhibiting ultra-high anomalous refraction[40] and the diffraction of different wavelengths to different diffraction angles.[41]

TO-based metasurfaces have also been extended to the design of wavelength-scale scatterers with defined scattering directions and phases, as shown in **Figure 2**b.[42] The final devices comprise single crystal silicon,[43] scattering light at visible wavelengths with strong directionality in the desired direction, in a process that is mediated by strong near-field interactions between nanostructures. These wavelength-scale building blocks can be stitched together to produce high efficiency aperiodic metasurfaces, such as metalenses. Aperiodic metasurfaces have also been considered in refs. [4, 36], in which efficient metalens devices and photonic fibers for nonlinear frequency conversion are optimized using TO. However, near field coupling is crucial and highly affects the performance of the final optimized design. In Ref. [44], the authors introduced an overlap technique to their topology optimization approach to take into account the near field coupling from neighbouring unit cells. They applied their technique to optimize a large-scale metalens, demonstrating higher efficiency with respect to the solution obtained with the classical local approximation technique.[45]

The influence of the initial guess (initial geometry) on the overall performance of the optimized design has been discussed in Ref. [35]. It is shown that conventional metasurface devices serving as starting points for optimization do not produce highly efficient topology optimized devices. Instead, random initial guess geometries have the potential to yield final devices with ultrahigh efficiencies (see right column in **Figure 2**).

3. Gradient-Free Optimization Techniques

3.1. Genetic Algorithm

A Genetic Algorithm (GA) is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). GAs are commonly used to generate high quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover, and selection. In a GA, a population of candidate solutions to an optimization problem, called individuals, creatures, or phenotypes, is evolved toward better solutions. Each candidate solution has a set of properties (i.e., its chromosomes or genotype) that is iteratively

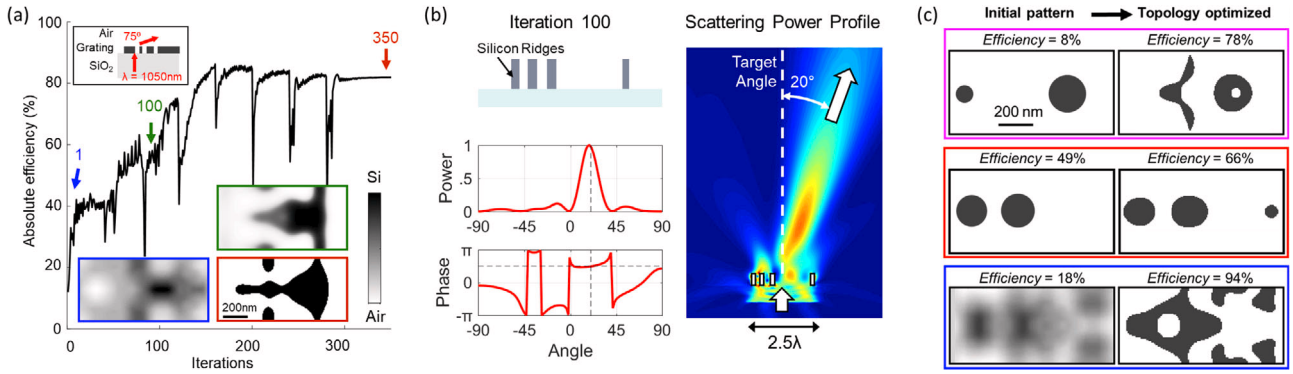


Figure 2. Metasurface design using topology optimization. a) Metagrating efficiency as a function of iteration number. The insets show the dielectric distribution in the course of the optimization process. Reproduced with permission.[37] Copyright 2017, American Chemical Society. b) Design of a wavelength-scale scatterer with a scattering direction of 20 degrees and phase response of 45 degrees. Reproduced with permission.[42] Copyright 2019, Springer Nature. c) Topology-optimized metagratings designed using different initial dielectric distributions. Reproduced with permission.[35] Copyright 2017, Optical Society of America.

mutated and altered. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encoding methods such as mixed-integer heuristics, have been proposed.[46,47] GAs can effectively deal with multiobjective optimization problems and multimodal functions.[48] The capabilities of GAs for optimization problems involving electromagnetic waves have been clearly demonstrated in several works.[10,48–51] More recently, GAs have been applied to metasurface design.[11,12,14]

Here, we summarize the mains steps of GA and review some recent works on the development and application of GAs for designing efficient and practical metasurface devices. A GA algorithm can be summarized as follows:

- **Initial generation**. A GA starts by creating a number of random candidate solutions (e.g., different metasurface designs). Each design (solution) is characterized by a chromosome that comprises the optimization parameters, such as the device width, height, period of the gratings, etc. Each parameter in the chromosome is coded by genes. Very often, this coding is binary value-based (0 or 1), but alternative approaches are possible as well. Each chromosome in the initial generation is associated with a value of a so-called fitness function (for example, the transmission or reflection coefficient). Then, the chromosomes are ranked according to their fitness function value.
- **Selection**. After ranking each of the chromosomes in the current generation according to their fitness function values, one needs to select the most promising chromosomes to be used for producing the next generation (i.e., survival of the fittest). For instance, one can decide that only 50% of the chromosomes that are closer to the target fitness value are kept while the rest of the chromosomes are discarded. The selected chromosomes are considered as parents and are used to obtain the next generation of devices. The next step consists in mating the parents to generate new children. Different mating techniques can be applied, for instance, a mating between devices ordered within the ranking list or random mating between devices.
- **Crossover and mutation**. The objective of these two operators is to generate two children from two parents. The question is: how do we generate the children, which correspond to new designs? One method is the basic one-point crossover operator, in which a random chromosome location is first chosen. Then, the chromosome of child 1 consists of an initial chromosome segment from parent 1 spanning the start of the chromosome to the random location, followed by the chromosome segment from parent 2 spanning the random location to the end of the chromosome. The chromosome of child 2 has a similar structure, except that it starts with a chromosome segment from parent 2 followed by one from parent 1. The the probability of crossover should lie between 0.6 and 0.8).[48] The next step is to apply a mutation operator, in which each gene in the chromosome of an offspring is randomly changed. In case of a binary gene representation, each "1" becomes "0" and each "0" becomes "1" upon mutation. This mutation operation should occur at low probability, between 0.01 and 0.08.[48] **Figure 3**, represents an example for the crossover mechanism using a binary representation of the parameters in the chromosomes.
- Now the children replace the parents and one has the same number of chromosomes (devices) as in the previous generation. Then, one evaluates the fitness function for each chromosome in the new generation using an electromagnetic solver. The GA can then continue by selecting the survivors in the new generation. The termination of the algorithm can depend either on an appropriate convergence threshold or a number of iterations (generations).

General considerations to implementing the algorithm include the following:

- **The representation of each parameter in the chromosome**. To operate effectively, GAs require the use of properly defined coding schemes that map metasurface parameters to genes.[48] Note that with respect to binarization of the dielectric values obtained after TO, as discussed previously, the purpose of using binary encoding in GA algorithm is to assign the physical information to a sequence of bits in the chromosome, and do not readily correspond to the value of the material parameters. The most common method is to use a binary representation in which each of the parameter values is represented by a binary analog, such that a string of bits represents each parameter. The evaluation of the cost function, which is determined by

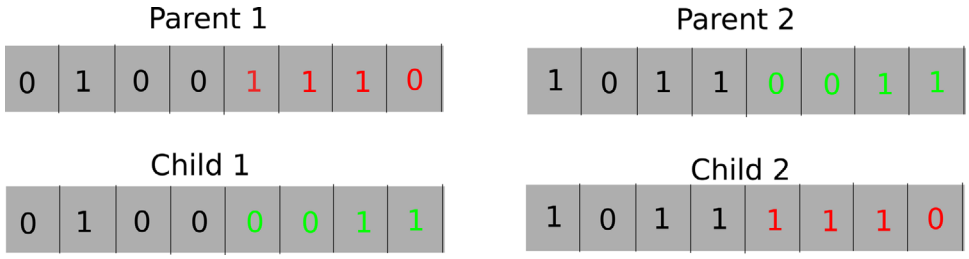


Figure 3. Illustration of the crossover mechanism in a GA for two chromosomes with a binary representation of the parameters (genes).

performing a fullwave electromagnetic simulation, requires a protocol in which the binary representation of the parameters is converted into real physical values. According to Ref. [48], one can consider the following decoding method for any parameter p in the optimization problem:

$$p = \left(\frac{p_{max} - p_{min}}{2^N - 1} \sum_{n=0}^{N-1} 2^n b_n + p_{min} \right), \quad (4)$$

p_{max} and p_{min} represent the upper and lower limits of p, respectively, and represent boundaries to parameters (e.g., feature size, refractive index, etc.) describing the physical device. N is the number of bits in the binary representation of the parameter p and b_n is the $n-1$th bit. The role of Eq. (4) is to scale the binary representation of the parameter p to a real value in a manner that takes into account its lower and upper limits.[48]

- **Number of chromosomes**. It is also important to have a sufficient number of chromosomes at each generation to ensure sufficient diversity in the gene pool. It has been shown recently that more than 100 bits can be used, as discussed in refs. [15, 17]
- **Generating the random list of bits**. During the initial generation of parent chromosomes, the bits in the chromosomes are randomly generated. In many cases, a Poisson distribution is used for this task.[10]
- **Convergence and probability of mutation**. Mutation is important to explore the parameter space and avoid local maxima. It has been shown that mutating 1% of the chromosome bits at each GA iteration is a reasonable choice of mutation rate that enables sufficient exploration of the design space.[48] In order to terminate the algorithm, one needs to either specify the number of total generations or specify a threshold value for the fitness function. For example, for a transmissive metasurface, one can set a target threshold value for the transmission coefficient at a given frequency. As the behaviour of a GA algorithm is stochastic, it is necessary to run the algorithm several times to confirm its convergence to a global maximum. For cases where different solutions are obtained for different optimization runs, decreasing the number of mutations or adding some physical constraints based on the problem at hand can improve the convergence.[10]

More resources on the implementation of GAs to electromagnetics problems are discussed in refs. [10, 48, 49, 52].

Figure 4 highlights important works that have used GAs for the design of metasurfaces. The GA has been applied to enhance the transmission and increase the deflection angle for all-dielectric metasurfaces made of Si nanodisks by optimizing two parameters: the radius and thickness of the nanodisk[12] (see **Figure 4**a). The authors show theoretically that transmission efficiencies up to 87.2% are obtained in visible spectrum (580 nm) and up to 82% at the telecommunication wavelength (1550 nm).

Another application of GAs is discussed in Ref. [14] for the design of a highly efficient beam deflector in the visible regime using an extended unit cell approach. The building block cell is made of elliptical nanoantennas, and the optimization is performed for five parameters: the minimum and maximum radii of the ellipses, the x and y position for the center of the ellipses, and the orientation angle of the major axes (see **Figure 4**b). The authors have compared their GA variant with another bio-inspired algorithm called the Artificial Bee Colony (ABC), which is a global optimization technique based on the concept of swarm intelligence.[14] They conclude that the two methods provide nearly the same efficiency; however, the ABC method converges faster to the global optimal solution.

Recently, another class of metasurfaces, referred to as binary metasurfaces, has been introduced. Binary metasurfaces are described in Ref. [59] and are based on a binary coding of the constituent meta-atoms (see **Figure 4**c,d). As systems that explicitly utilize a binary coding scheme, they are naturally amenable to GA design methods. These concepts have been applied generally to the tailoring of scattering patterns and specifically to the reduction of device radar cross sections. With this class of devices, the optical response is specified by a meta-atom coding sequence.[59] For example, a 1-bit coding describes a sequence of binary coding particles that utilize elements represented by "0" and "1." The "0" and "1" represent meta-atoms with "0" and "π" phase shift responses, respectively. A 2-bit binary representation can also be considered: "00," "01," "10," and "11" represent meta-atoms with "0," "$\pi/4$," "π," and "$3\pi/4$" phase shifts, respectively.[58,59] The material choice and dimensions of each meta-atom are individually optimized prior to GA optimization of the array coding sequence,[58,59] as shown in **Figure 4**c,d. In this case, this approach does not properly account for near-field coupling between meta-atoms, which limits the overall device performance.

GAs have also been used to enhance light-matter interactions such as magnetic effects in the visible regime. In Ref. [60], the authors identified the optimal configuration of a binary configuration, made of Si and air, which maximizes the magnetic field intensity. The geometry and its binary representation are shown in **Figure 5**a, in which each "1" represents Si rectangular meta-atoms and each "0" represents an air void. As

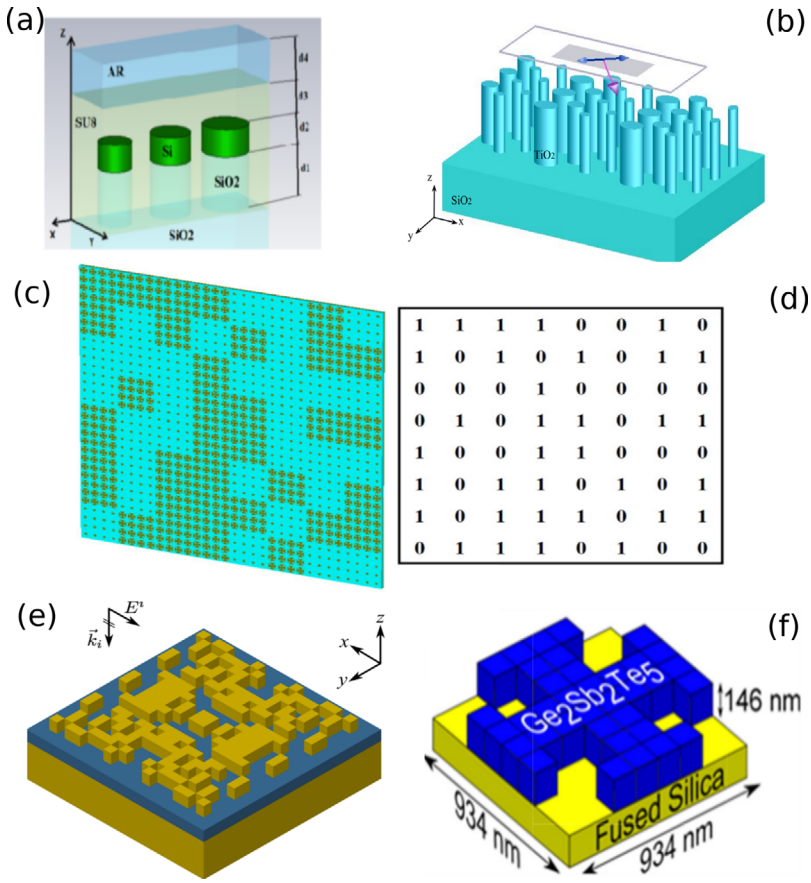


Figure 4. Examples of optimized metasurface devices using a GA or its adaptive form. The readers can refer to refs. [21–23] for a broad overview of optimized configurations based on GA or, more generally, on advanced evolutionary strategies as in refs. [53–57]. a) An all-dielectric metasurface configuration made of Si nanodisks. The parameters to be optimized were the radii and the thicknesses of the nanodisks. The authors demonstrated an optimized beam deflector metasurface with nearly 82% at the telecommunication wavelength. b) A 3D sketch of a beam deflector metasurface made of TiO_2 elliptical nanoantennas. (a,b) Reproduced with permission.[35] Copyright 2017, Optical Society of America. The parameters to be optimized were the minimum radius, maximum radius, *x* and *y* coordinates of the ellipse in the unit-cell, and the orientation angle of the major axes. The authors showed that a 60% efficiency can be obtained for beam steering with an angle as large as 50 degree at $\lambda = 520$ nm). c,d) example of an optimized metasurface geometry with its corresponding binary representation. (c,d) Reproduced with permission.[58] Copyright 2018, Optical Society of America. e) A plasmonic metasurface comprising a backside Au mirror (yellow bottom region) and Si spacer (blue region) and a binary gold pattern (top yellow regions). This geometry has been used to optimize a beam reflector metasurface with nearly 92% of performance efficiency. Reproduced with permission.[11] Copyright 2019, Springer Nature. f) A reconfigurable metasurface pattern designed using a GA in order to switch from highly transmissive mode with efficiency (80%) to highly absorptive modes with efficiency as high as (76%). Reproduced with permission.[13] Copyright 2017, Optical Society of America.

an initial step, a random population consisting of 20 geometries is considered, each simulated using an electromagnetic solver to compute the magnetic intensity enhancement at the center of the geometry (see red point in **Figure 5**a). The five best geometries, providing the highest magnetic enhancement, are kept and used to generate the new population. This process is repeated until an optimized geometry is obtained. For a GA run of nearly 350 generations, with each generation coding 20 different geometries, 7000 independent simulations are performed using an electromagnetic solver. The authors conclude that the magnetic power density obtained using the optimized geometry (see **Figure 5**) can be increased by a factor of five, compared to state-of-the-art dielectric nanoantennas. In similar circumstances, the authors in Ref. [15], employed GA to improve the near field intensity in a plasmonic configuration. It has been revealed that the optimized geometry exceeds the state of the art reference plasmonic geometry by more than a factor two (see **Figure 5**d–f). For more details about the implementation of the inverse design, the readers can refer to Ref. [15].

Various devices have been recently optimized using GA or using some advanced evolutionary strategies as shown in refs. [53–56, 61]. To highlight the basic concepts of the numerical optimization methods and mention briefly some possible applications, the readers can refer to recent review works about metasurfaces and nanophotonics (refs. [21–23]), discussing in details different applications relying on inverse design.

So far, our discussion has focused on the optimization of only a single target in the objective function. A more challenging

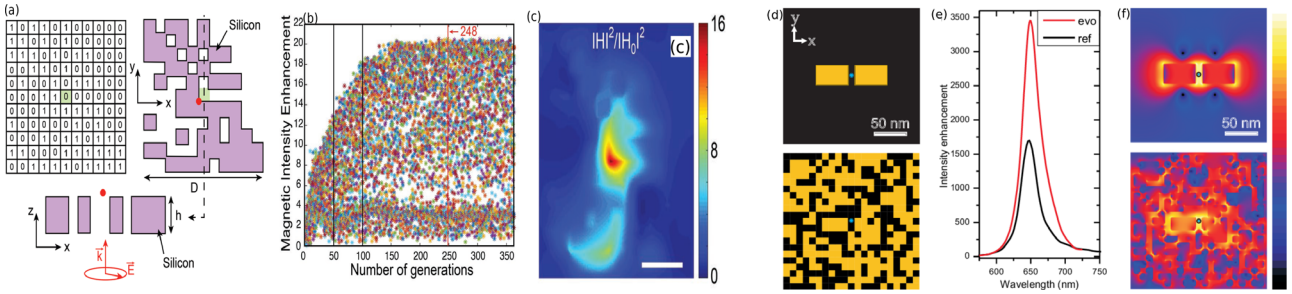


Figure 5. a) Binary matrix representation of a metasurface geometry that enhances the magnetic field intensity at the center of the structure (red dot in right and bottom figure). Each "1" represents a rectangular nanostructure made of Si while each "0" corresponds to an air void. b) Magnetic intensity enhancement for the distribution of parent devices as a function of generation number. Each generation contains 20 individual devices. c) The magnetic field distribution of the optimized geometry after 248 generations. (a–c) Reproduced with permission.[60] Copyright 2019, Wiley–VCH. d–f): another example to illustrate the usefulness of utilizing evolutionary approach based technique in enhancing near field intensity. d) refers to the reference slot gold antenna configuration (top), and the bottom shows the optimized geometry. The small point refers to where the field enhancement is calculated. e): represents the comparison between the reference geometry (black curve) and the optimized one (red curve) as indicated by the field maps in (f). (d–f) Reproduced with permission.[15] Copyright 2012, American Physical Society.

question is: how can we deal with multiple targets? such as multiple optical functionalities or optimization at different operating wavelengths? In the following, we present a quick survey of the attempts to include multiobjective optimization in the framework of GAs for metasurface designs. We refer to Ref. [62] for more information about multiobjective optimization for evolutionary strategies.

The most straightforward way is to combine the targets into a single objective function and weigh each target with a normalization term based on the prioritization of the target. However, according to Ref. [11], this is sub-optimal. A more effective method to perform this multi-objective optimization is to consider an adaptive form of GAs, which has been applied to design binary metasurfaces with two targets in the objective function (see **Figure 4**e). The main idea behind the adaptive approach is to initially perform the GA that considers only one target in the objective function. After this first round of optimization is done, the objective function is modified to account for the second target and the GA is applied again to the whole problem until it reaches a satisfactory solution for both targets simultaneously. This adaptive GA is used to theoretically optimize different metasurface devices, including binary pattern reflect-arrays and dual beam aperiodic leaky wave antennas.[11]

There have been other attempts to apply GAs to the multi-objective optimization of metasurface devices, including a reconfigurable metasurface device that has been demonstrated experimentally[13] (see **Figure 4**f). The main goal was to achieve a tunable metasurface configuration, using a specific composition of the Chalcogeneide glass, which changes its response from being highly transparent to being highly absorptive at $\lambda = 1.55\ \mu m$ as a function of temperature. The optimization of the device is done using an adaptive GA together with a full-wave electromagnetic solver based on the periodic finite element boundary integral method.[63] Another study aiming at optimizing the design of colour pixels based on Si nanostructures used an evolutionary algorithm coupled to a frequency-domain Maxwell solver to treat a multi-objective function.[64] Recently, some advanced evolutionary strategies have also been extended to design multifunctional metasurfaces.[53,65,66]

For the GA approaches presented here, all require fullwave calculations of Maxwell's equations for all devices in each generation, which adds up to at least a few thousand of simulations. Therefore, this technique is computationally expensive and has to be used with efficient fullwave solvers, especially when considering the optimization of 3D structures. We note that the previously mentioned ABC evolutionary algorithm has been used to optimize metasurface designs, and when compared with a classical GA,[14] the ABC method is 35% faster while producing comparable results to GA.

3.2. Particle Swarm Optimization for Metasurfaces

The Particle Swarm Optimization (PSO) algorithm is an iterative global optimization technique in which the population (swarm) consists of a predefined number of small particles, each of which are coordinates in the search space.[1] These particles try to improve their location in the search space by remembering their best location and sharing this information with the other members of the population. The PSO has been used to optimize different photonic devices including diffraction grating structures,[67] photonic crystal waveguides,[68] and to optimize metal nanoparticles to obtain broadband plasmonic field enhancement over the entire visible regime.[69]

In Ref. [70], the PSO algorithm was coupled to an FDTD solver to realize metasurfaces consisting of etched features within extended slab waveguides. This metasurface architecture, which is also discussed theoretically in Ref. [14], is complementary to more traditional metasurface layouts based on structurally-isolated nanostructures. With PSO, the radii of ten nanoholes and their relative positions are initially optimized to maximize light deflection at the wavelength $\lambda = 4.2\ \mu m$, using a predefined number of iterations. At the end of this step, the best device within the population is identified and further locally optimized using a gradient descent-based technique (see **Figure 6**). This work showed for the first time the connection of the high forward scattering efficiency of a cell with the well-known Kerker conditions that exist for isolated scatterers (see also Ref. [71]).

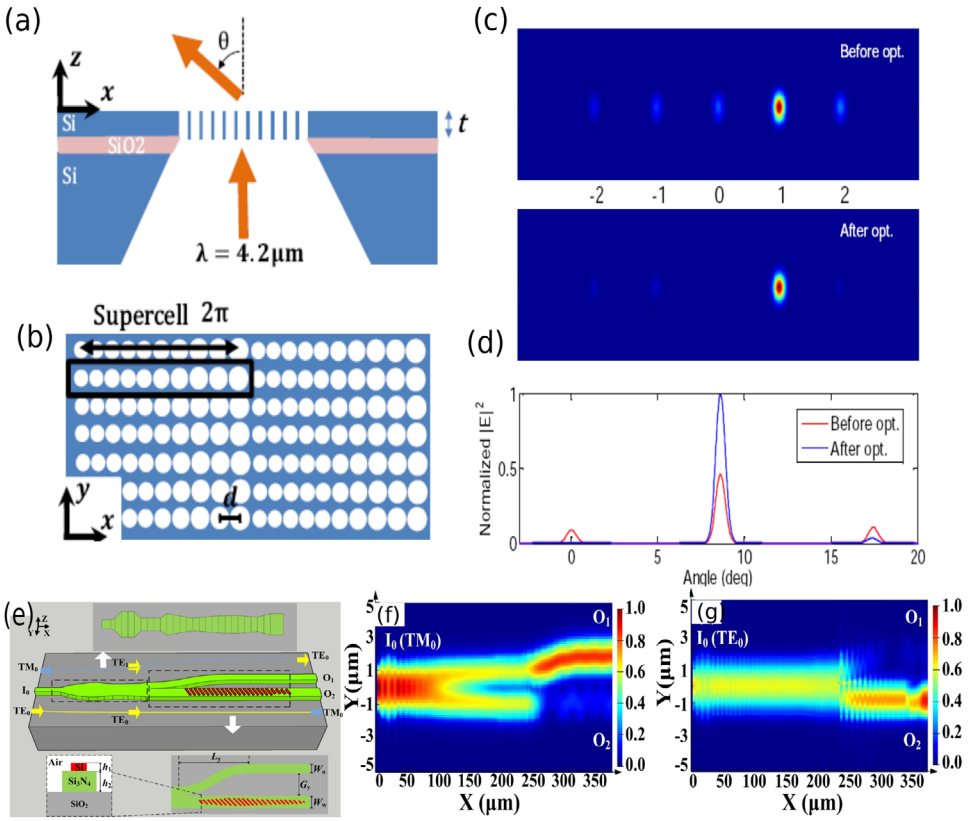


Figure 6. a) Beam deflector Si metasurface made of spherical nanoholes. b) Top view of the structure, in which the supercell is delineated by the black rectangle. the parameter *d* in the figure represents center-to-center distance. c,d) Normalized far field radiation pattern before and after optimization. (a-d) Reproduced with permission.[70] Copyright 2017, Optical Society of America. e) A polarization beam splitter based on metasurface-assisted silicon nitride Y-junction. The input port is I_0, TM_0 can be converted to TE_0 through the path $I_0 - O_1$ similarly, TE_0 can be converted to TM_0 through the path $I_0 - O_2$. The metasurface (red part) needs to be optimized to enhance the efficiency of the device. f,g): represent the simulated TM_0 and TE_0 injected modes and their output responses, respectively. (e–g) Reproduced with permission.[72] Copyright 2019, Elsevier.

PSO has also been applied to optimize a polarization beam splitter based on metasurface-assisted silicon nitride Y-junction for mid-infrared wavelengths[72] (see **Figure 6**e. The main objective of this device is to convert the fundamental TM_0 mode to the TE_0 and vice versa with high efficiency. In order to maximize the mode conversion efficiency, the Y-junction is patterned with silicon metasurfaces (see red parts in **Figure 6**e). The optimization results are validated using numerical simulations, as indicated in **Figure 6**f,g.

3.3. Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

With the evolutionary optimization strategies presented thus far, we have seen that the internal optimization parameters, such as the number of generations and mutation protocol, must be carefully chosen. The convergence of these methods can be accelerated by tuning these internal parameters, but this task is usually tedious in practice and is computationally costly. While it is possible to automatically adjust these internal parameters during optimization, most classical evolutionary strategies operate with fixed parameters during the optimization process.[10–12,52]

The Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is an alternative evolutionary strategy that is capable of adapting its internal optimization parameters during the optimization process. CMA-ES is population-based and operates by iteratively evolving a Gaussian distribution of design candidates within the search space in order to find the global maxima.[1,73,74] This Gaussian distribution is fully defined by its mean and covariance matrix, the latter describing the shape of the distribution. This advanced evolutionary strategy uses several automatically adjustable parameters that allow the covariance matrix to adapt to the local characteristics of the function to be optimized. Starting from an initial random guess, the algorithm searches for the global maxima by reshaping and resizing its Gaussian sampling automatically every few iterations.

Recently, CMA-ES has been applied to the design of phase gradient metasurfaces operating at the visible light regime.[75] The CMA-ES algorithm has also been used to optimize several metasurface devices such as infrared broadband quarter-wave plate metasurfaces,[76] metasurface absorbers,[77] and apochromatic singlets metasurface-augmented GRIN lenses.[78]

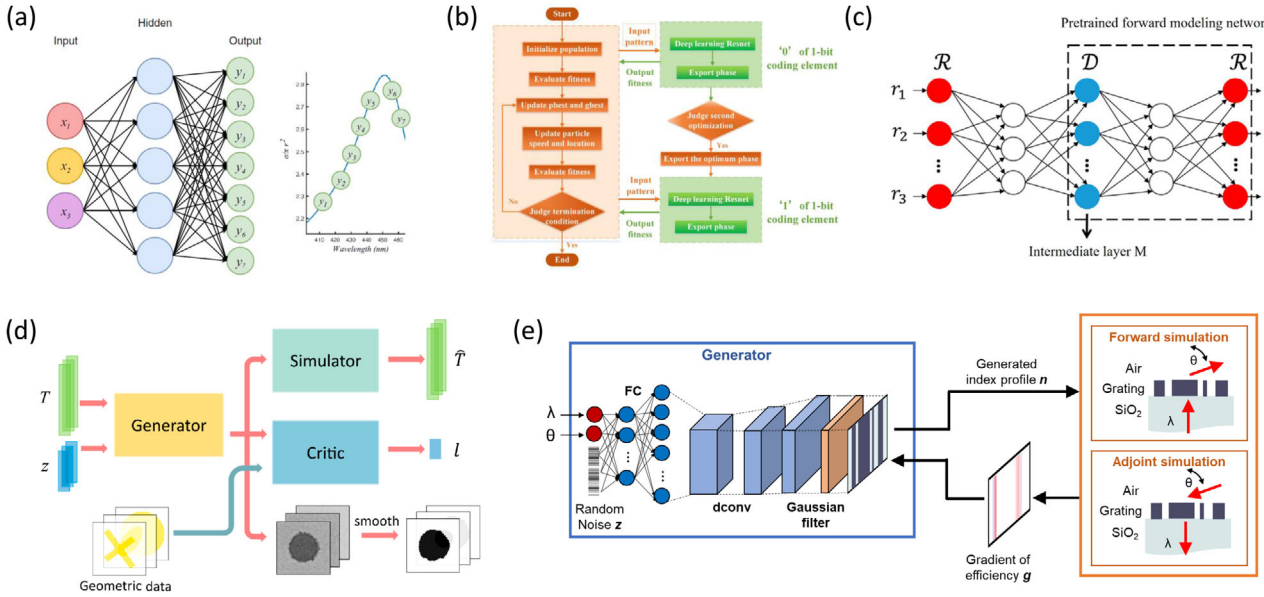


Figure 7. Neural network architectures for the inverse design of nanophotonics systems. a) deep, fully connected neural networks as a surrogate electromagnetic simulator. Reproduced with permission.[83] Copyright 2018, American Association for the Advancement of Science. b) optimizer that combines neural network simulators with particle swarm algorithms. Reproduced with permission.[87] Copyright 2019, Wiley-VCH. c) tandem neural network combining an inverse and forward deep network. Reproduced with permission.[84] Copyright 2018, American Chemical Society. d) freeform nanostructure inverse design using generative adversarial networks. Reproduced with permission.[88] Copyright 2019, American Chemical Society. e) global topology optimization networks (GLOnets) are global, population-based optimizers that train using simulations from the adjoint variables method. Reproduced with permission.[89] Copyright 2019, American Chemical Society.

Algorithm 1 Simple illustration for the CMA-ES steps

1. Start with an initial guess (mean, and shape of the distribution)
   - m: mean of the distribution
   - C (matrix): shape of distribution ($C = I$ is the identity matrix for the initial guess)
2. Generate population: $x_i = m + N_i(\sigma^2, C)$
3. Evaluate the objective function using the electromagnetic solver for all the individuals of the population
4. Sort the generation (choose the most fitted individuals)
5. Update mean, variance, and the covariance matrix
6. Repeat steps (2 – 5) until convergence is achieved

4. Inverse Designs Using Artificial Neural Networks

The machine learning revolution has transformed the way large datasets are handled and processed in all fields of technology and science. The electromagnetic device geometries and their responses can be treated as large datasets, making the use of machine learning highly applicable and relevant. The last two years have witnessed rapid growth in applying deep learning in the field of nanophotonics.[24,25,79–82] However, in this section, we will examine how machine learning can aid in the inverse design and optimization of metasurface structures. We will focus on two types of deep learning architectures, discriminative and generative networks.

For electromagnetic design problems described by a small number of geometric parameters, discriminative networks can accurately map the explicit relationship between a geometry and its electromagnetic response. Discriminative networks are supervised learning algorithms that learn from a training set, and the learning process can be mathematically described as mini batch gradient descent. In basic form, these algorithms are deep neural networks[83,84] that comprise multiple layers of interconnected nodes, called neurons, which perform non-linear mathematical operations on a set of weighted inputs to produce an output value (**Figure 7**a). More sophisticated algorithms utilize convolutional neural networks,[85] in which convolution operations are performed by neurons. The weights and convolutional kernels are learnable parameters that are determined from the network training process. Given the non-linear responses of individual neurons and their ensembles, the highly non-linear relationship between geometry and response can be properly captured. To date, discriminative networks have been demonstrated to accurately model a wide range of nanoscale electromagnetic systems, including the scattering and chiral properties of plasmonic structures,[83,85] silicon photonic devices,[86] and metasurfaces.[19]

Trained discriminative networks can be used to optimize electromagnetic systems in a variety of ways. One way is to treat the discriminative network as a high speed electromagnetic solver and embed it into classical iterative optimization schemes, such as genetic[90] and particle swarm[87] algorithms (**Figure 7**b). Compared to conventional electromagnetic solvers, a trained discriminative network can model the electromagnetic response of a system with order-of-magnitude faster times. Another way is to directly optimize the electromagnetic device using iterative

backpropagation, which performs optimization using gradient descent.[83] The idea is to first start with a random geometric input, calculate its optical response with the neural network, and iteratively backpropagate the difference between the actual and desired optical response. During backpropagation, the network weights are fixed and the input geometry is modified to reduce the optical response error. In a third approach, discriminative networks can be configured to directly solve the inverse problem without requiring an iterative process by specifying the inputs of the network to be the desired optical response and the outputs to be the device layout. To help ensuring the stability and accuracy of this inverse network, tandem architectures that combine an inverse network with a forward solver network (**Figure 7**c) are effective at producing properly trained networks.[84] In this algorithm, the forward solver network is first trained, using supervised learning with a training set, to create a high speed surrogate solver. An inverse network is then attached to the trained forward solver network, which has fixed weights, and the inverse network is trained in the framework of the tandem network. Tandem networks have been used to design wavelength filters[84] and topological photonic devices.[91]

The main drawback of discriminative networks is the extreme amount of required training data. Empirically, for electromagnetic problem settings that are described by ten geometric parameters, approximately 10 000 system configurations need to be considered as inputs for the network to accurately converge. As a typical example, a metagrating described by sixteen geometric parameters requires 90 000 devices in the training set for proper network training.[19] As such, even though a trained neural network can serve as a fast electromagnetic solver, it requires a very computationally costly off-line training phase. Furthermore, strategies based on such networks cannot practically scale to complex electromagnetic geometries due to the curse of dimensionality. This concept, which is a long-standing problem in machine learning, states that the design space and training set scale exponentially with the complexity of the system being modeled. As such, discriminative network approaches cannot practically applied to structures described by freeform geometric layouts.

Generative neural networks are an alternative type of network architecture that can be used in the design of complex electromagnetic devices described by tens to hundreds of geometric parameters. Typically, these networks have desired optical parameters as inputs and high resolution images of the devices as outputs. A key feature of generative networks is that, in addition to the desired optical parameters, a high dimensional latent random variable is also used as an input to the network. As such, for a given desired optical parameter, a wide range of devices can be generated, each of them mapping onto a unique latent random variable value. This mapping of an optical parameter to a distribution of devices in the design space is fundamentally different from the mechanics of discriminative networks, and allows for very high dimensional structures to be modeled and generated.

There are various methods to train generative networks. One way is to use a generative adversarial network (GAN), in which a generative neural network is trained against a discriminative neural network using a training set (**Figure 7**d). During training, the generative network generates devices and feeds them into the discriminator network. Its objective is to fool the discriminator network. The discriminator is a classifier with the goal of accurately differentiating between the generated devices and those from the training set. Upon the completion of GAN training, the generator will be able to generate devices that match the distribution of training set devices. GANs have been used to realize freeform geometries with tailored reflection and transmission spectra.[88,92] They can also learn from images of topology-optimized dielectric metasurfaces to generate topologically complex devices with high performance.[93]

Generative networks can also be trained directly, without a training set, using calculations based on the adjoint variables method. These global topology optimization networks, termed GLOnets, use the training of a generative neural network to perform global topology optimization.[89,94] The concept is outlined in **Figure 7**e for metagratings as a model system. In one iteration of the optimization process, the generator produces a batch of devices, from which the efficiencies are calculated using an electromagnetic solver and the efficiency gradients are evaluated using the adjoint variables method. These efficiencies and efficiency gradients are then used to update the network weights through backpropagation. Upon training completion, the generative network maps the desired optical parameters and latent noise vectors to an ensemble of high performance devices. Benchmarking of GLOnets with iterative-only topology optimizers for metagratings indicates that GLOnets can generate ultra-high efficiency devices, with efficiencies higher than those produced from many instances of iterative-only topology optimization. While it is not possible to determine whether these devices are globally optimal, due to the non-convexity of our optimization problem, they clearly have exceptional performance metrics. We anticipate that hybrid machine learning concepts, which properly incorporate physical knowledge into neural networks through the use of physics-based calculations and that can even directly solve physics-based differential equations,[95] will play a large role in the future of electromagnetics inverse design.

5. Bayesian Optimization

A central benchmarking standard for any inverse design technique is the computational cost. For most photonics problems, these techniques require rigorous and expensive electromagnetic solvers to accurately compute the objective function at each optimization step. The computational cost of the electromagnetic solver can be mitigated by using an Artificial Neural Network (ANNs) as a surrogate solver, but a significant number of electromagnetic simulations still must be performed to train the network prior to optimization.[19,84] In this section, we discuss an alternative optimization strategy based on surrogate modeling, named Efficient Global Optimization (EGO),[96,97] which has been introduced recently in the context of the design of nanophotonic devices.[98] The EGO algorithm is a global optimization algorithm that substitutes the complex and costly iterative electromagnetic evaluation process with a simpler and cheaper model.[96,97] Its aim is to maximize a specific statistical criterion related to the optimization target, which is referred to as the merit function. EGO involves two main phases. The first one is called the Design Of Experiment (DOE), in which an initial database is generated using a sampling of photonic devices within the design space. These devices are simulated using an

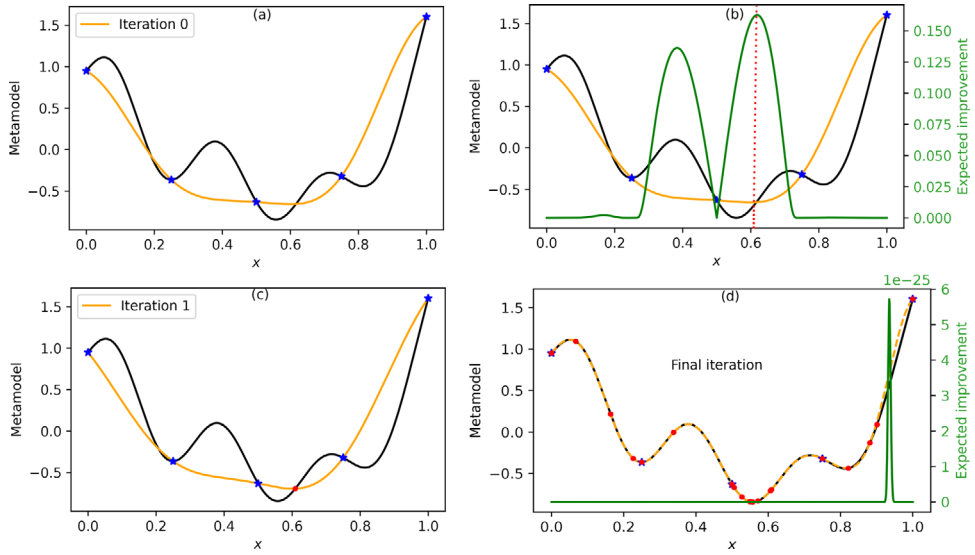


Figure 8. Illustration of the EGO algorithm using a simple 1D example. The black curve is the exact analytical function to be minimized. a) The initial step. The blue points are the DOE database elements, and the orange curve is the first surrogate model that fits the blue points. b) Similar to (a), except that the merit function is computed for each point on the orange curve (each parameter value). It is represented by the green curve and the right label. As can be seen, at $x \approx 0.62$, the merit function is maximized (see red dashed line). It means that this value of x corresponds to the highest probability to improve the results, i.e obtain a better minimum. c) The objective function is computed at $x \approx 0.62$ using the solver and the new value is added to the database (red point). A new model is constructed (see orange curve in (c)). d) One computes the merit function based on this new model and all the steps are repeated until the value of the merit function approaches zero, shown here. In this example, the surrogate model (orange curve) coincides with the exact analytical function (black curve). For this simple example, only 20 iterations are required for the system to converge.

electromagnetic solver to compute their corresponding objective function values, which are stored in the database. In the second phase, a Gaussian Process (GP) model is constructed by interpolating the database points. Internal model parameters are calibrated according to a maximum likelihood principle.[99] Once this GP model is defined, one can obtain an estimation of the objective function at any point of the design space, which represents the model mean, and an estimate of the prediction uncertainty, which is the model variance. These quantities are then used to define a statistical merit function, called the Expected Improvement (EI), whose maximum corresponds to the next design parameters to be evaluated using an electromagnetic solver. After simulating this new point, this new data is added to the database. This process in the second phase is repeated until convergence. In **Figure 8**, we present a simple 1D example that illustrates the basic mechanisms of the EGO algorithm.

In Ref. [98], EGO was applied to optimize 3D nanoparticle shapes to design the morphology of metal nanoparticles. The main target was to maximize the average electric fields on their surfaces. The optimization is performed by changing the shapes of the particles and the excitation wavelength. Several plasmonic materials were considered in the optimization, including gold and silver. We would like to mention that a comparison between five benchmarking global optimization methods have been performed in Ref. [100], in which it is shown that the Bayesian-based optimization techniques require less simulation time compared to other techniques. The main limitation of the Bayesian optimization is the cost to construct the model when a large number of observations is included. It is usually related to the difficulty in handling a large parameter space. Indeed, this requires a large database in the design of experiments phase, nearly ten times the number of parameters, while the increased number of iterations requires costly simulations for the metamodel adaptation after each iteration. Moreover, the model training becomes computationally expensive when too many points are fitted.

Recently in Ref. [75], we applied the EGO algorithm to the optimization of metasurfaces. The objective was to maximize the light deflection efficiency at $\lambda = 600$ nm using metasurface designs based on rectangular and spherical nanopillars. Optimizing up to eight parameters describing arrays of cylindrical nanopillars, one can obtain more than 85% efficiency for the deflection of both TM and TE polarizations. In addition, in using rectangular-shaped antennas and optimizing twelve parameters, one obtains more than 88% efficiency for incident TM polarized waves, as indicated in **Figure 9**. Moreover, in Ref. [75], we have shown that several optima may exist for this problem, and that the use of EGO allows for the identification of all the physically relevant global optima related to the geometry under consideration.

This optimization approach, based on the iterative construction of a database and an associated model, can be considered as a statistical learning strategy. The main feature of EGO is the use of internal uncertainty estimation (i.e., variance) to drive both the search for the optimum and the improvement of the model accuracy simultaneously. This concept is different from ANNs, which aim to construct an accurate model within the whole design space prior to optimization, which is very expensive when the parameter space is large. On the contrary, EGO focuses on the most promising areas of the design space. It is therefore far less expensive in terms of electromagnetic solver calls, and in practice, only a few hundred electromagnetic simulations are typically required for EGO. For instance, it was demonstrated in Ref. [75]

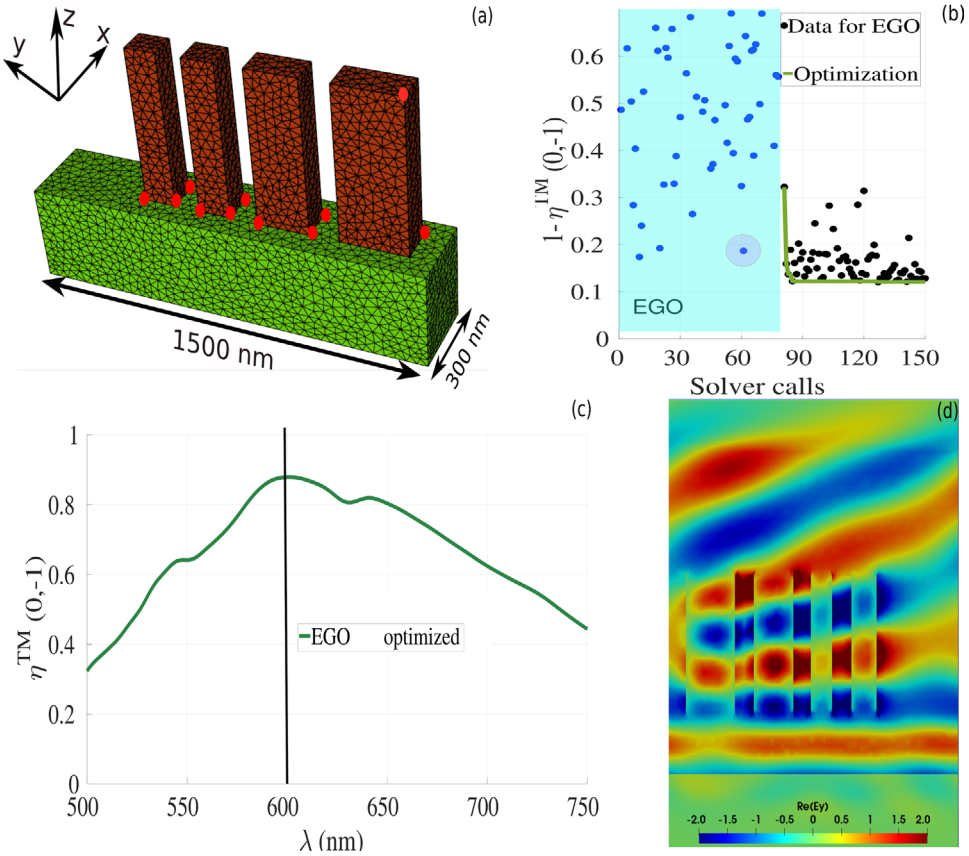


Figure 9. Optimization of rectangular-shaped nanoantennas using the EGO algorithm. a) The geometry under consideration consists of rectangular nanoridges made of GaN (dark red) on top of a semi-infinite substrate made of Al_2O_3 (green region). The twelve red circles represent the optimization parameters. Under normal illumination, we aim at maximizing the diffraction efficiency for the first order mode $\eta^{TM}(0,-1)$ at $\lambda = 600$ nm. b) Optimization process using EGO as a function of the number of fullwave solver calls. The blue points represent the DOE (shaded region), the black points represent the value of the objective function at each iteration, and the green solid line indicates the best performance obtained so far. c) Diffraction efficiency as a function of the wavelength for the optimized design. We notice that it is maximal at the desired wavelength indicated by the black line. d) Field map at $\lambda = 600$ nm for the optimized design. (a–c) Reproduced with permission.[75] Copyright 2019, Springer Nature.

that one needs in total 150 solver calls for both the iterative enrichment and DOE (80 points) in order to optimize a structure with twelve parameters.

6. Robustness

This burgeoning field of research in computational nanophotonics applied to the design of metasurfaces is currently booming, offering new design perspectives. It is nevertheless important to point out that the uncertainties related to fabrication errors are not considered for most of the optimization methods discussed so far, and that they play an important role to addressing errors and uncertainties during the fabrication process.

In gradient-based topology optimization, geometric robustness can be incorporated by considering the performance of the eroded and dilated versions of the device throughout the optimization process.[101] By incorporating the performance of these geometric variants into the objective cost function, the final devices become relatively insensitive to geometric perturbations. Experimental characterization of devices designed with robustness criteria show that robust devices are relatively insensitive to differing levels of over- and underexposure from the lithography fabrication process. We note that there exists a trade off between robustness and maximum device performance, which does place practical limits on metasurface performance. It is worth mentioning that, if to the best of our knowledge very few papers are discussing this issue for metasurface designs, several other works have been reported, applying sensitivity analysis in optimizing various electromagnetic devices.[28,102,103]

An alternative methodology for the optimization of robust metasurfaces leverages a concept termed Uncertainty Quantification (UQ).[104] In this reference, the authors optimize a 2D (i.e., periodic only in one direction), high contrast, subwavelength grating comprising gallium nitride (GaN) in order to maximize the light deflection at a given fixed angle at visible regime. The results are summarized in **Figure 10**. With UQ, the authors optimize the influence of the manufacturing process in order to obtain robust structures that are insensitive to small manufacturing imperfections.

As a first step, the authors optimize several designs by computing the electromagnetic response of arrays of 2D subwavelength

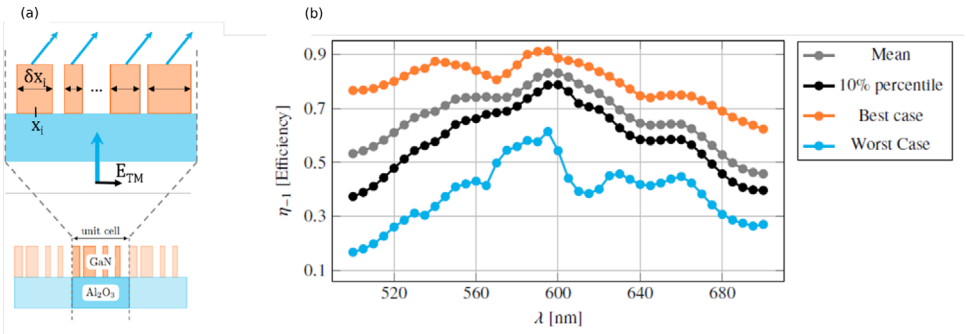


Figure 10. a) Illustration of the 2D phase gradient metasurface made of GaN nano-resonators (orange ridges) placed over a substrate made of Al_2O_3. The height is fixed to 1000 nm. The main goal is to optimize the thicknesses (δX_i) and the positions of the nanoridges X_i. b) UQ results for uniform input distributions. (a,b) Reproduced with permission.[105] Copyright 2019, Optical Society of America.

ridges using the rigorous coupled wave analysis method, together with a gradient-free pattern search algorithm,[104] implemented in the Matlab optimization toolbox. In a second step, they perform a UQ analysis and explore the impact of geometrical variations in technologically relevant parameters such as ridge widths δX_i and relative ridge positions X_i. A Monte Carlo ensemble of a million random device realizations are numerically simulated to evaluate the sensitivity of the optimal design with respect to manufacturing imperfections through statistical indicators, including the mean deflection efficiency, standard deviation to the mean value, and confidence intervals. The results show that the number of elements per phase gradient period plays a considerable role in the reliability of the structures with respect to small ±5 nm uncertainties in the widths and positions.[105]

7. Conclusion

As a general conclusion of this work, we gave an up-to-date overview of optimization techniques used in the field of metasurface designs. We focused on the most general and common inverse design techniques used in the literature. We hope that the discussion provided herein will be helpful and useful even for readers that are non-experienced in the field of inverse design.

First, we introduced gradient-based optimization techniques, including the objective-first algorithm and topology optimization. These algorithms can produce freeform geometrical shapes, and can deal with a very large number of design variables in an efficient manner. They can lead to sophisticated, complex, and non-intuitive designs, yet with high efficiencies. Nevertheless, there are still some challenges in fabricating the produced non-intuitive designs. In addition, these techniques depend strongly on the initial design.

Second, we presented genetic algorithms (GA) and the covariance matrix adaptation evolution strategy (CMA-ES), which are widely used global optimization techniques. These methods are iterative and based on a population of designs that represent the optimization parameters. In general, GAs and CMA-ESs can deal with large parameter space at both the continuous and the discrete levels. However, they require expensive electromagnetic solver calls, especially when dealing with large parameter spaces. CMA-ES has the particular advantage in that it is a self-adapted global optimization method. Unlike GA and other global techniques, CMA-ES tunes its internal parameters during the optimization process, making it a suitable global optimization strategy for complex problems with large parameter spaces.

Third, we discussed the utilization of artificial neural networks (ANNs) and deep learning architectures, namely discriminative and generative networks, for the inverse design of metasurface devices. For systems described with a small number of parameters, discriminative networks can accurately determine the relationship between the optimization parameters and their electromagnetic responses, thereby serving as computationally efficient surrogate electromagnetics solvers. However, this type of ANN requires a high computational overhead in order to effectively train the network and it cannot be practically applied to complex problems. With generative neural networks, a high dimensional latent random variable is used as an input to the network, meaning that a wide range of devices can be generated and outputted. This mapping of a latent space to a distribution of devices allows the amount of training data to be reduced and high dimensional structures to be modeled and generated. This concept has been used recently in the framework of global topology optimization.

Finally, we discussed the concept of Bayesian Optimization as an alternative approach for the inverse design of metasurfaces. More precisely, we focused on a widely used approach, which is called Efficient Global Optimization (EGO). The EGO algorithm is a global optimization algorithm based on a surrogate model, and it replaces the costly evaluation process by a simpler and computationally cheaper model. EGO uses internal statistical criteria in order to choose correctly the new evaluations that will enrich the model to miminize/maximize of the objective function and improve its accuracy as well. EGO uses trained data in its initial phase, however the number of trained data is far smaller than the ones used in the ANNs framework. We also touch on the importance of optimizing robust devices the implementation of Uncertainty Quantification to capture the sensitivity of manufacturing imperfections in the device design process. Among all of these optimization methods, a great deal of attention will soon be devoted to the simultaneous optimisation of multifunctional metasurfaces, i.e. metasurface designs involving more than one objective function, in particular to find trade-off designs able to resolve the poor efficiency of broadband metasurfaces. Methods such as multiobjective programming, multicriteria optimization, multiattribute optimization, vector optimization or Pareto-front optimization, which have already been applied to many fields of

science and engineering, would lead realistic metadevice designs with increased performance and capabilities.

Acknowledgements

S.L., R.D., and P.G. acknowledge support from French defence procurement agency under the ANR ASTRID Maturation program, grant agreement number ANR-18-ASMA-0006-01. P.G. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement no. 639109, 874986). J.F. is supported by the U.S. Air Force under Award Number FA9550-18-1-0070 and the David and Lucile Packard Foundation.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

artificial intelligence, bayesian optimization, inverse design, metasurface, optimization methods

Received: December 18, 2019
Revised: June 30, 2020
Published online: August 13, 2020

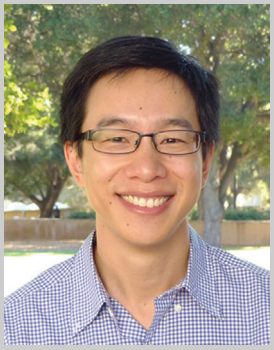
[1] *Numerical Methods for Metamaterial Design, Topics in Applied Physics* (Ed: K. Diest), Vol. *127*, Springer Netherlands, Dordrecht **2013**.
[2] P. Lalanne, S. Astilean, P. Chavel, E. Cambril, H. Launois, *Opt. Lett.* **1998**, *23*, 1081.
[3] P. Lalanne, S. Astilean, P. Chavel, E. Cambril, H. Launois, *JOSA A* **1999**, *16*, 1143.
[4] Z. Lin, B. Groever, F. Capasso, A. W. Rodriguez, M. Lončar, *Phys. Rev. Appl.* **2018**, *9*, 044030.
[5] F. Callewaert, V. Velev, P. Kumar, A. V. Sahakian, K. Aydin, *Scientific Reports* **2018**, *8*.
[6] F. Callewaert, S. Butun, Z. Li, K. Aydin, *Scientific Reports* **2016**, *6*.
[7] A. Y. Piggott, J. Lu, T. M. Babinec, K. G. Lagoudakis, J. Petykiewicz, J. Vučkovíc, *Scientific Reports* **2015**, *4*.
[8] D. Sell, J. Yang, S. Doshay, J. A. Fan, *Adv. Opt. Mater.* **2017**, *5*, 1700645.
[9] J. Lu, S. Boyd, J. Vučkovíc, *Opt. Express* **2011**, *19*, 10563.
[10] R. Haupt, *IEEE Antennas and Propagation Magazine* **1995**, *37*, 7.
[11] S. Jafar-Zanjani, S. Inampudi, H. Mosallaei, *Scientific Reports* **2018**, *8*, 11040.
[12] V. Egorov, M. Eitan, J. Scheuer, *Opt. Express* **2017**, *25*, 2583.
[13] A. V. Pogrebnyakov, J. A. Bossard, J. P. Turpin, J. D. Musgraves, H. J. Shin, C. Rivero-Baleine, N. Podraza, K. A. Richardson, D. H. Werner, T. S. Mayer, *Opt. Mater. Express* **2018**, *8*, 2264.
[14] K. D. Donda, R. S. Hegde, *Prog. Electromagn. Res. M* **2017**, *60*, 1.
[15] T. Feichtner, O. Selig, M. Kiunke, B. Hecht, *Phys. Rev. Lett.* **2012**, *109*, 127701.
[16] M. D. Huntington, L. J. Lauhon, T. W. Odom, *Nano Lett.* **2014**, *14*, 7195.
[17] T. Feichtner, O. Selig, B. Hecht, *Opt. Express* **2017**, *25*, 10828.
[18] P. R. Wiecha, C. Majorel, C. Girard, A. Cuche, V. Paillard, O. L. Muskens, A. Arbouet, *Opt. Express* **2019**, *27*, 29069.
[19] S. Inampudi, H. Mosallaei, *Appl. Phys. Lett.* **2018**, *112*, 241102.
[20] Z. Liu, D. Zhu, S. Rodrigues, K. T. Lee, W. Cai, *Nano Lett.* **2018**.
[21] K. Yao, R. Unni, Y. Zheng, *Nanophotonics* **2019**, *8*, 339.
[22] S. D. Campbell, D. Sell, R. P. Jenkins, E. B. Whiting, J. A. Fan, D. H. Werner, *Opt. Mater. Express* **2019**, *9*, 1842.
[23] S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vuckovic, A. W. Rodriguez, *arXiv preprint arXiv:1801.06715* **2018**.
[24] K. Yao, R. Unni, Y. Zheng, *Nanophotonics* **2019**, *8*, 339.
[25] P. R. Wiecha, O. L. Muskens, *Nano Lett.* **2019**.
[26] L. Su, R. Trivedi, N. V. Sapra, A. Y. Piggott, D. Vercruysse, J. Vučkovíc, *Opt. Express* **2018**, *26*, 4023.
[27] J. Lu, J. Vučković, *Opt. Express* **2013**, *21*, 13351.
[28] A. Y. Piggott, J. Lu, K. G. Lagoudakis, J. Petykiewicz, T. M. Babinec, J. Vučković, *Nat. Photonics* **2015**, *9*, 374.
[29] S. Boyd, S. P. Boyd, L. Vandenberghe, *Convex optimization*, Cambridge university press, **2004**.
[30] J. Lu, *Nanophotonic Computational Design*, PhD thesis, University of Stanford, **2013**.
[31] J. S. Jensen, O. Sigmund, *Laser Photonics Rev.* **2011**, *5*, 308.
[32] M. P. Bendsøe, O. Sigmund, in *Topology Optimization*, Springer, **2004**, pp. 1–69.
[33] J. S. Jensen, O. Sigmund, *Appl. Phys. Lett.* **2004**, *84*, 2022.
[34] M. Burger, S. Osher, E. Yablonovitch, *IEICE TRANSACTIONS ON ELECTRONICS* **2004**, *E87C*, 258.
[35] J. Yang, J. A. Fan, *Opt. Lett.* **2017**, *42*, 3161.
[36] C. Sitawarin, W. Jin, Z. Lin, A. W. Rodriguez, *Photonics Res.* **2018**, *6*, B82.
[37] D. Sell, J. Yang, S. Doshay, R. Yang, J. A. Fan, *Nano Lett.* **2017**, *17*, 3752.
[38] J. Yang, J. A. Fan, *Opt. Express* **2017**, *25*, 23899.
[39] J. Yang, D. Sell, J. A. Fan, *Annalen der Physik* **2018**, *530*, 1700302.
[40] D. Sell, J. Yang, E. W. Wang, T. Phan, S. Doshay, J. A. Fan, *ACS Photonics* **2018**, *5*, 2402.
[41] D. Sell, J. Yang, S. Doshay, J. A. Fan, *Adv. Opt. Mater.* **2017**, *5*, 1700645.
[42] T. Phan, D. Sell, E. W. Wang, S. Doshay, K. Edee, J. Yang, J. A. Fan, *Light: Science & Applications* **2019**, *8*, 48.
[43] D. Sell, J. Yang, S. Doshay, K. Zhang, J. A. Fan, *ACS Photonics* **2016**, *3*, 1919.
[44] Z. Lin, S. G. Johnson, *Opt. Express* **2019**, *27*, 32445.
[45] R. Pestourie, C. Pérez-Arancibia, Z. Lin, W. Shin, F. Capasso, S. G. Johnson, *Opt. Express* **2018**, *26*, 33732.
[46] M. Schlüter, J. A. Egea, J. R. Banga, *Computers & Operations Research* **2009**, *36*, 2217.
[47] R. Li, M. T. Emmerich, J. Eggermont, T. Bäck, M. Schütz, J. Dijkstra, J. H. Reiber, *Evolutionary computation* **2013**, *21*, 29.
[48] J. Johnson, V. Rahmat-Samii, *IEEE Antennas and Propagation Magazine* **1997**, *39*, 7.
[49] R. Haupt, *IEEE Transactions on Antennas and Propagation* **1994**, *42*, 993.
[50] C. Forestiere, A. J. Pasquale, A. Capretti, G. Miano, A. Tamburrino, S. Y. Lee, B. M. Reinhard, L. Dal Negro, *Nano Lett.* **2012**, *12*, 2037.
[51] A. Mirzaei, A. E. Miroshnichenko, I. V. Shadrivov, Y. S. Kivshar, *Appl. Phys. Lett.* **2014**, *105*, 011109.
[52] R. L. Haupt, D. H. Werner, *Genetic Algorithms in Electromagnetics*, John Wiley & Sons, **2007**.
[53] D. Z. Zhu, E. B. Whiting, S. D. Campbell, D. B. Burckel, D. H. Werner, *ACS Photonics* **2019**, *6*, 2741.
[54] C. Liu, S. A. Maier, G. Li, *ACS Photonics* **2020**, *7*, 1716.
[55] M. J. Wallace, S. T. Naimi, G. Jain, R. McKenna, F. Bello, J. F. Donegan, *Opt. Express* **2020**, *28*, 8169.
[56] Z. Li, L. Stan, D. A. Czaplewski, X. Yang, J. Gao, *Opt. Lett.* **2019**, *44*, 114.
[57] Z. Jin, S. Mei, S. Chen, Y. Li, C. Zhang, Y. He, X. Yu, C. Yu, J. K. Yang, B. Luk'yanchuk, S. Xiao, C.-W. Qiu, *ACS nano* **2019**, *13*, 821.

[58] S. Sui, H. Ma, Y. Lv, J. Wang, Z. Li, J. Zhang, Z. Xu, S. Qu, *Opt. Express* **2018**, *26*, 1443.
[59] T. J. Cui, S. Liu, L. Zhang, *J Mater Chem C* **2017**, *5*, 3644.
[60] N. Bonod, S. Bidault, G. W. Burr, M. Mivelle, *Adv. Opt. Mater.* **2019**, *7*, 1900121.
[61] H. Li, G. Wang, L. Zhu, X. Gao, H. Hou, *Optics Communications* p. 124601 (**2020**).
[62] K. Deb, *Multi-objective optimization using evolutionary algorithms*, John Wiley & Sons, **2001**.
[63] T. F. Eibert, J. L. Volakis, D. R. Wilton, D. R. Jackson, *IEEE Transactions on Antennas and Propagation* **1999**, *47*, 843.
[64] P. R. Wiecha, A. Arbouet, C. Girard, A. Lecestre, G. Larrieu, V. Paillard, *Nature Nanotechnology* **2017**, *12*, 163.
[65] D. Tang, L. Chen, J. Liu, *Opt. Express* **2019**, *27*, 12308.
[66] E. B. Whiting, S. D. Campbell, D. H. Werner, P. L. Werner, in *2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting*, **2019**, pp. 1815–1816.
[67] M. Shokooh-Saremi, R. Magnusson, *Opt. Lett.* **2007**, *32*, 894.
[68] S. M. Mirjalili, K. Abedi, S. Mirjalili, *Optik* **2013**, *124*, 5989.
[69] C. Forestiere, M. Donelli, G. F. Walsh, E. Zeni, G. Miano, L. Dal Negro, *Opt. Lett.* **2010**, *35*, 133.
[70] J. R. Ong, H. S. Chu, V. H. Chen, A. Y. Zhu, P. Genevet, *Opt. Lett.* **2017**, *42*, 2639.
[71] Q. Yang, S. Kruk, Y. Xu, Q. Wang, Y. K. Srivastava, K. Koshelev, I. Kravchenko, R. Singh, J. Han, Y. Kivshar, I. Shadrivov, *Adv. Funct. Mater.* **2020**, *30*, 1906851.
[72] B. Zhang, W. Chen, P. Wang, S. Dai, H. Li, H. Lu, J. Ding, J. Li, Y. Li, Q. Fu, T. Dai, Y. Wang, J. Yang, *Opt. Commun.* **2019**, *451*, 186.
[73] N. Hansen, A. Ostermeier, *Evolutionary Computation* **2001**, *9*.
[74] N. Hansen, S. Muller, P. Koumoutsakos, *Evolutionary Computation* **2003**, *11*, 1.
[75] M. M. R. Elsawy, S. Lanteri, R. Duvigneau, G. Brière, M. S. Mohamed, P. Genevet, *Scientific Reports* **2019**, *9*.
[76] P. E. Sieber, D. H. Werner, *Opt. Express* **2014**, *22*, 32371.
[77] I. Martinez, A. H. Panaretos, D. H. Werner, G. Oliveri, A. Massa, in *2013 7th European Conference on Antennas and Propagation (EuCAP)*, IEEE, **2013**, pp. 1843–1847.
[78] J. Nagar, S. Campbell, D. Werner, *Optica* **2018**, *5*, 99.
[79] S. So, T. Badloe, J. Noh, J. Rho, J. Bravo-Abad, *Nanophotonics* **2020**.
[80] Y. Kiarashinejad, M. Zandehshahvar, S. Abdollahramezani, O. Hemmatyar, R. Pourabolghasem, A. Adibi, *Advanced Intelligent Systems* **2020**, *2*, 1900132.
[81] Z. A. Kudyshev, A. V. Kildishev, V. M. Shalaev, A. Boltasseva, *arXiv preprint arXiv:1910.12741* **2019**.
[82] Y. Kiarashinejad, S. Abdollahramezani, A. Adibi, *npj Comput. Mater.* **2020**, *6*, 1.
[83] J. Peurifoy, Y. Shen, L. Jing, Y. Yang, F. Cano-Renteria, B. G. DeLacy, J. D. Joannopoulos, M. Tegmark, M. Soljačić, *Sci. Adv.* **2018**, *4*, eaar4206.
[84] D. Liu, Y. Tan, E. Khoram, Z. Yu, *ACS Photonics* **2018**, *5*, 1365.
[85] W. Ma, F. Cheng, Y. Liu, *ACS Nano* **2018**, *12*, 6326.
[86] D. Gostimirovic, W. N. Ye, *IEEE J. Sel. Top. Quantum Electron.* **2019**, *25*, 1.
[87] Q. Zhang, C. Liu, X. Wan, L. Zhang, S. Liu, Y. Yang, T. J. Cui, *Adv. Theory Simul.* **2019**, *2*, 1800132.
[88] Z. Liu, D. Zhu, S. P. Rodrigues, K. T. Lee, W. Cai, *Nano Lett.* **2018**, *18*, 6570.
[89] J. Jiang, J. A. Fan, *Nano Lett.* **2019**, *19*, 5366.
[90] Y. Liu, T. Lu, K. Wu, J. M. Jin, in *2018 IEEE 27th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)*, IEEE, **2018**, pp. 261–263.
[91] Y. Long, J. Ren, Y. Li, H. Chen, *Appl. Phys. Lett.* **2019**, *114*, 181105.
[92] S. So, J. Rho, *Nanophotonics* **2019**, *8*, 1255.
[93] J. Jiang, D. Sell, S. Hoyer, J. Hickey, J. Yang, J. A. Fan, *ACS Nano* **2019**, *8*, 8872.
[94] J. Jiang, J. A. Fan, *Nanophotonics* **2019**, *Ahead of Print*.
[95] M. Raissi, P. Perdikaris, G. E. Karniadakis, *J. Comput. Phys.* **2019**, *378*, 686.
[96] D. Jones, *J Global Optimization* **1998**, *13*.
[97] D. Jones, *J Global Optimization* **2001**, *21*, 345.
[98] C. Forestiere, Y. He, R. Wang, R. M. Kirby, L. Dal Negro, *ACS Photonics* **2015**, *3*, 68.
[99] D. J. MacKay, *Neural Computation* **1991**, *4*.
[100] P. I. Schneider, X. Garcia Santiago, V. Soltwisch, M. Hammerschmidt, S. Burger, C. Rockstuhl, *ACS Photonics* **2019**, *6*, 2726.
[101] E. W. Wang, D. Sell, T. Phan, J. A. Fan, *Opt. Mater. Express* **2019**, *9*, 469.
[102] Z. Ren, M. T. Pham, C. S. Koh, *IEEE Trans. Magn.* **2012**, *49*, 851.
[103] J. Jung, *IEEE Photonics Technol. Lett.* **2015**, *28*, 756.
[104] J. P. Hugonin, P. Lalanne, "Reticolo software for grating analysis," www.lp2n.institutoptique.fr. accessed: January 2014.
[105] N. Schmitt, N. Georg, G. Brière, D. Loukrezis, S. Héron, S. Lanteri, C. Klitis, M. Sorel, U. Römer, H. D. Gersem, S. Vézian, P. Genevet, *Opt. Mater. Express* **2019**, *9*, 892.

Mahmoud M. R. Elsawy received his Ph. D. degree at University of Aix-Marseille, France in 2017 with specialization in optics, photonics, and image processing. The topic of the Ph. D. was related to modeling and improvements of complex nonlinear plasmonic waveguides. He spent one year as a postdoctoral researcher at Institut Fresnel, Marseille, France. His research activities were dedicated to the field of nonlinear plasmonics and modeling of hollow-core negative curvature optical fibers. Since October 2018, he is working as a postdoctoral visitor researcher at Inria Sophia Antipolis. His research focuses on numerical optimization of metasurfaces and complex nanophotonic devices.

Stéphane Lanteri received his Ph.D. degree in engineering sciences at University of Nice Sophia Antipolis, France, in 1991. After a postdoc in the Aerospace Engineering Department of the University of Colorado at Boulder, he was appointed as a research scientist at Inria Sophia Antipolis, in 1993.He is currently a senior research scientist and heading research activities on innovative mathematical methods for the numerical modeling of nanoscale light-matter interactions, including high order finite element methods (Discontinuous Galerkin methods) and high performance multiscale solvers of full-wave differential models. He also coordinates the development of the DIOGENeS software suite (https://diogenes.inria.fr/) dedicated to computational nanophotonics.

Régis Duvigneau received his Ph. D. degree at Ecole Centrale de Nantes (France) in 2002, whose topic concerned shape optimization algorithms in fluid mechanics. After a post-doc period at Ecole Polytechnique in Montreal (Canada), he became permanent researcher at Inria Sophia-Antipolis (France) in 2005. His research activity was then focused on numerical methods for design optimization for systems governed by partial differential equations, with main applications in aerodynamics.



Jonathan A. Fan received his Ph. D. degree at Harvard University in 2010. After a postdoc at the University of Illinois at Urbana-Champaign, he became an assistant professor in the Department of Electrical Engineering at Stanford University in 2014. His research activities have focused around the growth, assembly, and computational design of new photonic materials.

Patrice Genevet received his Ph.D. degree at the université Côte d'Azur, France in 2009 on localized spatial solitons in semiconductor lasers. He did five years as a research fellow (2009–2014) in the Capasso group (Harvard University) in collaboration with Prof. Scully (Texas A&M University). In 2014, he worked as senior research scientist in A*STAR, Singapore, to join CNRS (France) in 2015. His research activities concern the development of metamaterials, metasurfaces and their applications.