Inverse design of free-form nanophotonic devices

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc. Yannick Augenstein

Tag der mündlichen Prüfung: 17. Mai 2024

Referent:	Prof. Dr. Carsten Rockstuhl
Korreferent:	Prof. Dr. Otto Muskens

Selbstständigkeitserklärung

Eidesstattliche Versicherung gemäß §13 Absatz 2 Ziffer 3 der Promotionsordnung des Karlsruher Instituts für Technologie (KIT) für die KIT-Fakultät für Physik:

1. Bei der eingereichten Dissertation zu dem Thema „Inverse design of free-form nanophotonic devices“ handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich wie bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum	Unterschrift

Acknowledgements

It has been over five years since I began this journey, initially as a master's student, in Prof. Carsten Rockstuhl's group. Throughout this time, his support and the freedom he has granted me to explore a broad spectrum of topics (sometimes maybe even too generously), all while offering invaluable scientific insights, have been wonderful. I sincerely thank him for his mentorship and guidance, both of which have been instrumental in my technical and professional development.

Further, I would like to thank Prof. Otto Muskens for taking the time to act as a second referee for this thesis, and for being so accommodating with shifting around the schedule for the defense.

I also owe a great deal of thanks to the Cluster of Excellence "3D Matter Made to Order" and its members. Their interdisciplinary approach has provided me with fresh perspectives that have allowed me to recontextualize my research. Similarly, my gratitude extends to the HEiKA Graduate School on "Functional Materials", which has offered me significant opportunities for development through its scientific and management modules.

It has been a pleasure to be part of the wonderful group that Carsten has built at TFP, and I am thankful for everyone I have had the opportunity to get to know and work with over the years. Although the COVID-19 pandemic unfortunately cut short much of the time we could have spent together, I believe we still made the most of it. And speaking as the last remaining member of the TFP bike team: It has been an honor and a privilege to hold down the fort – what a ride it has been!

My heartfelt gratitude extends to my close friends and family for their love and support over the years. I am profoundly fortunate to have such solid social support; without it, my journey would surely have taken a very different path. I am eternally indebted to my wife, Patricia, for her extraordinary support, especially in taking care of our daughter this last year. While I would dedicate this thesis to them, that seems like a terrible gift all around, so I guess I will have to come up with something better.

Lastly, I would like to extend my gratitude to Mr. Schrimpf, my tenth-grade physics teacher, for showing me that physics can be fun. Without that initial spark, it seems unlikely I would have ever pursued physics at all. It underscores the profound, yet often overlooked, impact teachers can have on their students. More than a decade later, acknowledging his impact here feels like the very least I can do.

Abstract

Inverse problems are ubiquitously encountered throughout science and engineering. Where the forward problem answers the question of what the output for a given input looks like, the inverse problem tries to answer the opposite: given a set of outputs, what were the inputs? While the forward problem is typically uniquely defined and can be solved through numerical modeling, the inverse problem is generally ill-posed, making its direct solution intractable. Inverse design is a class of methods that aim to solve the inverse problem, at least to a "good enough" approximation, by computational optimization of a mathematically defined objective function.

Topology optimization, in particular, is a gradient-based method for inverse design. The method has gained popularity in photonics in the past decade and has led to the creation of devices with non-intuitive designs and exceptional performance. This thesis applies topology optimization to designing various nanophotonic devices, from two-dimensional structures that manipulate and guide surface waves to fully free-form and three-dimensional devices such as fiber-to-chip and grating couplers. We find that while topology optimization and additive manufacturing via methods such as 3D laser nanoprinting ideally complement each other, creating fabrication-ready free-form nanophotonic devices presents unique challenges. We identify the issue of structural integrity and develop a method for coupled mechanical and electromagnetic inverse design, demonstrating that this approach can yield more feasible devices for fabrication.

Lastly, we focus on the issue of computational cost – topology optimization typically involves many iterations of computationally expensive numerical simulations, which can limit the extent to which the design space can be explored. We develop a framework for the inverse design of nanophotonic devices via a neural operator-based surrogate solver and apply it to optimize free-form electromagnetic scatterers. As these surrogate solvers are trained on data obtained from numerical simulations, we discuss the trade-offs in terms of generality and accuracy and examine the problem settings in which such trade-offs can be feasibly made.

Contents

List of Figures

List of Tables

List of publications

Published peer-reviewed articles

As first author

[P1] Y. Augenstein and C. Rockstuhl. “Inverse design of nanophotonic devices with structural integrity”. In: *ACS Photonics* 7.8 (2020), pp. 2190–2196.

[P2] Y. Augenstein, M. Roussey, T. Grosjean, E. Descrovi, and C. Rockstuhl. “Inverse design of cavities for Bloch Surface Waves interfaced to integrated waveguides”. In: *Photonics and Nanostructures - Fundamentals and Applications* 52 (2022), p. 101079.

[P3] Y. Augenstein, T. Repän, and C. Rockstuhl. “Neural operator-based surrogate solver for free-form electromagnetic inverse design”. In: *ACS Photonics* 10.5 (2023), pp. 1547–1557.

As contributing author

T. Repän, Y. Augenstein, and C. Rockstuhl. “Exploiting geometric biases in inverse nano-optical problems using artificial neural networks”. In: *Optics Express* 30.25 (2022), pp. 45365–45375.

R. Venkitakrishnan, Y. Augenstein, B. Zerulla, F. Z. Goffi, M. Plum, and C. Rockstuhl. “On the physical significance of non-local material parameters in optical metamaterials”. In: *New Journal of Physics* 25.12 (2023), p. 123014.

M. R. Whittam, A. G. Lamprianidis, Y. Augenstein, and C. Rockstuhl. “Identifying regions of minimal backscattering by a relativistically moving sphere”. In: *Physical Review A* 108.4 (2023), p. 043510.

P. Dhawan, L. Schulte, P. Piechulla, Y. Augenstein, M. Gaudig, A. Sprafke, R. B. Wehrspohn, and C. Rockstuhl. “On the reliability of the collective coordinate method to simulate metasurfaces with correlated disorder used for light management”. In: *Journal of the Optical Society of America B* 40.3 (2023), B8–B18.

Conference contributions

Talks given

Y. Augenstein and C. Rockstuhl. “Adjoint-based Topology Optimization Applied to Inverse Design of Nanophotonic Materials”. In: Metamaterials 2019 (poster). Rome, Italy.

Y. Augenstein and C. Rockstuhl. “Material reparametrization for topology optimization of 3D photonic nanostructures”. In: Workshop on Theoretical and Numerical Tools for Nanophotonics 2020. Berlin, Germany.

Y. Augenstein and C. Rockstuhl. “Inverse design of nanophotonic devices with structural integrity”. In: Metamaterials 2020. Virtual.

Y. Augenstein and C. Rockstuhl. “A perspective on parametrizations in topology optimization”. In: Metamaterials 2021. Virtual.

Talks given by others

C. Rockstuhl, Y. Augenstein, T. Repän, X. Garcia-Santiago, and S. Burger. “Inverse Design of Structured Dielectric Materials and Devices”. In: Smart NanoMaterials 2020. Virtual.

T. Repän, Y. Augenstein, and C. Rockstuhl. “Improved solutions to optical inverse problems by neural networks and prior assumptions”. In: Metamaterials 2021. Virtual.

Y. Augenstein, T. Repän, R. Venkitakrishnan, L. Kuhn, X. Garcia-Santiago, and C. Rockstuhl. “Solving Inverse Problems in the Field of Computational Nanophotonics”. In: IEEE COMCAS 2021. Tel Aviv, Israel.

T. Repän, Y. Augenstein, and C. Rockstuhl. “Improved solutions to optical inverse problems by neural networks and prior assumptions”. In: SPIE Photonics Europe 2022. Strassbourg, France.

Y. Augenstein, L. Kuhn, T. Repän, and C. Rockstuhl. “Artificial Neural Networks for Selected Challenges in Nanophotonics”. In: META 2023 (invited). Paris, France.

M. Whittam, A. Lamprianidis, Y. Augenstein, and C. Rockstuhl. “Minimizing the Back-Scattering by a Relativistically-Moving Sphere”. In: Metamaterials 2023 (poster). Chania, Greece.

Y. Augenstein, L. Kuhn, T. Repän, and C. Rockstuhl. “Exploring Multiple Network Architectures to Solve Selected Challenges in Computational Nanophotonics”. In: Metamaterials 2023 (invited). Paris, France.

Conference papers

T. Repän, Y. Augenstein, and C. Rockstuhl. “Improved solutions to optical inverse problems by neural networks and prior assumptions”. In: *Metamaterials XIII*. Vol. 12130. SPIE. 2022, pp. 35–38.

1 Introduction

Nanophotonics, as the investigation of light-matter interactions on the nanoscale, plays a crucial role in advancing various modern technologies such as photolithography, optical communication, and optoelectronics [1]. Most nanophotonic device development follows a structured process that starts with conceptual design and is followed by verification through numerical simulations, fabrication, and detailed characterization. The design process often begins with a basic model inspired by similar devices in related fields, like microwave engineering, adjusted through iterative simulations to optimize performance for the intended application. This iterative approach has led to the creation of an extensive collection of nanophotonic device templates. These templates integrate principles of material resonances, bandgap engineering, and index guiding, enabling unparalleled control over light propagation across a broad range of frequencies.

Computational techniques like the finite-difference time-domain (FDTD) [2, 3], finite-difference frequency domain (FDFD), and the finite element method (FEM) [4, 5] are crucial for simulating device responses to various inputs. Addressing a given set of inputs to predict device outcomes is referred to as solving the direct, or forward, problem. However, this direct method often does not correspond with practical scenarios where the actual query is about designing a device capable of producing specific outputs, thus necessitating the resolution of what is known as the “inverse problem”. The inverse problem focuses on deducing the necessary device parameters to achieve a desired outcome. Traditionally, this has involved manual adjustments of device parameters through trial and error – often guided by the designer’s physical intuition – which can be seen as a form of exhaustive brute-force optimization. Nevertheless, as nanophotonics continues to expand, embracing an increasingly wide array of phenomena and fabrication techniques, the complexity of the functional requirements for nanophotonic devices grows. Numerous applications require the simultaneous optimization of several independent features, such as device functionality at particular frequencies, minimization of the overall device size, and tolerance to fabrication imperfections.

As nanophotonic devices become increasingly more complex and need to meet specific functional requirements, the traditional manual optimization process often becomes impractical and inefficient. This evolution underscores the necessity for advanced computational tools to support, enhance, or fully automate this design process. Developing and integrating such tools into the workflow of nanophotonic device design would streamline the optimization process and enable the exploration of a broader design space with higher precision and efficiency. Consequently, there has been a significant push towards leveraging computational methods, such as machine learning and computational inverse design techniques [6]. These techniques can intelligently navigate the vast parameter space to identify optimal device configurations that meet the desired performance criteria without the exhaustive manual tuning previously required.

Gradient-free optimization methods, such as iterated searches [7, 8], genetic algorithms [9, 10], and particle swarm optimization [11, 12], have been applied extensively and successfully as a computational alternative to the traditional design approach. These algorithms adjust the design variables, compute the forward problem for each variant, and retain the variable set that most closely aligns with the design objectives. The distinction among these techniques primarily lies in the manner of perturbing design variables and the strategy for selecting solutions for subsequent iterations. A key advantage of gradient-free optimization is their inherent ability to converge to a solution within the search space irrespective of the space's differentiability or the problem's convexity. This trait makes them particularly useful in scenarios where gradient-based methods may not be applicable or efficient, *e.g.*, when the objective function is non-differentiable or the problem is combinatorial. Moreover, these methods are versatile and capable of handling a wide range of optimization problems without necessitating gradient information, which can be difficult or computationally expensive to compute in certain contexts. However, the trade-off for this flexibility is that these methods require many function evaluations to find a satisfactory solution, leading to increased computational costs, especially for high-dimensional or complex design spaces. The fundamental nature of gradient-free optimization methods, being essentially combinatorial, significantly impacts their scalability with complexity. The necessity to compute the forward problem multiple times is directly tied to the number of design variables, leading to poor scalability as the complexity of the device increases. In the context of nanophotonic device optimization, design variables often represent geometric parameters that define the distribution of materials. Consequently, limiting the number of degrees of freedom in the search space also limits the range of possible device topologies that can be explored through these optimization methods. A significant disparity exists between the complexity and variety of topologies that could theoretically be manufactured using advanced fabrication techniques and the relatively constrained search space accessible through these optimization strategies. This gap implies that the potential of nanophotonic device design, enabled by modern fabrication methods such as photolithography or 3D laser nanoprinting, may not be fully realized or explored using conventional gradient-free optimization approaches.

Gradient-based optimization techniques offer a more efficient approach to exploring the parameter space by leveraging gradient information to determine the search direction. These methods can optimize many design parameters – often on the order of millions – simultaneously. In theory, gradient-based optimization is straightforward, and implementing such algorithms is generally not labor-intensive because of their broad applicability and the availability of performant implementations in the public domain [13, 14, 15]. The challenge, however, lies in computing the gradients for a specific optimization problem. While obtaining numerical gradients through finite differences is conceptually simple, it necessitates solving the forward problem for each variation in the design variables, thereby incurring computational complexities akin to those encountered in gradient-free techniques. Adopting more efficient methods for gradient computation, such as the adjoint method, is the key to overcoming these complexities. The adjoint method computes the gradient of the objective function with respect to any number of design variables through only two full-wave simulations. This efficiency makes the adjoint method particularly attractive for problems involving many design variables, as is often the case in nanophotonic device optimization.

1.1 Inverse design in nanophotonics

Inverse design has seen a rapid rise in popularity in nanophotonic within the past decade. Pioneered by Jensen and Sigmund [16], who carried the concept over from mechanical engineering to photonics, and later popularized by Piggott *et al.* [17], countless studies have been dedicated to the exploration of inverse design in nanophotonics since. Much, if not most, of this work has been dedicated to realizing optical components for on-chip photonic integrated circuits (PICs) such as compact power splitters [18], polarization splitters [19], mode converters [20], and wavelength demultiplexers [17]. The strength of these devices has typically been their previously unparalleled performance at small footprints, with the drawback of lower tolerances for fabrication imperfections and oftentimes relatively narrow-band operation.

While earlier works focused on proof-of-concept designs [21, 6], much effort has gone into resolving issues with manufacturability, with the development of schemes for incorporating feature size constraints into the optimization [18, 20]. While important, such earlier schemes generally failed to fully capture the design restrictions imposed by commercial foundries with CMOS-compatible silicon photonics processes. Many of these foundries and their associated fabrication processes have design rules that are notoriously difficult to deal with, even in “conventional” PIC design. Because of this, much focus in recent developments has been on the incorporation of such design rules [22, 23, 24], marking an important step toward the commercialization of inverse design in the context of integrated photonics.

These developments have led inverse design in silicon photonics to a point where we are now seeing initial steps toward industry adoption, with commercial electromagnetic simulation and design software such as Lumerical FDTD and Tidy3D incorporating tools for inverse design. The demand for high-bandwidth interconnects due to the ever-rising requirements of datacenters and more recently, machine learning, coupled with key players in the chip industry announcing roadmaps for co-packaged optics, means that it is now only a matter of time until miniaturized photonics – and with it, inverse design – reaches a point of widespread adoption. In other words: it is a good time to be a photonics engineer.

The research community is – as always – forward-looking and, in the context of integrated photonics, has started exploring alternative material platforms with more “exotic” optical properties, where suitable methods have been developed for the inverse design of, *e.g.*, nonlinear optics [25, 26, 27, 28]. Some of these new material platforms include lithium niobate [29], diamond photonics [30], chalcogenide glass [31], and silicon carbide [32, 33].

Another area where inverse design has gained popularity is in the design of metasurfaces [34, 35, 36], particularly in the context of flat lenses ("metalenses") [37, 38] and other imaging systems [39, 40]. While being a different application and problem setting, most of these designs have stayed within the confines of two-dimensional patterning of single-layer structures by traditional photolithography techniques. However, there has been some investigation into structures made of multiple patterned layers [41]. Nevertheless, outside of on-chip applications, restricting designs to a 2D configuration severely limits the functionality that could, in principle, be attainable through 3D patterning of devices.

A technology with which such patterning is possible is two-photon lithography [42], sometimes also called 3D laser nanoprinting. 3D laser nanoprinting is a high-resolution additive manufacturing technique that uses focused laser beams to precisely fabricate structures at the nanoscale. It has seen rapid developments in terms of scalability and printable feature sizes in recent years [43, 44]. While this technology has, of course, not gone unnoticed by the inverse design community, work on the development and application of inverse design techniques tailored for such fabrication has been sparse [45, 46], and most have restricted the geometries to quasi-planar designs [47, 48].

1.2 Thesis outline

With 2D inverse design on the way to commercialization, a logical next step is the development and application of inverse design to fully free-form devices, which will be the focus of this thesis from Chapter 4 onward. We will explore the application of gradient-based inverse design to a series of increasingly more complex nanophotonic design challenges, where we will uncover some novel device designs and design strategies.

Following this introduction, **Chapter 2** will introduce the theory for the two cornerstones of large-scale, gradient-based inverse design in nanophotonics, namely the *adjoint method* and *topology optimization*, which will serve as the lens through which we will view and approach the design challenges introduced in the subsequent chapters.

In **Chapter 3**, we will introduce our first nanophotonic inverse design challenge – enhancing and steering the emission of a dipole emitter that excites Bloch surface waves into a waveguide. We will first discuss the theory of wave propagation in stratified media, which leads to the appearance of such surface modes. These modes are weakly guided and experience an extremely low effective index contrast, rendering intuitive design challenging. A unique property of this system is that it can be modeled accurately through 2D simulations, making it feasible to systematically study variations in the refractive index contrast through multiple optimizations.

Moving on to 3D simulations, we will proceed to design a device that couples light from an optical single-mode fiber into a photonic wire bond in **Chapter 4**. This system is characterized by its rotational symmetry. We will develop an efficient parameterization for topology optimization that constrains the design space accordingly. Using this parameterization, we design a device that offers high coupling efficiencies at a much smaller device footprint than typical couplers.

In **Chapter 5**, we will then proceed to tackle a truly large-scale inverse design challenge: optimizing a polarization-independent grating coupler. We design the device with additive manufacturing via 3D laser nanoprinting in mind, thereby obviating the need for planar or symmetry restrictions used in the previous chapters. This leads to a fully free-form geometry with millions of degrees of freedom in the optimization. The optimized device achieves some of the highest reported coupling efficiencies for both polarizations but presents a challenge for experimental realization.

Consequently, **Chapter 6** will highlight some of the unique challenges that arise when using topology optimization to design free-form nanophotonic devices. The chapter will focus on the aspects of *connectivity* and *structural integrity*, which are aspects of inverse design that have largely been neglected in the literature due to the focus on the optimization of planar (on-chip) devices. We develop a framework for co-designing devices' photonic and structural properties by combining optical and mechanical topology optimization and systematically study the tradeoffs on the optical objective associated with imposing structural penalties, where we discover that such co-designed devices can achieve both structural integrity and only marginally diminished optical performance

Chapter 7 will illuminate the challenge of nanophotonic inverse design from an entirely different perspective – the computational cost associated with large-scale inverse design. In particular, we will explore the use of neural network-based surrogate solvers to replace conventional full-wave solvers such as FDTD. Our work will focus on so-called *neural operators*, a recently developed class of machine learning models and compare their performance to other state-of-the-art models used in nanophotonics, where we show drastically improved accuracy and data-efficiency for our approach. We will then demonstrate the inverse design of fully free-form three-dimensional devices with the developed surrogate model.

Finally, **Chapter 8** will conclude with a summary of our findings, place the work presented in a broader scientific context, and outline possible future paths for development.

2 Theory

Two main ingredients are necessary to perform gradient-based inverse design: a way to obtain the gradients of some objective function with respect to a set of optimization variables and a suitable parameterization such that these parameters can represent a physical design. In this chapter, we will introduce the adjoint method in its general form in Section 2.1 as an efficient way of obtaining these gradients, which we will then apply to Maxwell's equations. Following this, we will show how these gradients are connected to *topology optimization* in Section 2.2 and introduce the three-field parameterization scheme.

2.1 Adjoint sensitivity analysis

Various approaches can generally tackle inverse design, but in the case of problems with a very large number of degrees of freedom, such as the ones we discuss here, gradient-free methods become infeasible as the problem's dimensionality is simply too large. Instead, we rely on local, gradient-based optimization, and to do this, we need first to establish a way of obtaining gradients of some figure of merit with respect to the optimization parameters. In particular, obtaining these gradients needs to be as efficient as possible, as Maxwell's equations are generally computationally expensive. Therefore, derived gradients by finite-differencing schemes are infeasible, as this would entail running at least one additional simulation per degree of freedom in the optimization. Instead, we will employ the so-called adjoint method, or adjoint sensitivity analysis, which allows for efficient gradient computation for linear systems with many free parameters.

Adjoint sensitivity analysis has a long history in optimal control theory as a method of finding a certain optimality criterion for a given system. It has been used extensively in the engineering disciplines in the context of computational fluid dynamics and structural engineering [49, 50].

This analysis technique draws its theoretical basis from Pontryagin's maximum principle, a cornerstone in control theory developed by Kopp [51], which facilitates the identification of the most effective control strategies for guiding a dynamical system from one state to another. Thus, adjoint sensitivity analysis can be seen as an extension of the calculus of variations, offering two principal approaches for establishing an adjoint system: one can either derive the gradient of the objective function from a continuous design problem before discretizing this gradient for computational implementation or discretize the problem initially and then derive the gradient from this discretized model. These approaches converge in outcomes as the discretization becomes infinitely fine, yet exhibit minor differences otherwise.

In this section, we will introduce the discrete adjoint formalism in its general form and then outline its application to Maxwell's equations.

2.1.1 The discrete adjoint method

Numerous physical systems, encompassing fields such as thermal flow, structural mechanics, and electromagnetics – as governed by Maxwell's equations – can be modeled as linear equations that encapsulate the underlying physics of the form

$$\mathrm{A}\boldsymbol{x} = \boldsymbol{b} \quad , \tag{2.1}$$

where A is the system matrix, $\boldsymbol{x}$ is the solution field vector and $\boldsymbol{b}$ are the sources. Resolution of this system can be achieved through direct methods, such as LU or Cholesky decomposition, or iterative techniques, such as the Richardson or Jacobi methods, depending on the problem's nature and computational considerations.

When designing physical structures, be it load-bearing elements in mechanical engineering or integrated optical circuits in nanophotonics, one is typically interested in optimizing some real-valued scalar objective function $F(\boldsymbol{x})$ of the solution vector $\boldsymbol{x}$. This can be expressed as a general optimization problem of the form

$$\begin{aligned} &\min_{\boldsymbol{\rho}} && F(\boldsymbol{x}) \\ &\text{subject to} && \mathrm{A}(\boldsymbol{\rho})\, \boldsymbol{x} = \boldsymbol{b}(\boldsymbol{\rho}) \quad , \end{aligned} \tag{2.2}$$

with the parameter vector $\boldsymbol{\rho}$. While $\boldsymbol{\rho}$ can, in principle, be any vector of coefficients in the linear system, it typically represents the material parameters that are discretized on the computational grid in our case. Since we are interested in the sensitivity of the objective function F with respect to a change in these parameters $\boldsymbol{\rho}$, we are ultimately interested in evaluating the following expression

$$\frac{\partial F}{\partial \boldsymbol{\rho}} = \frac{\partial F}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\rho}} \quad . \tag{2.3}$$

Obtaining $\partial F / \partial \boldsymbol{x}$ only requires the solution vector $\boldsymbol{x}$, which is simply the result of solving the linear system once. The derivative of the objective function F with respect to its inputs is usually either known analytically or easily derived through other means, such as automatic differentiation [52, 53].

Assuming that A is invertible, we can find for the derivative $\partial \boldsymbol{x} / \partial \boldsymbol{\rho}$:

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\rho}} = \frac{\partial \mathrm{A}^{-1}}{\partial \boldsymbol{\rho}} \boldsymbol{b} + \mathrm{A}^{-1} \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\rho}} \tag{2.4a}$$

$$= -\mathrm{A}^{-1} \frac{\partial \mathrm{A}}{\partial \boldsymbol{\rho}} \mathrm{A}^{-1} \boldsymbol{b} + \mathrm{A}^{-1} \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\rho}} \tag{2.4b}$$

$$= \mathrm{A}^{-1} \left(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\rho}} - \frac{\partial \mathrm{A}}{\partial \boldsymbol{\rho}} \boldsymbol{x} \right) \quad , \tag{2.4c}$$

leading to:

$$\frac{\partial F}{\partial \boldsymbol{\rho}} = \frac{\partial F}{\partial \boldsymbol{x}} \mathrm{A}^{-1} \left(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\rho}} - \frac{\partial \mathrm{A}}{\partial \boldsymbol{\rho}} \boldsymbol{x} \right) \quad . \tag{2.5}$$

Assuming that we have already solved the linear system once, it is straightforward to obtain both $\partial \boldsymbol{b} / \partial \boldsymbol{\rho}$ and $\partial \mathrm{A} / \partial \boldsymbol{\rho}\ \boldsymbol{x}$ – these terms have an analytical relationship with $\boldsymbol{\rho}$, and thus can be easily calculated. But how do we calculate $\partial F / \partial \boldsymbol{x}\ \mathrm{A}^{-1}$? In large systems, we never have direct access to A^{-1}, both due to numerical accuracy [54] as well as memory constraints. The latter is related to the fact that the system matrix of discretized PDE systems is generally sparse and banded, which means that even for very large systems, most sparse matrix formats can store them in $\mathcal{O}(3n)$ space, with n being the number of nonzero entries in the matrix. However, the inverse of that matrix can be dense and therefore require $\mathcal{O}(n^2)$ memory, with much larger n than before. The crux now lies in realizing that instead, $\partial F / \partial \boldsymbol{x}\ \mathrm{A}^{-1}$ can be expressed as another linear system:

$$\frac{\partial F}{\partial \boldsymbol{x}} \mathrm{A}^{-1} = \boldsymbol{x}_{\mathrm{aj}}^{T} \quad , \tag{2.6}$$

or

$$\boldsymbol{x}_{\mathrm{aj}} = \mathrm{A}^{-T} \frac{\partial F}{\partial \boldsymbol{x}^{T}} \quad , \tag{2.7}$$

where we defined the so-called *adjoint* solution $\boldsymbol{x}_{\mathrm{aj}}$. Interestingly, this new linear system uses the *transposed* system matrix of the original system, with a source term that depends on the derivative of the objective function with respect to the original solution. Putting everything together, we obtain

$$\frac{\partial F}{\partial \boldsymbol{\rho}} = \boldsymbol{x}_{\mathrm{aj}}^{T} \left(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\rho}} - \frac{\partial \mathrm{A}}{\partial \boldsymbol{\rho}} \boldsymbol{x} \right) \quad , \tag{2.8}$$

with the direct solution $\boldsymbol{x}$ and the adjoint solution $\boldsymbol{x}_{\mathrm{aj}}$, both which require solving one linear system. Thus, solving two linear systems makes it possible to obtain the gradient of a scalar-valued objective function with respect to an arbitrary number of degrees of freedom $\boldsymbol{\rho}$, and is the fundamental property that underpins all gradient-based optimization discussed herein.

2.1.2 Adjoint method & Maxwell's equations

Establishing a suitable set of equations for electromagnetics is essential before applying the adjoint formalism described previously. Our starting point will be the time-domain Maxwell equations, which underpin all macroscopic electromagnetic phenomena:

$$\begin{aligned} \boldsymbol{\nabla}\cdot\boldsymbol{B}(\boldsymbol{r},t) &= 0 & \boldsymbol{\nabla}\times\boldsymbol{E}(\boldsymbol{r},t) &= -\frac{\partial\boldsymbol{B}(\boldsymbol{r},t)}{\partial t} \\ \boldsymbol{\nabla}\cdot\boldsymbol{D}(\boldsymbol{r},t) &= \rho_{\text{ext}}(\boldsymbol{r},t) & \boldsymbol{\nabla}\times\boldsymbol{H}(\boldsymbol{r},t) &= \boldsymbol{j}_{\text{macr}}(\boldsymbol{r},t) + \frac{\partial\boldsymbol{D}(\boldsymbol{r},t)}{\partial t} \quad , \end{aligned} \tag{2.9}$$

where the macroscopic current density $\boldsymbol{j}_{\text{macr}}(\boldsymbol{r},t)$ represents the source of the electromagnetic field in the absence of free charges, which is generally the case in optics. The constitutive relations are defined as

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{r},t) &= \epsilon_0\boldsymbol{E}(\boldsymbol{r},t) + \boldsymbol{P}(\boldsymbol{r},t) \\ \boldsymbol{B}(\boldsymbol{r},t) &= \mu_0\boldsymbol{H}(\boldsymbol{r},t) + \boldsymbol{M}(\boldsymbol{r},t) \quad , \end{aligned} \tag{2.10}$$

which link the electric and magnetic fields to their respective auxiliary fields. In Eq. (2.9) and Eq. (2.10) we have used SI units and employed the following convention:

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{r},t) &= \text{electric field} \\ \boldsymbol{H}(\boldsymbol{r},t) &= \text{magnetic field} \\ \boldsymbol{D}(\boldsymbol{r},t) &= \text{electric displacement field} \\ \boldsymbol{B}(\boldsymbol{r},t) &= \text{magnetic displacement field} \\ \boldsymbol{j}_{\text{macr}}(\boldsymbol{r},t) &= \text{macroscopic current density} \\ \rho_{\text{ext}}(\boldsymbol{r},t) &= \text{external charge density} \\ \boldsymbol{P}(\boldsymbol{r},t) &= \text{electric polarization} \\ \boldsymbol{M}(\boldsymbol{r},t) &= \text{magnetic polarization} \quad . \end{aligned}$$

We will focus on single-frequency steady-state solutions to Maxwell's equations, which are obtained via a Fourier transform of Eq. (2.9):

$$\begin{aligned} \boldsymbol{\nabla}\cdot\tilde{\boldsymbol{B}}(\boldsymbol{r},\omega) &= 0 & \boldsymbol{\nabla}\times\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) &= i\omega\tilde{\boldsymbol{B}}(\boldsymbol{r},\omega) \\ \boldsymbol{\nabla}\cdot\tilde{\boldsymbol{D}}(\boldsymbol{r},\omega) &= \rho_{\text{ext}}(\boldsymbol{r},\omega) & \boldsymbol{\nabla}\times\tilde{\boldsymbol{H}}(\boldsymbol{r},\omega) &= -i\omega\tilde{\boldsymbol{D}}(\boldsymbol{r},\omega) + \tilde{\boldsymbol{j}}_{\text{macr}}(\boldsymbol{r},\omega) \quad . \end{aligned} \tag{2.11}$$

Further, we will limit our considerations in this thesis to dispersive, linear, local, and isotropic media, for which the material equations in the frequency domain are:

$$\tilde{\boldsymbol{P}}(\boldsymbol{r},\omega) = \epsilon_0\chi(\boldsymbol{r},\omega)\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) \tag{2.12a}$$

$$\tilde{\boldsymbol{D}}(\boldsymbol{r},\omega) = \epsilon_0\underbrace{(1+\chi(\boldsymbol{r},\omega))}_{\epsilon_r}\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) \tag{2.12b}$$

$$\tilde{\boldsymbol{M}}(\boldsymbol{r},\omega) = 0 \tag{2.12c}$$

$$\tilde{\boldsymbol{B}}(\boldsymbol{r},\omega) = \mu_0\tilde{\boldsymbol{H}}(\boldsymbol{r},\omega) \quad , \tag{2.12d}$$

with the susceptibility $\chi(\boldsymbol{r},\omega)$:

$$\chi(\boldsymbol{r},\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(\boldsymbol{r},t)\, e^{i\omega t}\,\mathrm{d}t \quad , \tag{2.13}$$

where $R(\boldsymbol{r},t)$ is the response function related to the polarization induced in the material in the time domain for a delta-type excitation. In the following, we will only consider non-dispersive materials with $\epsilon(\boldsymbol{r},\omega) = \epsilon_r$. To arrive at a linear equation system to which we can apply the adjoint formalism, we will first derive the wave equation for the electromagnetic field.

By taking the curl of Faraday's law and utilizing the constitutive relations Eq. (2.12b) and Eq. (2.12d), we get

$$\nabla \times \nabla \times \tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) = i\omega\mu_0 \nabla \times \tilde{\boldsymbol{H}}(\boldsymbol{r},\omega) \tag{2.14a}$$

$$= \omega^2\mu_0\epsilon_0\epsilon_r\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) + i\omega\mu_0\tilde{\boldsymbol{j}}_{\mathrm{macr}}(\boldsymbol{r},\omega) \quad , \tag{2.14b}$$

and therefore

$$\nabla \times \nabla \times \tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) - \omega^2\mu_0\epsilon_0\epsilon_r\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) = i\omega\mu_0\, \tilde{\boldsymbol{j}}_{\mathrm{macr}}(\boldsymbol{r},\omega) \quad . \tag{2.15}$$

To implement the adjoint method as referenced earlier, we recast the wave equation in the form of a linear system, akin to the structure introduced in Section 2.1:

$$\underbrace{[\nabla \times \nabla \times - \omega^2\mu_0\epsilon_0\boldsymbol{\rho}]}_{\mathrm{A}(\boldsymbol{\rho})}\, \underbrace{\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega)}_{\boldsymbol{x}} = \underbrace{i\omega\mu_0\, \tilde{\boldsymbol{j}}_{\mathrm{macr}}(\boldsymbol{r},\omega)}_{\boldsymbol{b}} \quad , \tag{2.16}$$

introducing a notation change $\epsilon_r \longrightarrow \boldsymbol{\rho}$ to align with our established convention for design variables $\boldsymbol{\rho}$. Note that a slight subtlety in the derivation of the adjoint equation for this system arises from the fact that $\tilde{\boldsymbol{E}}$ is complex-valued.

Assuming a complex variable of the form $z = x + iy$, one can define the Wirtinger derivative operators [55]:

$$\frac{\partial}{\partial z} := \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right) \tag{2.17}$$

$$\frac{\partial}{\partial \bar{z}} := \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right) \quad , \tag{2.18}$$

which leads to the differential $\mathrm{d}f$ of a complex-valued function via the chain rule:

$$\mathrm{d}f = \frac{\partial f}{\partial z}\mathrm{d}z + \frac{\partial f}{\partial \bar{z}}\mathrm{d}\bar{z} \quad . \tag{2.19}$$

For a real-valued function with complex arguments, this is equivalent to

$$\mathrm{d}f = 2\,\mathrm{Re}\left\{\frac{\partial f}{\partial z}\mathrm{d}z\right\} \quad . \tag{2.20}$$

Applying this identity to Eq. (2.8), we get

$$\frac{\partial F}{\partial \boldsymbol{\rho}} = 2\,\mathrm{Re}\left\{ \boldsymbol{x}_{\mathrm{aj}}^{T} \left(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\rho}} - \frac{\partial \mathrm{A}}{\partial \boldsymbol{\rho}} \boldsymbol{x} \right) \right\} \quad . \tag{2.21}$$

We can now substitute $\boldsymbol{x} \longrightarrow \tilde{\boldsymbol{E}}$ and further assume the source term to be independent of the material distribution, *i.e.* $\partial \boldsymbol{b} / \partial \boldsymbol{\rho} = 0$, and we arrive at the “adjoint rule” for Maxwell’s wave equation in its most common form:

$$\frac{\partial F}{\partial \boldsymbol{\rho}} = -2\,\mathrm{Re}\left\{ \tilde{\boldsymbol{E}}_{\mathrm{aj}}^{T} \frac{\partial \mathrm{A}}{\partial \boldsymbol{\rho}} \tilde{\boldsymbol{E}} \right\} \quad . \tag{2.22}$$

The derivative $\partial \mathrm{A} / \partial \boldsymbol{\rho}$ is, in the simplest case where $\boldsymbol{\rho} = \epsilon_r$, apparent from Eq. (2.16), with $\partial \mathrm{A} / \partial \boldsymbol{\rho} = -\omega^2 \mu_u \epsilon_0$. But, as we will see in the following section, much of the challenge lies in finding good mappings from $\boldsymbol{\rho} \longrightarrow \epsilon_r$.

2.2 Topology optimization

Inverse design approaches that exploit the adjoint method can generally be divided into two main categories: *shape* and *topology* optimization. Both derive their gradients from Eq. (2.22); however, they differ in what the parameters $\boldsymbol{\rho}$ represent. In the case of shape optimization, $\boldsymbol{\rho}$ represents the boundary of a level set function that evolves according to the Hamilton-Jacobi equation [56]:

$$\frac{\partial \varphi}{\partial t} = v |\nabla \varphi| \quad , \tag{2.23}$$

where v is the normal velocity of the boundary $\boldsymbol{\rho}$. While shape optimization seems like a natural choice for problems where the boundary of the structure under investigation is known, the implementation in the context of electromagnetics is somewhat tricky, primarily because the normal component of the electric field is discontinuous across the boundary $\boldsymbol{\rho}$, so $\partial \mathrm{A} / \partial \boldsymbol{\rho}$ must be obtained through, *e.g.*, perturbation theory [57, 58]. While not an insurmountable challenge, this makes shape optimization somewhat unattractive for many practical purposes, especially when additional constraints are added to the optimization and when, at least to a good approximation, similar results can also be achieved through appropriately chosen parameterization in topology optimization.

In topology optimization, $\boldsymbol{\rho}$ represents the numerical value of the permittivity, either directly or through a series of transformations. This means that $\partial \mathrm{A} / \partial \boldsymbol{\rho}$ represents a continuous change in the material at each point in space or each pixel/voxel in a discrete setting. At each step in an optimization, this derivative is evaluated and used to decide whether the material at each pixel should have a higher or a lower permittivity at that point to maximize some objective. The designs are then updated accordingly, and the process is repeated until convergence is achieved. Topology optimization can be regarded as a “zero knowledge” approach to inverse design, in the sense that it requires no, and in fact often works best without, prior initialization of the device geometry. The optimization is generally free to create whatever device topology best maximizes the objective function, which is where the method derives its name from. This approach to inverse design has some particular implications:

- The number of elements in $\boldsymbol{\rho}$ is determined directly by the size of the design region and its spatial discretization.
- The device geometry is represented as an image (either in 2D or 3D), and its feature sizes are on the scale of single pixels.
- The parameters $\boldsymbol{\rho}$ are continuous, leading to a continuous material distribution in the optimized devices.
- The values of $\boldsymbol{\rho}$ are unbounded, which can lead to unphysical materials.

While the first point is merely an observation, it does imply that these optimization problems can become very large, on the order of millions of degrees of freedom. Arguably, this can lead to an over-parameterization of the design, as the "real" design space of such devices is generally much smaller, *i.e.*, not on the pixel scale. However, in practice, these large parameter spaces are typically not a problem, and in cases where such over-parameterization is undesired, it can be avoided through mappings that reduce the dimensionality of $\boldsymbol{\rho}$, as we will see later.

The latter points, however, present real challenges in topology optimization, and in the following section, we will outline the most common way of addressing them.

2.2.1 Three-field parameterization

Three-field topology optimization is an established parameterization scheme that maps a design density $\boldsymbol{\rho} \in [0, 1]$ to the physical material via a series of transformations [59] to introduce length-scale and binarization constraints.

Features on the scale of a single discretization step in the simulation can generally not be fabricated and must, thus, be avoided. This can be achieved by correlating neighboring pixels in the design through a convolution, canonically referred to as the "filtering" step [60], expressed as

$$\tilde{\boldsymbol{\rho}} = \boldsymbol{\rho} * \boldsymbol{\kappa} \quad , \tag{2.24}$$

with the design density $\boldsymbol{\rho}$, the convolution kernel $\boldsymbol{\kappa}$, and the resulting filtered density $\tilde{\boldsymbol{\rho}}$.

The effect of this filtering step is illustrated in Fig. 2.1, where we have chosen to represent the design density $\boldsymbol{\rho}$ as a randomly initialized 2D array with values in the range $[0, 1]$. Note that we assume periodic boundaries for the convolution here, which preserves the intensity at the edges, but this choice is problem-dependent and zero-padding is also often used. This step is essentially an image processing technique, and the computer vision literature often inspires the kernels used for filtering. Because the feature sizes are achieved by "smoothing" of $\boldsymbol{\rho}$, the most common kernels used for this purpose are radial kernels ("conic

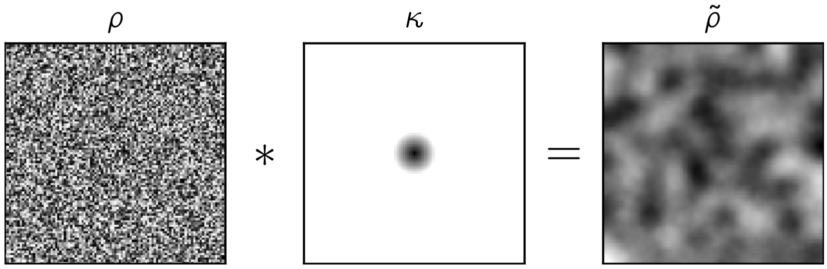

Figure 2.1: Illustration of the filtering step in the three-field topology optimization scheme. A design density $\boldsymbol{\rho}$ (left), represented as random noise, is convolved with a radial kernel $\boldsymbol{\kappa}$ (middle), which yields a filtered density $\tilde{\boldsymbol{\rho}}$ (right) with feature sizes approximately on the scale of the radius of the kernel. The boundaries are treated as periodic for the purposes of this illustration.

filters"), such as illustrated in Fig. 2.1, uniform kernels ("mean filters"), and Gaussian filters. Gaussian filters in particular can be advantageous in scenarios where the design domain, and hence $\boldsymbol{\rho}$, is very large, as they are *separable* filters, *i.e.*, they can be applied to each dimension individually in the form of a one-dimensional convolution. Although we focus here on the filtering step as a way to apply simple feature size constraints, it should be noted that such filtering schemes can be applied much more generally to yield designs with specific geometric features through careful design of these kernels, and they can be tailored to particular fabrication techniques [23, 24]. Further, although we focus on the 2D case for illustration, these filters can be applied in exactly the same way in 1D or 3D settings, both of which we will encounter in later chapters of this thesis.

After filtering, we are typically left with a washed-out "gray-scale" image of the design. This is generally undesirable for fabrication, as most devices are made of discrete materials. Therefore, designs with continuous material distributions should be avoided. To address this, the three-field approach introduces a so-called "projection" step, which thresholds the filtered density field.

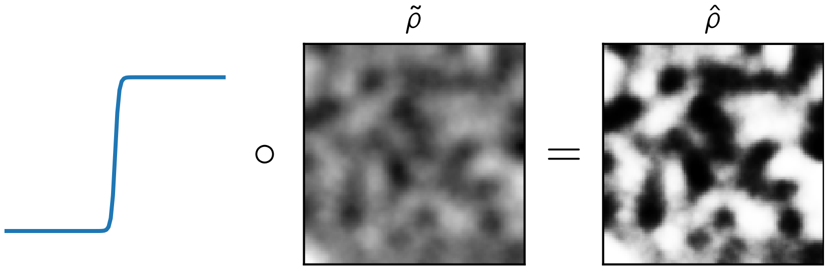

Figure 2.2: Illustration of the projection step. A soft-thresholding function in the form of a sigmoid (left) is applied to the filtered design density $\tilde{\boldsymbol{\rho}}$ (middle) to yield the "projected" density field $\hat{\boldsymbol{\rho}}$ (right). The projected field shows a much stronger contrast but still contains some non-binary material.

Because thresholding is not a differentiable operation, a similar behavior is typically achieved through a "soft-thresholding" function such as

$$\hat{\boldsymbol{\rho}} = \frac{\tanh(\alpha\beta) + \tanh\big(\alpha\left(\tilde{\boldsymbol{\rho}} - \beta\right)\big)}{\tanh(\alpha\beta) + \tanh(\alpha\,(1 - \beta))} \quad , \tag{2.25}$$

which represents a normalized sigmoid function with outputs between 0 and 1 for inputs in the same range [61]. The parameter α controls the steepness of this function and β its center, and it is generally assumed that $\beta = 0.5$. Figure 2.2 illustrates this projection scheme. We can see that the projected density $\hat{\boldsymbol{\rho}}$ exhibits much higher contrast than the filtered field $\tilde{\boldsymbol{\rho}}$ while maintaining the same overall features. However, it is apparent that $\hat{\boldsymbol{\rho}}$ is not fully binary, owing to the smooth nature of the projection function. A common strategy to remedy this is to increase the value of α during the optimization, continuously or at specific steps so that the final optimized structure is binary. Note that while the projection defined in Eq. (2.25) is the most common one found in the literature for topology optimization in nanophotonics, it is certainly not the only possible choice. In principle, any function that maps a linear input to an output with stronger contrast will do the trick, be it a sigmoid, a piecewise function, or something entirely different. A discussion of different projection functions and their tradeoffs can be found in Hooten *et al.* [62]. The only requirement is that these functions are differentiable to calculate $\partial \mathrm{A} / \partial \boldsymbol{\rho}$. However, if one is willing to sacrifice some accuracy in the gradients, it is possible to use hard thresholding by approximating the gradient with that of a smooth projection function, as shown in Schubert *et al.* [24].

After projection, $\hat{\boldsymbol{\rho}}$ represents the geometry of the design to be simulated, although the values are still within the range $[0, 1]$, so they are mapped to the actual physical material values, in our case generally the permittivity, via linear interpolation:

$$\boldsymbol{\epsilon}_r = \epsilon_{\mathrm{min}} + \hat{\boldsymbol{\rho}} \left(\epsilon_{\mathrm{max}} - \epsilon_{\mathrm{min}} \right) \quad , \tag{2.26}$$

where ϵ_{min} and ϵ_{max} represent the minimum and maximum permittivity values, respectively.

With this, we have fully outlined the three-field parameterization most commonly used in topology optimization in nanophotonics. It is important to stress that the filter-and-project scheme should be considered a template for parameterizing topology optimization problems. However, many variations exist and proper parameterization should always be considered in tandem with the problem at hand.

3 Cavity-enhanced Bloch surface waves coupled to integrated waveguides

For our first application of topology optimization to a nanophotonic inverse design problem, we will explore the efficient extraction of light from a dipole emitter coupled to a waveguide attached to the design region. This contribution's unique aspect is that we consider Bloch surface waves (BSWs) to be the platform used to steer the light in the integrated system. BSWs are supported at the interface between a one-dimensional photonic crystal (1DPC) and some isotropic ambient material. On top of the 1DPC, a defect layer is structured, which locally controls the refractive index the surface wave perceives. The physical properties of this system allow us to model it accurately using 2D simulations. After a short introduction in Section 3.1, Section 3.2 will outline the theory behind propagating surface waves, particularly BSWs, on which we will focus here. We then describe the problem setting and optimization setup in Section 3.3 and demonstrate the results of our optimization in Section 3.4, followed by concluding statements in Section 3.5.

The work presented in this chapter is primarily based on [P2] and can be regarded as a continuation of the work done during my master's thesis [63], which led to the publication of an article on subwavelength focusing of BSWs [8]. Because these works share the same underlying theory, much of Section 3.2 was adapted from [63].

3.1 Introduction

Bloch surface waves are unique solutions to Maxwell's equations that are evanescently confined to and propagating along the surface of a truncated 1DPC [64, 65]. Notably, these waves reach their maximum intensity at the surface of the 1DPC and can possess either transverse electric (TE) or transverse magnetic (TM) polarization [66]. The structure of a 1DPC is generally formed by stacking layers of dielectric materials, which alternate in their permittivity. Specifically, BSWs appear within the frequency range of the photonic crystal's band gap. This guarantees that the field decays exponentially in the homogeneous medium (such as air or water) and exhibits an oscillating amplitude with an exponentially decaying envelope within the layered medium, enabling it to propagate over extended distances with relatively low loss.

While BSWs and other guided surface modes such as surface plasmon polaritons [67, 68] differ in their physical principles, they share the characteristic of being localized at the interface. The degree of field confinement in BSWs is less pronounced because the effective index of these guided surface modes can not exceed the material indices involved, which are generally small in real-world BSW-sustaining platforms. Nevertheless, BSWs present the remarkable advantage of achieving propagation distances from hundreds of micrometers to even millimeters, with the field predominantly extending into the surrounding medium [69, 70]. This extended propagation length is possible due to the negligible dissipation losses within the all-dielectric structure, where the propagation length is typically constrained by the photonic crystal's limited layer count rather than inherent material losses [71]. Although a finite number of layers may result in some leakage radiation through evanescent tunneling to the underlying substrate, this issue can be regarded as a technical challenge that can be mitigated with engineering solutions.

Guiding light across macroscopic lengths within integrated systems opens the door to numerous applications to address societal challenges [72, 73]. Past endeavors in this area have included the development of devices for on-chip data processing [74] and various sensing technologies [75, 76, 77]. These sensing devices, in particular, have been designed with the flexibility to attach directly to fiber ends [78, 79]. A key factor for incorporating BSWs into integrated circuits involves confining them within a second dimension; that is, across the plane perpendicular to their direction of travel and parallel to the substrate surface [80]. Typically, this confinement is achieved by structuring the last layer of the one-dimensional photonic crystal, often referred to as the functional layer. The design ensures that BSWs encounter slightly different effective indices at the operational frequency between the patterned and non-patterned sections of the functional layer [81]. Such structuring enables the creation of waveguides and resonators with lateral confinement [82, 83] and facilitates the construction of more complex devices, taking advantage of the minor contrast in indices [84].

However, the relatively small difference in refractive index, often merely around $\Delta n \approx 0.1$, represents a significant challenge in the design process [85, 86]. This minor index contrast implies that the straightforward application of traditional optical elements, such as lenses, into these systems is not as simple as one might expect. This complexity is due to the fact that the effectiveness of these conventional components typically depends on a much more substantial contrast in refractive index. Therefore, designing functional components within an integrated BSW framework emerges as an intriguing area for applying innovative computational techniques in inverse design [6]. Through these advanced design methodologies, viable configurations that achieve specific operational goals can be discovered while navigating the limitations brought about by slight differences in refractive index.

We utilize topology optimization to develop an integrated photonic architecture based on BSWs. Our investigation focuses on facilitating the extraction of light from an embedded source into an integrated circuit on a BSW platform. Here, we consider only a waveguide instead of a more complex circuit, as this can be "attached" to any suitable application downstream. The source, modeled as an electric dipolar emitter, is designed to be externally stimulated, with the aim of maximizing energy transfer into the waveguide. The specific structure under analysis is a defined area around the emitter, which can be patterned to guide as much light as possible into the waveguide efficiently.

Two crucial considerations complicate straightforward solutions and necessitate sophisticated methodologies for tackling this inverse design challenge. Firstly, there is a need to optimize the coupling efficiency, which represents the proportion of emitted light successfully channeled into the waveguide. Secondly, the structure that supports this process can significantly enhance the interaction between the dipole emitter and the waveguide through Purcell enhancement. This phenomenon reflects an increase in the local density of optical states, facilitating more efficient light extraction from the emitters. Purcell enhancement is quantified by comparing the power extracted from the emitter near the supporting photonic structure to that extracted in free space or against another baseline structure. Achieving a high Purcell enhancement [87] while achieving fast light extraction means the emitter can be re-excited more quickly and emit additional photons. The research challenge we aim to address revolves around defining the optimal characteristics of a supporting photonic structure that maximizes light transfer from a quantum emitter to a waveguide by coupling with the BSWs, especially under conditions of dipole excitation in the saturation regime. We focus particularly on examining the impact of index contrast within the BSW platform, aiming to illuminate the complex relationship between coupling efficiency and Purcell enhancement. To achieve this, topology optimization is employed in the inverse design process of the supporting photonic structure, providing a systematic approach to discovering configurations that effectively bridge these two critical aspects.

3.2 Wave propagation in stratified media

Before discussing device optimization in the subsequent section, it is crucial to lay down a theoretical foundation for understanding the propagation of electromagnetic waves in periodic media and examine surface states. This process entails two primary steps: Initially, we must scrutinize the eigenmodes present within the periodic medium, paying close attention to their dispersion relation and field distributions. Following this, an analysis of the dispersion relation associated with the surface states is required. The concepts discussed in this section rely on the studies conducted by Yeh, Yariv, and Hong [65] and Yeh [88].

Our initial examination centers on a periodically stratified dielectric medium that extends infinitely, structured in layers from two materials differing in refractive indices. The refractive index n of an isotropic and homogeneous material is given by $\sqrt{\mu\epsilon/\mu_0\epsilon_0}$. Given that most transparent materials are non-magnetic, with $\mu = \mu_0$, this expression simplifies to $n = \sqrt{\epsilon/\epsilon_0}$. It is critical to note that ϵ, and consequently n, are functions of frequency. This relationship between n and frequency introduces *dispersion*, indicating that the phase velocity $v = c/n$ of a light wave varies with frequency.

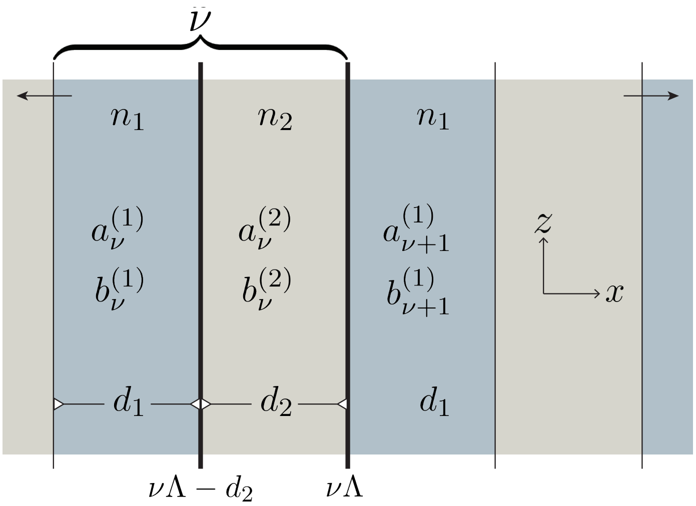

Figure 3.1: A segment of an infinitely extended periodic stratified medium composed of alternating layers of two distinct dielectric materials, characterized by refractive indices n_1 and n_2, and thicknesses d_1 and d_2, respectively. Bold lines indicate the interfaces included in the analysis, with the overbrace marking the ν-th unit cell.

The study here focuses on a layered structure consisting of two materials, each with specific thicknesses d_1 and d_2, and respective refractive indices n_1 and n_2, as illustrated in Fig. 3.1. The profile of the refractive index for this arrangement is defined by:

$$n(x) = \begin{cases} n_1 & 0 < x \leq d_1 \\ n_2 & d_1 < x \leq \Lambda \end{cases} \quad \text{with} \quad n(x + \nu\Lambda) = n(x) \quad . \tag{3.1}$$

Here, $\Lambda = d_1 + d_2$ denotes the periodicity of the layered structure, and x is the axis perpendicular to the interfaces of the layers. The analysis initially focuses on the transverse electric (TE) polarization, where the field is characterized by the components E_y, H_x, and H_z.

All components mentioned follow the wave equation, with our detailed examination directed towards E_y:

$$\frac{\partial^2 E_y}{\partial z^2} + \frac{\partial^2 E_y}{\partial x^2} + \frac{\omega^2}{c^2} n^2(x) E_y = 0. \tag{3.2}$$

To identify the eigenmodes and their dispersion relations, we utilize a plane wave ansatz in the frequency domain along the z-direction, reflecting the uniformity of the material in this direction:

$$E_y(x, z) = E(x) e^{i\beta z}. \tag{3.3}$$

Within each layer, the x-dependent solutions $E(x)$ comprise both forward and backward propagating plane waves:

$$E(x) = a_\nu^{(\alpha)} e^{ik_x^{(\alpha)}(x - \nu\Lambda)} + b_\nu^{(\alpha)} e^{-ik_x^{(\alpha)}(x - \nu\Lambda)} \quad \text{with} \quad k_x^{(\alpha)} = \sqrt{\left(\frac{n_\alpha \omega}{c}\right)^2 - \beta^2}, \tag{3.4}$$

where $a_\nu^{(\alpha)}$ and $b_\nu^{(\alpha)}$ represent the amplitudes of the forward and backward propagating waves in the α layer ($\alpha \in 1, 2$) of the nth unit cell.

For an electric field vector perpendicular to the considered plane of propagation (TE mode), E and $\partial E/\partial x$ have to be continuous across the material interface:

$$a_\nu^{(1)} e^{ik_x^{(1)}(x-\nu\Lambda)} + b_\nu^{(1)} e^{-ik_x^{(1)}(x-\nu\Lambda)} = a_\nu^{(2)} e^{ik_x^{(2)}(x-\nu\Lambda)} + b_\nu^{(2)} e^{-ik_x^{(2)}(x-\nu\Lambda)} \tag{3.5a}$$

$$ik_x^{(1)}\left(a_\nu^{(1)} e^{ik_x^{(1)}(x-\nu\Lambda)} + b_\nu^{(1)} e^{-ik_x^{(1)}(x-\nu\Lambda)}\right) = ik_x^{(2)}\left(a_\nu^{(2)} e^{ik_x^{(2)}(x-\nu\Lambda)} + b_\nu^{(2)} e^{-ik_x^{(2)}(x-\nu\Lambda)}\right) \quad . \tag{3.5b}$$

Considering these interface conditions at two adjacent layers ν and $\nu + 1$ (shown in bold in Fig. 3.1) with $x_\nu = \nu\Lambda - d_2$ and $x_{\nu+1} = \nu\Lambda$, we obtain the following continuity relations:

$$\begin{pmatrix} e^{-ik_x^{(1)}d_2} & e^{ik_x^{(1)}d_2} \\ e^{-ik_x^{(1)}d_2} & -e^{ik_x^{(1)}d_2} \end{pmatrix}\begin{pmatrix} a_\nu^{(1)} \\ b_\nu^{(1)} \end{pmatrix} = \begin{pmatrix} e^{-ik_x^{(2)}d_2} & e^{ik_x^{(2)}d_2} \\ \frac{k_x^{(2)}}{k_x^{(1)}} e^{-ik_x^{(2)}d_2} & -\frac{k_x^{(2)}}{k_x^{(1)}} e^{ik_x^{(2)}d_2} \end{pmatrix}\begin{pmatrix} a_\nu^{(2)} \\ b_\nu^{(2)} \end{pmatrix} \tag{3.6a}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} a_\nu^{(2)} \\ b_\nu^{(2)} \end{pmatrix} = \begin{pmatrix} e^{-ik_x^{(1)}\Lambda} & e^{ik_x^{(1)}\Lambda} \\ \frac{k_x^{(1)}}{k_x^{(2)}} e^{-ik_x^{(1)}\Lambda} & -\frac{k_x^{(1)}}{k_x^{(2)}} e^{ik_x^{(1)}\Lambda} \end{pmatrix}\begin{pmatrix} a_{\nu+1}^{(1)} \\ b_{\nu+1}^{(1)} \end{pmatrix} \quad . \tag{3.6b}$$

From these relations, we can obtain an expression relating $a_\nu^{(1)}$ and $b_\nu^{(1)}$ to $a_{\nu+1}^{(1)}$ and $a_{\nu+1}^{(1)}$:

$$\begin{pmatrix} a_\nu^{(1)} \\ b_\nu^{(1)} \end{pmatrix} = \underbrace{\begin{pmatrix} A_{\mathrm{TE}} & B_{\mathrm{TE}} \\ C_{\mathrm{TE}} & D_{\mathrm{TE}} \end{pmatrix}}_{T}\begin{pmatrix} a_{\nu+1}^{(1)} \\ b_{\nu+1}^{(1)} \end{pmatrix} \quad , \tag{3.7a}$$

with the matrix elements given by

$$A_{\mathrm{TE}} = e^{-ik_x^{(1)}d_1}\left(\cos\left(k_x^{(2)}d_2\right) - \frac{i}{2}\left(\frac{k_x^{(1)}}{k_x^{(2)}} + \frac{k_x^{(2)}}{k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.8a}$$

$$B_{\mathrm{TE}} = e^{ik_x^{(1)}d_1}\left(\frac{i}{2}\left(\frac{k_x^{(1)}}{k_x^{(2)}} - \frac{k_x^{(2)}}{k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.8b}$$

$$C_{\mathrm{TE}} = e^{-ik_x^{(1)}d_1}\left(-\frac{i}{2}\left(\frac{k_x^{(1)}}{k_x^{(2)}} - \frac{k_x^{(2)}}{k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.8c}$$

$$D_{\mathrm{TE}} = e^{ik_x^{(1)}d_1}\left(\cos\left(k_x^{(2)}d_2\right) + \frac{i}{2}\left(\frac{k_x^{(1)}}{k_x^{(2)}} + \frac{k_x^{(2)}}{k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \quad . \tag{3.8d}$$

The transfer matrix T, as defined in Eq. (3.8), serves to correlate the amplitudes of forward and backward propagating plane waves, specifically $a_\nu^{(1)}$ and $b_\nu^{(1)}$, within a single layer of the unit cell, to those, $a_{\nu+1}^{(1)}$ and $b_{\nu+1}^{(1)}$, in the corresponding layer of the subsequent unit cell.

For transverse magnetic (TM) polarized waves, the continuity conditions differ slightly and yield

$$A_{\mathrm{TM}} = e^{-ik_x^{(1)}d_1}\left(\cos\left(k_x^{(2)}d_2\right) - \frac{i}{2}\left(\frac{n_2^2k_x^{(1)}}{n_1^2k_x^{(2)}} + \frac{n_1^2k_x^{(2)}}{n_2^2k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.9a}$$

$$B_{\mathrm{TM}} = e^{ik_x^{(1)}d_1}\left(\frac{i}{2}\left(\frac{n_2^2k_x^{(1)}}{n_1^2k_x^{(2)}} - \frac{n_1^2k_x^{(2)}}{n_2^2k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.9b}$$

$$C_{\mathrm{TM}} = e^{-ik_x^{(1)}d_1}\left(-\frac{i}{2}\left(n_2^2\frac{k_x^{(1)}}{n_1^2k_x^{(2)}} - \frac{n_1^2k_x^{(2)}}{n_2^2k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.9c}$$

$$D_{\mathrm{TM}} = e^{ik_x^{(1)}d_1}\left(\cos\left(k_x^{(2)}d_2\right) + \frac{i}{2}\left(\frac{n_2^2k_x^{(1)}}{n_1^2k_x^{(2)}} + \frac{n_1^2k_x^{(2)}}{n_2^2k_x^{(1)}}\right)\sin\left(k_x^{(2)}d_2\right)\right) \tag{3.9d}$$

for the matrix elements. Here, we have focused on an infinitely periodic multilayer structure characterized by only two distinct refractive indices and layer thicknesses. The general transfer matrix, which links the electric field amplitudes of the mth layer in two adjacent cells of a multilayer slab comprising M layers per unit cell, is detailed in Yeh, Yariv, and Hong [65]:

$$T^{(M)} = \frac{1}{2^M}\prod_{\alpha=m+1}^{M+m}\begin{pmatrix}\left(1+C^{(\alpha)}\right)e^{-ik_x^{(\alpha)}t^{(\alpha)}} & \left(1-C^{(\alpha)}\right)e^{k_x^{(\alpha)}t^{(\alpha)}} \\ \left(1-C^{(\alpha)}\right)e^{-ik_x^{(\alpha)}t^{(\alpha)}} & \left(1+C^{(\alpha)}\right)e^{k_x^{(\alpha)}t^{(\alpha)}}\end{pmatrix} \quad , \tag{3.10}$$

where $t_m = x_m - x_{m-1}$ is the thickness of the mth layer and

$$C^{(\alpha)} = \begin{cases} k_x^{(\alpha)}/k_x^{(\alpha-1)} & \text{TE waves} \\ n_{(\alpha)}^2k_x^{(\alpha-1)}/n_{(\alpha-1)}^2k_x^{(\alpha)} & \text{TM waves}\end{cases} \quad . \tag{3.11}$$

Given the wave's propagation within a periodic medium, the Floquet-Bloch theorem establishes that the wave's amplitude will mimic the lattice's periodicity, with solutions characterized by

$$E_K(x,z) = E_K(x)e^{iKx}e^{i\beta z} \quad \text{with} \quad E_K(x+\nu\Lambda) = E_K(x) \quad . \tag{3.12}$$

where K represents the Bloch wave number, a crucial wave vector component. To ascertain the values for K and $E_K(x)$, we apply the periodicity condition to the Bloch wave:

$$\begin{pmatrix}a_\nu \\ b_\nu\end{pmatrix} = e^{iK\Lambda}\begin{pmatrix}a_{\nu+1} \\ b_{\nu+1}\end{pmatrix} \quad , \tag{3.13}$$

which, in conjunction with Eq. (3.7a), formulates an eigenvalue problem:

$$\begin{pmatrix}A & B \\ C & D\end{pmatrix}\begin{pmatrix}a_\nu \\ b_\nu\end{pmatrix} = e^{-iK\Lambda}\begin{pmatrix}a_\nu \\ b_\nu\end{pmatrix} \tag{3.14}$$

leading to the condition for eigenvalues:

$$\begin{vmatrix} A - e^{-iK\Lambda} & B \\ C & D - e^{-iK\Lambda} \end{vmatrix} = 0 \quad . \tag{3.15}$$

The eigenvalues are thus determined by

$$0 = AD - e^{-iK\Lambda}A - e^{-2iK\Lambda}D - BC \tag{3.16a}$$

$$= e^{-iK\Lambda}A - e^{-2iK\Lambda}D + 1 \tag{3.16b}$$

$$\Rightarrow e^{-iK\Lambda} = \frac{1}{2}(A + D) \pm \sqrt{\frac{1}{4}(A + D)^2 - 1} \quad , \tag{3.16c}$$

assuming the transfer matrix T is unimodular (*i.e.*, $AD - BC = 1$) when considering nonabsorbing media. The eigenvectors related to these eigenvalues are given by

$$\begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} B \\ e^{-iK\Lambda} - A \end{pmatrix} \quad . \tag{3.17}$$

The expression in Eq. (3.16c) establishes the dispersion relation linking K, ω, and β:

$$K(\beta, \omega) = \frac{1}{\Lambda} \cos^{-1}\left(\frac{1}{2}(A + D)\right) \quad . \tag{3.18}$$

When $\frac{1}{2}|A+D| < 1$, the values of K are real, indicating that the Bloch waves are propagating. Conversely, if $\frac{1}{2}|A+D| > 1$, the solutions for K become complex, introducing an imaginary component K_i to the wavenumber $K = \frac{m\pi}{\Lambda} + iK_i$, thereby rendering the wave evanescent. This delineation between propagating and evanescent waves demarcates the forbidden bands within the periodic medium, with the band edges specified by $\frac{1}{2}|A + D| = 1$.

The comprehensive solution for a Bloch wave within the first layer of the nth unit cell, representing the distribution of the eigenmodes' fields, is described by:

$$E_K(x)e^{iKx} = \left(\left(a_0 e^{ik_x^{(1)}(x-\nu\Lambda)} + b_0 e^{-ik_x^{(1)}(x-\nu\Lambda)}\right) e^{-iK(x-\nu\Lambda)}\right) e^{iKx} \quad , \tag{3.19}$$

with the coefficients a_0 and b_0 as defined in (3.17). With the analysis of Bloch waves concluded the discussion will now shift toward the propagation of Bloch surface waves.

Surface waves are confined to the interface between two semi-infinite systems. In this context, we will explore BSWs, which are electromagnetic surface modes that exist at the interface between a semi-infinite periodic multilayer and a dielectric medium, such as air. These waves are capable of propagating along these interfaces. The refractive index profile of this system is characterized as follows:

$$n(x, z) = \begin{cases} n_a & x \leq 0 \\ n_1 & m\Lambda \leq x < m\Lambda + d_1 \\ n_2 & m\Lambda + d_1 \leq x < (m+1)\,\Lambda \end{cases} \qquad \text{with} \quad m \in \mathbb{N} \quad . \tag{3.20}$$

This analysis will focus on TE polarization, specifically looking for waves that propagate in the z-direction along the interface. Employing the same ansatz, $E_y(x, y, z) = E(x)e^{i\beta z}$, the wave equation simplifies to:

$$\frac{\partial^2 E(x)}{\partial x^2} + \left(\frac{\omega^2}{c^2} n^2(x) - \beta^2\right) E(x) = 0 \quad , \tag{3.21}$$

considering solutions of the form:

$$E(x) = \begin{cases} \alpha e^{q_a x} & x \leq 0 \\ E_K(x) e^{iKx} & x \geq 0 \end{cases} \quad \text{with} \quad q_a = \sqrt{\beta^2 - \frac{\omega^2}{c^2} n_a^2} \quad \text{and} \quad \alpha = \text{const.} \quad , \tag{3.22}$$

where we incorporate the previously derived general Bloch wave solutions in the periodic medium and introduce a new solution for the dielectric medium. To ensure a mode is localized at the interface, it must exhibit a decay within the dielectric medium and the periodic multilayer. This requirement implies that the mode must be evanescent on both sides of the interface, positioning it within the photonic crystal's forbidden band gap while also lying below the dielectric medium's light line. This gives rise to the condition for the surface modes:

$$q_a = -ik_x^{(1)} \frac{e^{-iK\Lambda} - A - B}{e^{-iK\Lambda} - A + B} \quad . \tag{3.23}$$

The electromagnetic field experiences an exponential decay outside the dielectric multilayer, whereas within the photonic crystal, it diminishes following a $(-1)^m e^{-iK_i\Lambda}$ pattern for each repeating unit, where m represents the integer denoting the mth forbidden gap.

Incorporating a slim terminal layer, referred to as the functional layer, characterized by an appropriate thickness and dielectric constant atop the final layer of the configuration enables deliberate alteration of the BSW's local dispersion relation. The BSW's effective index, denoted as $n_{\mathrm{BSW}} = \frac{c\beta}{\omega}$, is obtained through the propagation constant of the surface mode, both in the presence and absence of the functional layer. An extensive empirical study exploring the effect of the functional layer on the BSW's positioning within the band gap is detailed in Yu *et al.* [73].

The photonic band gap of the crystal contains the guided surface modes, situated beneath the air light line. The origin of their separation lies in the modestly altered index profile induced by the functional layer atop the multilayer arrangement. For any specific ω, the refractive index of index-guided modes surpasses that of the light line. The discrepancy in the effective refractive indices of the dual BSW modes elucidates the effective refractive index contrast intrinsic to the BSW framework, which fundamentally facilitates the surface guidance of propagating modes along the multilayer stack. Modifying the spatial configuration of the functional layer affords control over the BSW's propagation, laying the groundwork for the optimization strategies discussed in the following.

3.3 Optimization setup

We aim to improve functional photonic devices that can extract light emitted by a dipole-like source that couples into a surface mode sustained by a dielectric multi-layer stack into a waveguide, which could be used for on-chip light sources. The proposed setup is shown in Fig. 3.2, with a change in coordinates from $x \rightarrow z$ from the discussion in Section 3.2. The system under consideration is optically large, making optimizing using full-wave 3D simulations impractical. Instead, we use an effective index method to transform the problem into a 2D one. This method has been successfully applied to design functional devices for manipulating BSWs because of the low refractive index contrast of such systems [89, 8, 90].

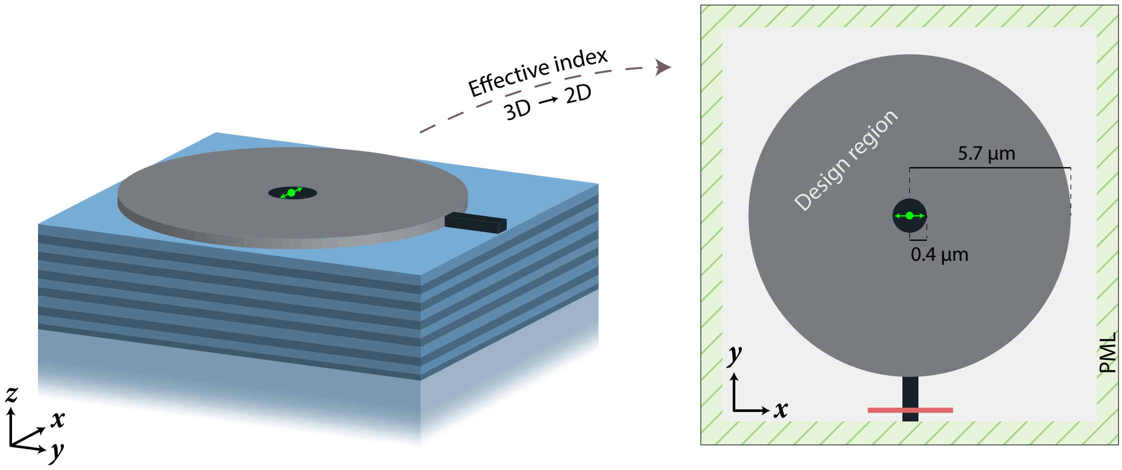

Figure 3.2: A depiction of a standard material setup conducive to Bloch Surface Waves (BSWs) is presented on the left. In this schematic, a functional layer is patterned atop a dielectric multi-layer structure, which is itself placed on a base substrate. The framework for optimizing the 2D effective index simulation is shown on the right. It features a design area shaped as a circle with a 5.7 µm radius, encompassing a central spacer zone with a 0.4 µm radius and a predetermined material configuration. Positioned at the center of this spacer zone is an x-polarized dipole source. A waveguide protrudes from the design area's bottom edge, measuring 0.3 µm in width. Reprinted from [P2], licensed under the Creative Commons license CC BY 4.0.

In the outlined configuration, we position a dipole source to emit along the x-axis, with its emission characterized by a Gaussian spectrum peaking at a wavelength of $\lambda_0 = 570\,\text{nm}$. This source is located at the heart of the spacer area, denoted by a black circle. The design zone, circular in shape as depicted in Fig. 3.2, has a radius of 5.7 µm. The objective function F is expressed as

$$F = |\alpha_0|^2 \quad , \tag{3.24}$$

with α_0 being the mode coefficient for the primary TM-polarized mode (signifying the in-plane electric field) moving forward in the waveguide. This coefficient is normalized so that $|\alpha_0|^2$ corresponds to the total power.

Our approach uses the filter-and-project parameterization technique [91, 61, 92] alongside the method of moving asymptotes (MMA) [93] and integrated within the nlopt software [14] (version 2.7.1) for executing constrained, nonlinear topology optimization. The design process utilizes a 2D grid for discretization with a resolution of100 $\frac{\text{px}}{\mu\text{m}}$. For design parameterization, we employ the three-field scheme described in Subsection 2.2.1, with a Gaussian filtering kernel and the standard projection function from Eq. (2.25). The parameter α controls the binarization intensity, incrementally raised throughout the optimization procedure to achieve ultimately binary designs. To prevent reverting to less binary designs when the optimization is restarted with a higher α, we compute the "gray indicator" by summing over n pixels:

$$c_{\mathrm{g}}(\hat{\boldsymbol{\rho}}) = \frac{1}{n} \sum_{i=1}^{n} 4\hat{\rho}_i \left(1 - \hat{\rho}_i\right) \quad . \tag{3.25}$$

This indicator function is a parabola with a maximum of 1 at $\rho_i = 0.5$ and 0 at $\rho_i = 0$ and $\rho_i = 1$. It approaches zero the more the values of $\hat{\boldsymbol{\rho}}$ are shifted towards 0 and 1, *i.e.*, it measures how "gray" the design is. By evaluating this function directly after increasing α and constraining its value to be smaller than a threshold γ (given by the computed value) in the subsequent optimization.

Following the projection step, the design variables are subjected to a circular mask to ensure the device maintains a circular shape. The geometric configurations of the devices are subsequently mapped linearly to their specific refractive indices for simulation purposes, conducted using Meep [3]. The objective function's sensitivities with respect to the parameterization are derived through Meep's adjoint module [94], and the sensitivities of the parameterization with respect to the optimization variables ρ are calculated using [95].

The full optimization problem can thus be stated as:

$$\max_{\boldsymbol{\rho}} \quad |\alpha_0|^2 \tag{3.26a}$$

$$\text{s.t.} \quad \alpha_0 = \int_S \left[\tilde{\mathbf{E}}^*(\boldsymbol{r}) \times \tilde{\mathbf{H}}_0(\boldsymbol{r}) + \tilde{\mathbf{E}}_0(\boldsymbol{r}) \times \tilde{\mathbf{H}}^*(\boldsymbol{r})\right] \cdot \hat{\mathbf{n}} \; \mathrm{d}A \tag{3.26b}$$

$$c_{\mathrm{g}}(\hat{\boldsymbol{\rho}}) - \gamma \leq 0 \tag{3.26c}$$

$$0 \leq \boldsymbol{\rho} \leq 1 \quad , \tag{3.26d}$$

with the Fourier-transformed fields $\tilde{\mathbf{E}}$, $\tilde{\mathbf{H}}$ in Eq. (3.26b) at the target wavelength of 570 nm at the mode monitor (c.f. Fig. 3.2) and $\tilde{\mathbf{E}}_0$, $\tilde{\mathbf{H}}_0$, which are the fields corresponding to the fundamental mode in the waveguide (obtained using the eigenmode solver MPB [96]).

The base effective index for the surface mode, before adding the functional layer, is set at $n_{\mathrm{eff},0} = 1.019$. The optimization process explores various effective index differences, Δn_{eff}, starting from 0.01 up to 0.13, with a step increase of 0.01. In the scenario with the maximum index contrast of $\Delta n_{\mathrm{eff}} = 0.13$, the surface mode, when integrated with the functional layer, achieves an effective index of $n_{\mathrm{eff},1} = 1.149$, aligning with the specifications of the material framework discussed in previous research [90], which comprises a stack sequence that sits on top of a glass coverslip with ten repetitions of alternating $\mathrm{Ta_2O_5}$ ($n_{\mathrm{Ta_2O_5}} = 2.08$, $d_{\mathrm{Ta_2O_5}} = 95\,\mathrm{nm}$) and $\mathrm{SiO_2}$ ($n_{\mathrm{SiO_2}} = 1.46$, $d_{\mathrm{SiO_2}} = 137\,\mathrm{nm}$) layers, a top layer of $\mathrm{SiO_2}$ with $d_{\mathrm{SiO_2,top}} = 127\,\mathrm{nm}$, and a final layer of PMMA ($n_{\mathrm{PMMA}} = 1.48$, $d_{\mathrm{PMMA}} = 75\,\mathrm{nm}$), which serves as the patterned functional layer.

The optimization's 2D simulations are conducted using the MPI-parallel implementation of Meep (version 1.23.0) on a desktop equipped with a 10-core Intel Core i9-10900 CPU operating at 2.80 GHz, with a simulation resolution of 50 $\frac{\text{px}}{\mu\text{m}}$.

3.4 Results and discussion

This section will discuss the results obtained using the previously described optimization setup. These results are further validated using 3D full-wave simulations in Subsection 3.4.2.

3.4.1 Optimization and evaluation in 2D

By conducting a series of optimization processes across a spectrum of effective refractive index contrasts, specifically within the range of $\Delta n_{\text{eff}} = 0.01$ to 0.13, we engage in a detailed numerical exploration into how variations in the refractive index contrast of different potential material platforms for sustaining surface waves can impact the performance capabilities of the devices under consideration.

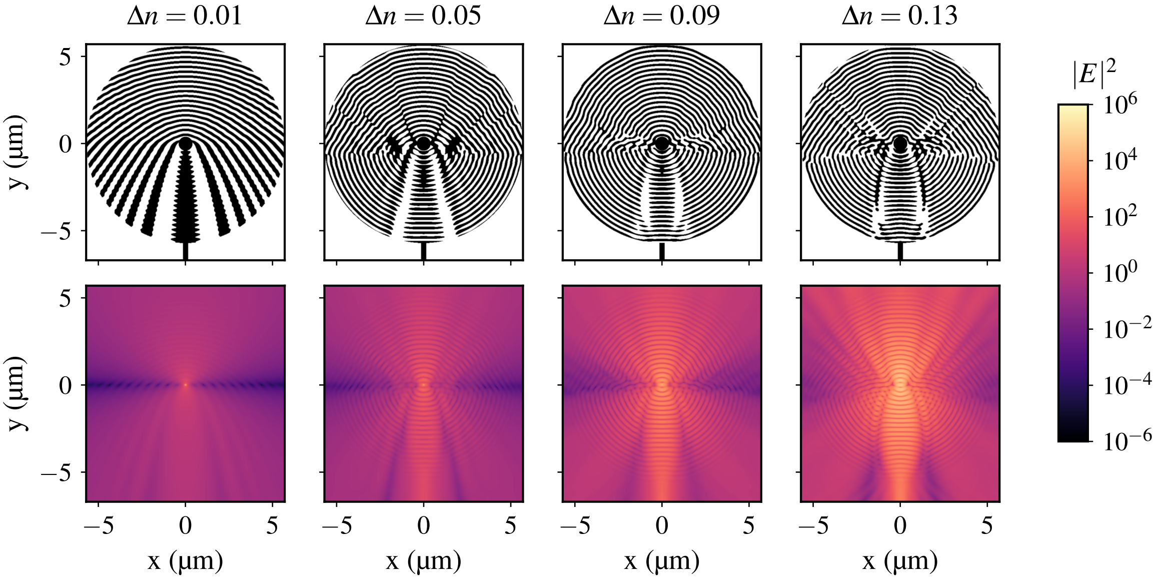

Figure 3.3: Optimized device designs (top row) and corresponding electric field intensities (bottom row) in log scale at the target wavelength of 570 nm. Four selected devices are designed for different effective index contrasts (columns). Note that a short length of waveguide is included for clarity but it is not part of the design region. Reprinted from [P2], licensed under the Creative Commons license CC BY 4.0.

The optimization results for four selected devices and the corresponding electric field intensities (log scale) are shown in Fig. 3.3.

Across all the device designs evaluated, a recurring outcome of the optimization is the emergence of a structure resembling a distributed Bragg reflector (DBR), predominantly manifested in the design's upper half-space. This outcome can be considered quite expected and logical, given that the amount of power funneled into the waveguide is intrinsically linked to the power emitted by the source. From this perspective, it becomes clear why the optimization process gravitates towards configurations that, to some extent, create an electromagnetic cavity: it is a straightforward and effective means to enhance the coupling of emitted power into the waveguide.

However, the actual coupling of power from the source into the waveguide is influenced by more than the total emitted power; it also hinges on how adeptly the device can direct this emitted field into the waveguide. The refractive index constrains this guiding efficiency contrast the device can leverage to manipulate the electromagnetic fields. Notably, as the refractive index contrast increases, the device gains a greater capacity to exploit subtle spatial variations within the patterned layer for more effective light guidance. This phenomenon becomes increasingly apparent upon examining the variations among the devices, as shown in Fig. 3.3, where the influence of different refractive index contrasts on device performance is visually and quantitatively illustrated.

For the device with a minimal refractive index contrast of $\Delta n_{\mathrm{eff}} = 0.01$, the design transitions smoothly from a Distributed Bragg Reflector (DBR) configuration in the upper portion to a structure that resembles a tapered waveguide at the bottom. This gradual transition suggests a design intent to reflect light within the upper half while primarily focusing on enhancing light coupling efficiency into the waveguide by appropriately directing the light. The potential for emission enhancement due to cavity effects is minimally observed, which aligns with expectations given the low index contrast contributing to cavities with limited quality factors.

In contrast, the device designed for a higher refractive index contrast of $\Delta n_{\mathrm{eff}} = 0.13$ showcases a complex and less predictable configuration. Its central design is characterized by several sub-wavelength features, diverging from the simpler waveguide-like structure seen in devices with lower Δn_{eff}. The upper section still functions in a manner reminiscent of a DBR, yet the lower section – specifically the area linking the source to the waveguide – demonstrates a sophisticated balance. This design meets the dual objectives of achieving high coupling efficiency and significantly enhancing the power extracted.

To comprehensively assess the performance of these optimized devices, our evaluation focuses on two key metrics: the Purcell enhancement experienced by the source within the spacer region, and the efficiency of coupling into the waveguide. These measurements allow us to quantitatively understand the impact of the device's design on its ability to enhance light emission and guide it effectively into the waveguide, thereby providing a detailed insight into the practical efficacy of each device configuration in optimizing light manipulation.

For all optimized devices, spectra within the range of (570 ± 30) nm are acquired, as presented in Fig. 3.4, while the values at the specific target wavelength of 570 nm are detailed separately in Fig. 3.5. The computation of Purcell enhancement involves determining the power flow ratio from the source within the device (termed as "cavity") compared to that in a uniform medium, identified by $n_{\mathrm{eff},1}$. This measurement is obtained by encircling the

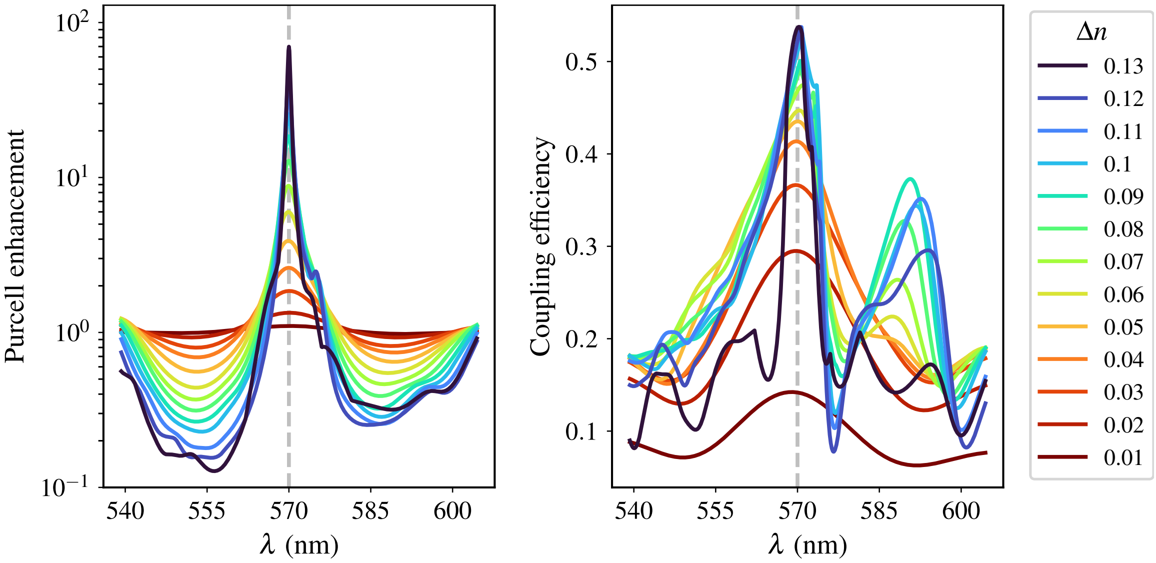

Figure 3.4: Depiction of both Purcell enhancement (on the left) and coupling efficiencies (on the right) for each design, derived from the 2D effective index method used for device optimization. Reprinted from [P2], licensed under the Creative Commons license CC BY 4.0.

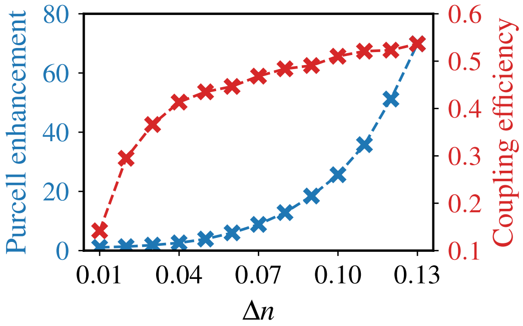

Figure 3.5: Data on Purcell enhancement and coupling efficiency at the specified target wavelength of 570 nm, based on 2D effective index simulations. Reprinted from [P2], licensed under the Creative Commons license CC BY 4.0.

source with a cubic flux monitor having a 0.5 µm edge, subsequently gathering the relevant Fourier-transformed fields. In a similar manner, the coupling efficiency is measured as the power ratio within the waveguide's fundamental mode, expressed as

$$P_{\text{mode}} = |\alpha_0|^2 \tag{3.27}$$

to the power radiated by the source.

The devices, once optimized, demonstrate a notable Purcell enhancement at the target wavelength. Specifically, the device with an index contrast of $\Delta n_{\text{eff}} = 0.01$ showcases a Purcell factor approximately equal to 2. This factor grows exponentially with an increase in Δn_{eff}, as highlighted in Fig. 3.5), peaking with the device at $\Delta n_{\text{eff}} = 0.13$, which exhibits a Purcell enhancement near 70. Additionally, it is observed that designs with a greater index contrast tend to have improved coupling efficiency, with the $\Delta n_{\text{eff}} = 0.13$ design achieving the highest at 54 %.

Designs featuring higher index contrasts also demonstrate a markedly greater specificity to the target wavelength than those with lower index contrasts, which display a broader response spectrum (cf. Fig. 3.4). This observation underscores the earlier discussion regarding the advantage of high index contrasts in optimization, which allows for more precise tuning of the device's "guiding performance" towards the desired wavelength. Conversely, at very low index contrasts, the benefits are relatively minimal.

Moreover, a substantial increase in coupling efficiency is noticed as Δn_{eff} grows from 0.01 to 0.04, after which the efficiency gradually approaches saturation, leveling around 0.5 for higher contrasts. This phenomenon can be attributed to the fact that devices optimized for extremely low ($\Delta n_{\text{eff}} \approx 0.01$) to moderate ($\Delta n_{\text{eff}} \approx 0.1$) index contrasts exhibit a significant performance disparity. Nevertheless, the index contrast across all optimized devices remains on the lower end, leading to a considerable loss of dipole-emitted radiation in the design's upper half-space. These findings reflect the theoretical discussions and observations regarding the variance in device performances across different index contrasts, further validating the empirical data gathered from optimizing these devices.

3.4.2 Comparison between 2D and 3D

We will now transition from assessing the performance of optimized devices within a two-dimensional framework, specifically under the guise of an effective index approximation, to scrutinizing the fidelity of these approximations against the envisioned three-dimensional scenarios. To undertake this examination, we integrate a dielectric multi-layer stack as outlined by [90]. This construct features a combination of Silica and Tantalia layers atop a glass substrate, distinguished by a refractive index of $n_{\text{glass}} = 1.5$, capped with a PMMA layer in which the device geometry is etched. The arrangement comprises ten layers alternating between Ta_2O_5 (with a refractive index of 2.08 and thickness of 95 nm) and SiO_2 (with a refractive index of 1.46 and thickness of 137 nm), followed by an additional Ta_2O_5 layer, a 127 nm thick SiO_2 layer, and concluding with a PMMA layer (refractive index 1.48, thickness 75 nm. Assuming a dispersion-free wavelength range of (570 ± 30) nm, the effective indices at the 570 nm wavelength without and with the PMMA layer are 1.019 and 1.149, respectively. These indices align with the design parameters for the device optimized for a $\Delta n_{\text{eff}} = 0.13$.

The 2D optimized design is extruded vertically by 75 nm to form the functional PMMA layer, incorporating an x-polarized dipole source within the PMMA, 35 nm below its surface. The 3D simulations employ finite-difference time-domain (FDTD) methodology via Meep, with a computational resolution of 80 μm^{-1}, leveraging the computational power of 76 CPU cores across two Intel Xeon Platinum 8368 processors, clocked at 2.4 GHz.

These simulations, spanning 16.5 hours, mimic the 2D analysis by tracking power flux from the dipole in every spatial direction using flux monitor planes, each with a 0.5 μm edge. These measures permit the computation of Purcell enhancement and coupling efficiency analogously to the 2D scenario, thus offering a direct comparative insight between the 2D effective index and 3D full-wave simulations.

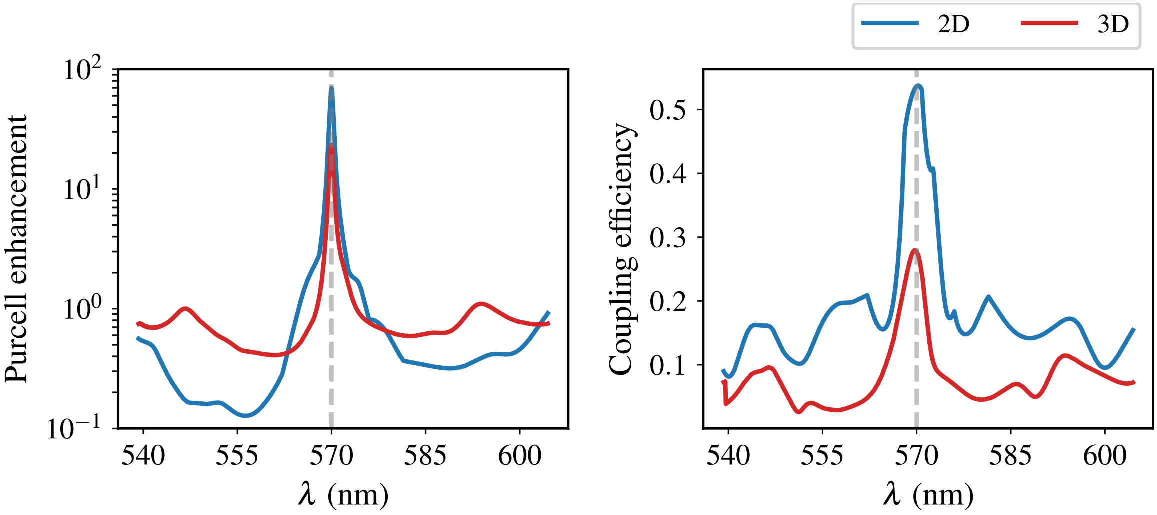

Figure 3.6: A juxtaposition of Purcell enhancement (left) and coupling efficiency (right) derived from 2D effective index simulations against those from a comprehensive 3D full-wave simulation. The optimization targeted an effective index contrast of $\Delta n_{\text{eff}} = 0.13$, in line with the material configuration detailed by Stella *et al.* [90]. Reprinted from [P2], licensed under the Creative Commons license CC BY 4.0.

The juxtaposition of outcomes from the comprehensive 3D full-wave simulation with those derived from the corresponding 2D effective index simulation, as depicted in Fig. 3.6, reveals notable discrepancies between the two methodologies. Despite these differences, the 3D simulation outcomes underscore a pronounced improvement in both the Purcell enhancement and coupling efficiency at the optimization's target wavelength. Specifically, whereas the 2D simulation estimated a Purcell factor of 70, the 3D simulation demonstrates a reduced Purcell enhancement factor of 23. This reduction aligns with expectations, given that 2D effective index simulations overlook out-of-plane scattering losses, which account for a significant portion of radiation emitted by the source. Such losses are neither captured by the patterned PMMA layer in 3D nor contemplated in the 2D simulations. Nonetheless, the Purcell factor achieved by the analytically devised Distributed Bragg Reflector (DBR) structure, as per Stella *et al.* [90], stands at 32. Our optimized device, designed to balance Purcell enhancement and coupling efficiency, naturally exhibits a diminished enhancement, reflective of the ongoing energy extraction from the system.

This disparity between 2D and 3D simulation results, albeit significant, does not undermine the utility of the effective index method; the optimized devices still demonstrate commendable performance in a 3D setting. The coupling efficiency observed mirrors this trend, with values of 0.54 in 2D simulations and 0.28 in 3D, attributable to substantial out-of-plane scattering losses, limiting the power that can be effectively coupled into the waveguide.

Given the PMMA layer's exclusive patterning, the design's degrees of freedom in the 3D setup remain essentially two-dimensional, lacking any z-direction patterning. Consequently, the additional losses encountered in the 3D scenario are likely consistent across different devices when transitioning from the 2D effective index framework to the full 3D geometry. This consistency suggests that the optimized structures remain near-optimal even in a 3D context, validating the effective index method's applicability and effectiveness for device optimization within this specific design domain.

Practicality is crucial in device optimization, especially when factoring in the computational demands of 3D full-wave simulations. In the abovementioned case, a single high-resolution 3D simulation can take approximately 16 hours on a high-performance computing setup. Utilizing the adjoint method for sensitivity calculations doubles this timeframe to around 32 hours for each gradient-based update, making a full optimization, which typically requires several hundred iterations, practically unfeasible.

Contrastingly, effective index simulations in 2D are considerably more time-efficient, each taking about 30 seconds on a standard mid-range computer. These simulations still lead to device designs exhibiting satisfactory performance. A potential middle ground could involve optimizing devices via the effective index method and refining the solution with a handful of full-wave simulation steps. Yet, even a mere ten such refinement steps could extend the process to several weeks, exceeding the scope of this investigation.

In recent years, significant progress has been made in exploring alternative design strategies for photonic devices, both in 2D and 3D. Nonetheless, the primary bottleneck in our optimization process remains the computational expense of performing simulations. While various optimization techniques exist, none circumvent the necessity for numerical simulations. Gradient-free methodologies, like swarm or evolutionary optimization, demand extensive solution space exploration through numerous simulations [97, 98, 99, 100]. These strategies can excel in scenarios where gradients are inaccessible, such as discrete optimization problems, but they inherently require more objective function evaluations due to the absence of gradient information.

Data-driven approaches, including machine learning, effectively shift the computational burden of simulations towards the creation of a dataset, which is then used to train a rapid surrogate model [101, 102, 103]. This strategy can be advantageous for repetitive tasks or when a dataset is versatile enough for multiple applications. However, the initial dataset generation often demands more simulations than would be necessary for solving a singular task or a limited number of tasks directly, as we will show in later chapters.

Therefore, our proposition to leverage 2D effective index simulations instead of more computationally intensive 3D full-wave simulations seeks to expedite the optimization process substantially. This approach incurs a trade-off in accuracy for significantly faster simulation times, a compromise we consider worthwhile for enhancing the feasibility and efficiency of photonic device optimization within this framework.

3.5 Conclusion

In this chapter, we have demonstrated how to apply topology optimization to design devices capable of efficiently guiding light from a dipole source into a waveguide within a dielectric system with low index contrast. Our method utilizes an effective index approach to simplify the representation of the three-dimensional system, facilitating a direct comparison between the performance of the optimized device and its expected functionality based on well-established experimental results. This comparison demonstrates notable improvements in Purcell enhancement and coupling efficiency in the three-dimensional context.

The devices developed through this optimization process not only enable the effective integration of spontaneous light emission into integrated photonic circuits but also open new possibilities for controlling the flow of Bloch Surface Waves. This development represents a considerable advancement in the ability to manipulate and control light on the nanoscale, with potential implications for optical communication, sensing, and other applications. Our work bridges an important gap in coupling spontaneous emission to waveguides in environments of low index contrast, marking a significant step towards more efficient and adaptable photonic device architectures.

4 Inverse-designed photonic wire bond couplers

In this chapter, we explore the development of compact, high-performance couplers for fiber-to-chip interfaces through inverse design, focusing on photonic wire bonding (PWB). We begin with an introduction in Section 4.1 that outlines the challenges in fiber-to-chip coupling and the potential of PWBs. Following this, Section 4.2 details the system configuration and parameterization we employ to define the coupler's geometry. We then demonstrate the efficacy of our design in Section 4.3, showcasing significant improvements over traditional designs. Finally, we conclude with Section 4.4 and summarize our work's achievements and potential applications in the context of PWB couplers, emphasizing the broader implications for nanophotonic device integration and efficiency.

4.1 Introduction

Fiber-to-chip interfaces pose a significant bottleneck in the design of nanophotonic devices due to the high loss associated with coupling light into and out of photonic chips. This challenge is crucial as efficient coupling is essential for the overall performance of photonic integrated circuits (PICs). Coupling approaches fall into two main categories: in-plane and out-of-plane. Out-of-plane coupling, often employed for its ability to directly interface with fibers or free-space beams, typically utilizes diffraction gratings etched onto the PIC surface. These gratings are devices that diffract the light from a usually close-to-normal incidence into the chip's plane. While they can achieve relatively low losses, typically ranging between 1 dB and 2 dB [104, 105, 106, 107], their performance is inherently wavelength-specific. The efficiency of these gratings peaks around a narrow wavelength band centered at their target wavelength. This narrow-band performance directly results from their resonant nature, making them unsuitable for use in systems where broadband performance is critical.

Addressing this issue, broadband fiber-to-chip coupling primarily relies on in-plane (edge) coupling techniques, where light from an optical fiber is directly coupled into a side of the photonic chip [108, 109, 110], an approach frequently called "butt-coupling". This method stands out for its potential to achieve broadband coupling efficiency. Still, it introduces the complexity of matching the fiber's mode profile – which is considerably larger than that of any single waveguide on the PIC – with the edge coupling device on the chip. Efficiently bridging this mode size discrepancy poses a significant design challenge. One strategy to mitigate this involves collimating the fiber mode by printing microlenses directly onto the fiber [111, 112]. While this can improve the situation, it does not fully address the fundamental challenge of mode matching. On-chip edge couplers often necessitate elaborate designs, incorporating multiple chip layers and extensive propagation lengths to optimize the match between the incoming mode and the chip's waveguides. During these lengths, the mode undergoes an adiabatic tapering process to the dimensions required on the chip.

Though such design and fabrication approaches are widely adopted, they are not without drawbacks. The intricacy of these devices renders them susceptible to fabrication inaccuracies and occupies significant chip estate, a valuable resource in PIC design. Additionally, the scalability of these coupling mechanisms is constrained by the physical dimensions of the couplers, limiting the density of fibers that can be interfaced with a PIC. This scalability challenge presents a crucial bottleneck, particularly in applications requiring many input/output waveguides, underscoring the need for innovative design solutions that can streamline the coupling process while minimizing the footprint and enhancing the scalability of fiber-to-chip interfaces.

Photonic wire bonding [113, 114, 115] emerges as a compelling solution to the limitations encountered in traditional fiber-to-chip coupling methods. This technique involves the 3D printing of transparent polymer waveguides that interface directly to the chip. These waveguides facilitate chip-to-chip interconnections and serve as efficient interfaces for fiber-to-chip coupling. The flexibility in shaping PWBs allows them to adapt to the chip's local topology, thus supporting the dense integration of photonic components. This adaptability is advantageous in applications like connecting multi-core fibers directly to a PIC [116], showcasing the technique's potential for enhancing device integration density.

A key attribute of PWBs is their relatively small diameter, typically around 2 µm, which results in mode profiles significantly smaller than those of fibers. This characteristic makes PWBs inherently more compatible with the mode profiles prevalent in low-index materials employed in CMOS processes, such as silicon nitride, facilitating efficient coupling to PIC waveguides. Despite the advantage of smaller mode profiles for seamless integration with PICs, a critical consideration remains: the mode profile of the PWB must be matched to that of the optical fiber on one end and converted to align with the structures fabricated on the PIC on the other [117].

On the fiber side of the system, these structures are typically relatively large linear tapers [116] (on the order of 100 µm). This chapter is dedicated to investigating the application of inverse design in developing coupling devices for fiber-to-chip interfaces utilizing PWBs. We will demonstrate that with this technique, we can design couplers that exhibit high performance and significantly reduce the spatial footprint.

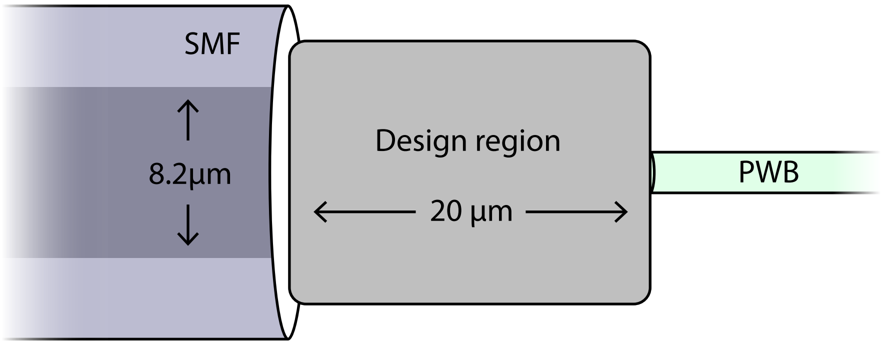

Figure 4.1: Illustration depicting the configuration for optimizing the coupling between an SMF and a PWB. The optimized coupling interface, referred to as the "design region," is positioned over the SMF's core (represented by the shaded area within the SMF) to efficiently channel the light into the PWB. It is important to note that the fiber's geometry is not depicted to scale in this visualization. The SMF's cladding extends significantly beyond the core and the diameter of the design region is twice the diameter of the SMF's core.

4.2 Setup

In this section, we will outline the general setting of our optimization problem, describe the physical layout, and describe the simulation and optimization setup. We will then detail the parameterization used to create the geometries of our designs, as our approach here significantly deviates from typical schemes used in topology optimization.

4.2.1 Problem statement

We will first need to specify the system's physical setup to optimize a device for coupling light from a single-mode optical fiber (SMF) into a PWB. An illustration of this setup is shown in Fig. 4.1.

The SMF's core has a diameter of $d_{\text{SMF,core}} = 8.2\,\mu\text{m}$ with a refractive index of $n_{\text{SMF,core}} = 1.449$. The core is surrounded by a cladding with an index of $n_{\text{SMF,clad}} = 1.444$ with a spatial extent much larger than the core. The design region for the optimization is attached directly to the SMF, with a spatial extent of 16.4 μm along the SMF's radius (twice the size of the core) and a length of 20 μm (the target length of the coupler). On the right side of the coupling region, the PWB is attached, with a diameter of $d_{\text{PWB}} = 2.2\,\mu\text{m}$. The PWB and the optimized coupler are made from a polymer with a refractive index of $n_{\text{resist}} = 1.53$, as they are envisioned to be 3D-printed. The PWB and the coupler are embedded in a low-index cladding material with a refractive index of $n_{\text{clad}} = 1.4$. The target operational wavelength for the device is $\lambda_0 = 1.55\,\mu\text{m}$.

Given these parameters, the simulation setup is relatively simple - an eigenmode source is placed in the SMF, exciting its fundamental mode. A mode monitor is then placed in the PWB, which records the power in the fundamental mode of the PWB. The optimization target is to maximize the power coupled into this mode. Note that since the SMF's cladding is much larger than the core diameter, the cladding boundaries are not included explicitly in the simulation. Instead, we truncate the simulation within the cladding. This assumption is valid as long as the simulation domain is large enough to capture the guided mode carried in the SMF fully. The simulation is truncated with PMLs on all sides.

The figure of merit to be maximized in the optimization is the normalized power transmitted into the forward-propagating fundamental mode of the PWB and can be written as

$$F = T = \frac{\left|\alpha_{0,\mathrm{out}}\right|^2}{\left|\alpha_{0,\mathrm{in}}\right|^2} \quad , \tag{4.1}$$

where $\alpha_{0,\mathrm{in}}$ is the mode coefficient of the fundamental mode of the SMF and $\alpha_{0,\mathrm{out}}$ is the mode coefficient of the fundamental mode as measured in the PWB. These coefficients are normalized such that $|\alpha_n|^2$ represents the total power carried in the n-th mode.

4.2.2 Parameterization

A notable aspect of addressing the optimization challenge is the inherent rotational symmetry of the SMF and the PWB. This symmetry is reflected in their modes, naturally suggesting that a similar symmetry should be adopted within the design region. This approach is underpinned by the anticipation that an optimally designed device would inherently conform to this symmetry, rendering exploring design spaces that diverge from this principle counterproductive. Furthermore, the device's intended role as a bridging component between the SMF and the PWB necessitates a unified, singularly connected structure. Considering the envisioned manufacturing process via two-photon lithography, it is essential that the design accommodates feasibly large feature sizes – specifically, lateral dimensions on the order of 500 nm and vertical dimensions approximating 1 µm. This consideration ensures the device's structural integrity and manufacturability, precluding designs that might incorporate unconnected voids potentially filled with unpolymerized resist. Given these criteria, a strategic approach to the design would prioritize the *shape* of the design region over detailed patterning within its confines, tailoring the optimization process to the specific constraints and symmetries intrinsic to the system.

In the context of our simulations, which utilize a discretized FDTD grid, every component including fields, geometrical structures, and gradients, is mapped onto this grid. This discrete nature poses a challenge for shape optimization, where the device's boundary is explicitly defined, for example, through splines, due to the necessity of gradient propagation through such a parameterization. To navigate this complexity, we have developed a variation of the filter-and-project parameterization scheme. The details of this parameterization are illustrated in Fig. 4.2, showcasing how it effectively enables shape optimization within the constraints of the FDTD grid.

In our modified projection approach, the initial setup involves defining the design density, denoted as $\boldsymbol{\rho}_{xy}$, across a two-dimensional grid reflecting the discretized dimensions of the device's length and radius. This grid is populated with a linear gradient that transitions from 0 to 1 along the radial direction, or the y-axis. The design density is then assumed

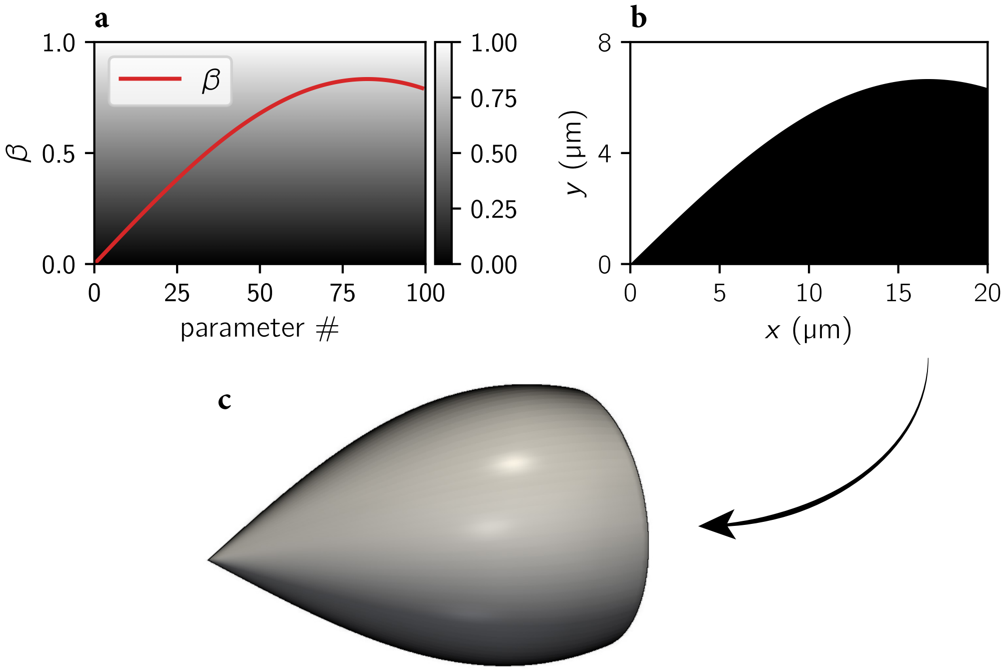

Figure 4.2: Overview of the parameterization scheme for the coupler optimization. Illustration **a** shows the parameters in the optimization, which represent the offset values β of a soft thresholding function (projection). The projection is then applied row-wise to an array filled with a linear gradient, mapping it to a binary design. The exemplary design is shown in **b**. This binary design undergoes a rotation about its central axis to achieve a three-dimensional, rotationally symmetric design shown in **c**.

to be fixed and is not modified during the optimization. Instead, we let $\boldsymbol{\beta}$ be an array of parameters instead of a scalar, and apply the projection along the x-axis of the design density, with a different β_x for each x along the length of the device. Remember that the projection is essentially a (differentiable) thresholding function, with $\boldsymbol{\beta}$ parameterizing the level at which the threshold is applied. Additionally, we apply a Gaussian filter to $\boldsymbol{\beta}$, which correlates neighboring threshold levels and creates a smooth boundary, with the size of the filter implicitly imposing different minimum radii of curvature, analogous to how the filtering step in the standard three-field parameterization imposes a minimum feature size.

By assigning distinct β_x values to each position along the x-axis within $\boldsymbol{\rho}_{xy}$, this method introduces varying threshold levels across the density field, kept constant throughout the optimization process. Consequently, the array $\boldsymbol{\beta}$ becomes the principal agent for defining the geometry's boundary post-threshold application to $\boldsymbol{\rho}_{xy}$. In this optimization framework, the design variables are the individual β_x elements within $\boldsymbol{\beta}$, which effectively delineate the geometry's contours by intersecting the static density field $\boldsymbol{\rho}_{xy}$ at varying heights. A significant advantage of this technique is its efficiency in decreasing the optimization's degrees of freedom from the total grid points, $n_x \times n_y$, to merely n_x, the count of grid points along the x-axis. Following the projection phase, the modified density field $\boldsymbol{\rho}_{xy}$ undergoes a rotational operation around the x-axis to achieve a rotationally symmetric design. Subsequently, the processed density, now represented as $\hat{\boldsymbol{\rho}}_{xy}$, undergoes linear

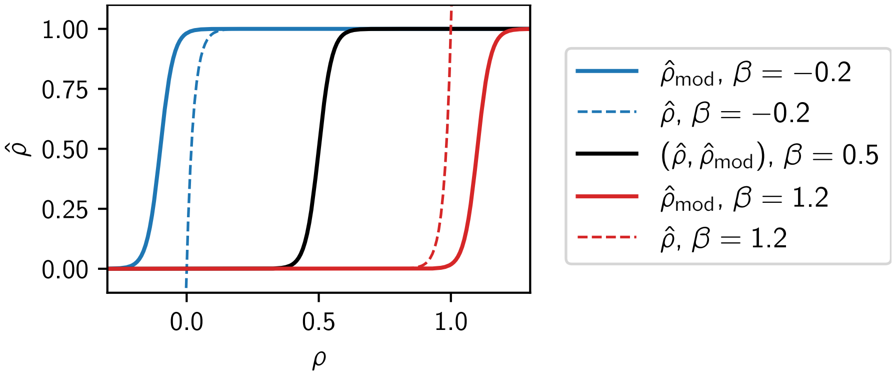

Figure 4.3: Illustration of the tailored projection method for optimizing the coupler. Solid lines represent the adapted soft threshold function across various β offsets, contrasting with the dashed lines that depict the original projection approach at equivalent offsets. The two projections converge for $\beta = 0.5$, yet exhibit significant differences towards the interval's extremes.

interpolation to assign the final permittivity values, completing the transformation from a discretized density field to a geometrically and materially defined device.

A slight subtlety of this approach lies in the regular projection scheme's assumption that the design density always lies within $[0, 1]$, with β generally assumed to be equal or close to 0.5. However, this standard approach to projection exhibits a non-linear reaction in relation to β within this bounded interval and tends towards divergence for values extending beyond these confines. To address this characteristic, a revised projection formula is employed here, expressed as:

$$\hat{\boldsymbol{\rho}}_{\mathrm{mod}}(\boldsymbol{\beta}) = \frac{1}{2}\left(\tanh\left(\alpha\left(\boldsymbol{\rho}_{xy} - \boldsymbol{\beta}\right)\right) + 1\right) \quad , \tag{4.2}$$

where α denotes the standard projection strength, $\boldsymbol{\rho}_{xy}$ represents the consistent density field, and $\boldsymbol{\beta}$ signifies the array of threshold levels. This adjusted function ensures a linear threshold level adjustment across the entire spectrum of $\boldsymbol{\beta}$ values, a feature illustrated in Fig. 4.3.

Initially, we start with a projection strength of $\alpha = 10$ and a vector of $\boldsymbol{\beta}$ uniformly initialized to 0.5, which leads to a straight device with “washed out” edges. The value of α is then doubled every 30 iterations until the optimization converges, *i.e.*, the design is sufficiently binary, and the objective function does not improve further. For the optimization, we choose Adam [118] as our optimizer, mainly for its insensitivity towards discontinuous gradients when α is increased.

With this, the optimization problem and its parameterization are comprehensively defined, setting the stage for the subsequent phase of actual device optimization.

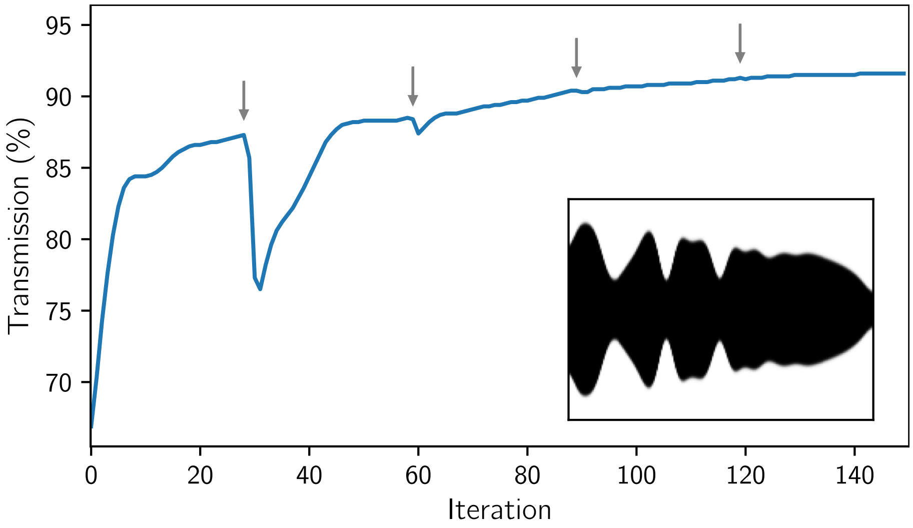

Figure 4.4: Evolution of the figure of merit during optimization. The gray arrows indicate the iterations at which the optimization was restarted with a larger α value in the projection, leading to stronger binarization and dips in the objective function. A cross-section of the final optimized coupler design is shown in the inset.

4.3 Results

After running the optimization for 150 iterations, it converges to a final transmission in the PWB of around 92 %. The optimization history is shown in Fig. 4.4. We can observe dips in the objective function when the projection strength α is increased, and the initial adjustment, increasing α from 10 to 20, is marked by the most pronounced drop in the objective function, indicating a substantial modification to the design's boundary. Subsequent increments in the binarization strength yield progressively milder impacts on the objective, as the design's boundary approaches a nearly binary state post the first adjustment. The following steps further refine the design, enhancing the boundary's definition towards an optimal binary configuration.

The inset in Fig. 4.4 showcases a cross-sectional view of the final design, while a complete three-dimensional rendering of the system is presented in Fig. 4.5. Upon examination, it becomes evident that the device's design intricately tapers from a dimension approximately equivalent to the SMF's core to a width matching that of the PWB. This transition is characterized by oscillating boundaries that collectively define a gradual tapering profile, culminating in a notably sharper taper near the device's endpoint.

To elucidate the mechanism through which the coupler channels light into the PWB, we present the absolute electric fields in Fig. 4.6. This illustration encompasses the fields of the fundamental modes of both the SMF (left) and the PWB (right), alongside a cross-sectional view of the electric field within the coupler (center). The simulation extends from within the SMF, where the SMF mode is initially excited, to a point just beyond the optimized coupler, terminating within the PWB. The simulation results show that the device efficiently transitions the light from the SMF mode on the left to the PWB mode on

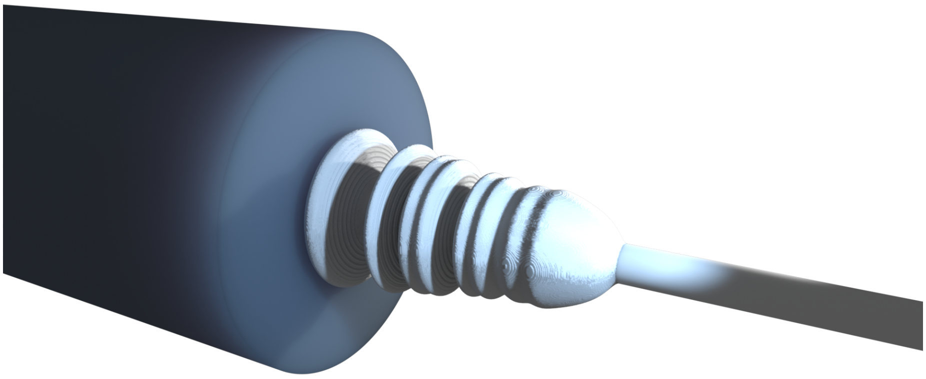

Figure 4.5: Rendering of the optimized SMF to PWB coupling system. The optimized coupler has a total length of 20 µm.

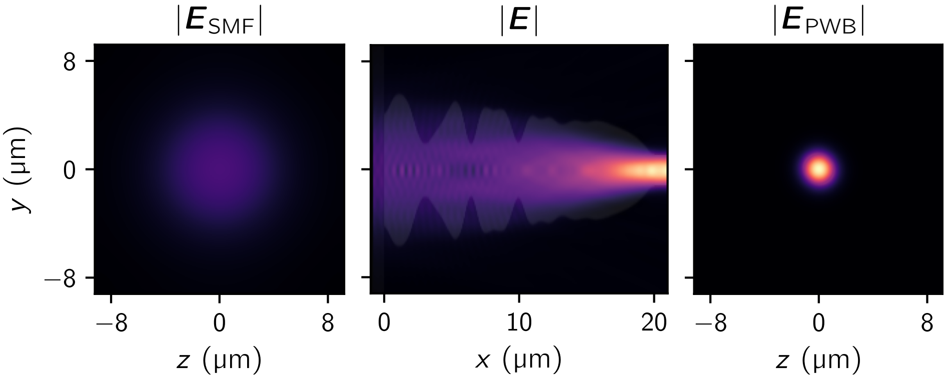

Figure 4.6: Mode profiles and field distribution for the optimized coupler. The SMF mode (left) is injected into the coupler (middle), which couples it into the fundamental mode of the PWB (right). Note that the SMF mode appears dimmer than the PWB mode because it carries the same power spread over a larger area. The geometry of the coupler is overlaid on top of the field. The device can be seen to guide the light from the SMF into the PWB. efficiently

the right. The oscillating peaks and valleys along the coupler's boundary, which might initially appear unconventional, effectively serve as a series of "lenses". They sculpt the beam profile to match the target mode profile at the output. This unique guiding effect can likely be attributed to the system's minimal index contrast, indicated by a mere $\Delta_n = 0.13$ difference between the resist ($n_{\text{resist}} = 1.53$) and the cladding ($n_{\text{clad}} = 1.4$). Given this slight index contrast, the light experiences only weak confinement, suggesting that directing the light along the structure's boundaries, akin to traditional waveguide behavior, may not be viable over such a short device.

To further illustrate this point, we examine the transmission spectrum of the optimized coupler in comparison to that of a linear taper, the latter spanning a length of 20 µm, as depicted in Fig. 4.7. Observations indicate that both devices exhibit a broadband response, demonstrating a remarkable level of insensitivity to wavelength fluctuations

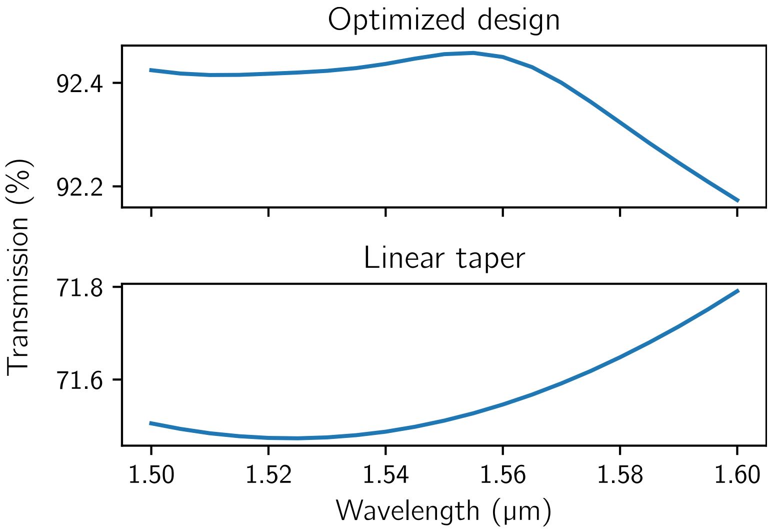

Figure 4.7: Transmission spectra of the optimized PWB coupler (top) and a 20 µm long linear taper (bottom), measured as the power coupled into the fundamental mode of the PWB, normalized to the input power from the SMF mode. The optimized device shows good broadband coupling performance, with transmission being above 92 % over a 100 nm bandwidth. In contrast, the linear taper shows a transmission of only 72 % over the same wavelength range.

across a 100 nm spectrum, specifically within the range $\lambda \in [1.5\,\mu\text{m}, 1.6\,\mu\text{m}]$. Notably, the optimized design achieves a substantially enhanced transmission, reaching 92 % across the spectrum, in stark contrast to the linear taper's transmission rate of merely 72 %. This comparative analysis underscores the efficacy of the optimized design, especially when considering that the performance metrics of the linear taper are consistent with findings reported in existing literature [116]. Critically, to match the performance of our optimized coupler, one would require a linear taper extending approximately 100 µm in length. This comparison not only illustrates the advanced capability of the optimized design but also underscores the pivotal role inverse design plays in enhancing device performance and reducing physical dimensions, thereby advancing the development of compact and efficient photonic systems.

4.4 Conclusion

In this chapter, we have demonstrated the inverse design of a coupling device that connects a single-mode fiber to a photonic wire bond. A modified parameterization scheme, which defines the coupler's shape through a static, radially graded density field, was central to optimizing such a coupler. We reformulated the standard three-field topology optimization parameterization by optimizing the levels at which the density field is thresholded instead of directly modifying it. This approach enables the implicit parameterization of the device's boundary, which we subsequently optimize.

A comparative analysis with a conventional linear taper revealed a significant improvement in performance for our optimized design, yielding a transmitted power of 92 % compared to 72 % for the linear taper. This comparison underscored the inefficiency of traditional designs and showcased the optimized coupler's ability to achieve superior performance within a considerably reduced length.

While our focus was primarily on coupling from a SMF to a PWB, the potential applications of these couplers extend to linking multi-core fibers with chips via PWBs similar to the application showcased in Lindenmann *et al.* [116]. Our design methodology, characterized by its emphasis on rotational symmetry and gradient-based inverse design, harbors broader applicability. For instance, it could be adapted for crafting free-form lenses that are 3D-printed directly onto fibers. Such applications could, for example, include modifying the fiber's mode profile to achieve a target Gaussian beam for free-space edge coupling.

To conclude this chapter, our work has successfully demonstrated the design of boundary-parameterized coupling elements that leverage enforced rotational symmetry for gradient-based inverse design. The resulting devices not only exhibit superior performance in comparison to conventional linear tapers but also facilitate efficient optical coupling across previously unattainable length scales.

5 Inverse design of polarization-independent free-form vertical couplers

Having investigated the minimization of device footprints for broadband coupling in an in-plane setting in Chapter 4, In this chapter, we will investigate the application of inverse design, coupled with additive manufacturing, for the design of a device for an out-of-plane coupling scheme. We will begin with a motivation in Section 5.1, which we will keep relatively brief as it mirrors the discussion in Section 4.1 to an extent. We then introduce the problem setting and optimization setup in Section 5.2, which we will follow with a discussion of the optimization results in Section 5.3. The chapter concludes with Section 5.4, summarizing our findings and discussing possible future paths.

5.1 Introduction

Efficient coupling from optical fibers to photonic integrated circuits (PICs) is a major bottleneck in integrated optics. It is typically where the most significant chunk of loss is incurred in such a system, as we have previously discussed in Chapter 4. A key challenge in designing such coupling devices is the mode mismatch arising from the low refractive index contrast in optical fibers and the typically much higher refractive index contrast available on the chip. The two dominant schemes for coupling light from fibers into photonic integrated circuits are edge couplers (in-plane) and diffraction gratings (out-of-plane). This chapter will focus on the latter, where we will design a free-form device for surface-normal coupling from an optical fiber into a photonic chip.

The performance of grating couplers is usually highly polarization selective, as standard designs use slab-shaped waveguides. However, many applications require polarization-insensitive devices, such as, *e.g.*, optical interconnects. Here, the polarization may shift during propagation in a fiber, and non-uniform coupling across different polarizations will lead to signal errors. A common approach to remedy this is to design polarization-*splitting* grating couplers that split the incoming light into two (usually orthogonal) output waveguides [119, 120, 121]. However, these designs are more difficult to integrate on-chip as they require additional routing. Moreover, they may fundamentally have limitations in terms of coupling efficiency for off-normal incidence due to symmetry constraints.

As such, polarization-insensitive coupling into a single waveguide that supports both polarization states is a desirable property for many applications and has been the subject of multiple prior studies [122, 123, 124]. However, these couplers typically suffer from lower coupling efficiencies than their polarization-sensitive counterparts due to the additional complexity of supporting two polarizations. To remedy this, we explore the feasibility of designing free-form polarization-insensitive grating couplers using adjoint optimization in this chapter.

5.2 Setup

The setup for our inverse design challenge is straightforward – we aim to couple light at a wavelength of $\lambda_0 = 1.55\,\mu\text{m}$, originating from an optical fiber at normal incidence, into a photonic integrated circuit via a bespoke, 3D-printed coupling structure placed atop a substrate. This configuration is depicted in Fig. 5.1, where a gray box, specifying the design region with dimensions of $20\,\mu\text{m} \times 20\,\mu\text{m}$ and a height of $5\,\mu\text{m}$, is illustrated.

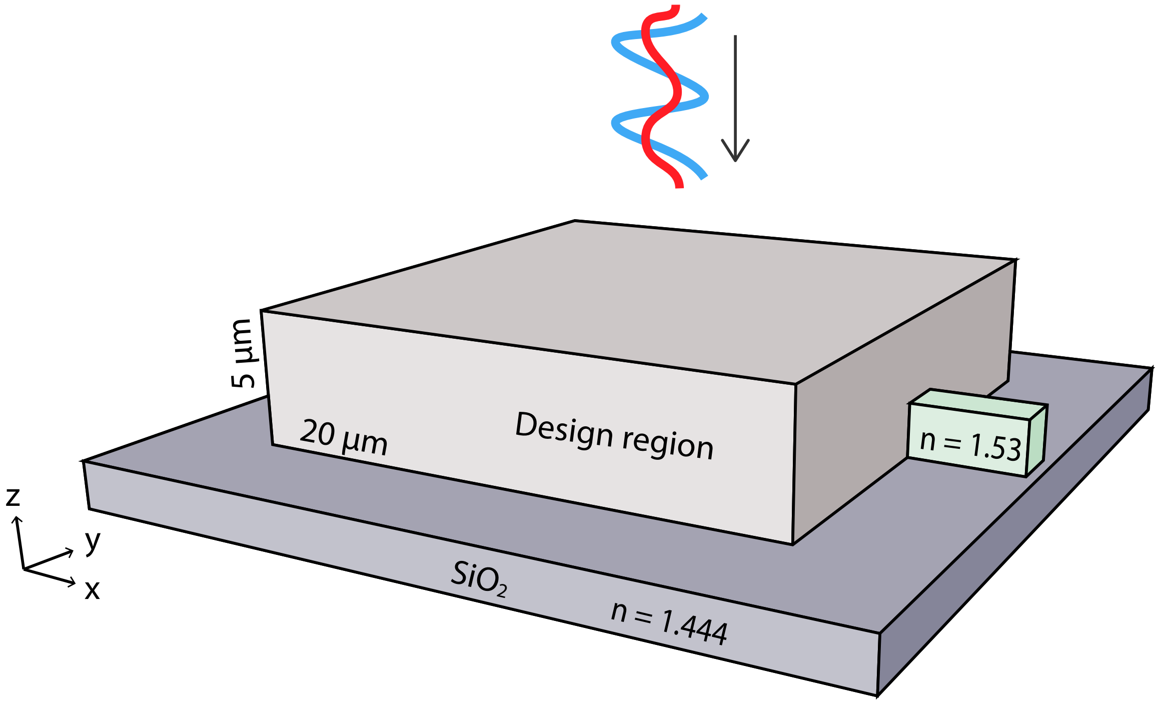

Figure 5.1: Setup of the coupling problem. The incoming x- or y-polarized (indicated by blue and red waves) with a target wavelength of $\lambda_0 = 1.55\,\mu\text{m}$ should be coupled into a 3D-printed waveguide at normal incidence. Attached to the waveguide and positioned beneath the source, a designated area is reserved for an optimized coupling structure, denoted as the design region. The device is placed on a substrate of silicon dioxide.

Given the anticipated manufacturing method of laser nanolithography, the material within the design area is presumed to possess a refractive index identical to that of a typical resist, such as IP-Dip, noted as $n_{\text{Resist}} = 1.53$. The device integrates a 3D-printed rectangular waveguide, having a cross-sectional area of $2\,\mu\text{m} \times 2\,\mu\text{m}$, which is conceptually designed to smoothly transition into the waveguides on the chip, for example, those made of silicon nitride, to facilitate low-index contrast coupling. The entire assembly is mounted on a buried oxide cladding, characterized by a refractive index of $n_{\mathrm{BOX}} = 1.444$, and the entire setup is encased in air with $n = 1$.

The incoming light originates from an optical single-mode fiber, characterized by a mode field diameter (MFD) of $d_{\mathrm{MFD}} = 10.4\,\mu\mathrm{m}$. This specification informs the selection of the design region's dimensions, intentionally chosen to be approximately twice that size, to fully encompass the incoming beam and afford ample space to transform this light into a waveguide mode at the device's output. In our simulations, the optical fiber's geometry is omitted. Instead, the fiber mode is represented by a Gaussian beam, with its beam waist equating to the fiber mode's MFD. The beam is directed to propagate in free space in the negative z direction. Although it will show divergence propagating in free space after exiting the fiber, we assume this divergence to be negligible since the fibers are typically placed very close to the coupling device.

To fulfill our design goal of optimizing the coupling from the LP_{01}^{x} fiber mode, modeled as an x-polarized Gaussian beam, into the fundamental transverse electric (TE) mode, as well as from the LP_{01}^{y} fiber mode, represented as a y-polarized Gaussian beam, into the fundamental transverse magnetic (TM) mode of the output waveguide, we establish the following objective function:

$$F_{\mathrm{obj}} = \frac{\left|\alpha^{+}_{\mathrm{TE}_{00}}\right|^{2}_{x} + \left|\alpha^{+}_{\mathrm{TM}_{00}}\right|^{2}_{y}}{P_{\mathrm{in},x} + P_{\mathrm{in},y}} \quad , \tag{5.1}$$

where $P_{\mathrm{in},x}$ and $P_{\mathrm{in},y}$ signify the input power of the incident x- and y-polarized Gaussian beams, respectively. Moreover, $|\alpha^{+}_{\mathrm{TE00}}|^{2}_{x}$ and $|\alpha^{+}_{\mathrm{TM}_{00}}|^{2}_{y}$ represent the power transmitted into the forward propagating fundamental TE and TM modes within the waveguide, respectively. This objective function quantitatively measures the efficiency of the designed coupling structure in channeling the incoming light into the desired waveguide modes, thus serving as a crucial metric for evaluating and guiding the optimization process toward our target of maximizing mode-specific transmission efficiencies.

Note that the x- and y-polarized input beams may also partially couple into the waveguide's TM and TE modes, respectively, which means that both input beams need to be simulated with two independent simulations to ensure accurate mode coefficients for $\alpha^{+}_{\mathrm{TE}_{00}}$ and $\alpha^{+}_{\mathrm{TM}_{00}}$.

The design region has a size of $20\,\mu\mathrm{m} \times 20\,\mu\mathrm{m} \times 5\,\mu\mathrm{m}$, which we discretize at $10\,\mathrm{px}\,\mu\mathrm{m}^{-1}$, leading to a total of 2×10^{6} degrees of freedom in the optimization. The simulations are performed at a spatial resolution of $20\,\mathrm{px}\,\mu\mathrm{m}^{-1}$ (twice that of the design region) using the Meep [3] FDTD package. For design region parameterization, we choose a standard three-field approach where design density ρ is initialized as a three-dimensional array representing the discretized design region, which is then filtered using a Gaussian filter to impose minimum feature size, and projected and linearly interpolated to the corresponding permittivity values for simulation. To account for the physical dimensions of the focal spot of the laser during printing, we choose a prolate Gaussian filter size of 500 nm laterally and 800 nm vertically, which is a parameterization scheme that leads to somewhat elongated features in the z direction. Note that this does not perfectly correspond to the laser's actual spot size, so a Lorentzian filter shape along the z direction would probably be more suitable. We choose a purely Gaussian filter here as a proof of concept and because of its favorable computational properties. It is a linearly separable filter, making the parameterization step in the optimization significantly more efficient. The projection is initialized with a strength of $\alpha = 3$, which is doubled every 20 optimization iterations, starting at iteration 15, until the optimization converges to a binary design. As in Section 4.2, we use Adam as the optimization algorithm, as it is robust to changes in projection and suitable for large optimization problems such as this one.

5.3 Results

The optimization is run for 160 iterations, comprising 460 simulations in total – two per polarization, with one forward and adjoint simulation each. The evolution of the objective function during this optimization is shown in Fig. 5.2.

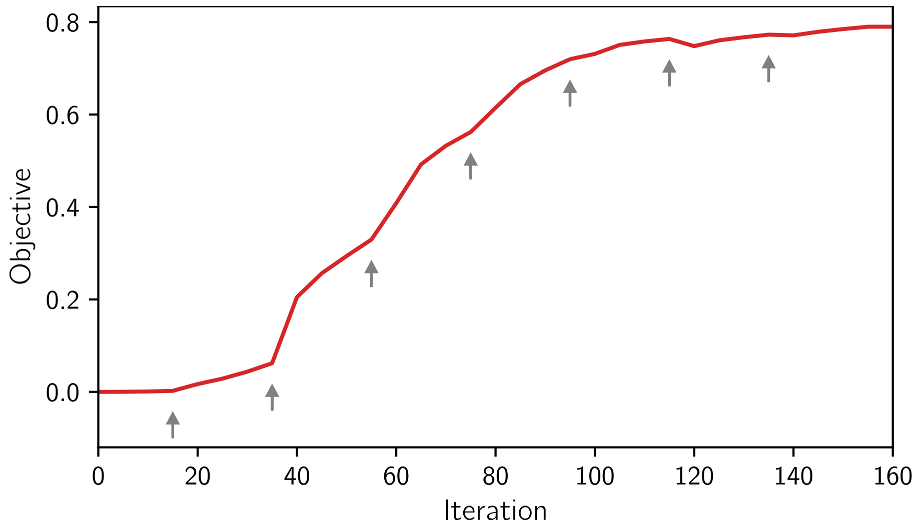

Figure 5.2: History of the objective function for the grating coupler optimization. The gray arrows indicated iterations when the optimizer was restarted with a higher binarization factory α in the projection.

Interestingly, the increase in binarization strength (gray arrows) does not lead to a drop in the figure of merit from which the optimization needs to recover, as observed in the previous chapter. Instead, we observe that the earlier increases in binarization strength benefit the overall objective function. We posit that this is due to the filter's interaction with the projection step and the optimization problem's physical nature. The first thing to realize is that inherently, we are designing a resonant device here, in the sense that to couple light at normal incidence via a (approximately) planar device layer, the structure's mode of operation will be mostly based on reflection – as opposed to creating an adiabatic transition between incident light and waveguide mode, which would necessitate a significantly more elongated design dimension z. Reflection, however, is most pronounced at interfaces with strong contrast in refractive index, a property that is exponentiated with grating-like structures. So, fundamentally, a device with continuous transitions between materials can not be a good reflector, *i.e.*, the performance of such a device with respect to the design problem at hand will perform sub-optimally. A large filtering radius, coupled with a low projection strength, essentially removes such binary designs from the design space in the optimization, which is likely the reason why we see only a very slight improvement in the objective within the first 40 iterations in the optimization, which is when the projection strength α has increased four-fold from 3 to 12. Essentially, the optimization moves as close as possible to a feasible design within the design space it can operate in. This design then already exhibits the overall topology of the "intended" device but with soft boundaries, which is then immediately made more efficient by a stronger projection, which preserves the overall topology of the device but makes the material interfaces more pronounced.

This can be observed by investigating the device at different points in the optimization as illustrated in Fig. 5.3. The figure shows top views of the design density in the x-y-plane and a slice along the center of the design density in the x-z-plane at different points in the optimization. At first glance, we can see how the design progresses from a "grayscale image" into a binary one throughout optimization – not surprising, given that we increase the strength of the projection throughout. Interestingly, though, and in support of our previous argument about binarization, we can observe that the overall layout of the device is already established after the first iteration of the optimization. In particular, when looking at the device after the first design variable update in the first column of Fig. 5.3, we already see a structure that resembles a conventional grating coupler from the top and a series of "Bragg mirrors" at an approximately 45 deg angle when viewed from the side. These general features remain until the end of the optimization and are still present in the final design shown in Fig. 5.4.

Apparently, this is the dominant mode of operation of the optimized device. Subsequent design iterations then enhance the index contrast and add additional, finer features, the purpose of which is more difficult to elucidate. It seems that the most intricate features emerge on the right side of the device, towards the output waveguide. Particularly in the x-y-plane, it looks like a light-guiding structure is created whose purpose is to funnel as much light as possible from the fringes of the design region toward the waveguide. In the x-z-plane, this structure entails something akin to a tapered extension of the output waveguide, effectively increasing the area over which light can couple into the waveguide.

To observe this behavior, we simulate the design using the commercial FDTD solver Tidy3D [125] and plot slices along the center of the device for both input polarizations in Fig. 5.5, with the top row showing a slice in the x-y-plane at $z = 2.5\,\mu\text{m}$ and the bottom row showing a slice in the x-z-plane. Overall, we see that the field profiles look similar for both polarizations. The electric field seems to extend a bit further to the left in the case of x-polarized input, but overall, the device's behavior seems to match its intended design goal, polarization-independent coupling. Qualitatively, it seems like the mode is coupled relatively well into the output waveguide, with the field profile looking approximately

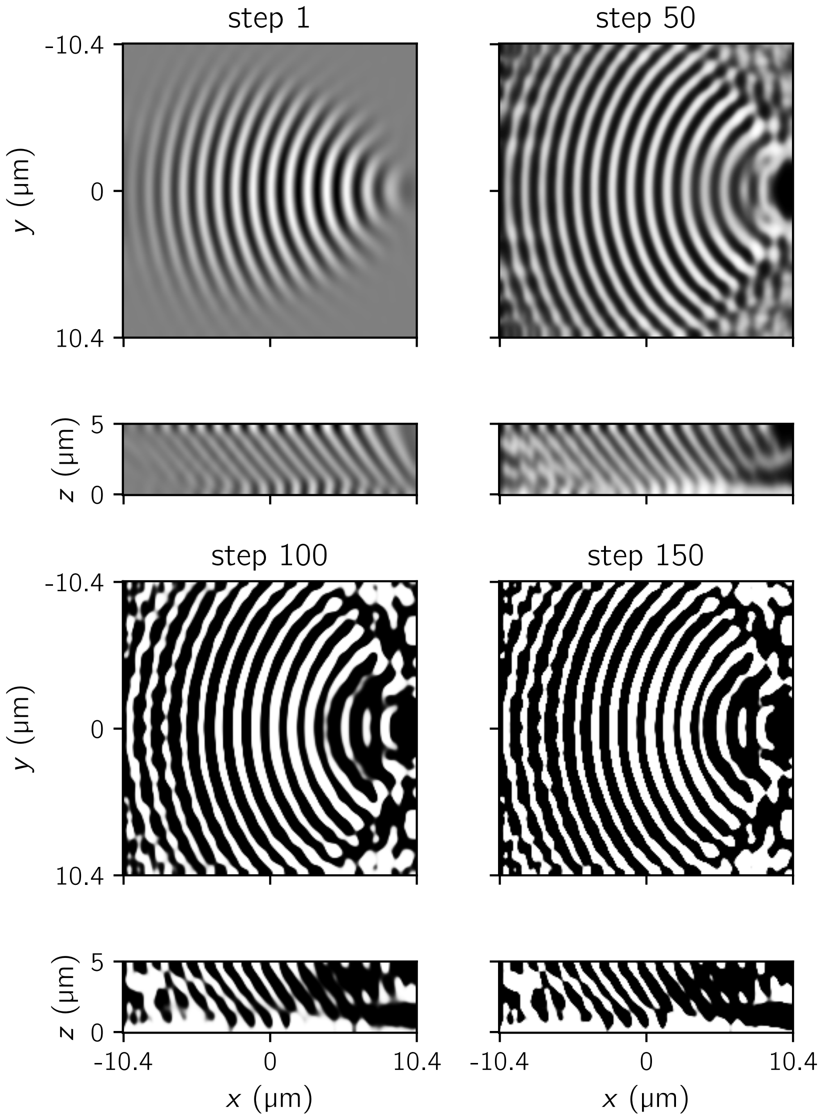

Figure 5.3: Slices of the design region in the x-y-plane and x-z plane at different optimization steps. The optimization starts by forming a basic grating structure, which is enhanced with more intricate features as the optimization goes on. The initial design features a continuous permittivity distribution, which is transformed into a fully binarized design by the end of the optimization.

Gaussian, without any apparent beating patterns or ripples. However, some portion of the light seems to radiate into the cladding, with the tails of the fields on the right side of the device extending far beyond the waveguide.

This is best illustrated by comparing the normalized power carried in the fundamental TE/TM modes of the waveguide (our design target) to the flux that passes through the waveguide, as illustrated in Fig. 5.6. The power is normalized to the incident (source) power. We first observe that the flux passing through the output side of the coupler for both polarizations carries around 88 % of the incoming power, with a maximum centered around the target wavelength of $\lambda_0 = 1.55$ µm and dropping down to a minimum of around 50 % for x-polarization and 72 % for y-polarization. We notice a slight shift in wavelength

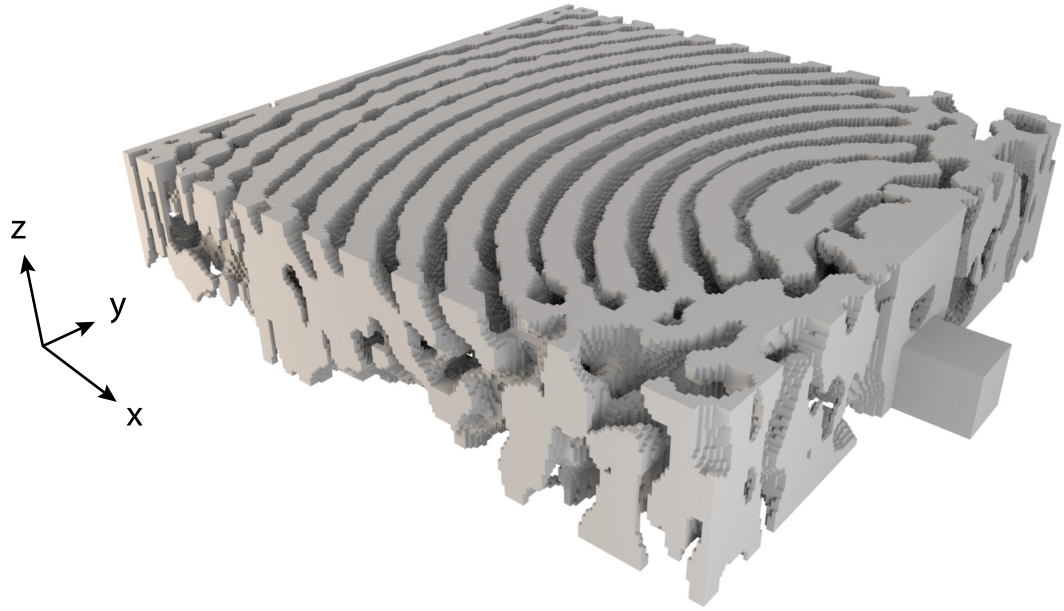

Figure 5.4: 3D rendering of the optimized polarization independent coupler. The device exhibits a Bragg-like grating on the top, with intricate features along the z-axis.

for the peak flux for the two polarizations of around 20 nm. The power carried in the fundamental modes shows essentially the same line shape as the respective flux, but they carry around 10 % less power, with both TE and TM modes peaking at around 78 %. Upon further investigation, we find that only around 1.5 % of this difference is due to coupling to higher order waveguide modes, while the remaining discrepancy can be explained by radiation losses into the substrate, which is also apparent from the previously discussed field tails in Fig. 5.5.

It proves tricky to compare the coupling efficiency of our design to that of existing literature because, to the best of our knowledge, only one variation on the concept of 3D-printed vertical couplers has been reported in literature [126], where coupling elements are designed that adiabatically taper down in the z-direction, with a reported coupling efficiency of 66 %. While these couplers share the same material platform, they are based on adiabatic coupling and are, therefore, much larger than our reported design, featuring a length of around 100 µm and a height of around 60 µm. Numerous studies have been conducted on both polarization-splitting [119, 120, 121], where two polarization are separated into different output waveguides (usually at a 90 deg angle) and polarization-independent [122, 123, 124] grating couplers, where both polarizations enter the same waveguide such as here, for standard CMOS platforms, which is another comparison we can draw. We find that perhaps unsurprisingly, polarization-splitting grating couplers, where the coupling efficiency is typically in the range from 60 % to 80 %, are generally more efficient than polarization-independent ones, where the reported coupling efficiencies generally lie between 25 % to 50 %. While our design handily outperforms the reported polarization-independent couplers and can keep up with the reported performance of

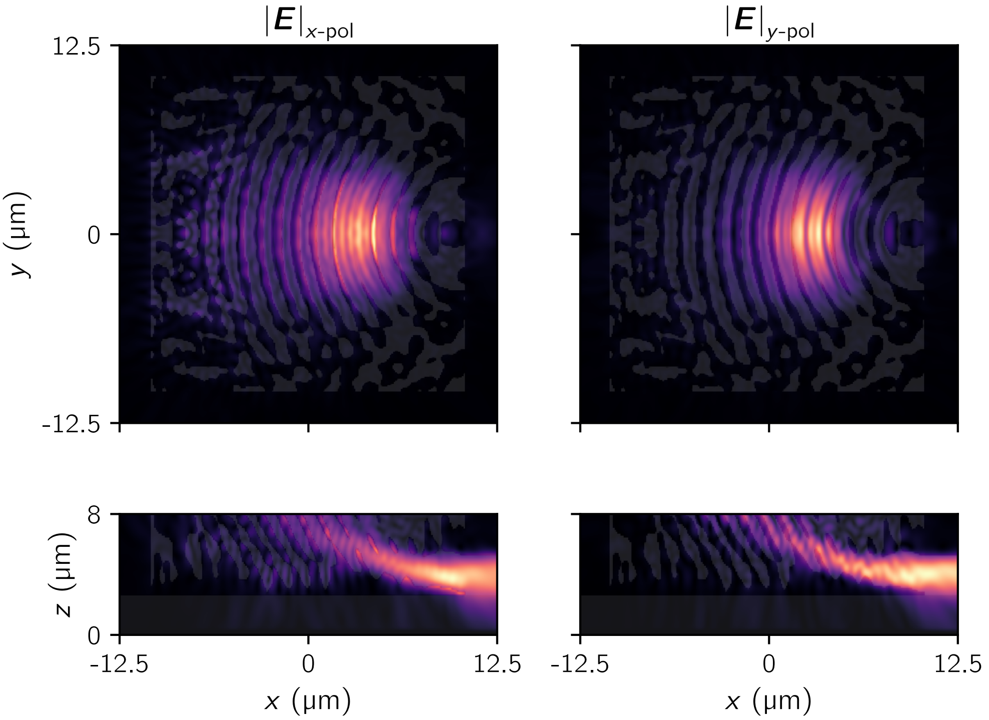

Figure 5.5: Amplitude of the absolute electric field distribution of the optimized grating upon illumination with an x-polarized (left) and a y-polarized (right) Gaussian beam. The top row depicts a slice of the fields in the x-y-plane at $z = 2.5\,\mu\text{m}$ and the bottom row shows a slice in the x-z-plane at $y = 0\,\mu\text{m}$, where the funneling of the light into the waveguide can be seen. A slice of the structure in the corresponding plane is overlaid on top of the fields.

polarization-splitting designs, we stress that this is not a straightforward comparison as our design is incompatible with standard CMOS processes.

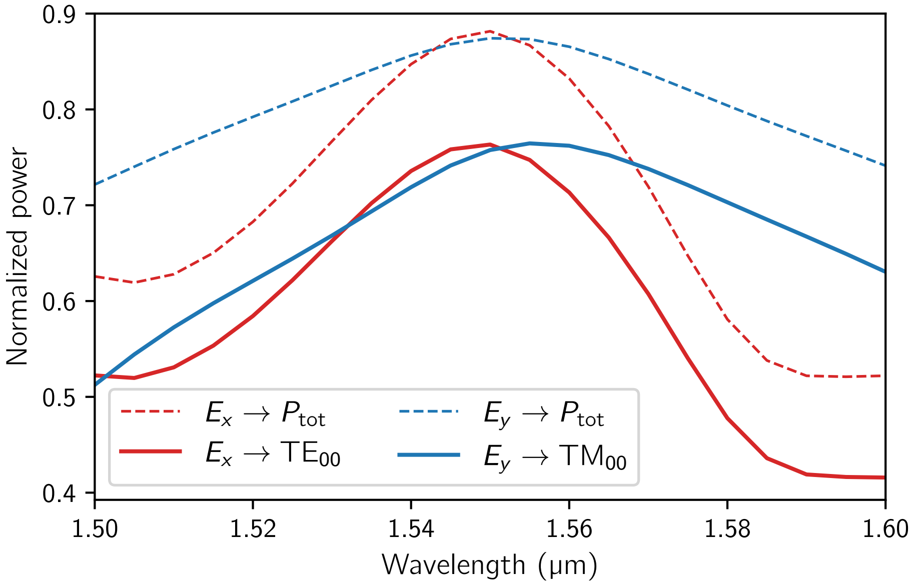

Figure 5.6: Source-normalized power coupled into the waveguide for an x-polarized (red) and a y-polarized (blue) source. The solid lines indicate the power carried in a waveguide mode. For x-polarization, the light is coupled mainly into the fundamental TE_{00} mode of the waveguide, while for y-polarization it is coupled into the fundamental TM_{00} mode. The dashed lines indicate the flux passing through a flux monitor placed in the output waveguide, indicating that some light is also coupled into higher-order and/or radiating modes.

5.4 Conclusion

In this chapter, we have explored the inverse design of a fully free-form, polarization-independent coupling element that resembles an on-chip grating coupler. Due to the device's comparatively large volume of 20 μm × 20 μm × 5 μm and free-form nature, the optimization comprises a very large number of degrees of freedom (2×10^6 parameters), making this truly a large-scale optimization problem that would not be possible to tackle without the adjoint method.

The optimized couplers achieve polarization-independent vertical coupling with peak coupling efficiencies at the target wavelength $\lambda_0 = 1.55\,\mu\text{m}$ of 78 %, on par with state-of-the-art grating couplers designed for a single polarization [106, 127, 128] and more efficient than comparable designs designed with multiple polarizations in mind [122, 123, 124].

The optimized device is free-form and should be regarded as a proof of concept for the potential performance of coupling elements created using additive manufacturing techniques. Here, we have outlined the general optimization process and a path toward 3D-printed couplers; however, further efforts are needed to ensure the manufacturability of the proposed (or alternate) designs. There are three main aspects regarding this challenge: feature sizes, sensitivity to perturbations, and mechanical stability.

The first two points can be tackled via more advanced parameterization and additional optical simulations. In this chapter, we adopted a rather crude approximation to feature size constraints by filtering with a prolate Gaussian kernel. A more accurate alternative would be to develop an approximate fabrication model that is integrated directly into the optimization. For example, instead of filtering with a Gaussian kernel, the point spread function of the focal spot could be used directly, leading to an implicit integration of the writable feature sizes of the system. This could be further enhanced by modeling the resist's polymerization threshold and the degrees of freedom in the 3D printing process, such as focal spot intensity and writing speed. The design's sensitivity to perturbations in the geometry can be incorporated through erosion/dilation schemes [18, 24] or, more implicitly, by targeting multiple wavelengths in the optimization.

Lastly, mechanical stability presents a unique challenge in free-form electromagnetic structures designed for additive manufacturing. This aspect has largely been neglected, as the focus of nanophotonic inverse design has been on planar structures for on-chip applications. This is an aspect that we will explore in more detail in the next chapter.

6 Nanophotonic devices with mechanical constraints

In Chapter 5, we have demonstrated the inverse design of a fully three-dimensional device. This chapter will outline some of the challenges associated with fabricating such devices and propose a solution to address them – co-designing for both electromagnetic and structural design objectives. After an introduction in Section 6.1, we will discuss the required theory for the following sections in Section 6.2; in particular, we will introduce structural topology optimization. Section 6.3 will then showcase some results for devices optimized using our proposed method, demonstrating structural integrity while maintaining high optical performance. After a brief summary, Section 6.4 will discuss possible future developments related to this work. The results presented in this chapter are based on [P1].

6.1 Introduction

Most studies in inverse design have so far focused on planar devices [129, 130, 131, 30, 132], with three-dimensional geometries receiving significantly less attention [133, 45]. This limited application in three-dimensional designs is primarily due to the constraints of existing manufacturing technologies. However, recent advancements in additive manufacturing, mainly 3D laser nanolithography [134, 135], now enable the production of complex three-dimensional microstructures. This evolution in manufacturing technologies paves the way for creating optical devices with customizable functionalities, provided that precise design blueprints are available. In nanophotonics, the ability to explore topology-optimized free-form geometries introduces new challenges that must be addressed to realize their potential fully.

This chapter explores an important issue that arises with additive manufacturing techniques, such as 3D laser nanolithography. Specifically, it addresses the challenge of maintaining the structural integrity of free-form structures created through such methods. In contrast to traditional planar devices produced through subtractive processes, free-form geometries lack inherent support from a substrate, so it is necessary to ensure that the material layout remains fully connected to prevent collapse.

Interestingly, the importance of structural integrity in the context of electromagnetic topology optimization has been largely overlooked until now. However, it is a critical consideration when it comes to designing free-form geometries using additive manufacturing techniques.

To address this challenge, we propose a new approach to topology optimization in nanophotonics, which considers compliance minimization as part of the optimization framework. This methodology simultaneously addresses the structural and electromagnetic optimization challenges and enables the design of devices that integrate optical and mechanical functionalities.

As an example, we demonstrate the design of a photonic nanolens and a waveguide-integrated mode converter, both of which showcase enhanced structural integrity. By broadening the scope of topology optimization in this way, we open up new possibilities for creating functional elements in nanophotonics.

6.2 Topology optimization for nanophotonics and structural mechanics

In nanophotonic device optimization, the inclination towards feature sizes that align with the wavelength scale is a natural consequence of an optimization trying to utilize interference to maximal effect. However, this attribute often compromises the structural integrity of the devices. To navigate the inverse design of devices that do not suffer from such structural issues, we explore insights from a field where topology optimization is a cornerstone: structural mechanics, as referenced in seminal works [136, 137, 138]. The core principle in mechanical topology optimization is the inherent mechanical stability of the optimized structures, which mandates the formation of interconnected structures as solutions.

Leveraging this foundational principle, we introduce a holistic optimization framework combining electromagnetic and mechanical topology optimization. This dual-focus framework is designed to refine the structural robustness and optical performance of devices concurrently. It incorporates considerations for external forces acting upon the structure, thereby ensuring the structural integrity of the device under specified loads. The method dictates the pattern of material connectivity based on the forces' distribution, which could be imposed either as hypothetical constructs to guide material connectivity along desired axes or as representations of actual physical stresses expected on the device. The formulation supports both abstract and realistic force applications, and illustrative examples are provided for each scenario in this chapter.

The convergence of structural and electromagnetic topology optimization within this multi-physics methodology involves delineating these two optimization problems separately. Following this, we introduce a unified optimization challenge that encapsulates compliance minimization (for structural integrity) and electromagnetic performance enhancement. This combined approach allows for a detailed examination of each optimization independently before merging them into a single objective. This integrated strategy paves the way for designing nanophotonic devices that excel in their optical functionality and are engineered to withstand physical demands.

Here, we employ a three-field parameterization scheme as discussed in Subsection 2.2.1 for structural and electromagnetic optimization. To prevent the emergence of minor features within the optimized design, we implement a Gaussian filter on the design variables:

$$\tilde{\rho}_i = \frac{\sum_{j\in\mathcal{D}_i} w_{ij}\rho_j}{\sum_{j\in\mathcal{D}_i} w_{ij}} \quad \text{with} \quad w_{ij} = \begin{cases} r_{\min}\exp\left(-\frac{|\boldsymbol{r}_i-\boldsymbol{r}_j|^2}{2\sigma^2}\right) & \forall \boldsymbol{r}_j \in \mathcal{D}_i \\ 0 & \text{otherwise} \end{cases} \quad , \tag{6.1}$$

wherein $\mathcal{D}i$ denotes the collection of elements situated within the vicinity, defined by the filter radius $r_{\min}$, from element i. The weights w_{ij} correspond to the weights in the filtering kernel. The standard deviation, σ, is set as $\sigma = {}^{r_{\min}}/\sqrt{3}$, mirroring the behavior observed in the widely utilized cone filter [139], with a chosen minimum feature size of $r_{\min} = 100\,\text{nm}$. To obtain the projected density $\hat{\boldsymbol{\rho}}$, we use the standard projection scheme presented in Eq. (2.25), where we set the parameters $\alpha = 30$ and $\beta = 0.5$.

Following the application of the filtering and projection, the resulting design densities are subjected to linear interpolation to determine the permittivities and Young's modulus necessary for electromagnetic and structural analyses, respectively. This is mathematically represented for the permittivities as:

$$\boldsymbol{\epsilon}_r = \epsilon_{\min} + \hat{\boldsymbol{\rho}}\left(\epsilon_{\max} - \epsilon_{\min}\right) \quad , \tag{6.2}$$

where $\boldsymbol{\epsilon}_r$ represents the relative permittivity distribution used in the electromagnetic simulations. Similarly, for Young's modulus, the relationship is given by:

$$\boldsymbol{Y} = Y_{\min} + \hat{\boldsymbol{\rho}}\left(Y_{\max} - Y_{\min}\right) \quad , \tag{6.3}$$

where $\boldsymbol{Y}$ denotes the spatial distribution of Young's modulus utilized in the structural analysis. It is important to note that, aside from the final material interpolation step, all other aspects of the parameterization remain the same for both simulations. This ensures that a single design density leads to corresponding material distributions for both the electromagnetic and structural simulations, thus avoiding discrepancies in geometry between the two types of analyses.

In the electromagnetic optimization, the optical response of the structure is analyzed through simulations conducted using the finite-difference frequency-domain (FDFD) method, as implemented in [140]. The goal of the optimization is to minimize an electromagnetic objective, denoted as $F_{\text{EM}}(\boldsymbol{\rho})$, with regard to the design variables $\boldsymbol{\rho}$. This can be formulated as follows:

$$\min_{\boldsymbol{\rho}} \quad F_{\text{EM}}(\boldsymbol{\rho}) \tag{6.4a}$$

$$\text{s.t.} \quad \nabla \times \frac{1}{\mu_0}\nabla \times \boldsymbol{E} - \omega^2\mu_0\epsilon_0\boldsymbol{\epsilon}_r(\boldsymbol{\rho})\boldsymbol{E} = -i\omega\,\boldsymbol{j} \tag{6.4b}$$

$$0 \leq \boldsymbol{\rho} \leq 1 \quad , \tag{6.4c}$$

where $\boldsymbol{E}$ represents the electric field, $\boldsymbol{j}$ stands for the electromagnetic current source, and $\boldsymbol{\epsilon}_r$ is the relative permittivity derived from the design variables. The analysis assumes a uniform relative permeability $\mu_r = 1$ across the entire domain. The definition of the electromagnetic objective function $F_{\mathrm{EM}}(\boldsymbol{\rho})$ varies depending on the specific requirements of the optimization problem. The gradients of this objective function with respect to the design variables are calculated using adjoint sensitivity analysis, as detailed in [16, 141].

The structural optimization focuses on minimizing compliance and is executed through a custom implementation of the direct stiffness method, as highlighted in Sigmund [142]. The formulation of this optimization problem is presented as follows:

$$\begin{aligned}
\min_{\boldsymbol{\rho}} \quad & F_{\mathrm{C}}(\boldsymbol{\rho}) = \boldsymbol{U}^{\top}\mathbf{K}\boldsymbol{U} = \sum_{e=1}^{N} Y_e(\rho_e)\boldsymbol{u}_e^{\top}\mathbf{k}_0\boldsymbol{u}_e & (6.5\text{a}) \\
\text{s.t.} \quad & \mathbf{K}\boldsymbol{U} = \boldsymbol{F} & (6.5\text{b}) \\
& 0 \leq \boldsymbol{\rho} \leq 1 \quad , & (6.5\text{c})
\end{aligned}$$

where $F_{\mathrm{C}}(\boldsymbol{\rho})$ signifies the compliance objective, $\boldsymbol{K}$ denotes the global stiffness matrix, $\boldsymbol{U}$ is the global displacement vector, $\boldsymbol{k}_0$ represents the stiffness matrix for a unit element, $\boldsymbol{u}_e$ is the displacement vector for an element, and $\boldsymbol{F}$ symbolizes the external mechanical force vector applied. The Young's modulus $Y_e(\rho_e)$, which is dependent on the design variable, specifies the stiffness for each of the N elements.

While a volume constraint, which limits the material volume in the design domain Ω to not exceed a predetermined volume fraction V, is commonly applied in such optimizations, it is not enforced in this scenario. This decision is made to avoid unnecessarily limiting the design space for the optical optimization, as the electromagnetic design objectives naturally prevent trivial solutions. Furthermore, volume constraints are typically not a concern in developing nanophotonic devices.

Knowledge of compliance sensitivity concerning the design variables is crucial for effective topology optimization. This sensitivity is obtained as:

$$\frac{\partial F_{\mathrm{C}}}{\partial \rho_e} = -\frac{\partial \hat{\rho} e}{\partial \rho_e} \times (Y_{\max} - Y_{\min}) \times \boldsymbol{u}_e^{\top}\boldsymbol{k}_0\boldsymbol{u}_e \quad , \tag{6.6}$$

where $\frac{\partial \hat{\rho}_e}{\partial \rho_e}$ is determined by the selected filtering and projection method.

The structural and electromagnetic problems are discretized identically for a cohesive approach to optimization. This allows for using a unified set of design variables to define the material geometry for both simulations without the need for spatial interpolation on two disparate grids. This strategy enables the combination of both the compliance and optical objectives into a singular objective function, thereby creating a comprehensive optimization problem. This is represented as:

$$
\begin{aligned}
\min_{\boldsymbol{\rho}} \quad & F(\boldsymbol{\rho}) = (1 - \omega_\mathrm{C})\, F_\mathrm{EM}(\boldsymbol{\rho}) - \omega_\mathrm{C}\, F_\mathrm{C}(\boldsymbol{\rho}) + F_\mathrm{B}(\boldsymbol{\rho}) && \text{(6.7a)}\\
\text{s.t.} \quad & \boldsymbol{\nabla} \times \frac{1}{\mu_0} \boldsymbol{\nabla} \times \boldsymbol{E} - \omega^2 \mu_0 \epsilon_0 \boldsymbol{\epsilon}_r(\boldsymbol{\rho}) \boldsymbol{E} = -i\omega\, \boldsymbol{j} && \text{(6.7b)}\\
& \mathrm{K}(\boldsymbol{\rho})\, \boldsymbol{U} = \boldsymbol{F} && \text{(6.7c)}\\
& \mathbf{0} \le \boldsymbol{\rho} \le \mathbf{1} \quad , && \text{(6.7d)}
\end{aligned}
$$

where $\omega_\mathrm{C} \in [0, 1]$ serves as the compliance factor, allowing for the adjustment of the emphasis placed on each term within the objective function. This flexibility in prioritization facilitates the balancing of structural and electromagnetic performance criteria within the optimization process, enabling the achievement of a design that meets specified requirements in both domains.

Furthermore, to enhance the binarization of the structure, we include a binarization penalty $F_\mathrm{B}(\boldsymbol{\rho})$. This metric is formulated as:

$$
F_\mathrm{B}(\boldsymbol{\rho}) = \min\left(-\log\left(\frac{\sum_{i=1}^{n} 4\hat{\rho}_i(1 - \hat{\rho}_i)}{n}\right), \gamma\right) \tag{6.8}
$$

and is introduced late in the optimization process – specifically when the optimization is nearing convergence, indicated by the relative change in the total figure of merit dropping below a certain threshold (*e.g.*, 1×10^{-3}). The γ parameter caps the binarization objective's maximum value, with a value of $\gamma = 2$ found effective for achieving desirable binarization while maintaining numerical stability.

The calculation of gradients for each term within the objective function and those related to the material parameterization is facilitated by automatic differentiation [95]. This methodological choice permits modifications to the objective function and material parameterization without requiring the manual derivation of new gradients for each alteration.

While the resultant material distributions are inclined towards connected features, they embody a compromise between the mechanical and optical optimization goals, hence not unequivocally precluding the presence of isolated elements. To address this, a multi-step optimization strategy is employed: initially, the optimization is executed until convergence; after this, a post-processing step involving connected component labeling [15] eliminates any isolated elements from the material distribution. The refined material distribution is then subjected to a second optimization phase, culminating in final designs exclusively composed of connected material structures.

6.3 Example applications

Here, we showcase the application of the previously detailed methodology to the inverse design of two distinct nanophotonic devices. By systematically varying the compliance factor ω_{C} within the range $[0, 1]$, we conduct multiple optimizations for each device. Here, a compliance factor of 0 signifies an optimization exclusively focusing on electromagnetic topology optimization (serving as our baseline), whereas a value of 1 entirely omits the optical design objective, emphasizing structural considerations. The structural simulation's driving forces are calibrated such that the compliance's lower boundary aligns in magnitude with the electromagnetic figure of merit $F_{\mathrm{EM}}(\boldsymbol{\rho})$ from the baseline optimization at $\omega_{\mathrm{C}} = 0$. The chosen parameters for structural analysis are a minimum element stiffness of $Y_{\min} = 1 \times 10^{-6}$ and a maximum stiffness of $Y_{\max} = 1$. For electromagnetic simulations, the selected operational wavelength is $\lambda = 1\,\mu\mathrm{m}$, with potential materials being polymers, typical for 3D laser nanolithography with a refractive index of $n = 1.5$, and air with $n = 1$. The simulation domain extends over $12 \times 8\,\mu\mathrm{m}^2$, with a resolution of $30\,\mu\mathrm{m}^{-1}$ and is surrounded by perfectly matched layers (PML) of $1\,\mu\mathrm{m}$ thickness on each side. For simplicity, this investigation is confined to 2D simulations, although it is noted that the underlying equations remain applicable in 3D scenarios.

The primary aim here is the inverse design of functional photonic elements that also possess structural integrity, thus making them suitable for production via additive manufacturing techniques. While the electromagnetic simulation sources are predefined based on each device's requirements, defining structural loads can be less straightforward. Considering that mechanical forces typically do not significantly impact nanophotonic devices, and no load-bearing optical elements are recognized, the optimization's main target is the creation of connected structures. To achieve this, fictitious forces may be employed to steer the optimization towards designs with connectivity, underscoring the method's adaptability to arbitrary loading scenarios, including real physical forces.

These simulations are implemented using Python, with the L-BFGS-B [143] algorithm from the nlopt [14] library employed for local gradient-based optimization. When run on a computer equipped with an Intel Core i7-7700 processor, each optimization process is estimated to converge within approximately 30 minutes.

6.3.1 Photonic nanolens

In our first demonstration, we target the inverse design of a photonic nanolens by employing our optimization framework. The configuration for this design challenge is visualized in Fig. 6.1, delineating the setup for optimizing a nanophotonic focusing element. The photonic nanolens comprises two sections of solid material flanking a central design area, destined for the optimized material layout. A plane wave, linearly polarized perpendicular to the depicted plane and incident normally on the design's upper boundary, illuminates the structure. The optical design objective focuses on concentrating this light into a small focal spot beneath the design zone.

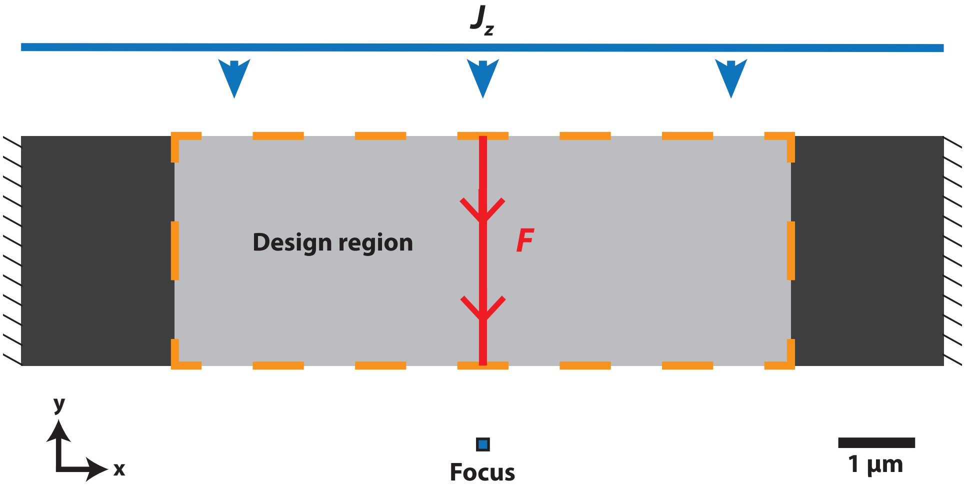

Figure 6.1: Illustration of the optimization setup for a photonic focusing element. The design domain, outlined by the dashed red rectangle, spans an area of 8 × 3 µm^2, nestled between two solid slabs of material (n = 1.5) secured at both ends of the simulation domain and enveloped by air (n = 1). An E_z-polarized plane wave source, with an operational wavelength of 1 µm, is positioned at the domain's upper boundary to illuminate the structure. The objective in the electromagnetic design is to maximize the electric field intensity at a focal spot situated 1.5 µm below the structure's lower edge. The structural challenge is articulated as reducing the material's compliance within the design region under a vertically applied load at its center. Adapted with permission from [P1]. Copyright 2020 American Chemical Society.

We define the electromagnetic figure of merit, F_{EM}, as the intensity ratio of the electric field E_z within the focal region $\mathcal{M}$ to that within the design region $\mathcal{D}$:

$$F_{\text{EM}} = \frac{\int_{\mathcal{M}} |E_z|^2 \mathrm{d}A}{\int_{\mathcal{D}} |E_z|^2 \mathrm{d}A} \quad , \tag{6.9}$$

where the division by the field intensity in the design area $\mathcal{D}$ aims to discourage designs that leverage intense field amplification due to resonances within the design space, as such configurations are highly sensitive to manufacturing inaccuracies. Prior studies, such as [8], have illustrated that optimizing for elevated electric field intensities within a compact spatial region, specifically a square measuring 60 × 60 nm^2, effectively yields sharp focal points in inverse design endeavors.

For the design's structural stability, the photonic structure is presumed to be mechanically secured at both the left and right edges of the domain. This could be practically achieved by incorporating structural support pillars, to which the end blocks are connected and positioned outside the confines of the simulation area. A vertical load is applied along the centerline of the design region, representing the weight exerted by the lens element itself. The objective of the structural optimization is to enhance the device's rigidity in response to this vertical load. Achieving optimal stiffness necessitates the device's anchorage to the solid material at the domain's extremities.

Although framed as a hypothetical scenario, it is important to recognize that when considering rotationally symmetric configurations, this setup bears a resemblance to the design of free-form fiber coupling elements and microlens systems as explored in previous research [144, 145, 146, 147]. These studies highlight the practical applications of such optimization problems in developing photonic devices that meet specific optical performance criteria and are structurally sound.

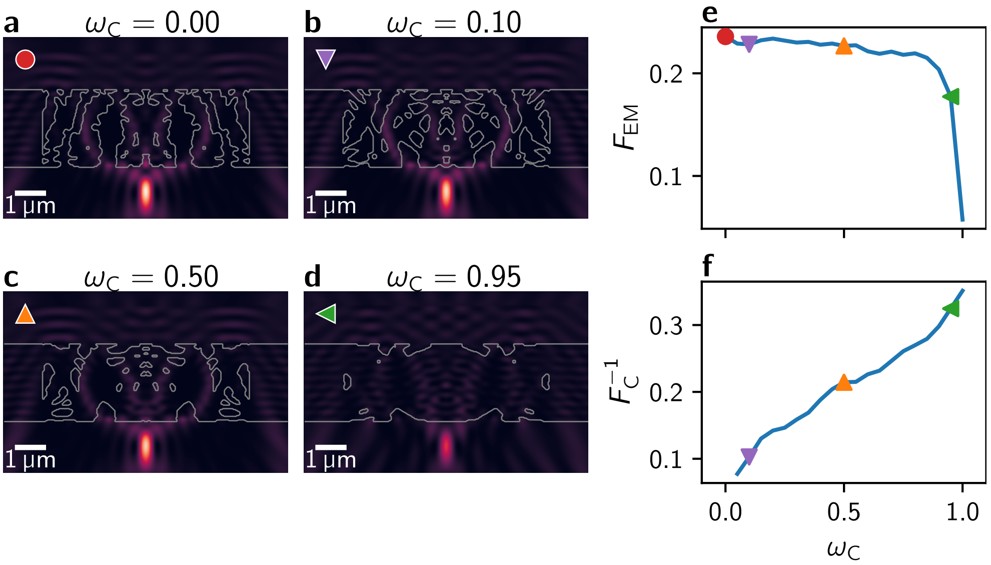

Figure 6.2: Optimization outcomes for a photonic nanolens. Panels **a–d** depict the evolution of optimized lens configurations, with a contour of each design being presented alongside the normalized intensity of the electric field, $|E_z|^2$, for compliance factors ω_C of 0.0, 0.2, 0.5 and 0.9. Panel **e** visualizes the electromagnetic figure of merit, F_{EM}, and panel **f** the structural stiffness, F_C, across a range of compliance factors from 0 to 1 for the optimized designs. The specific designs shown in panels **a–d** are marked on the graph. It is important to note that the values for structural stiffness at $\omega_C = 0$ and for the electromagnetic figure of merit at $\omega_C = 1$ are excluded, reflecting the focus on solely optical or mechanical optimizations at these particular compliance factors, respectively. Adapted with permission from [P1]. Copyright 2020 American Chemical Society.

The initial design, showcased in Fig. 6.2(a) for a compliance factor of $\omega_C = 0$ presents a configuration with several disconnected elements within the design space. Although this purely electromagnetic topology optimization yields a lens with an efficient focal point, the design is not feasible for production through 3D laser nanolithography due to its disjoint nature. However, the refined designs illustrated in Fig. 6.2(b)–(d), corresponding to higher compliance factors, demonstrate progressively more interconnected structures with enhanced stiffness. These designs, devoid of any isolated elements, closely emulate the field profile observed in the baseline design. As the compliance factor increases, prioritizing structural considerations leads to a noticeable decline in focal region intensity.

The illustration in Fig. 6.2(e) plots the electromagnetic figure of merit $F_{EM}(\boldsymbol{\rho})$, and Fig. 6.2(f) the inverse of material compliance $F_C^{-1}(\boldsymbol{\rho})$ for the generated designs, revealing that a higher compliance factor tends to compromise the optical performance objectives. Nonetheless, the design optimized at a compliance factor of $\omega_C = 0.2$ nearly matches the base design's field intensity, falling short by only 0.8 %, while forming a coherent, single-structure design. This example demonstrates that the proposed design methodology can yield photonic nanolenses that are structurally sound and perform nearly optimally.

6.3.2 Mode converter

Mode converters are a fundamental example of integrated photonic devices [148], widely utilized in on-chip optical systems [21, 22]. The capacity to craft such devices for free-form geometries could enable their integration into, for example, photonic wire bonds [113, 116]. However, a critical design criterion for such applications is eliminating unattached elements within the device layout.

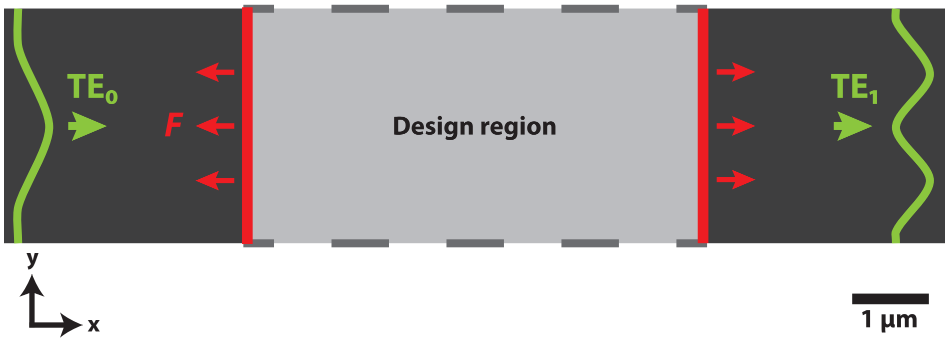

Figure 6.3: Configuration for waveguide mode converter optimization. Encapsulated within a dashed red rectangle, the design area spans $4 \times 6\,\mu\text{m}^2$ and represents a section of a waveguide with a width of $3\,\mu\text{m}$ (refractive index $n = 1.5$) surrounded by air ($n = 1$). The task involves injecting the fundamental TE$_0$ mode at an operational wavelength of $1\,\mu\text{m}$ from the waveguide's left side. The aim of the electromagnetic design is to enhance the mode coupling with the TE$_1$ mode toward the waveguide's right terminus. Structurally, the challenge is to minimize compliance within the design space under the influence of opposing forces applied at its lateral boundaries, utilizing a material with a maximum stiffness ($Y_{\text{max}} = 1$). Adapted with permission from [P1]. Copyright 2020 American Chemical Society.

In this subsequent case study, we design a mode converter that facilitates the transition from the fundamental to the second-order TE mode within a waveguide, as showcased in the depicted optimization setup in Fig. 6.3. The fundamental TE mode is excited at the waveguide's entrance, located on the left side of the design space. Our goal is to maximize the coupling between the outgoing electric field $\boldsymbol{E}$ and the targeted second-order TE mode field $\boldsymbol{E}_{\text{TE1}}$, as measured across the surface S of a field monitor $\mathcal{M}$ positioned on the simulation's right side. This objective is quantitatively captured by the electromagnetic figure of merit F_{EM}, calculated as:

$$F_{\text{EM}} = \left| \int_{\mathcal{M}} \boldsymbol{E}^*_{\text{TE1}} \boldsymbol{E}\, \mathrm{d}S \right| \quad . \tag{6.10}$$

The optimization process employs the finite-difference frequency-domain (FDFD) method to simulate and iteratively refine the design. Subsequently, the finite-difference time-domain (FDTD) method is utilized to assess the mode conversion efficiency [3] precisely.

The initial design, as showcased in Fig. 6.4(a), manages to achieve a mode conversion efficiency of 98.6 %. This design, however, includes multiple disconnected elements, notably a significant ridge-like feature that channels the field through the bottom half of the design space, rendering it unsuitable for incorporation into a free-form waveguide. As the compliance factor is incremented, we observe a transition towards designs that exhibit

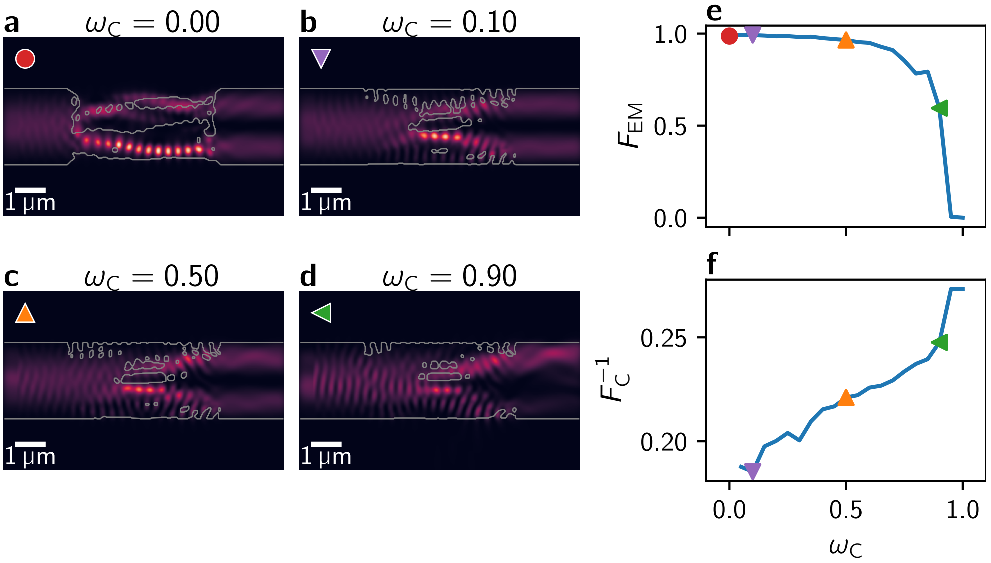

Figure 6.4: Results of the mode converter optimization. Panels **a–d** show the evolution of design configurations alongside their associated normalized field intensities $|E_z|^2$, for varying compliance factors of ω_C values 0.0, 0.2, 0.5 and 0.9. A contour of the design is overlaid on top of the fields. Panel **e** shows the electromagnetic figure of merit F_{EM}, with the structural robustness F_C show in panel **f** across a continuum of compliance factors from 0 to 1. The designs in **a–d** are highlighted with markers. It is noted that the F_C values at $\omega_C = 0$ and the F_{EM} values at $\omega_C = 1$ are excluded as the optimization targets are exclusively optical or mechanical at these specific factors, respectively. Adapted with permission from [P1]. Copyright 2020 American Chemical Society.

enhanced structural cohesion (Fig. 6.4(b)–(d)), though this improvement often comes at the expense of mode coupling efficiency, particularly at higher compliance values as illustrated in Fig. 6.4(e)–(f).

Interestingly, for designs with compliance factors up to 0.3, the coupling efficiency remains on par with or even surpasses the baseline design, peaking at a 99.2 % efficiency at a compliance factor of 0.1. This finding suggests a scenario where structural enhancement does not detract from optical performance up to a certain threshold. Thus, a mode converter design that maintains connectivity throughout its structure, such as the one presented in Fig. 6.4(b), could be fabricated using additive manufacturing processes, marrying structural solidity with high optical performance.

6.4 Conclusion & outlook

In this chapter, we introduced a topology optimization framework that seamlessly integrates optical and mechanical design goals. Through this approach, we successfully demonstrated the feasibility of generating well-structured devices amenable to additive manufacturing processes by factoring in mechanical constraints within the optimization phase. Notably, the integration of mechanical objectives does not inherently compromise the optical performance metrics and, in some instances, can even uncover superior solutions compared to approaches focused solely on electromagnetic topology optimization.

Although our exploration was limited to two-dimensional constructs for ease of demonstration, extending the methodology to three-dimensional geometries is conceptually direct, and even necessary. We did not explore strategies for enhancing manufacturing tolerance, such as erosion-dilation [149, 150] techniques commonly referenced in the literature, yet these could be readily incorporated into this framework.

One possible improvement of this work is tackling the somewhat arbitrary selection of forces applied to the structures under consideration. The process, as described, involves speculatively determining forces that would lead to a desired structural configuration, an approach that might not be universally insightful or practically applicable. A shift towards a more physically grounded methodology, such as incorporating self-weight under gravity into the optimization process, presents a more tangible and realistic strategy. This approach would aim to optimize the structure to be self-supporting under the influence of gravity, which is a natural and omnipresent force, as opposed to relying on artificial or "magic" forces.

This approach of applying body forces is complicated somewhat by the fact that these forces dynamically evolve with each iteration of the optimization. Specifically, the challenge lies in accounting for the derivatives of the objective function concerning changes in these forces, necessitating differentiation with respect to the right-hand side of the linear system governing the optimization. This requirement diverges from conventional practices in photonic inverse design, where source forces are typically static and unchanging throughout the optimization process. In practice, this could be as simple as incorporating the material density at each point into the total force calculation for each optimization iteration. This methodological pivot towards incorporating physical forces like gravity not only enriches the optimization process with more realistic constraints but also potentially simplifies the decision-making process regarding the application of external forces, aligning the optimization closer to practical and physically achievable designs.

The project embarked with the primary aim of incorporating connectivity into photonic devices, addressing a common issue in photonic inverse design where the resultant structures often feature disconnected segments, such as Bragg gratings. The incorporation of mechanical optimization, applying forces merely served as a mechanism to foster connectivity, without a genuine emphasis on ensuring mechanical robustness beyond the requirement for connectivity.

An intriguing alternative to achieving structural coherence could involve employing a thermal solver in lieu of the mechanical framework. This method would entail designing the device to produce heat, which is then dissipated at the simulation boundaries. The optimization goal is to minimize the device's overall temperature. Theoretically, unconnected elements would result in excess heat due to inadequate dissipation pathways, leading to their exclusion from the optimization process. Additionally, by simulating voids as heat generators in a parallel scenario, the approach could facilitate the formation of connected material domains alongside connected voids, thereby eliminating enclosed voids. Such a configuration is advantageous for manufacturing processes, particularly when considering the removal of unpolymerized resist from within enclosed voids.

This thermal-based strategy for ensuring connectivity is not novel, as highlighted by Li *et al.* [151], who applied a similar concept for structural topology optimization. Their work, however, focused solely on the connectivity of voids, given that structural integrity is inherently maintained to prevent the stiffness matrix from becoming singular. The challenge within optical design encompasses ensuring connectivity across both the material and voids, necessitating a nuanced optimization formulation that diverges slightly from the structural domain. This approach underscores the adaptability and innovation required in optimizing photonic devices. It moves beyond traditional methods to explore new avenues for achieving desired structural and functional outcomes.

Deciding whether to incorporate these connectivity and physical considerations as penalties within the objective function or as explicit constraints in the optimization problem presents another avenue of possible future investigation. In this work, we have included them as terms in the objective function for ease of implementation and flexibility; however, it only approximates the desired outcome without guaranteeing complete connectivity. On the other hand, formulating these considerations as constraints could provide more definitive assurances regarding the structure's connectivity but would complicate the optimization process, necessitating careful normalization and potentially limiting the choice of optimization algorithms. Establishing a benchmark temperature for a fully connected design could provide a clear constraint target for thermal constraints. In contrast, defining a reference compliance for self-weight considerations in a mechanical context is less straightforward and would require further deliberation.

In conclusion, the methods outlined in this chapter combine electromagnetic with structural topology optimization, thereby enabling the concurrent optimization of functional photonic devices' optical and mechanical properties. This enables the design of devices that rely critically on their structural and optical characteristics.

7 Machine learning-based surrogate solvers for simulation and inverse design

After a short introduction in Section 7.1, this chapter will explore the use of machine learning, specifically in the form of neural networks trained as surrogate models, for solving Maxwell's equations. We will present an overview of different model architectures in Section 7.2, their training in Section 7.3, and their application to model electromagnetic scattering problems in Subsection 7.4.1. Subsequently, the focus shifts to applying these trained models for the inverse design of scatterers, detailed in Section 7.5. The chapter concludes by reflecting on the overall viability of such models and envisaging potential future directions in the field, as outlined in Section 7.6. The insights and findings delineated in this chapter are primarily derived from [P3].

7.1 Introduction

In nanophotonics, tools for solving Maxwell's equations effectively play a pivotal role in simulating the interaction between light and matter at the wavelength scale and, consequently, in the innovation of novel optical devices. Prominent methods are the finite element method (FEM) [4, 5] in the frequency domain and the finite difference method (FDM) in the frequency (FDFD)[152, 153] and time domains (FDTD)[3, 154, 2], respectively. Solvers based on these methods only discretize Maxwell's equations suitably but do not impose additional approximations. Thus, these techniques constitute the most detailed and precise category of tools for electromagnetic system simulations. Nonetheless, achieving full-wave solutions frequently entails substantial computational resources and time, thereby setting practical constraints on the complexity of problems that can be addressed. This challenge becomes even more pronounced in scenarios like inverse design [6, 25, 155, P1, 39], where it is expected to undertake hundreds of simulations to derive satisfactory outcomes. Although there is continuous progress in developing more efficient full-wave solvers [156, 157], alternative semi-analytical methods [158, 159, 160, 161] exist, offering significant speed improvements by making specific physical presumptions, thus limiting their scope to particular types of problems.

A relatively new strategy involves the application of machine learning-based surrogate models [162, 163, 164, 165] for approximating solutions to partial differential equations

(PDEs). These models have the potential to surpass semi-analytical methods in speed during inference. They can be adapted for a broad spectrum of problems due to their inherent capability as universal function approximators [166, 164]. Recently, the integration of machine learning in nanophotonics has been accelerating [167, 168], with surrogate solvers finding applications in both forward modeling and inverse design tasks [169, 170, 171, 172, 103]. However, these models introduce their own challenges, notably regarding accuracy and data efficiency. The term “data efficiency” refers to the capability of using less data to achieve more, which usually inversely correlates with accuracy. A model can be trained to yield more accurate predictions and vice versa by utilizing more data. Given that high-quality training data is typically produced using traditional methods, like full-wave solvers, the balance between accuracy and data generation costs becomes a critical factor in employing surrogate solvers, as their advantages diminish when the data production becomes a bottleneck. Therefore, it is paramount to enhance the data efficiency of such models and identify scenarios where surrogate solvers can significantly reduce data generation costs.

A promising approach to enhancing the data efficiency of surrogate solvers is to leverage advancements in scientific machine learning by introducing models endowed with partial physical insights. This can be achieved by either directly incorporating governing equations into the model, as seen with physics-informed neural networks (PINNs) [173, 174], or by penalizing solutions that deviate from physical principles during the training process [102, 103], which can be broadly applied given known governing equations. Alongside embedding physics into machine learning, the exploration of novel model architectures has been crucial. Notably, models that learn operator mappings between function spaces, such as graph kernel networks (GKN) [175], deep operator networks (DeepONets) [164, 176], and Fourier Neural Operators (FNO) [177], have shown superior performance in solving PDE-constrained problems compared to their predecessors.

In this chapter, we develop an enhanced version of the FNO to act as a surrogate solver for electromagnetic scattering problems [P3]. Our model, trained on a varied dataset of electromagnetic scatterers, demonstrates improved data efficiency and accuracy compared to the conventional convolutional network architecture (UNet) in a two-dimensional setting. Further, we employ this model in the computationally intensive task of gradient-based inverse design of three-dimensional free-form scatterers starting from various initial conditions, showcasing the practical advantages of this approach over traditional full-wave solvers.

7.2 Neural network architectures

This section offers an introductory overview of the neural network architectures, which will be discussed in subsequent sections. It aims to outline the fundamental aspects of each architecture without engaging in an exhaustive exploration of their intricacies. For those interested in a more detailed understanding, references to relevant literature will be provided for further reading.

7.2.1 Convolutional neural networks

Convolutional Neural Networks (CNNs) are a fundamental part of deep learning and are used for solving problems in areas such as computer vision and image processing. They trace their origins back to pioneering work by LeCun *et al.* [178]. This early application of CNNs was notable for utilizing backpropagation for gradient descent, enabling the network to iteratively learn optimal filter weights directly from the input data. The later resurgence of interest in CNNs was catalyzed by the advent of the AlexNet architecture [179], which outperformed traditional computer vision methods by a large margin. This success highlighted the power of deep learning for image processing tasks, leading to a rapid evolution in CNN architectures. The introduction of the UNet [180] in 2015 marked another milestone, featuring several architectural innovations, such as its unique symmetric structure and skip connections that facilitate precise localization by preserving spatial information lost during downsampling. The general layout of a UNet architecture is shown in Fig. 7.1. This efficient design allowed it to excel on small datasets, making it a versatile and powerful tool for various applications beyond computer vision.

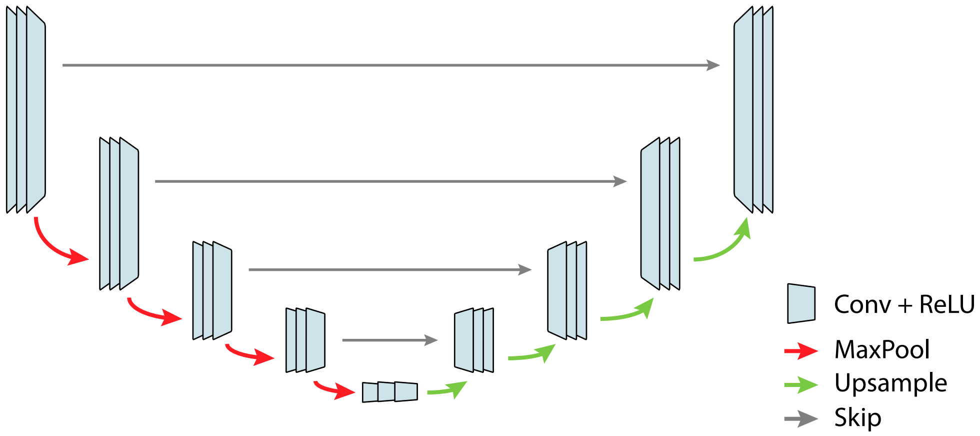

Figure 7.1: Illustration of the UNet architecture, showcasing its symmetrical structure with convolutional layers with ReLU activation (depicted as light blue boxes), followed by downsampling via maximum pooling (red arrows). After four levels of downsampling, the layers are symmetrically upsampled again, indicated by green arrows. Layers on opposite sides of the network are connected via skip connections.

Various applications in photonics leverage UNets, notably in approximating solutions to Maxwell's equations, thus serving as surrogate solvers as an alternative to full-wave solvers [181, 171]. Typically, these deep learning architectures ingest a grid-discretized distribution of physical properties, such as permittivity, akin to an image format. Their objective is to estimate specific electromagnetic field components. This estimation is accomplished through supervised learning, employing a dataset of scatterers and their resulting electromagnetic fields after illumination generated via traditional solvers, which facilitates the training of these networks. Acknowledging that these networks do not genuinely "solve" Maxwell's equations is crucial since they do not understand the underlying physical principles. Instead, they process the input similarly to an image, employing convolutions and nonlinear transformations to approximate the desired field distribution.

CNNs distinguish themselves by employing *local* convolutions on the input. In these operations, convolution kernels interact solely with a pixel's immediate vicinity in a sliding window manner, termed the *receptive field.* This field usually spans a width of 3 to 7 pixels. The convolution's output is referred to as a *feature map*, with the convolutional kernels' weights, responsible for producing these feature maps, being the degrees of freedom learned during the training phase.

An individual feature map F in a CNN can be formulated as follows:

$$F_{mn}(x) = \sum_{m} \sum_{n} x_{i-m,,j-n} * k_{mn} + b \quad , \tag{7.1}$$

where m and n represent the kernel's (k) dimensions, and i, j denote the input dimensions of the image x. Convolutional layers often include a bias term b, adding an extra parameter.

UNets are renowned for their unique architecture, featuring contracting and expanding paths linked by skip connections. The contracting path is realized through the application of max pooling following convolutional stages, effectively lowering the spatial dimensions of the input by selecting the maximum value from each distinct sub-region of the input feature map, usually within a 2×2 pixel window, as expressed by:

$$\text{MaxPool}(x_{mn}) = \max_{i \in m,, j \in n} x_{ij} \quad , \tag{7.2}$$

where m and n are the dimensions of the receptive field. These dimensions are usually chosen as 2×2, meaning that the spatial dimensions are halved after one such max pooling operation. Max pooling is not considered an actual network layer since it merely maps the input without any trainable parameters. With each downsampling step, the contracting path broadens the convolutional filters' field of view, enabling CNN to identify features across varied scales.

Conversely, the expanding path seeks to restore the input's spatial dimensions, essential for producing an output that matches the input image's size. This restoration is accomplished via feature map upsampling followed by convolutional stages, ensuring symmetry in the UNet architecture to match the downsampling layers' sizes. This symmetry facilitates the integration of skip connections, merging feature maps from the contracting path with their upsampled counterparts, aiding in retaining spatial details that are potentially lost during downsampling. Additionally, this approach can help to improve network training by mitigating the vanishing gradient issue.

UNets have demonstrated remarkable proficiency as surrogate solvers for electromagnetic challenges, establishing themselves as the benchmark architecture against which alternative methodologies are evaluated [102, 103]. The forthcoming subsection will explore an alternative method recently emerging as a promising contender.

7.2.2 Neural operators

Neural operators represent a novel category of neural network designs aimed at learning mappings between infinite-dimensional function spaces. They are specifically tailored to create surrogate models for solving partial differential equations (PDEs). These models are purpose-built for use as surrogate solvers, showing superior performance in solving PDEs over traditional methods like convolutional neural networks across various problem domains [182, 183, 184, 176, 185].

Specifically, neural operators aim to emulate an operator $\mathcal{O} : \mathcal{A} \to \mathcal{B}$ through a parametric mapping $\mathcal{O}(\theta) : \mathcal{A} \to \mathcal{B}$, formulated as

$$\mathcal{O}(\theta) = \mathcal{P} \circ \sigma(W_n + \mathcal{K}_n + b_n) \circ \cdots \circ \sigma(W_1 + \mathcal{K}_1 + b_1) \circ \mathcal{L} \quad , \tag{7.3}$$

where σ denotes a pointwise nonlinearity, W is a local operator typically instantiated by a fully connected layer, $\mathcal{K}$ is a learned kernel integral operator, and b is a bias function. The operators $\mathcal{L}$ and $\mathcal{P}$, serving as lifting and projection functions respectively, elevate the input into a higher-dimensional space – its dimensionality defined by a hyperparameter – before projecting it back to the output's solution space. These functions are often implemented as fully connected layers, with output channel dimensions matching the targeted dimensionality.

The defining feature of different neural operator architectures is the learnable kernel integral operator $\mathcal{K}$, leading to a variety of implementations such as Graph Kernel Operators [175], Fourier Neural Operators [177], DeepONets [164], and Spectral Neural Operators [186], among others. Given the rapid advancements in this area, a thorough exploration of these architectures is best found in existing literature [187, 188, 189]. This chapter will concentrate on employing FNOs for electromagnetic scattering problems.

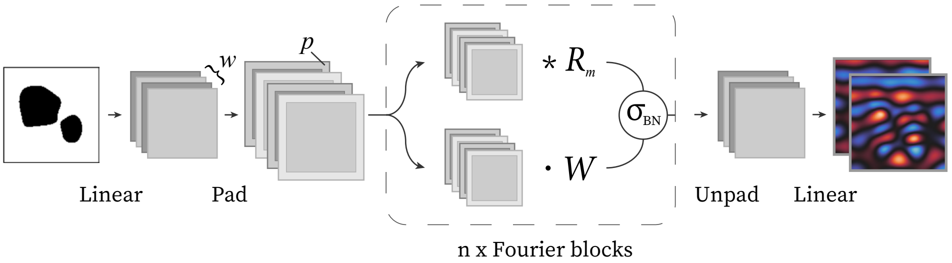

Figure 7.2: Illustration of the Fourier Neural Operator framework. Initially, the input undergoes expansion to w channels through a linear transformation, followed by an addition of zero-padding of p pixels. Subsequently, the data navigates through a sequence of n Fourier blocks, each comprising a linear transformation, Fourier space convolution with the learned kernel R_m, a batch normalization process, and applying a GELU activation function. The process concludes with removing zero-padding and reducing channel dimensions to the required output size via a linear transformation. Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

Here, we will first introduce some of the foundational aspects of the original FNO, as presented in [177], and highlight certain architectural modifications specific to this work. At its essence, the FNO's approach involves the development of a kernel defined in the Fourier domain, enabling each Fourier layer to execute a *global* convolution on the input. The procedure begins by elevating the input $v(x)$ – with x being a point on the computational grid – to a more complex representation $y(x)$ via:

$$y(x) = C_{\text{in}}(v(x)) \in \mathbb{R}^w \quad , \tag{7.4}$$

where $C_{\text{in}} : \mathbb{R}^{d_{\text{in}}} \longrightarrow \mathbb{R}^{w}$ denotes a linear transformation characterized by input dimension d_{in} and output dimension w. The input dimension, d_{in}, is typically limited and pre-defined by the specific problem. In contrast, the output dimension, w, is a hyperparameter known as the FNO's "width", indicating the number of kernels (or feature channels) within each Fourier layer.

Following this dimensional enhancement, the data traverses through a series of n "Fourier blocks". Each block encompasses a Fourier layer $\kappa_m(y)$ combined with a linear transformation, a batch normalization layer as described in [190], and a ReLU activation function σ_{BN}:

$$u(y) = \sigma_{\text{BN}}\left(\mathcal{F}^{-1}(R_m \cdot \mathcal{F}(y)) + Wy + b\right) \quad , \tag{7.5}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ signify the Fourier transform and its inverse, respectively, with R_m, W, and b as the adjustable parameters. The elements W and b constitute the linear layer adjustment, with W being a weight matrix that operates locally on y and b acting as the bias vector. The complex-valued tensor R_m functions as the convolution's kernel matrix in Fourier space. After the Fourier blocks, the transformation y is remapped to the output's required dimensionality d_{out} through a linear layer:

$$z(x) = C_{\text{out}}(y(x)) \in \mathbb{R}^{d_{\text{out}}} \quad . \tag{7.6}$$

A distinctive characteristic of FNOs is the Fourier layer $\kappa_m(y)$, which limits the Fourier series to the m-th coefficient, meaning that R_m comprises entries only up to that frequency mode's index. This selective inclusion acts as a built-in low-pass filter, inherently smoothing out high-frequency spatial details in the network's output, thus producing smoother results based on the selected m value. This feature proves particularly beneficial for modeling physical systems described by partial differential equations (PDEs) that yield wave-like solutions, such as Maxwell's equations, making FNOs especially apt for these scenarios. This inherent smoothing ability obviates the need for external smoothing techniques or loss function adjustments to achieve similar effects, methods that, while effective, add layers of complexity. Nonetheless, loss function modifications to enhance model predictions' consistency with Maxwell's equations, as seen in works like Chen *et al.* [103] and Lim and Psaltis [102], can still be integrated into FNO training if necessary. In Sections 7.3 and 7.4, we aim to outline a performance benchmark for FNOs in tackling electromagnetic scattering problems.

To improve accuracy, traditional FNO implementations usually incorporate a process known as *feature expansion* for the inputs [183, 177, 187]. A typical method involves augmenting the input with a Cartesian coordinate grid, effectively increasing the input dimensionality, for example, from $\mathbb{R}^{d}$ to $\mathbb{R}^{d+3}$ in three-dimensional spaces. Our findings indicate that comparable levels of accuracy can be achieved by applying a modest zero-padding ($p = 2$) to the inputs before their procession through the Fourier blocks, subsequently removing this padding before the network's final layer, thereby circumventing the need for explicit feature expansion. Additionally, our empirical analysis suggests the Gaussian Linear Error Unit (GELU) activation function [191] offers a slight improvement in reducing prediction errors over the traditionally used ReLU. The adapted FNO architecture employed here is visually detailed in Fig. 7.2.

7.2.3 Variational autoencoders

Autoencoders are a neural network architecture designed for unsupervised learning [192]. They primarily focus on encoding input data into a compressed, lower-dimensional representation called the latent space and then reconstructing it as closely as possible to the original input. This process is achieved through two main components: the encoder, which compresses the input into a latent space representation, and the decoder, which attempts to reconstruct the input from this latent representation. By training the network to minimize the reconstruction error, autoencoders learn to capture the most salient features of the data in the latent space. This capability makes them highly versatile, and they find applications in dimensionality reduction, feature learning, and denoising images, among others. Their simplicity and effectiveness in learning data representations without labeled data have made autoencoders a fundamental tool in machine learning and deep learning research.

A common variation of the autoencoder architecture employs convolutional layers, aptly named Convolutional Autoencoders (CAEs), which are particularly effective for tasks involving image data. In a CAE, the encoder comprises convolutional layers to progressively downsample the input image, capturing spatial hierarchies and features in a compressed latent space representation. Pooling layers typically follow these convolutional layers to reduce the spatial dimensions, thereby compressing the essential information. The decoder then uses convolutional layers in reverse, often supplemented by upsampling or transposed convolutional layers, to reconstruct the input image from the latent representation progressively. This configuration enables CAEs to effectively capture spatial correlations within images, aligning them with applications in image compression, denoising, and generative modeling [193]. While the architecture shares some similarities with that of a UNet, a notable distinction lies in the structural symmetry; unlike UNets, the encoder and decoder components of a CAE are not strictly required to mirror each other, provided the output dimension matches the input dimension. Additionally, the encoder and decoder operate more independently within a CAE framework, focusing on their respective tasks of compression and reconstruction without the direct feature-level communication facilitated by skip connections found in UNets.

While effective in learning compact data representations, traditional autoencoders face certain limitations. In particular, autoencoders tend to learn a dense, entangled latent space, making it challenging to interpret the encoded representations or perform controlled manipulations of the generated outputs. These limitations have spurred the development of advanced models like Variational Autoencoders (VAEs) [194] and Generative Adversarial Networks (GANs) [195], which address these issues by introducing probabilistic latent spaces and adversarial training regimes, respectively, to improve the quality and diversity of generated data. Unlike traditional autoencoders, which aim to compress data into a latent space and reconstruct it, VAEs introduce a statistical layer that models the data distribution in the latent space. This approach generates new data points similar to the original input data, making VAEs particularly useful for tasks involving interpolation in the latent space.

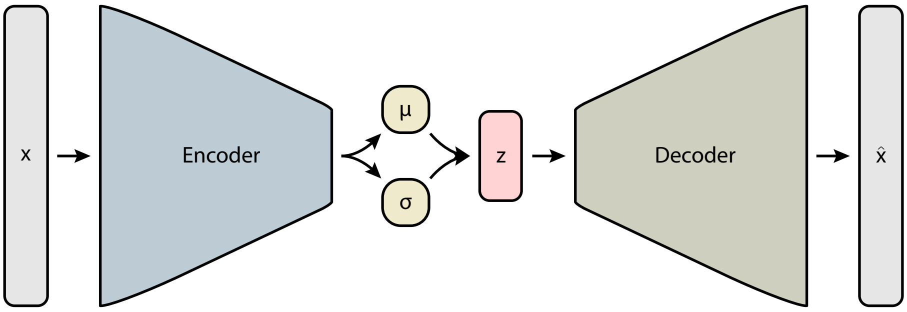

Figure 7.3: Diagram depicting a variational autoencoder architecture, where the input x is processed by the encoder to produce the parameters μ and σ, parameterizing a normal distribution. These parameters are then used to sample a latent vector z, from which the decoder generates $\hat{x}$, approximating the original input.

The architecture of a VAE is similar to that of a standard autoencoder, comprising an encoder, a latent space, and a decoder, shown in Fig. 7.3. The encoder maps the input data to a distribution over the latent space, typically characterized by mean and variance parameters. This distribution is then sampled to produce latent variables, which the decoder uses to reconstruct the input data. A key innovation of VAEs lies in their training process. It involves optimizing the reconstruction loss and a regularization term that encourages the learned distribution in the latent space to approximate some prior distribution, usually a Gaussian. This regularization is termed the Kullback-Leibler (KL) divergence, which ensures that the latent space is well-structured and enables the generation of new data points by sampling from the prior distribution.

For training, VAEs employ a loss function called the “evidence lower bound” (ELBO), which consists of the aforementioned KL divergence and an additional *reconstruction loss* term that quantifies how well the VAE can reconstruct the input in its output.

We will later (in Section 7.5) use such a VAE to parameterize geometries for inverse design using an FNO-based surrogate solver, particularly due to their smooth latent space, making it feasible to interpolate between different geometries freely.

7.3 Data generation & model training

As the training and hyperparameters are crucial in neural networks, we will detail the data generation procedure in Subsection 7.3.1 and discuss the training and validation setup for the surrogate solvers in Subsection 7.3.2 and for the VAE in Subsection 7.3.3, respectively.

7.3.1 Dataset generation

The methodology for creating an extensive array of random scatterer shapes follows the approach detailed in [S1] and is illustrated in Fig. 7.4. We initiate this process by distributing points randomly within the range $[0, 1[$ across a uniformly spaced square

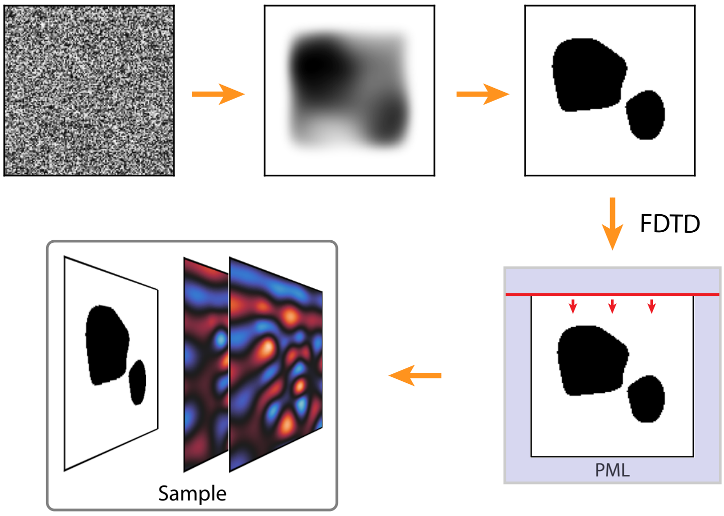

Figure 7.4: Procedure for generating training data. A binary image is initially created by applying smoothing and thresholding techniques to an image derived from a random uniform distribution. This image is subsequently treated as a distribution of materials, which is then simulated under plane-wave illumination using the finite-difference time-domain method. The resulting dataset for training comprises the scatterer (as input) alongside the individual components of the electromagnetic field (as the target). Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

grid (or cubic grid for 3D scenarios) with each side measuring 128 px. Subsequently, the entire grid is subjected to a Gaussian blur with zero-padding ($\sigma = 12\,\text{px}$), followed by a thresholding operation at 0.5. Zero-padding is crucial here to ensure that scatterer formations are confined within the simulation space, avoiding overlap with the domain boundaries. This technique yields smooth, randomly generated geometries, potentially featuring single or multiple scatterers. The intricacy of these generated shapes is primarily influenced by the Gaussian blur's kernel size, akin to the "filtering" strategy employed in topology optimization to define minimum feature sizes [91, 59].

These random geometries are then interpreted as variations in material distribution, with the designation of 1 representing material presence ($n_{\text{high}} = 1.5$) and 0 signifying air ($n_{\text{low}} = 1$). Each sample is illuminated by a plane wave at a wavelength of $\lambda_0 = 1\,\mu\text{m}$, choosing $n_{\text{high}} = 1.5$ to approximate the refractive index of standard polymers utilized in 3D laser nanoprinting. The simulations are conducted at a spatial fidelity of 25 px μm^{-1}, equating to a 5.12 µm span along each axis, with the simulation domain further encapsulated by perfectly matched layers (PMLs) extending 0.5 µm on every side to achieve a total domain length of 6.12 µm. This dataset includes pairs of scatterers with their corresponding steady-state electric fields, excluding the regions affected by PMLs. The generation of these samples, applicable to two-dimensional and three-dimensional datasets, utilizes Meep [3], with a summarized dataset overview provided in Table 7.1.

The 2D dataset encompasses a total of 16384 (2^{14}) training samples, alongside 256 validation samples and 400 test samples. This larger dataset is the source from which smaller 2D datasets, referred to later in Subsection 7.4.1, are selectively extracted. Each 2D sample was simulated on a single core of an Intel Xeon Platinum 8368 CPU, and the generation

Table 7.1: Summary of the datasets of random scattering geometries used for training all models.

Type	Samples	Input shape	Output shape	Output
2D	17040	128×128	$2 \times 128 \times 128$	E_z
3D	8720	$128 \times 128 \times 128$	$6 \times 128 \times 128 \times 128$	E_x, E_y, E_z

of the whole dataset was parallelized over one compute node (2 × 38 cores), with each sample taking a few seconds to simulate.

Conversely, the 3D dataset comprises 8192 training samples, with 128 designated for validation and 400 allocated for testing. The simulation of each 3D sample consumed approximately 20 minutes parallelized across 4 cores of an Intel Xeon Platinum 8368 CPU, albeit with minor variations attributed to the differing complexities of scatterers. The data generation was distributed across 40 nodes within the HoreKa HPC cluster, where each node simultaneously conducted simulations for 19 samples (2 × 38 cores per node, divided by 4 cores per sample). As a result, the comprehensive assembly of the 3D dataset was achieved in just under four hours.

7.3.2 FNO & UNet training

Hyperparameters for the FNO are detailed in Table 7.2. The UNet model has five downsampling (max pooling) stages and five corresponding upsampling segments. Each stage comprises six convolutional layers integrated with batch normalization and ReLU activation functions. Here, we reference the UNet architecture described by Chen *et al.* [103] without implementing it directly. However, we modify their design by substituting the periodic padding within the convolutional layers with zero padding. This adjustment has enhanced the UNet's performance on our dataset by approximately 1 % in all experimental iterations.

We configure the FNO such that its input comprises the distribution of dielectric materials discretized over a regular grid, leading to a single input feature, $d_{\text{in}} = 1$. The desired output is the complex electromagnetic field components, each split into two channels to represent the real and imaginary parts separately. The FNO architectures examined here are fully detailed by a set of hyperparameters listed in Table 7.2.

Table 7.2: FNO architecture hyperparameters with specified values for the networks investigated in this chapter, applicable to both 2D and 3D configurations.

Parameter	Description	Value
n	No. of Fourier blocks	10
m	Truncation order for Fourier modes	12
w	Width (hidden channels)	32
p	Zero-padding per spatial dimension	2

For the purpose of training the models, we construct datasets comprising scatterers and their corresponding fields using a comprehensive full-wave Maxwell solver, details of which are elaborated in Subsection 7.3.1. The evaluation and training of the models are conducted using the normalized L_p loss metric, defined as

$$L_p(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\|y_i - \hat{y}_i\|_p}{\|y_i\|_p} \right) \quad , \tag{7.7}$$

where $\boldsymbol{y}$ is the network's output and $\hat{\boldsymbol{y}}$ the target field distribution. Specifically, the models undergo training employing the normalized L_2 loss, essentially the normalized root mean square error, to refine their accuracy. However, we predominantly reference the normalized L_1 loss for discussions and evaluations due to its straightforward interpretation as the absolute discrepancy between two samples. It is important to note the minor variation between L_1 and L_2 losses in the context of this study, indicating their near-interchangeable applicability. Evaluation relies on a distinct test dataset, consisting of 400 samples not previously utilized in either training or validation phases, to ensure an unbiased assessment of model performance.

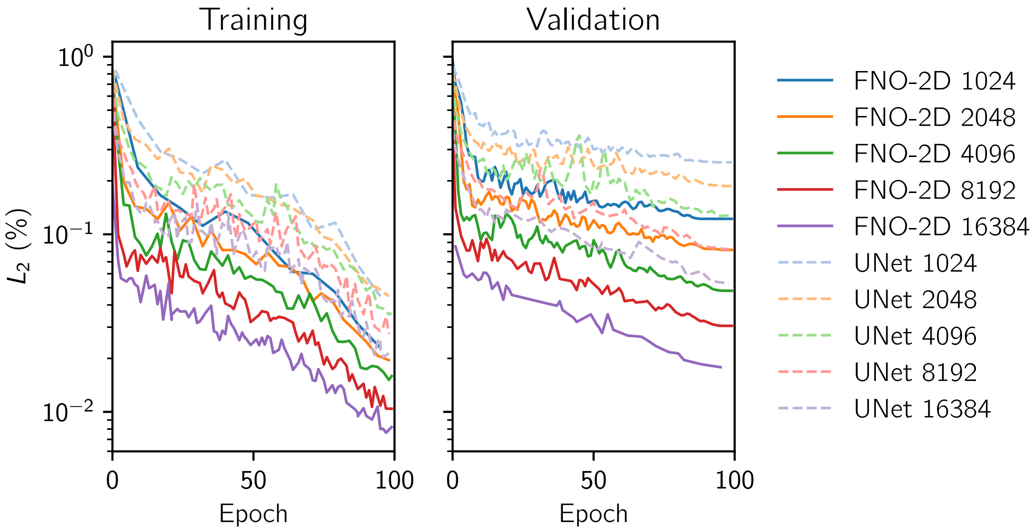

Figure 7.5: Illustration of L_2 loss curves for training and validation phases across all training sessions for UNet and FNO-2D, using various dataset sizes. Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

Each model undergoes training over 100 epochs, leveraging the AdamW optimizer [196] combined with a one-cycle learning rate strategy [197], and uses a minibatch size of 32. The relative L_2 error (refer to Eq. (7.7), $p = 2$) is employed as the loss function for training. In contrast, both relative L_1 and L_2 errors are evaluated during validation after every epoch, utilizing a validation subset comprising 256 samples in 2D. The curves depicting training and validation losses are presented in Fig. 7.5. Training sessions for both FNO-2D and UNet models are conducted on a single NVIDIA A100 SXM4 GPU.

We maintain the same effective batch size for the 2D variants to facilitate the training of the FNO-3D model. Due to the constraints posed by GPU memory capacity, direct training with this batch size is infeasible. Therefore, we adopt a parallel training approach, distributing the workload across two nodes, each equipped with 4 GPUs, wherein each GPU processes only four samples concurrently. To further alleviate memory demands, we employ activation checkpointing techniques [198] within the FNO blocks. This strategy allows for the temporary release of intermediate activations from memory. After each training iteration, gradient values are aggregated and averaged across all participating processes through the utilization of PyTorch's `DistributedDataParallel` framework [199], ensuring coherent and unified model updates. The entire training process for the FNO-3D model spanned approximately 11 hours, with specific details on the training hyperparameters provided in Table 7.3.

Table 7.3: Training hyperparameters used for the surrogate solver models.

Model	Parameters	Learning rate (min / max)	Batch size	Validation split
UNet	23 617 970	5×10^{-5} / 5×10^{-4}	32	256
FNO-2D	5 909 250	1×10^{-3} / 1×10^{-2}	32	256
FNO-3D	141 568 902	1×10^{-3} / 1×10^{-2}	2x4x4*	128

* Trained on two nodes with 4 GPUs each and four samples per GPU (NVIDIA A100 SXM4).

Various approaches were adopted to train the surrogate solver models, catering to the specific requirements and characteristics of the UNet and FNO architectures. A critical aspect of this training process involved adjusting learning rates, given that UNets tended to be unstable when trained at higher learning rates comparable to those suitable for FNO models. Extensive hyperparameter sweeps were conducted to ensure a balanced evaluation and optimize performance, focusing on learning rates and learning rate scheduling strategies specifically for the UNet models. This process aimed to identify the most effective configurations, leading to the selection of the best-performing models for each architecture, as presented in this work.

7.3.3 VAE setup & training

In this work, the Variational Autoencoder (VAE) is structured around a convolutional encoder-decoder architecture tailored explicitly for processing three-dimensional image data. The encoder component features five downsampling blocks designed to condense an input image – represented by a 128 × 128 × 128 grid depicting material distribution – into a compact latent space with 2048 elements. Each downsampling block comprises a 3D convolutional layer, succeeded by batch normalization and SELU activation [200]. These convolutional layers employ a kernel size of 5, stride of 2, and padding of 2, with a channel count that doubles sequentially through the layers. Conversely, the decoder reverses the encoder's structure, substituting convolutional layers with transposed convolutions for upsampling [201]. The entire VAE network comprises a total of 86 433 514 trainable parameters.

For the network's training, we minimize the evidence lower bound (ELBO) [194], augmented by a penalty for binarization, expressed as:

$$L_{\mathrm{VAE}}(x, \hat{x}) = \gamma_{\mathrm{KL}}(L_{\mathrm{KL}}(x, \hat{x}) - L_{\mathrm{R}}(x, \hat{x})) - \gamma_{\mathrm{B}} L_{\mathrm{B}} \quad , \tag{7.8}$$

where x represents the original image, $\hat{x}$ the reconstructed version, L_{KL} the Kullback-Leibler divergence, L_{R} the reconstruction loss, and L_{B} the binarization penalty. These components are detailed as follows:

$$L_{\mathrm{KL}} = \mathbb{E}_q[\log q(z|x) - \log p(z)] \tag{7.9}$$

$$L_{\mathrm{R}} = \mathbb{E}_q[\log p(x|z)] \tag{7.10}$$

$$L_{\mathrm{B}} = \frac{1}{3} \min \left[-\log_{10}\left(\frac{1}{N} \sum_i^N 4x_i(1 - x_i) \right), 3 \right] \quad , \tag{7.11}$$

with the expected value $\mathbb{E}_q$ under q, the normal distribution p, q the encoder-parameterized distribution given an input x, and z being a sample drawn from q. The introduction of annealing parameters γ_{KL} and γ_{B} serves to fine-tune the impact of the L_{KL} and L_{B} components throughout training.

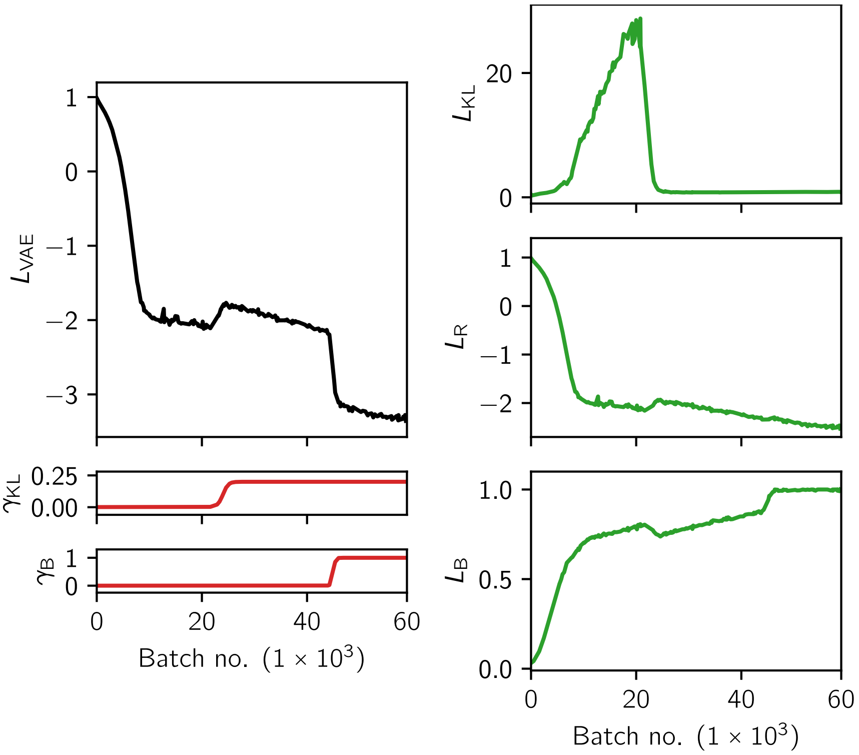

Figure 7.6: Depiction of the total VAE loss L_{VAE} during the training phase, alongside the evolution of the annealing weights γ_{KL} and γ_{B} (left), and the individual contributions from the components L_{KL}, L_{R}, and L_{B} to the overall VAE loss (right). Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

Distinctively, the VAE model does not rely on a pre-existing dataset for training. Instead, it employs the geometry sampling method described in Subsection 7.3.1 to dynamically generate images of random scatterers throughout the training process. This methodology eliminates the traditional concept of training epochs, as each sample is uniquely presented to the network simultaneously, rendering the dataset "infinite". This continuous generation of unseen data samples ensures that the network is consistently exposed to fresh examples, fostering a robust learning environment that mitigates the risk of overfitting and enhances the model's ability to generalize across a broad spectrum of input geometries. Training curves for L_{VAE} as well as all terms in the loss function are visualized in Fig. 7.6. Training is executed with a batch size of 64 over a total of 60 000 batches, culminating in a training duration of 48 hours on a single NVIDIA A100 SXM4 GPU.

7.4 Field inference

After training the surrogate solvers, we will now proceed to use them as a tool for solving electromagnetic scattering problems. In particular, we will systematically compare the performance of FNOs and UNets in Subsection 7.4.1 in 2D. We will then incorporate a physics-driven term in the loss function and examine its effects on network performance in Subsection 7.4.2. This section concludes with a further modification of the FNO architecture via a tensor decomposition, which we will explain in detail in Subsection 7.4.3.

7.4.1 Comparing FNO & UNet performance

In this first subsection, we explore the application of FNOs as surrogate solvers for Maxwell's equations, aiming to replace conventional full-wave methods like the finite-difference time-domain (FDTD) method for specific scattering problems. In our discussions, specific model designations are used to denote their dimensional training context: "FNO-2D" refers to models trained with two-dimensional data, whereas "FNO-3D" identifies those trained on three-dimensional datasets. We simply employ "FNO" for broader insights or principles encompassing both types.

There has been significant interest in using UNet-like convolutional architectures [180] for field inference [171, 103, 102], given their proficiency in handling complex spatial data. In our work, to enable a direct and fair comparison of the FNO's capabilities against UNet for electromagnetic field inference, we adopt the UNet architecture detailed by Chen *et al.* [103] as a benchmark. Both FNO and UNet models are subjected to training using a wide variety of simulated random scatterer datasets in two dimensions, with a specific focus on assessing and comparing their data efficiency – essentially evaluating the number of training samples required by each model to achieve a predefined level of predictive accuracy. This comparative analysis entails utilizing dual output channels for both models to represent the real and imaginary components of the electromagnetic field's z-component, E_z. Uniform training conditions are applied to both models, ensuring any observed performance disparities are attributable solely to the inherent architectural differences rather than external variables, with exact training details as discussed in Subsection 7.3.2.

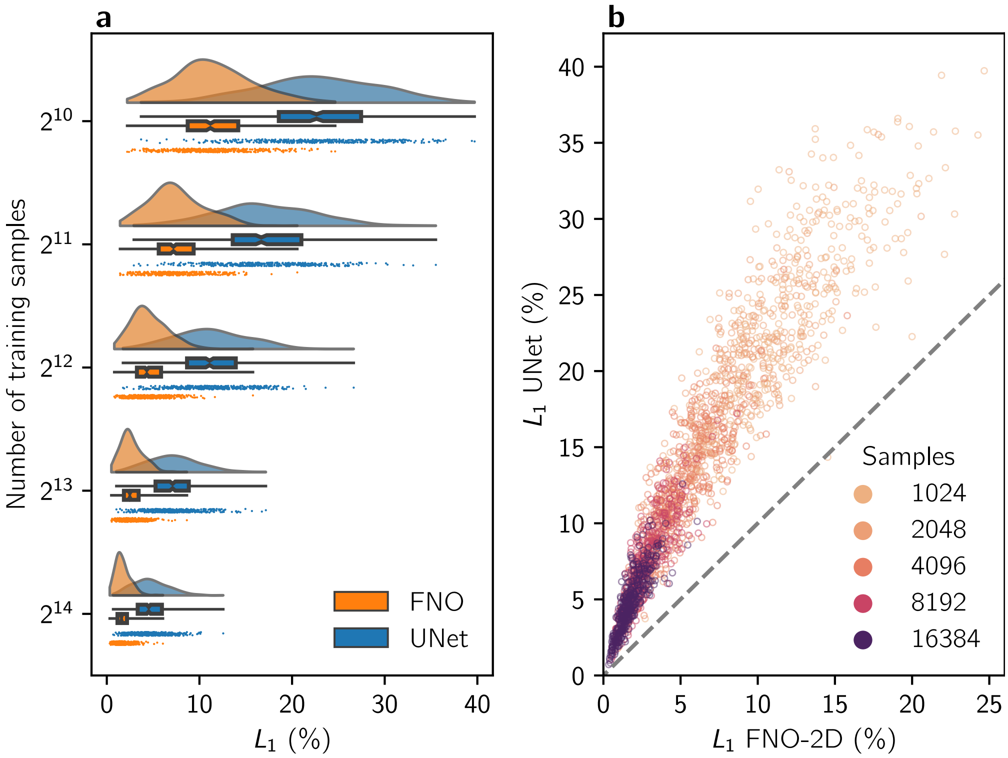

Figure 7.7: **a** Visualization via a raincloud plot [202] comparing normalized L_1 error distributions across a test dataset comprising 400 samples, for both FNO-2D and UNet models across varied sizes of training datasets. For each training sample size, the plot illustrates a kernel density estimate (left, bandwidth determined by Scott's rule [203]), a box plot highlighting the distribution's range and median (center), and a scatter plot displaying individual sample errors (right). **b** A side-by-side scatter plot analysis of the L_1 error for every test sample against the sizes of the training sets for both FNO-2D and UNet, with the dashed gray line demarcating identical error levels between the two models. Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

The evaluation outcomes performed on the test set, consisting of 400 samples, are graphically presented in Fig. 7.7, with the quantitative findings detailed in Table 7.4. These results offer insights into the comparative performance of the models under study, highlighting the effectiveness of each model in terms of accuracy and error metrics as they apply to the specific task at hand.

The analysis reveals a notable improvement in test error reduction for both FNO-2D and UNet with an increase in the number of training samples. However, a consistent trend emerges where FNO-2D surpasses UNet regarding prediction accuracy. Not only does FNO-2D achieve markedly higher accuracy, but its error distribution on the test set demonstrates a tighter concentration around the mean. This suggests that FNO-2D predicts with greater accuracy on average and shows superior generalization across diverse samples, including those significantly divergent from the dataset's mean. This efficiency is particularly impressive given FNO-2D's parameter count of only 5 909 250, starkly contrasting to UNet's 23 617 970 parameters, underscoring FNO-2D's better parameter efficiency.

Table 7.4: Summary of median normalized L_1 and L_2 errors across 400 test samples, with models trained using different sizes of training datasets. The minimal error recorded for each dataset size is highlighted in **bold**.

Samples	Model	L_1 (%)	L_2 (%)	$\sigma(L_2)$ (%)	Time per epoch (s)*
1024	FNO-2D	**11.18**	**12.07**	4.39	8.31
	UNet	22.64	24.42	6.80	10.15
2048	FNO-2D	**7.19**	**7.72**	3.15	10.10
	UNet	16.68	18.01	5.59	13.24
4096	FNO-2D	**4.32**	**4.64**	2.12	13.72
	UNet	11.09	11.92	4.15	18.96
8192	FNO-2D	**2.48**	**2.65**	1.30	21.13
	UNet	7.10	7.58	2.88	30.92
16384	FNO-2D	**1.59**	**1.68**	0.89	35.81
	UNet	4.52	4.84	1.99	55.26

* Timing was performed on a single NVIDIA A100 SXM4 GPU.

For FNO-2D to approximate a test error around 5 %, it necessitates 4096 training samples, whereas UNet demands a fourfold increase in training samples to achieve similar accuracy levels. This pattern is consistent across all examined training set sizes, as depicted in Fig. 7.7a, with UNet's performance trailing FNO-2D's by a substantial margin regarding the number of required training samples. These observations align with prior studies comparing FNO and UNet efficiency in solving PDE-constrained tasks [177, 187].

Moreover, FNO-2D's training process is approximately 30 % quicker than UNet's, a disparity that becomes more pronounced with larger training datasets. This speed advantage largely stems from FNO-2D's leaner parameter architecture. Notably, FNO-2D is slower than UNet on a per-parameter basis, highlighting its computational efficiency in leveraging a smaller parameter set for rapid training.

Exploring whether FNO-2D's superior test error is consistent across all samples is insightful in determining if its lower average error comes with exceptions where UNet might excel. This investigation could reveal instances where, despite FNO-2D's overall lower error, UNet provides more accurate predictions for specific sample geometries, suggesting a potential niche advantage for UNet in some cases. The comparative analysis of L_1 losses for each test sample across all training iterations between FNO-2D and UNet is depicted in the scatter plot within Fig. 7.7b, with the dashed line denoting the point of error equivalence between the two models. This visualization unmistakably demonstrates FNO-2D's dominance. It shows that it outperforms UNet across the board for every test sample, regardless of the size of the training set utilized. This uniform superiority indicates that FNO-2D's enhanced predictive accuracy extends universally across the dataset without exceptions, underscoring its robustness and wide-ranging applicability compared to UNet.

Figure 7.8 presents an illustrative example from the test dataset, evaluated using the FNO-2D and UNet models, each trained with 16k samples. This particular sample was

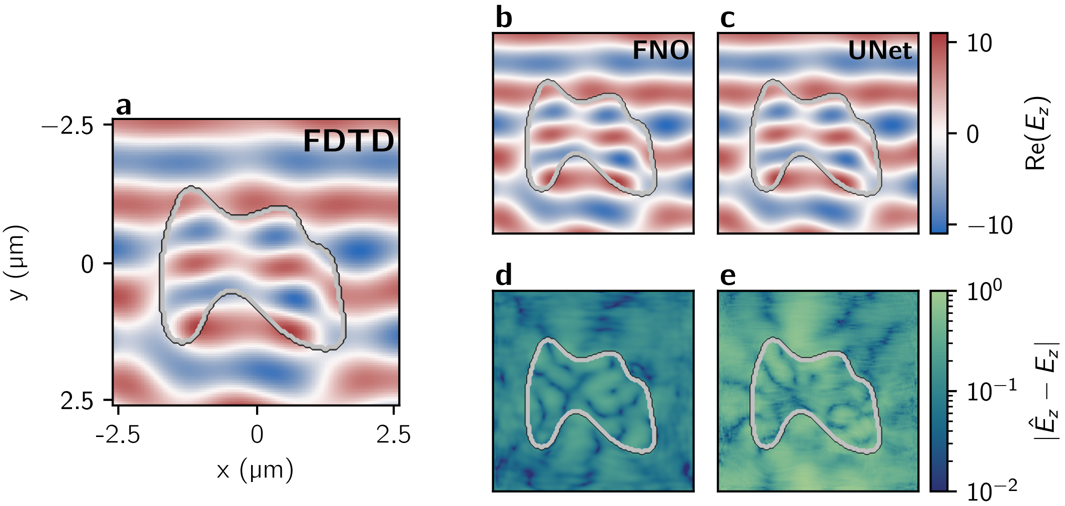

Figure 7.8: Visualization of the real component of the electric field E_z: **a** depicts the ground truth from a test set sample, while **b** and **c** show the fields as predicted by FNO-2D and UNet, respectively. The absolute error in the predictions compared to the ground truth is also presented for **d** FNO-2D and **e** UNet. The FNO-2D model demonstrates a normalized L_1 error of 1.58 %, in contrast to the UNet model's 4.29 % for this sample. Each network underwent training with a dataset comprising 16k samples. The scatterer's boundary is delineated in all figures, with illumination provided by a plane-wave source from the negative y-direction. Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

selected to closely match the median L_1 error observed in the FNO-2D's performance across the test set, registering a relative L_1 error of 1.58 % for FNO-2D and 4.29 % for UNet. A qualitative examination reveals a reasonably high level of agreement between the actual fields and those predicted by both models, as depicted in Figs. 7.8a to 7.8c. However, a deeper analysis of the absolute error maps highlighted in Figs. 7.8d and 7.8e uncovers a more pronounced deviation in the UNet predictions than FNO-2D, corroborating the numerical error data.

More notably, the error distribution for FNO-2D (Fig. 7.8d) demonstrates a smoother pattern relative to UNet's (Fig. 7.8e), suggesting reduced spatial noise. This outcome aligns with the inherent architectural distinctions between the models: UNet relies on localized, pixel-wise convolutions, whereas FNO engages in global convolutions, systematically omitting Fourier components beyond a certain threshold. This design choice ingrains FNO's propensity to learn and produce more continuous solutions, an attribute directly observable in its smoother error distribution. This inherent characteristic underscores FNO's architectural advantages in learning to predict complex field distributions with a more uniform error profile.

7.4.2 FNO with Maxwell loss

In this subsection, we study the impact of incorporating a physics-informed loss component into the training of the FNO, as illustrated on other network architectures by prior studies [103, 102]. The premise of this strategy involves either augmenting [103] or completely substituting [102] the conventional data-driven loss function with a physics-based one that seeks to minimize the residuals derived from the governing physical equations, specifically Maxwell's equations, in relation to the model's outputs.

It is important to delineate a physics-informed *loss* function, which we use here, from so-called physics-informed neural networks (PINNs), which are another class of neural networks altogether and have been gaining some popularity in recent years [204, 205, 174]. While PINNs also make use of the governing physical equations, they incorporate them much more directly into their architecture, *e.g.*, by parameterizing a PDE directly using a neural network. In particular, PINNs can be trained purely by minimizing the *residual* function, *i.e.*, they can be trained in an *unsupervised* manner and do not require any data in principle. In that sense, they are similar to iterative PDE solvers, both in terms of accuracy and speed. However, it should also be noted that this distinction is not always made clearly, and the term "physics-informed" neural network can refer to both PINNs in the original sense and other machine learning techniques that use the governing physical equations – in this subsection, we will be doing the latter.

Introducing a physics-informed loss does not aim to directly enhance the model's performance on the standard data-driven training loss. Instead, it serves as an additional regularization mechanism during the training process. This method nudges the model's outputs closer to solutions that are not only consistent with the input data but also adhere to the underlying physical laws. Such a refinement has been demonstrated to bolster the model's ability to generalize across different scenarios and to mitigate errors associated with subsequent computations on the predicted fields, like far-field transformations [103].

Implementing this physics-informed loss is relatively straightforward and does not necessitate alterations to the network's architecture or the overarching training methodology. The governing physical equations are incorporated as an extra term within the loss function. This approach underlines a harmonious blend of data-driven learning with physical theory, aiming to enhance the reliability and accuracy of surrogate models like FNO in simulating complex physical phenomena.

In this exploration, we enhance the existing data-driven loss function by incorporating an additional term that enforces adherence to the wave equation for a source-free region:

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \tilde{\boldsymbol{E}}(\boldsymbol{r}, \omega) - \omega^2 \mu_0 \epsilon_0 \epsilon(\boldsymbol{r}) \tilde{\boldsymbol{E}}(\boldsymbol{r}, \omega) = 0 \quad , \tag{7.12}$$

where $\tilde{\boldsymbol{E}}(\boldsymbol{r}, \omega)$, representing the network's output, is expected to conform to the wave equation's constraints. This formulation implies that the network's output should remain unchanged after undergoing a double-curl operation and scaled by appropriate constants. This principle is aptly captured in a loss function specifically designed for this purpose:

$$L_{\mathrm{MW}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\left\| \boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{y} - \omega^2 \mu_0 \epsilon_0 \, \boldsymbol{x} \odot \boldsymbol{y} \right\|_2}{\| \boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{y} \|_2} \quad , \tag{7.13}$$

where $\boldsymbol{y}$ symbolizes the network's output (the fields) and $\boldsymbol{x}$ represents its input (the permittivity distribution). Combining this with Eq. (7.7), the comprehensive loss function for guiding the network's training is formulated as

$$L = L_2(\boldsymbol{y}, \hat{\boldsymbol{y}}) + L_{\mathrm{MW}}(\boldsymbol{x}, \boldsymbol{y}) \quad , \tag{7.14}$$

integrating the standard data-driven loss with the physics-informed Maxwell wave (MW) equation loss. The outcomes of this training approach and its comparative analysis against a purely data-driven FNO model are detailed in Table 7.5. Here, "FNO-2D" refers to the original FNO model, while "FNO-2D MW" designates the variants that incorporate the Maxwell loss component, showcasing the impact of this physics-informed regularization on model performance.

Table 7.5: Summary of median normalized L_2 errors across 400 test samples, with FNO models trained using different sizes of training datasets and with and without the addition of a physics-driven loss term.

Samples	Model	L_1 (%)	L_2 (%)	L_{MW} (%)
1024	FNO-2D	11.18	12.07	48.55
	FNO-2D MW	8.33	8.85	3.39
2048	FNO-2D	7.19	7.72	37.90
	FNO-2D MW	5.50	5.84	2.47
4096	FNO-2D	4.32	4.64	27.21
	FNO-2D MW	3.51	3.72	1.92
8192	FNO-2D	2.48	2.65	19.62
	FNO-2D MW	2.48	2.63	1.57
16384	FNO-2D	1.59	1.68	14.11
	FNO-2D MW	1.52	1.62	0.78

The findings detailed in Table 7.5 reveal that incorporating the Maxwell loss term into the training process results in a modest, yet significant, enhancement in performance over the baseline training loss. This improvement tends to exhibit diminishing returns as the size of the training dataset increases. Importantly, the addition of the Maxwell loss (FNO-2D MW) consistently achieves better precision across various evaluation metrics (L_1, L_2, and L_{MW}) and all sizes of training datasets.

The most notable disparities between the models emerge in the L_{MW} errors. Models trained without the Maxwell loss term show a substantial initial error of 48.55 % with a smaller dataset, which significantly decreases to 14.11 % as the dataset size expands. This trend is expected since achieving greater accuracy, even solely in terms of the L_2 loss, inherently requires the generation of higher fidelity fields, yielding outputs that better adhere to Maxwell's equations and, thus, a lower Maxwell loss.

However, the difference in performance between models trained with and without the Maxwell loss term is particularly striking. Incorporating the L_{MW} term reduces the error dramatically from 48.55 % to as low as 3.38 %, even with the smallest dataset, further narrowing down to a mere 0.78 % for the largest dataset. This significant reduction underscores that the fields predicted by models trained with the physics-informed loss are substantially more self-consistent with physical laws than those trained using a purely data-driven approach. Consequently, this evidence strongly supports the premise that integrating a physics-based loss into the training regimen offers substantial benefits, promoting more physically accurate and self-consistent solutions, and should be considered a valuable practice in model training where applicable.

7.4.3 Tensorized FNO

In Subsection 7.2.2, we introduced the FNO as

$$\kappa_m(y) = \mathcal{F}^{-1}(R_m \cdot \mathcal{F}(y)) \quad , \tag{7.15}$$

where R_m is the operator's learnable kernel during training. In practice, this kernel is represented as a multi-dimensional array, commonly referred to as a *tensor* [206]. This tensor contains a total of $2^d n^d + 1)\ w_{\text{in}} w_{\text{out}} + w_{\text{out}}$ trainable parameters, where d represents the input dimensionality, n is the number of Fourier modes, and w_{in}, w_{out} represent the number of input and output channels of the Fourier block, respectively. It is easy to see that the number of trainable parameters can quickly explode due to the exponential scaling factor, depending on the dimensionality of the problem and the number of Fourier modes considered in the expansion.

To alleviate this problem, Kossaifi *et al.* [207] introduce "tensorized" neural operators, where the parameter tensor is parameterized efficiently through a low-rank tensor factorization using a Tucker decomposition. This decomposition acts as a low-rank constraint on the entire weight tensor, which regularizes the model. Additionally, and most importantly, it leads to a large reduction in the number of parameters in the model, depending on how the dimensionality of the factorization is chosen. A Tucker decomposition of a rank n tensor T can be written as

$$T = \mathcal{T} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 \dots \times_n U^{(n)} \quad , \tag{7.16}$$

where $\mathcal{T} \in \mathbb{C}^{d_1 \times d_2 \times \dots \times d_n}$ is the so-called *core tensor* containing the singular values of T and $U^{(i)}$ are unitary matrices in $\mathbb{C}^{d_i \times n_i}$, and $\times_n$ denotes the n-mode product defined as

$$(\mathcal{T} \times_n U)_{i_1 i_2 \dots i_{n-1} j_n j_{n+1} \dots j_N} = \sum_{i_n=1}^{I_n} t_{i_1 i_2 \dots i_{n-1} i_n i_{n+1} \dots i_N} \times u_{j_n i_n} \quad . \tag{7.17}$$

In the case of $d_i = n_i$, the Tucker decomposition is always an exact representation of T. However, by choosing $d_i < n_i$, T can be compressed and efficiently approximated. By choosing the ranks of d_i, different compression ratios for the number of model weights can be achieved.

In the following, we will investigate the predictive performance of a “tensorized” FNO (termed TFNO) and compare them with our baseline FNO-2D regarding accuracy and the number of model weights. To do this, we train two TFNO models with different compression ratios on a dataset of 2048 training samples, with all hyperparameters being exactly the same as the FNO-2D trained on the same number of samples. The results of this training, along with the final losses, are detailed in Table 7.6.

Table 7.6: Comparative results of validation losses for FNO and TFNO models trained on 2048 samples with different compression ratios.

Model	L_1 (%)	L_2 (%)	Parameters	Compression ratio
FNO-2D	7.19	7.72	5.90×10^6	1
TFNO	3.25	3.50	7.55×10^5	8
TFNO	3.76	4.06	1.60×10^5	37

The comparison between the baseline FNO-2D and the TFNO architectures presents a compelling case for factorization. The baseline FNO model, which employs 5.9 million parameters, completed its training cycle with a final L_1 loss of 7.5 %. In contrast, the TFNO model with a compression ratio of 8 minimizes the parameter count to 755 000 while simultaneously cutting the final L_1 loss in half, down to 3.25 %. Even more impressively, the TFNO with only 160 000 parameters still achieves almost the same low L_1 loss at 3.76 % while decreasing the model parameter count by a factor of almost 37. This substantial decrease in model complexity, alongside enhanced prediction accuracy, underscores the potential of TFNOs to make neural network models more efficient and effective. However, it is important to note the nuanced aspect of computational cost. Despite the significant parameter reduction leading to much smaller final model sizes, TFNOs do not necessarily equate to reduced computational demands during training. Due to the requirement of expanding the factorized weight matrix throughout the training process, the computational load can be similar to or slightly exceed that of the baseline FNO-2D. Consequently, while TFNOs offer a promising avenue for enhancing model efficiency and accuracy, they might not directly address challenges associated with handling larger problem sizes. For such scenarios, alternative strategies like domain decomposition [208] or multi-grid methods [207] present viable pathways for further exploration and optimization.

7.5 Inverse design

The deployment of data-driven surrogate solvers for Maxwell’s equations might initially appear inefficient, considering the significant investment in time and computational resources required to generate an extensive dataset via a full-wave Maxwell solver from which the surrogate model is trained. This investment may seem especially disproportionate if the surrogate is intended for a limited number of scattering problem solutions. Moreover, the utility of such surrogate models is inherently confined to the problem types encapsulated within the training dataset, thus limiting their applicability as a broad-spectrum solution for various scattering issues.

Nonetheless, the field of inverse design, particularly when leveraging gradient-based methodologies, stands to significantly gain from the expedited computational capabilities offered by surrogate models, with their limitations having potentially minimal impact, as evidenced in prior research [103, 209]. In gradient-based inverse design, the optimization process involves iteratively refining a device's geometry to enhance a specific performance metric, employing gradients of this metric with respect to geometrical modifications. Traditional approaches, such as topology optimization [59], typically require several hundred iterations for convergence. Each iteration necessitates two full-wave simulations – one to assess the performance metric and another to calculate its gradient using the adjoint method. In three-dimensional scenarios, a single optimization could last several days, and the starting conditions heavily influence the outcomes. Since gradient descent-based optimization finds *locally* optimal solutions, the starting point (*e.g.*, the initial permittivity distribution) is a deciding factor in *which* local optimum the solution will fall into, which necessitates an exploration of multiple initial conditions to evaluate the solution's quality. Moreover, the sequential nature of these updates limits the potential for parallelization, leaving the acceleration of individual simulations as the sole avenue for expediting the inverse design process.

Thus, while surrogate models may not universally supplant traditional simulation methods for direct problem-solving, their integration into gradient-based inverse design presents a compelling strategy for substantially reducing computation times. This acceleration is invaluable, particularly in applications where rapid prototyping or iterative design processes are critical, offering a practical and impactful use case for surrogate models in specialized engineering and design tasks.

Neural network-based surrogate solvers present an optimal solution for overcoming the challenges associated with traditional gradient-based optimization techniques. The key advantages of these surrogate solvers lie in their significantly reduced inference times compared to conventional full-wave simulations, the ability to generate training samples in parallel due to their independent nature, and their inherent differentiability. This last attribute, in particular, enables their seamless integration into existing gradient-based optimization pipelines as functional equivalents to differentiable Maxwell solvers, as highlighted in recent studies [153, 94].

Moreover, some of the limitations traditionally associated with neural network surrogates are less impactful in the context of inverse design. Since the simulation parameters remain constant throughout such optimizations, with only the device geometry undergoing variation, it becomes viable to employ specialized models trained on minimal data, ensuring operations remain well within the confines of the training distribution. Additionally, the demand for extreme numerical precision is often relaxed in many inverse design scenarios, where achieving a practically viable solution is more critical than attaining high numerical accuracy. This relaxation is further justified by the inherent uncertainties in fabrication processes, which typically introduce a margin of error on the order of a few percent, overshadowing minor inaccuracies in simulation results.

In situations where utmost precision is necessary, surrogate models' initial device designs produced with lower numerical accuracy could serve as starting points. These preliminary designs can be refined using more accurate simulation methods in just a few extra iterations, providing an efficient pathway to high-quality solutions. Therefore, neural network-based surrogate solvers expedite the optimization process and offer a pragmatic balance between computational efficiency and the precision demands of inverse design tasks, making them highly advantageous tools in the field of optical and electromagnetic device design.

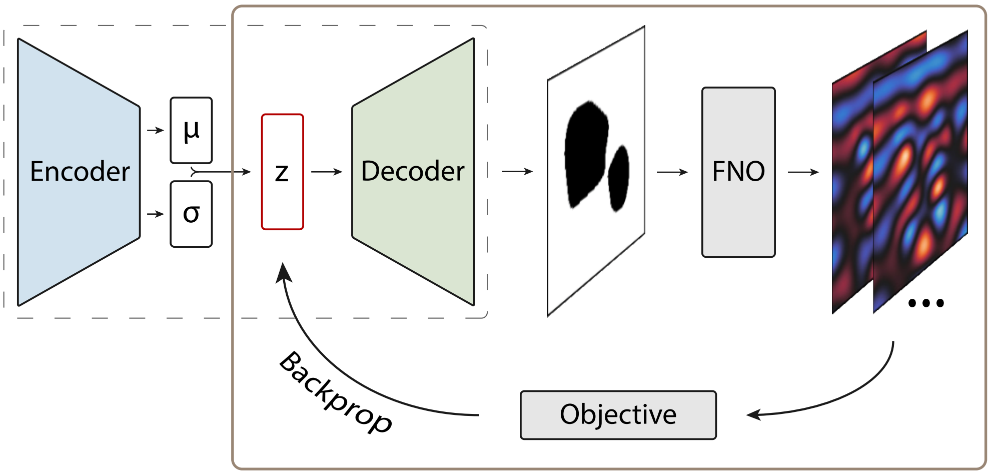

Figure 7.9: Diagram illustrating the process of iterative inverse design. This methodology employs the decoder segment of a previously trained variational autoencoder (highlighted by the dashed gray box) to create scatterer configurations from a given latent vector z (indicated by the red box). Subsequent to scatterer generation, the corresponding electric field is determined using the FNO model, facilitating the loss computation. The optimization process leverages backpropagation to derive gradients of the target objective relative to the latent variables, which are then progressively refined through iterative updates until convergence is achieved. Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

Here, we explore the inverse design of intricate, three-dimensional nanophotonic devices by applying an FNO model specifically trained on a dataset comprising 8192 volumetric scatterer-field pairings. The methodologies behind the data compilation and the training of the FNO-3D model are elaborated upon in Subsection 7.3.1 and Subsection 7.3.2. Following a training duration of 100 epochs, the FNO-3D model demonstrated a normalized L_1 loss of 5.05 % on a test set of 400 samples, showcasing its predictive capabilities.

The inverse design approach utilized herein deviates from traditional methods by incorporating FNO-3D as a surrogate for the conventional Maxwell solver. This substitution allows the objective function to remain an arbitrary differentiable function of the electromagnetic fields, maintaining the usual flexibility of inverse design. However, a distinctive aspect of our methodology lies in the parameterization of the scatterers. Rather than adjusting the scatterers' properties directly on the computational grid, we represent them within the latent space of a variational autoencoder (VAE) that has been pre-trained [194]. This strategy is motivated by one of the limitations of our surrogate solver – a direct grid-based parameterization could inadvertently cause permittivity values of the scatterer to vary continuously, diverging from the binary data the FNO-3D was trained on, potentially undermining the model's field inference accuracy for non-binary permittivity values. By leveraging the latent space of a VAE, we maintain the scatterer's representation's binary nature, aligning it with the data the FNO-3D model was trained on and ensuring the integrity of the inverse design process. This approach enhances the feasibility of designing complex devices and aligns with the model's training, ensuring more reliable and accurate design outcomes.

Incorporating a broader range of permittivity values into the training dataset is theoretically straightforward but practically challenging, as it necessitates dense sampling across the spectrum of possible permittivities for seamless inverse design. However, this approach would balloon the dataset's size, undermining the efficiency that a dedicated surrogate solver brings to *rapid* inverse design. To circumvent this, we employ a convolutional Variational Autoencoder capable of generating random scatterer configurations, mirroring the distribution of the initial training dataset. This enables the direct optimization of the desired device within the VAE's latent space. Notably, the latent space representation allows for continuous interpolation, with the VAE's decoder translating these modifications into binary geometries that FNO-3D can model accurately. It is worth noting that the VAE's encoder segment is utilized solely during the training phase and is not required for the inverse design tasks. Generating data for training the VAE is notably cost-effective, as it eliminates the need for simulations and allows for real-time data production during the training phase, as detailed in Subsection 7.3.3. It is also important to mention that employing a VAE to parameterize the inverse design task with FNO is not obligatory. Theoretically, any parameterization strategy that results in binary scatterers, such as direct geometrical or boundary parameterization, could be applied. A visual diagram illustrating the inverse design workflow is presented in Fig. 7.9.

In this example, we demonstrate the optimization of two nanophotonic devices using FNO-3D, leveraging a straightforward objective function based on the electric field intensity:

$$J(\tilde{E}) = \sum_{\mathcal{D}} \left|\tilde{E}(\boldsymbol{r})\right|^2 \quad \forall \boldsymbol{r} \in \mathcal{D} \quad , \tag{7.18}$$

where $\mathcal{D}$ denotes the set of spatial points where we aim to enhance the electric field intensity. The objective of the first device is to maximize the intensity at a single point located at the center of the x-y plane, effectively designing a basic nanophotonic lens. In the second scenario, the goal shifts to intensifying the electric field at four distinct focal points, positioned centrally within each quadrant of the x-y plane. The design considers plane wave illumination across the x-y plane, entering from the top at $z = 0\,\mu\text{m}$, with the focal plane positioned at $z = 4.8\,\mu\text{m}$. Both devices undergo a 300-iteration optimization process using the AdamW optimizer [196]. While alternative optimization strategies like L-BFGS-B [13] or MMA [210] are viable, AdamW is selected for its advantages in GPU compatibility and the facility to optimize several devices concurrently. The outcomes of these optimization endeavors are detailed in Fig. 7.10.

We conduct 64 separate trials, each starting from a unique initial condition represented by random vectors in the latent space of the VAE, leading to the creation of 64 distinct optimized devices for each design challenge. We subsequently select the highest-performing model from these as our final device (refer to Figs. 7.10a and 7.10e). Numerous attempts

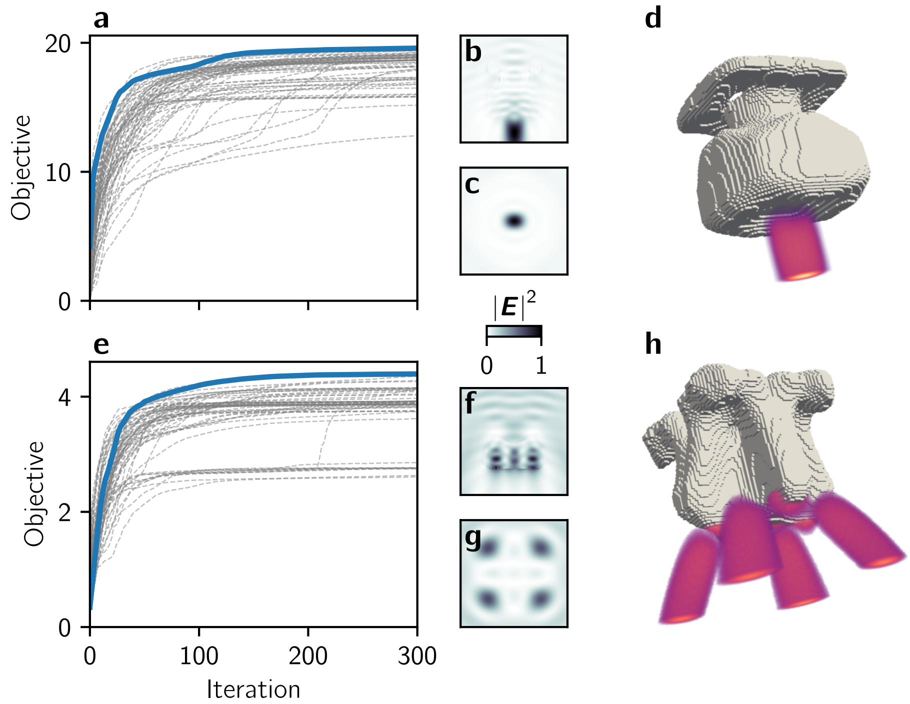

Figure 7.10: This figure illustrates the inverse design process using FNO-3D for two nanophotonic devices: one is designed to focus light into a single point (**a**-**d**) and another intended to create four distinct focal spots (**e**-**h**), both under plane-wave illumination. Panels (**a**, **e**) display the optimization trajectories for each device, conducted over 64 trials with varied initial conditions, where the blue line highlights the best-performing device. Panels (**b**, **f**) and (**c**, **g**) reveal slices of the electric field intensity, $|E|^2$, in the x-z and focal x-y planes, respectively, for the optimal devices as verified by comprehensive full-wave FDTD simulations. Finally, panels **d** and **h** provide three-dimensional representations of the devices, including volumetric renderings of the electric field intensity, emphasizing regions of the highest intensity. Adapted with permission from [P3]. Copyright 2023 American Chemical Society.

result in outcomes substantially inferior to the optimal one, indicating a pronounced sensitivity to initial parameter selection – a common challenge encountered in local optimization. Unlike the full-wave solver approach, which limits the feasibility of testing a wide array of initial conditions and often restricts it to only a handful of additional trials, our method stands out in its efficiency.

Specifically, an individual optimization with FNO-3D is completed in roughly 10 minutes, with each iteration taking about 2 seconds on an NVIDIA A100 SXM4 GPU. Moreover, conducting two parallel optimizations on a single GPU is feasible. By leveraging 32 GPUs to run all 64 trials simultaneously, the duration for each complete optimization sequence remains capped at 10 minutes. This efficiency is, however, contingent on the available computational resources. Even if the optimizations were to be executed sequentially, the total time required would extend to slightly over 5 hours – a stark contrast to the several days as are necessary for a single optimization using traditional full-wave solvers, where each simulation within an iteration consumes about 20 minutes, with two simulations per iteration as detailed in Subsection 7.3.1. This substantial difference underscores the impracticality of exploring numerous initial conditions with full-wave solver-based methods, starkly contrasting the feasibility demonstrated in our approach.

To ascertain the performance of the optimized devices, we conducted full-wave simulations. The lens depicted in Fig. 7.10d exhibited an L_1 error of 10.2 % between the simulated and predicted field intensities, $|\tilde{E}|^2$, whereas the four-point lens showcased in Fig. 7.10h demonstrated a marginally lower error of 9.1 %. Despite these errors being slightly elevated compared to those observed in the test dataset, they remain acceptably accurate, especially when considering the variability inherent in microfabrication technologies, such as 3D laser nanoprinting, which could be employed to manufacture these devices.

The observed increment in error, relative to the test dataset, is attributed to the inherently higher discrepancies associated with squared absolute field values. However, when evaluating the L_1 losses based on $|\tilde{E}|$, the results – 5.8 % for the single-point lens and 5.0 % for the four-point lens – are consistent with the average errors of the test set. Notably, the primary source of intensity error is the absolute numerical discrepancies in field components. However, the qualitative agreement between the FDTD and FNO-3D predicted fields remains strong. This qualitative alignment further underscores that both devices successfully achieve their design objectives. The fields depicted in Fig. 7.10 are derived from comprehensive full-wave FDTD simulations.

7.6 Conclusion

Utilizing surrogate models for addressing scattering problems presents a seemingly advantageous aspect regarding their significantly reduced inference times compared to conventional full-wave solvers. Nonetheless, this efficiency comes with notable trade-offs that warrant careful evaluation to determine their suitability for specific applications. Among the primary concerns are issues of generalization and accuracy. Surrogate solvers, particularly those trained on limited datasets, struggle to extend their applicability beyond the specific conditions and distributions they were trained under. Moreover, the minimization of inference errors is contingent upon the model's complexity and the comprehensiveness of the training dataset [211, 195].

In optical simulations, this limitation manifests in the model's capacity to accurately "solve" only a subset of scattering problems it has been explicitly trained on, such as certain types of source distributions, material properties, and geometrical configurations. While augmenting the model's architecture, diversifying training samples, and enlarging the dataset can mitigate these issues to an extent, thereby enhancing both generalization and accuracy, the effectiveness of this approach is bounded by practical considerations. Specifically, if the effort and resources required to amass an adequate training dataset surpass the expenses of traditional computational methods, the rationale for opting for a surrogate model diminishes.

As depicted in Fig. 7.11, we compare the cumulative durations required for FDTD and FNO-3D simulations alongside the time necessary for generating FNO-3D's training data. These

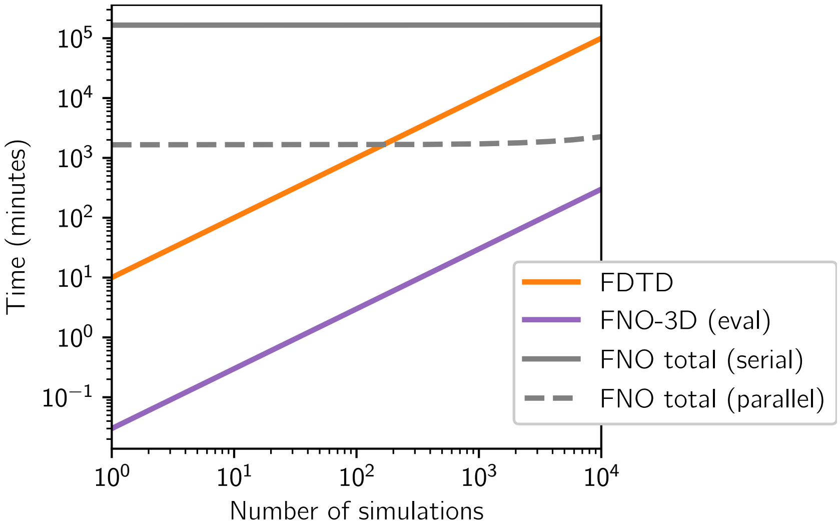

Figure 7.11: Illustration of a comparison between the time needed for completing full-wave simulations using FDTD and generating predictions with FNO-3D across various simulation counts, with each FDTD simulation estimated at 20 minutes. It also presents the cumulative duration for FNO-3D inferences, including the time for generating samples (8192 samples) and network training (approximately 24 hours). The comparison is depicted for scenarios of serial (solid gray line) and parallel (dashed gray line) data generation methodologies.

timings can significantly fluctuate based on the hardware and simulation configurations employed. Hence, the data presented in Fig. 7.11 should be interpreted specifically within the context of this study. Although FNO-3D demonstrates a remarkable speed advantage over FDTD during the inference phase – being faster by three orders of magnitude – this comparison in isolation does not fully capture the broader context when juxtaposed with traditional solvers. The predominant portion of time expenditure is attributed to data generation, dwarfing the time allocated for training and inference, which are comparatively minimal.

A notable benefit of data-driven methodologies is the independence of individual samples, allowing for parallel data generation constrained only by the availability of computational resources. To achieve a break-even point with a full-wave solver, a surrogate model must be utilized for inference at least as many times as the dataset size used in its training. Past this threshold, a surrogate model becomes more cost-effective regarding time and computational resources, including energy consumption. The benchmark for total time can be significantly reduced through parallel sample generation, as demonstrated in our approach, detailed in Subsection 7.3.1.

The reliance of data-driven methods on datasets produced by classical computational techniques underscores the continued indispensability and, in numerous instances, the superior efficacy of these traditional methods when a holistic assessment of all pertinent aspects is undertaken. This dynamic, however, may evolve over time, especially with the increasing availability of high-quality datasets suitable for training surrogate solvers. Reflecting a noteworthy and praiseworthy trend within the photonics community, there is a movement towards the publication of training datasets, an initiative that mirrors earlier progressions seen within the machine learning field, whether through individual contributions or collective endeavors [212].

Although this shift does not alter the fundamental premise that data must originate from conventional solvers, it is important to recognize that a significant portion of this data would be generated regardless. Thus, releasing it into the public domain represents a minimal extra burden. Importantly, the emergence of a comprehensive database of simulation results could lead to the development of more robust surrogate models. Such models could potentially bypass the need to generate new training data for each specific project, provided access is granted to an extensive repository of pre-existing simulation data. This possibility opens the door to creating more universally applicable models that, despite their dependency on vast quantities of training data, would not necessitate the production of this data anew for subsequent research.

In the broader landscape of scientific machine learning, especially in the surrogate solvers domain, the emphasis on data efficiency cannot be overstated. It demands a critical and open examination [213]. Within this work, we highlight three pivotal considerations for integrating deep learning techniques: *architecture*, *specialization*, and *application*. Selecting an appropriate model architecture can dramatically diminish the volume of training data needed to achieve sufficiently low error rates in test scenarios. This principle is exemplified in our electromagnetic scattering problem discussed in Subsection 7.4.1, with corroborating evidence found across various physical sciences [214, 176]. Moreover, physics-informed methodologies, which integrate fundamental equations into the training loss function, as seen in studies like [215, 184, 102, 103], markedly enhance data utilization efficiency.

Secondly, the concept of model specialization merits attention. Designing a surrogate solver to excel within a narrowly defined range of problem settings can be far more data-efficient than striving for universal applicability. Such broad-spectrum generalization often demands an impractical amount of data, underscoring the value of focused model development.

Minimizing data demands necessitates a strategic narrowing of the model's application domain. Moreover, deploying a surrogate solver should yield tangible benefits that outweigh the investment in development and training. This concept is exemplified in Section 7.5, where a surrogate solver facilitates the free-form, three-dimensional inverse design of electromagnetic scatterers. In this instance, the model is trained using a dataset comprising 8192 samples and is employed across 128 distinct optimizations, each involving 300 iterations. To draw a parallel, a similar optimization effort employing an adjoint-based technique, would necessitate approximately $128 \times 300 \times 2 = 76,800$ full-wave simulations. This figure starkly surpasses the simulation count required for dataset compilation, underscoring the efficiency of the surrogate approach in this particular case.

While surrogate models hold considerable promise and are poised to impact future research endeavors significantly, it is important to recognize the continuing relevance of more traditional methods. Specifically, when a problem's nature permits, fast "classical" strategies, including semi-analytical approaches, should remain the preferred option. Such methods not only compete in terms of speed but also boast superior accuracy and well-defined error metrics, offering a reliable and precise solution.

In this chapter, we have illustrated the capabilities of a neural operator-based model in addressing electromagnetic scattering problems, showcasing a performance that significantly exceeds the benchmarks set by contemporary state-of-the-art solutions. Despite the intrinsic trade-offs in accuracy and broad applicability accompanying the transition from conventional full-wave solvers to a data-driven surrogate model, our findings reveal the practicality of such an approach for specific applications, including the gradient-based, free-form inverse design of three-dimensional electromagnetic scatterers. Machine learning-based surrogate solvers show substantial promise for nanophotonic applications, conditional on the discovery and implementation of efficient model architectures alongside carefully selecting tasks to which they are best suited.

8 Summary & outlook

This work focused on shedding light on the inverse design process of nanophotonic devices, emphasizing the challenges associated with large-scale optimization problems and the creation of devices with free-form geometries from several perspectives. The foundational tools utilized in this endeavor were detailed in Chapter 2, where the adjoint method's application to Maxwell's equations and the principles of topology optimization were discussed. We commenced with optimizing a device to enhance the excitation and steering of Bloch Surface Waves in Chapter 3. This initial step allowed for comprehensive analyses of the optimized devices, particularly in terms of the effective index contrast of the surface modes, facilitated by the relatively lower complexity of the problem.

Progressing further, we explored the design of a compact and efficient coupler for interfacing optical fibers with photonic wire bonds in Chapter 4. This exploration led to developing an innovative boundary parameterization approach tailored for topology optimization of devices exhibiting rotational symmetry. The designs that emerged from this process matched the coupling efficiencies of prior couplers while occupying only one-fifth of their spatial footprint. This achievement underscores the potential of inverse design, particularly when applied to the domain of 3D laser nanoprinting.

Remaining in the realm of additive manufacturing while elevating the complexity of the design challenge, our efforts were directed towards optimizing a polarization-independent grating coupler, detailed in Chapter 5. Here, designed a device with state-of-the-art coupling efficiencies but also illuminated the complexities of free-form nanophotonic device design, a challenge which we further elaborated on in Chapter 6. We discussed the trade-offs between theoretical optimality and practical manufacturability, specifically considering additive manufacturing. We developed a framework for integrating structural considerations into the photonic design process to address the unique challenges of free-form device optimization. By co-designing devices' photonic and structural properties, we demonstrated the feasibility of achieving both structural integrity and high optical performance, highlighting the untapped potential of holistic design approaches in nanophotonics.

Lastly, we explored computational methods for reducing the cost associated with large-scale inverse design, where we designed and used a neural network-based surrogate solver in Chapter 7. We demonstrated the proposed architecture's effectiveness for solving electromagnetic scattering problems, achieving results that surpass the performance metrics of current state-of-the-art approaches to surrogate modeling. Further, we use this model for the gradient-based, free-form inverse design of three-dimensional electromagnetic scatterers.

Throughout this thesis, strides have been made in addressing the physical principles and computational strategies pertinent to the inverse design of nanophotonic devices tailored for additive manufacturing. Yet, numerous pathways for future research remain open. One significant limitation in achieving realizable structures is the lack of accurate, differentiable models for 3D laser nanoprinting fabrication processes. A critical issue in this context is dose accumulation within the material, where polymerization may occur under repeated illumination that is nominally below the threshold for polymerization. This effect means that the minimum feature size in actual devices is a function of the illumination's intensity and the surrounding voxel environment. Additionally, being a diffraction-limited system, closely positioned voxels tend to merge due to shared exposure, complicating the fabrication of small, intricate structures near the minimum feature size. Considering these challenges is crucial when designing devices for realization via 3D laser nanoprinting, as it is straightforward to optimize structures that perform well in simulations but are unfeasible in practical applications. The common approach of imposing a minimum feature size in topology optimization through low-pass filters does not adequately address these nuances.

Furthermore, the phenomenon of *shrinking* associated with this fabrication method presents another layer of complexity. In the direct laser writing process, a photoactive liquid polymer, or photoresist, is exposed to a high-intensity laser to polymerize and form the desired 3D structure selectively. The subsequent removal of unpolymerized resist results in shrinkage of the solid structure, partially due to the voids left by the removed liquid resist, among other factors. This shrinkage is typically non-uniform, constrained by the attachment to a substrate, leading to upward shrinkage. Given the sensitivity of inverse-designed structures, especially in optics, to their geometric configurations, non-uniform shrinkage can severely compromise their intended functionality. As such, integrating knowledge of this shrinkage phenomenon into the optimization process itself would be highly advantageous. This integration would enable the generation of devices that are pre-compensated for shrinkage or, at the very least, exhibit resilience to it. This would allow their printing with minimal manual adjustment without significantly detracting from their functional performance.

The research findings and developments documented in this thesis aim to provide insights into the free-form inverse design of nanophotonic devices and the particular challenges encountered when doing so in the context of 3D printing, and I hope it may inspire future research.

Bibliography

[1] A. F. Koenderink, A. Alù, and A. Polman. “Nanophotonics: Shrinking Light-Based Technology”. In: *Science* 348.6234 (May 2015), pp. 516–521. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1261243.

[2] A. Taflove, A. Oskooi, and S. G. Johnson, eds. *Advances in FDTD Computational Electrodynamics: Photonics and Nanotechnology*. Artech House Antennas and Propagation Series. Boston: Artech House, 2013. 623 pp. ISBN: 978-1-60807-170-8.

[3] A. F. Oskooi, D. Roundy, M. Ibanescu, P. Bermel, J. Joannopoulos, and S. G. Johnson. “Meep: A Flexible Free-Software Package for Electromagnetic Simulations by the FDTD Method”. In: *Computer Physics Communications* 181.3 (Mar. 2010), pp. 687–702. ISSN: 00104655. DOI: 10.1016/j.cpc.2009.11.008.

[4] J. L. Volakis, A. Chatterjee, L. C. Kempel, J. L. Volakis, and A. Chatterjee. *Finite Element Method for Electromagnetics: Antennas, Microwave Circuits, and Scattering Applications*. IEEE/OUP Series on Electromagnetic Wave Theory. New York, NY: IEEE Press [u.a.], 1998. 344 pp. ISBN: 978-0-7803-3425-0.

[5] J.-M. Jin. *The Finite Element Method in Electromagnetics*. Third edition. Hoboken. New Jersey: John Wiley & Sons Inc, 2014. 1 p. ISBN: 978-1-118-84198-3.

[6] S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez. “Inverse Design in Nanophotonics”. In: *Nature Photonics* 12.11 (Nov. 2018), pp. 659–670. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/s41566-018-0246-9.

[7] B. Shen, P. Wang, R. Polson, and R. Menon. “An Integrated-Nanophotonics Polarization Beamsplitter with $2.4 \times 2.4\ \mu\mathrm{m}^2$ Footprint”. In: *Nature Photonics* 9.6 (June 2015), pp. 378–382. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/nphoton.2015.80.

[8] Y. Augenstein, A. Vetter, B. V. Lahijani, H. P. Herzig, C. Rockstuhl, and M.-S. Kim. “Inverse Photonic Design of Functional Elements That Focus Bloch Surface Waves”. In: *Light: Science & Applications* 7.1 (Dec. 12, 2018), p. 104. ISSN: 2047-7538. DOI: 10.1038/s41377-018-0106-x.

[9] A. Håkansson and J. Sánchez-Dehesa. “Inverse Designed Photonic Crystal De-Multiplex Waveguide Coupler”. In: *Optics Express* 13.14 (July 11, 2005), p. 5440. ISSN: 1094-4087. DOI: 10.1364/OPEX.13.005440.

[10] M. Minkov and V. Savona. “Automated Optimization of Photonic Crystal Slab Cavities”. In: *Scientific Reports* 4.1 (May 30, 2014), p. 5124. ISSN: 2045-2322. DOI: 10.1038/srep05124.

[11] Y. Zhang, S. Yang, A. E.-J. Lim, G.-Q. Lo, C. Galland, T. Baehr-Jones, and M. Hochberg. “A Compact and Low Loss Y-junction for Submicron Silicon Waveguide”. In: *Optics Express* 21.1 (Jan. 14, 2013), p. 1310. ISSN: 1094-4087. DOI: 10.1364/OE.21.001310.

[12] A. Rahimzadegan, D. Arslan, D. Dams, A. Groner, X. Garcia-Santiago, R. Alaee, I. Fernandez-Corbaton, T. Pertsch, I. Staude, and C. Rockstuhl. “Beyond Dipolar Huygens’ Metasurfaces for Full-Phase Coverage and Unity Transmittance”. In: *Nanophotonics* 9.1 (Jan. 28, 2020), pp. 75–82. ISSN: 2192-8614, 2192-8606. DOI: 10.1515/nanoph-2019-0239.

[13] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. “Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization”. In: *ACM Transactions on Mathematical Software* 23.4 (Dec. 1997), pp. 550–560. ISSN: 0098-3500, 1557-7295. DOI: 10.1145/279232.279236.

[14] S. G. Johnson. *NLopt*. URL: https://github.com/stevengj/nlopt (visited on 04/05/2024).

[15] P. Virtanen *et al.* “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17.3 (Mar. 2, 2020), pp. 261–272. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[16] J. Jensen and O. Sigmund. “Topology Optimization for Nano-photonics”. In: *Laser & Photonics Reviews* 5.2 (Mar. 8, 2011), pp. 308–321. ISSN: 1863-8880, 1863-8899. DOI: 10.1002/lpor.201000014.

[17] 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”. In: *Nature Photonics* 9.6 (June 2015), pp. 374–377. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/nphoton.2015.69.

[18] A. Y. Piggott, J. Petykiewicz, L. Su, and J. Vučković. “Fabrication-Constrained Nanophotonic Inverse Design”. In: *Scientific Reports* 7.1 (May 11, 2017), p. 1786. ISSN: 2045-2322. DOI: 10.1038/s41598-017-01939-2.

[19] L. H. Frandsen and O. Sigmund. “Inverse Design Engineering of All-Silicon Polarization Beam Splitters”. In: SPIE OPTO. Ed. by A. Adibi, S.-Y. Lin, and A. Scherer. San Francisco, California, United States, Mar. 14, 2016, 97560Y. DOI: 10.1117/12.2210848.

[20] D. Vercruysse, N. V. Sapra, L. Su, R. Trivedi, and J. Vučković. “Analytical Level Set Fabrication Constraints for Inverse Design”. In: *Scientific Reports* 9.1 (June 21, 2019), p. 8999. ISSN: 2045-2322. DOI: 10.1038/s41598-019-45026-0.

[21] L. F. Frellsen, Y. Ding, O. Sigmund, and L. H. Frandsen. “Topology Optimized Mode Multiplexing in Silicon-on-Insulator Photonic Wire Waveguides”. In: *Optics Express* 24.15 (July 25, 2016), p. 16866. ISSN: 1094-4087. DOI: 10.1364/OE.24.016866.

[22] A. Y. Piggott, E. Y. Ma, L. Su, G. H. Ahn, N. V. Sapra, D. Vercruysse, A. M. Netherton, A. S. P. Khope, J. E. Bowers, and J. Vučković. "Inverse-Designed Photonics for Semiconductor Foundries". In: *ACS Photonics* 7.3 (Mar. 18, 2020), pp. 569–575. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.9b01540.

[23] A. M. Hammond, A. Oskooi, S. G. Johnson, and S. E. Ralph. "Photonic Topology Optimization with Semiconductor-Foundry Design-Rule Constraints". In: *Optics Express* 29.15 (July 19, 2021), p. 23916. ISSN: 1094-4087. DOI: 10.1364/OE.431188.

[24] M. F. Schubert, A. K. C. Cheung, I. A. D. Williamson, A. Spyra, and D. H. Alexander. "Inverse Design of Photonic Devices with Strict Foundry Fabrication Constraints". In: *ACS Photonics* 9.7 (July 20, 2022), pp. 2327–2336. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.2c00313.

[25] T. W. Hughes, M. Minkov, I. A. D. Williamson, and S. Fan. "Adjoint Method and Inverse Design for Nonlinear Nanophotonic Devices". In: *ACS Photonics* 5.12 (Dec. 19, 2018), pp. 4781–4787. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.8b01522.

[26] R. E. Christiansen, J. Michon, M. Benzaouia, O. Sigmund, and S. G. Johnson. "Inverse Design of Nanoparticles for Enhanced Raman Scattering". In: *Optics Express* 28.4 (Feb. 17, 2020), p. 4444. ISSN: 1094-4087. DOI: 10.1364/OE.28.004444.

[27] L. Raju, K.-T. Lee, Z. Liu, D. Zhu, M. Zhu, E. Poutrina, A. Urbas, and W. Cai. "Maximized Frequency Doubling through the Inverse Design of Nonlinear Metamaterials". In: *ACS Nano* 16.3 (Mar. 22, 2022), pp. 3926–3933. ISSN: 1936-0851, 1936-086X. DOI: 10.1021/acsnano.1c09298.

[28] S. A. Mann, H. Goh, and A. Alù. "Inverse Design of Nonlinear Polaritonic Metasurfaces for Second Harmonic Generation". In: *ACS Photonics* (Jan. 23, 2023), acsphotonics.2c01342. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.2c01342.

[29] C. Shang, J. Yang, A. M. Hammond, Z. Chen, M. Chen, Z. Lin, S. G. Johnson, and C. Wang. "Inverse-Designed Lithium Niobate Nanophotonics". In: *ACS Photonics* (Apr. 6, 2023), acsphotonics.3c00040. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.3c00040.

[30] C. Dory, D. Vercruysse, K. Y. Yang, N. V. Sapra, A. E. Rugar, S. Sun, D. M. Lukin, A. Y. Piggott, J. L. Zhang, M. Radulaski, K. G. Lagoudakis, L. Su, and J. Vučković. "Inverse-Designed Diamond Photonics". In: *Nature Communications* 10.1 (July 25, 2019), p. 3309. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11343-1.

[31] X. Lin, M. Wei, K. Lei, S. Yang, H. Ma, C. Zhong, Y. Luo, D. Li, J. Li, C. Lin, W. Zhang, S. Dai, X. Hu, L. Li, E. Li, and H. Lin. "Compact Mid-Infrared Chalcogenide Glass Photonic Devices Based on Robust-Inverse Design". In: *Laser & Photonics Reviews* 17.2 (Feb. 2023), p. 2200445. ISSN: 1863-8880, 1863-8899. DOI: 10.1002/lpor.202200445.

[32] M. A. Guidry, K. Y. Yang, D. M. Lukin, A. Markosyan, J. Yang, M. M. Fejer, and J. Vučković. "Optical Parametric Oscillation in Silicon Carbide Nanophotonics". In: *Optica* 7.9 (Sept. 20, 2020), p. 1139. ISSN: 2334-2536. DOI: 10.1364/OPTICA.394138.

[33] J. Yang, M. A. Guidry, D. M. Lukin, K. Yang, and J. Vučković. “Inverse-Designed Silicon Carbide Quantum and Nonlinear Photonics”. In: *Light: Science & Applications* 12.1 (Aug. 22, 2023), p. 201. ISSN: 2047-7538. DOI: 10.1038/s41377-023-01253-9.

[34] R. Pestourie, C. Pérez-Arancibia, Z. Lin, W. Shin, F. Capasso, and S. G. Johnson. “Inverse Design of Large-Area Metasurfaces”. In: *Optics Express* 26.26 (Dec. 24, 2018), p. 33732. ISSN: 1094-4087. DOI: 10.1364/OE.26.033732.

[35] Z. Lin, V. Liu, R. Pestourie, and S. G. Johnson. “Topology Optimization of Freeform Large-Area Metasurfaces”. In: *Optics Express* 27.11 (May 27, 2019), p. 15765. ISSN: 1094-4087. DOI: 10.1364/OE.27.015765.

[36] Z. Li, R. Pestourie, Z. Lin, S. G. Johnson, and F. Capasso. “Empowering Metasurfaces with Inverse Design: Principles and Applications”. In: *ACS Photonics* 9.7 (July 20, 2022), pp. 2178–2192. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.1c01850.

[37] 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”. In: *Optics Express* 28.23 (Nov. 9, 2020), p. 33854. ISSN: 1094-4087. DOI: 10.1364/OE.403192.

[38] Z. Li, R. Pestourie, J.-S. Park, Y.-W. Huang, S. G. Johnson, and F. Capasso. “Inverse Design Enables Large-Scale High-Performance Meta-Optics Reshaping Virtual Reality”. In: *Nature Communications* 13.1 (May 3, 2022), p. 2409. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29973-3.

[39] Z. Lin, C. Roques-Carmes, R. Pestourie, M. Soljačić, A. Majumdar, and S. G. Johnson. “End-to-End Nanophotonic Inverse Design for Imaging and Polarimetry”. In: *Nanophotonics* 10.3 (Jan. 22, 2021), pp. 1177–1187. ISSN: 2192-8614, 2192-8606. DOI: 10.1515/nanoph-2020-0579.

[40] Z. Lin, R. Pestourie, C. Roques-Carmes, Z. Li, F. Capasso, M. Soljačić, and S. G. Johnson. “End-to-End Metasurface Inverse Design for Single-Shot Multi-Channel Imaging”. In: *Optics Express* 30.16 (Aug. 1, 2022), p. 28358. ISSN: 1094-4087. DOI: 10.1364/OE.449985.

[41] M. Mansouree, H. Kwon, E. Arbabi, A. McClung, A. Faraon, and A. Arbabi. “Multifunctional 2.5D Metastructures Enabled by Adjoint Optimization”. In: *Optica* 7.1 (Jan. 20, 2020), p. 77. ISSN: 2334-2536. DOI: 10.1364/OPTICA.374787.

[42] V. Harinarayana and Y. Shin. “Two-Photon Lithography for Three-Dimensional Fabrication in Micro/Nanoscale Regime: A Comprehensive Review”. In: *Optics & Laser Technology* 142 (Oct. 2021), p. 107180. ISSN: 00303992. DOI: 10.1016/j.optlastec.2021.107180.

[43] V. Hahn, T. Messer, N. M. Bojanowski, E. R. Curticean, I. Wacker, R. R. Schröder, E. Blasco, and M. Wegener. “Two-Step Absorption Instead of Two-Photon Absorption in 3D Nanoprinting”. In: *Nature Photonics* 15.12 (Dec. 2021), pp. 932–938. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/s41566-021-00906-8.

[44] V. Hahn, P. Rietz, F. Hermann, P. Müller, C. Barner-Kowollik, T. Schlöder, W. Wenzel, E. Blasco, and M. Wegener. "Light-Sheet 3D Microprinting via Two-Colour Two-Step Absorption". In: *Nature Photonics* 16.11 (Nov. 2022), pp. 784–791. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/s41566-022-01081-0.

[45] P. Camayd-Muñoz, C. Ballew, G. Roberts, and A. Faraon. "Multifunctional Volumetric Meta-Optics for Color and Polarization Image Sensors". In: *Optica* 7.4 (Apr. 20, 2020), p. 280. ISSN: 2334-2536. DOI: 10.1364/OPTICA.384228.

[46] C. Roques-Carmes, Z. Lin, R. E. Christiansen, Y. Salamin, S. E. Kooi, J. D. Joannopoulos, S. G. Johnson, and M. Soljačić. "Toward 3D-Printed Inverse-Designed Metaoptics". In: *ACS Photonics* 9.1 (Jan. 19, 2022), pp. 43–51. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.1c01442.

[47] H. Wei, F. Callewaert, W. Hadibrata, V. Velev, Z. Liu, P. Kumar, K. Aydin, and S. Krishnaswamy. "Two-Photon Direct Laser Writing of Inverse-Designed Free-Form Near-Infrared Polarization Beamsplitter". In: *Advanced Optical Materials* 7.21 (Nov. 2019), p. 1900513. ISSN: 2195-1071, 2195-1071. DOI: 10.1002/adom.201900513.

[48] W. Hadibrata, H. Wei, S. Krishnaswamy, and K. Aydin. "Inverse Design and 3D Printing of a Metalens on an Optical Fiber Tip for Direct Laser Lithography". In: *Nano Letters* 21.6 (Mar. 24, 2021), pp. 2422–2428. ISSN: 1530-6984, 1530-6992. DOI: 10.1021/acs.nanolett.0c04463.

[49] J. C. Newman, A. C. Taylor, R. W. Barnwell, P. A. Newman, and G. J.-W. Hou. "Overview of Sensitivity Analysis and Shape Optimization for Complex Aerodynamic Configurations". In: *Journal of Aircraft* 36.1 (Jan. 1999), pp. 87–96. ISSN: 0021-8669, 1533-3868. DOI: 10.2514/2.2416.

[50] D. A. Tortorelli and P. Michaleris. "Design Sensitivity Analysis: Overview and Review". In: *Inverse Problems in Engineering* 1.1 (Oct. 1994), pp. 71–105. ISSN: 1068-2767, 1029-0281. DOI: 10.1080/174159794088027573.

[51] R. E. Kopp. "Pontryagin Maximum Principle". In: *Mathematics in Science and Engineering*. Vol. 5. Elsevier, 1962, pp. 255–279. ISBN: 978-0-12-442950-5. DOI: 10.1016/S0076-5392(08)62095-0.

[52] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. "Automatic Differentiation in Machine Learning: A Survey". In: *Journal of Machine Learning Research* 18.153 (2018), pp. 1–43. URL: http://jmlr.org/papers/v18/17-468.html.

[53] C. C. Margossian. "A Review of Automatic Differentiation and Its Efficient Implementation". In: *WIREs Data Mining and Knowledge Discovery* 9.4 (July 2019), e1305. ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1305.

[54] A. Druinsky and S. Toledo. *How Accurate Is Inv(A)*b?* Jan. 29, 2012. arXiv: 1201.6035 [cs, math]. URL: http://arxiv.org/abs/1201.6035 (visited on 04/05/2024). preprint.

[55] W. Wirtinger. "Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen". In: *Mathematische Annalen* 97.1 (Dec. 1927), pp. 357–375. ISSN: 0025-5831, 1432-1807. DOI: 10.1007/BF01447872.

[56] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces.* Red. by S. S. Antman, J. E. Marsden, and L. Sirovich. Vol. 153. Applied Mathematical Sciences. New York, NY: Springer New York, 2003. ISBN: 978-1-4684-9251-4. DOI: `10.1007/b98879`.

[57] S. G. Johnson, M. Ibanescu, M. A. Skorobogatiy, O. Weisberg, J. D. Joannopoulos, and Y. Fink. "Perturbation Theory for Maxwell's Equations with Shifting Material Boundaries". In: *Physical Review E* 65.6 (June 20, 2002), p. 066611. ISSN: 1063-651X, 1095-3787. DOI: `10.1103/PhysRevE.65.066611`.

[58] C. Kottke, A. Farjadpour, and S. G. Johnson. "Perturbation Theory for Anisotropic Dielectric Interfaces, and Application to Subpixel Smoothing of Discretized Numerical Methods". In: *Physical Review E* 77.3 (Mar. 25, 2008), p. 036611. ISSN: 1539-3755, 1550-2376. DOI: `10.1103/PhysRevE.77.036611`.

[59] O. Sigmund and K. Maute. "Topology Optimization Approaches: A Comparative Review". In: *Structural and Multidisciplinary Optimization* 48.6 (Dec. 2013), pp. 1031–1055. ISSN: 1615-147X, 1615-1488. DOI: `10.1007/s00158-013-0978-6`.

[60] O. Sigmund. "On the Design of Compliant Mechanisms Using Topology Optimization". In: *Mechanics of Structures and Machines* 25.4 (Jan. 1997), pp. 493–524. ISSN: 0890-5452. DOI: `10.1080/08905459708945415`.

[61] M. Zhou, B. S. Lazarov, F. Wang, and O. Sigmund. "Minimum Length Scale in Topology Optimization by Geometric Constraints". In: *Computer Methods in Applied Mechanics and Engineering* 293 (Aug. 2015), pp. 266–282. ISSN: 00457825. DOI: `10.1016/j.cma.2015.05.003`.

[62] S. Hooten, P. Sun, L. Gantz, M. Fiorentino, R. G. Beausoleil, and T. Van Vaerenbergh. *Automatic Differentiation Accelerated Shape Optimization Approaches to Photonic Inverse Design on Rectilinear Simulation Grids.* Nov. 7, 2023. arXiv: `2311.05646 [physics]`. URL: `http://arxiv.org/abs/2311.05646` (visited on 04/05/2024). preprint.

[63] Y. Augenstein. "Inverse Design of Nanophotonic Devices That Focus Bloch Surface Waves". MA thesis. Institute for Theoretical Solid State Physics, Karlsruhe Institute of Technology, 2019.

[64] R. D. Meade, K. D. Brommer, A. M. Rappe, and J. D. Joannopoulos. "Electromagnetic Bloch Waves at the Surface of a Photonic Crystal". In: *Physical Review B* 44.19 (Nov. 15, 1991), pp. 10961–10964. ISSN: 0163-1829, 1095-3795. DOI: `10.1103/PhysRevB.44.10961`.

[65] P. Yeh, A. Yariv, and C.-S. Hong. "Electromagnetic Propagation in Periodic Stratified Media I General Theory". In: *Journal of the Optical Society of America* 67.4 (Apr. 1, 1977), p. 423. ISSN: 0030-3941. DOI: `10.1364/JOSA.67.000423`.

[66] J. Chen, D. Zhang, P. Wang, H. Ming, and J. R. Lakowicz. "Strong Polarization Transformation of Bloch Surface Waves". In: *Physical Review Applied* 9.2 (Feb. 9, 2018), p. 024008. ISSN: 2331-7019. DOI: `10.1103/PhysRevApplied.9.024008`.

[67] A. V. Zayats, I. I. Smolyaninov, and A. A. Maradudin. “Nano-Optics of Surface Plasmon Polaritons”. In: *Physics Reports* 408.3-4 (Mar. 2005), pp. 131–314. ISSN: 03701573. DOI: 10.1016/j.physrep.2004.11.001.

[68] W. L. Barnes. “Surface Plasmon–Polariton Length Scales: A Route to Sub-Wavelength Optics”. In: *Journal of Optics A: Pure and Applied Optics* 8.4 (Apr. 1, 2006), S87–S93. ISSN: 1464-4258, 1741-3567. DOI: 10.1088/1464-4258/8/4/S06.

[69] R. Wang, Y. Wang, D. Zhang, G. Si, L. Zhu, L. Du, S. Kou, R. Badugu, M. Rosenfeld, J. Lin, P. Wang, H. Ming, X. Yuan, and J. R. Lakowicz. “Diffraction-Free Bloch Surface Waves”. In: *ACS Nano* 11.6 (June 27, 2017), pp. 5383–5390. ISSN: 1936-0851, 1936-086X. DOI: 10.1021/acsnano.7b02358.

[70] B. Vosoughi Lahijani, N. Descharmes, R. Barbey, G. D. Osowiecki, V. J. Wittwer, O. Razskazovskaya, T. Südmeyer, and H. P. Herzig. “Centimeter-Scale Propagation of Optical Surface Waves at Visible Wavelengths”. In: *Advanced Optical Materials* 10.10 (May 2022), p. 2102854. ISSN: 2195-1071, 2195-1071. DOI: 10.1002/adom.202102854.

[71] R. Dubey, E. Barakat, M. Häyrinen, M. Roussey, S. K. Honkanen, M. Kuittinen, and H. P. Herzig. “Experimental Investigation of the Propagation Properties of Bloch Surface Waves on Dielectric Multilayer Platform”. In: *Journal of the European Optical Society-Rapid Publications* 13.1 (Dec. 2017), p. 5. ISSN: 1990-2573. DOI: 10.1186/s41476-016-0029-1.

[72] A. Sinibaldi, N. Danz, E. Descrovi, P. Munzert, U. Schulz, F. Sonntag, L. Dominici, and F. Michelotti. “Direct Comparison of the Performance of Bloch Surface Wave and Surface Plasmon Polariton Sensors”. In: *Sensors and Actuators B: Chemical* 174 (Nov. 2012), pp. 292–298. ISSN: 09254005. DOI: 10.1016/j.snb.2012.07.015.

[73] L. Yu, E. Barakat, T. Sfez, L. Hvozdara, J. Di Francesco, and H. Peter Herzig. “Manipulating Bloch Surface Waves in 2D: A Platform Concept-Based Flat Lens”. In: *Light: Science & Applications* 3.1 (Jan. 3, 2014), e124–e124. ISSN: 2047-7538. DOI: 10.1038/lsa.2014.5.

[74] L. L. Doskolovich, E. A. Bezus, D. A. Bykov, and V. A. Soifer. “Spatial Differentiation of Bloch Surface Wave Beams Using an On-Chip Phase-Shifted Bragg Grating”. In: *Journal of Optics* 18.11 (Nov. 1, 2016), p. 115006. ISSN: 2040-8978, 2040-8986. DOI: 10.1088/2040-8978/18/11/115006.

[75] P. Rivolo, F. Michelotti, F. Frascella, G. Digregorio, P. Mandracci, L. Dominici, F. Giorgis, and E. Descrovi. “Real Time Secondary Antibody Detection by Means of Silicon-Based Multilayers Sustaining Bloch Surface Waves”. In: *Sensors and Actuators B: Chemical* 161.1 (Jan. 2012), pp. 1046–1052. ISSN: 09254005. DOI: 10.1016/j.snb.2011.12.006.

[76] Y. Kuai, Z. Xie, J. Chen, H. Gui, L. Xu, C. Kuang, P. Wang, X. Liu, J. Liu, Joseph. R. Lakowicz, and D. Zhang. “Real-Time Measurement of the Hygroscopic Growth Dynamics of Single Aerosol Nanoparticles with Bloch Surface Wave Microscopy”. In: *ACS Nano* 14.7 (July 28, 2020), pp. 9136–9144. ISSN: 1936-0851, 1936-086X. DOI: 10.1021/acsnano.0c04513.

[77] R. Wang, X. Lei, Y. Jin, X. Wen, L. Du, A. Wu, A. V. Zayats, and X. Yuan. “Directional Imbalance of Bloch Surface Waves for Ultrasensitive Displacement Metrology”. In: *Nanoscale* 13.25 (2021), pp. 11041–11050. ISSN: 2040-3364, 2040-3372. DOI: 10.1039/D1NR01251G.

[78] M. Scaravilli, G. Castaldi, A. Cusano, and V. Galdi. “Grating-Coupling-Based Excitation of Bloch Surface Waves for Lab-on-Fiber Optrodes”. In: *Optics Express* 24.24 (Nov. 28, 2016), p. 27771. ISSN: 1094-4087. DOI: 10.1364/OE.24.027771.

[79] M. Scaravilli, A. Micco, G. Castaldi, G. Coppola, M. Gioffrè, M. Iodice, V. La Ferrara, V. Galdi, and A. Cusano. “Excitation of Bloch Surface Waves on an Optical Fiber Tip”. In: *Advanced Optical Materials* 6.19 (Oct. 2018), p. 1800477. ISSN: 2195-1071, 2195-1071. DOI: 10.1002/adom.201800477.

[80] R. Wang, H. Xia, D. Zhang, J. Chen, L. Zhu, Y. Wang, E. Yang, T. Zang, X. Wen, G. Zou, P. Wang, H. Ming, R. Badugu, and J. R. Lakowicz. “Bloch Surface Waves Confined in One Dimension with a Single Polymeric Nanofibre”. In: *Nature Communications* 8.1 (Feb. 3, 2017), p. 14330. ISSN: 2041-1723. DOI: 10.1038/ncomms14330.

[81] L. Yu, E. Barakat, W. Nakagawa, and H. P. Herzig. “Investigation of Ultra-Thin Waveguide Arrays on a Bloch Surface Wave Platform”. In: *Journal of the Optical Society of America B* 31.12 (Dec. 1, 2014), p. 2996. ISSN: 0740-3224, 1520-8540. DOI: 10.1364/JOSAB.31.002996.

[82] T. Perani, D. Aurelio, and M. Liscidini. “Bloch-Surface-Wave Photonic Crystal Nanobeam Cavity”. In: *Optics Letters* 44.21 (Nov. 1, 2019), p. 5133. ISSN: 0146-9592, 1539-4794. DOI: 10.1364/OL.44.005133.

[83] T. Perani and M. Liscidini. “Long-Range Bloch Surface Waves in Photonic Crystal Ridges”. In: *Optics Letters* 45.23 (Dec. 1, 2020), p. 6534. ISSN: 0146-9592, 1539-4794. DOI: 10.1364/OL.412625.

[84] R. Dubey, B. Vosoughi Lahijani, M. Häyrinen, M. Roussey, M. Kuittinen, and H. P. Herzig. “Ultra-Thin Bloch-surface-wave-based Reflector at Telecommunication Wavelength”. In: *Photonics Research* 5.5 (Oct. 1, 2017), p. 494. ISSN: 2327-9125. DOI: 10.1364/PRJ.5.000494.

[85] M.-S. Kim, B. Vosoughi Lahijani, N. Descharmes, J. Straubel, F. Negredo, C. Rockstuhl, M. Häyrinen, M. Kuittinen, M. Roussey, and H. P. Herzig. “Subwavelength Focusing of Bloch Surface Waves”. In: *ACS Photonics* 4.6 (June 21, 2017), pp. 1477–1483. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.7b00245.

[86] M.-S. Kim, A. Vetter, C. Rockstuhl, B. V. Lahijani, M. Häyrinen, M. Kuittinen, M. Roussey, and H. P. Herzig. “Multiple Self-Healing Bloch Surface Wave Beams Generated by a Two-Dimensional Fraxicon”. In: *Communications Physics* 1.1 (Oct. 3, 2018), p. 63. ISSN: 2399-3650. DOI: 10.1038/s42005-018-0065-9.

[87] L. Novotny and B. Hecht. *Principles of Nano-Optics.* 2nd ed. Cambridge University Press, Sept. 6, 2012. ISBN: 978-1-107-00546-4. DOI: 10.1017/CBO9780511794193.

[88] P. Yeh. *Optical Waves in Layered Media.* Wiley Series in Pure and Applied Optics. Hoboken, NJ: Wiley-Interscience, 2005. 406 pp. ISBN: 978-0-471-73192-4.

[89] H. K. Baghbadorani, D. Aurelio, J. Barvestani, and M. Liscidini. "Guided Modes in Photonic Crystal Slabs Supporting Bloch Surface Waves". In: *Journal of the Optical Society of America B* 35.4 (Apr. 1, 2018), p. 805. ISSN: 0740-3224, 1520-8540. DOI: 10.1364/JOSAB.35.000805.

[90] U. Stella, L. Boarino, N. De Leo, P. Munzert, and E. Descrovi. "Enhanced Directional Light Emission Assisted by Resonant Bloch Surface Waves in Circular Cavities". In: *ACS Photonics* 6.8 (Aug. 21, 2019), pp. 2073–2082. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.9b00570.

[91] F. Wang, B. S. Lazarov, and O. Sigmund. "On Projection Methods, Convergence and Robust Formulations in Topology Optimization". In: *Structural and Multidisciplinary Optimization* 43.6 (June 2011), pp. 767–784. ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s00158-010-0602-y.

[92] B. S. Lazarov, F. Wang, and O. Sigmund. "Length Scale and Manufacturability in Density-Based Topology Optimization". In: *Archive of Applied Mechanics* 86.1-2 (Jan. 2016), pp. 189–218. ISSN: 0939-1533, 1432-0681. DOI: 10.1007/s00419-015-1106-4.

[93] K. Svanberg. "A Class of Globally Convergent Optimization Methods Based on Conservative Convex Separable Approximations". In: *SIAM Journal on Optimization* 12.2 (Jan. 2002), pp. 555–573. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/S1052623499362822.

[94] 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". In: *Optics Express* 30.3 (Jan. 31, 2022), p. 4467. ISSN: 1094-4087. DOI: 10.1364/OE.442074.

[95] *autograd*. Harvard Intelligent Probabilistic Systems Group. URL: https://github.com/HIPS/autograd (visited on 04/05/2024).

[96] S. Johnson and J. Joannopoulos. "Block-Iterative Frequency-Domain Methods for Maxwell's Equations in a Planewave Basis". In: *Optics Express* 8.3 (Jan. 29, 2001), p. 173. ISSN: 1094-4087. DOI: 10.1364/OE.8.000173.

[97] J. Baxter, A. Calà Lesina, J.-M. Guay, A. Weck, P. Berini, and L. Ramunno. "Plasmonic Colours Predicted by Deep Learning". In: *Scientific Reports* 9.1 (May 30, 2019), p. 8074. ISSN: 2045-2322. DOI: 10.1038/s41598-019-44522-7.

[98] Z. Jin, S. Mei, S. Chen, Y. Li, C. Zhang, Y. He, X. Yu, C. Yu, J. K. W. Yang, B. Luk'yanchuk, S. Xiao, and C.-W. Qiu. "Complex Inverse Design of Meta-optics by Segmented Hierarchical Evolutionary Algorithm". In: *ACS Nano* 13.1 (Jan. 22, 2019), pp. 821–829. ISSN: 1936-0851, 1936-086X. DOI: 10.1021/acsnano.8b08333.

[99] M. A. Barry, V. Berthier, B. D. Wilts, M.-C. Cambourieux, P. Bennet, R. Pollès, O. Teytaud, E. Centeno, N. Biais, and A. Moreau. "Evolutionary Algorithms Converge towards Evolved Biological Photonic Structures". In: *Scientific Reports* 10.1 (July 21, 2020), p. 12024. ISSN: 2045-2322. DOI: 10.1038/s41598-020-68719-3.

[100] P. Bennet, P. Juillet, S. Ibrahim, V. Berthier, M. A. Barry, F. Réveret, A. Bousquet, O. Teytaud, E. Centeno, and A. Moreau. “Analysis and Fabrication of Antireflective Coating for Photovoltaics Based on a Photonic-Crystal Concept and Generated by Evolutionary Optimization”. In: *Physical Review B* 103.12 (Mar. 16, 2021), p. 125135. ISSN: 2469-9950, 2469-9969. DOI: 10.1103/PhysRevB.103.125135.

[101] M. Makarenko, Q. Wang, A. Burguete-Lopez, F. Getman, and A. Fratalocchi. “Robust and Scalable Flat-Optics on Flexible Substrates via Evolutionary Neural Networks”. In: *Advanced Intelligent Systems* 3.11 (Nov. 2021), p. 2100105. ISSN: 2640-4567, 2640-4567. DOI: 10.1002/aisy.202100105.

[102] J. Lim and D. Psaltis. “MaxwellNet: Physics-driven Deep Neural Network Training Based on Maxwell’s Equations”. In: *APL Photonics* 7.1 (Jan. 1, 2022), p. 011301. ISSN: 2378-0967. DOI: 10.1063/5.0071616.

[103] M. Chen, R. Lupoiu, C. Mao, D.-H. Huang, J. Jiang, P. Lalanne, and J. A. Fan. “High Speed Simulation and Freeform Optimization of Nanophotonic Devices with Physics-Augmented Deep Learning”. In: *ACS Photonics* 9.9 (Sept. 21, 2022), pp. 3110–3123. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.2c00876.

[104] S. Scheerlinck, J. Schrauwen, F. Van Laere, D. Taillaert, D. Van Thourhout, and R. Baets. “Efficient, Broadband and Compact Metal Grating Couplers for Silicon-on-Insulator Waveguides”. In: *Optics Express* 15.15 (2007), p. 9625. ISSN: 1094-4087. DOI: 10.1364/OE.15.009625.

[105] D. Vermeulen, S. Selvaraja, P. Verheyen, G. Lepage, W. Bogaerts, P. Absil, D. Van Thourhout, and G. Roelkens. “High-Efficiency Fiber-to-Chip Grating Couplers Realized Using an Advanced CMOS-compatible Silicon-On-Insulator Platform”. In: *Optics Express* 18.17 (Aug. 16, 2010), p. 18278. ISSN: 1094-4087. DOI: 10.1364/OE.18.018278.

[106] X. Chen, C. Li, C. K. Y. Fung, S. M. G. Lo, and H. K. Tsang. “Apodized Waveguide Grating Couplers for Efficient Coupling to Optical Fibers”. In: *IEEE Photonics Technology Letters* 22.15 (Aug. 2010), pp. 1156–1158. ISSN: 1041-1135, 1941-0174. DOI: 10.1109/LPT.2010.2051220.

[107] Y. Tang, Z. Wang, L. Wosinski, U. Westergren, and S. He. “Highly Efficient Nonuniform Grating Coupler for Silicon-on-Insulator Nanophotonic Circuits”. In: *Optics Letters* 35.8 (Apr. 15, 2010), p. 1290. ISSN: 0146-9592, 1539-4794. DOI: 10.1364/OL.35.001290.

[108] S. McNab, N. Moll, and Y. Vlasov. “Ultra-Low Loss Photonic Integrated Circuit with Membrane-Type Photonic Crystal Waveguides”. In: *Optics Express* 11.22 (Nov. 3, 2003), p. 2927. ISSN: 1094-4087. DOI: 10.1364/OE.11.002927.

[109] M. Papes, P. Cheben, D. Benedikovic, J. H. Schmid, J. Pond, R. Halir, A. Ortega-Moñux, G. Wangüemert-Pérez, W. N. Ye, D.-X. Xu, S. Janz, M. Dado, and V. Vašinek. “Fiber-Chip Edge Coupler with Large Mode Size for Silicon Photonic Wire Waveguides”. In: *Optics Express* 24.5 (Mar. 7, 2016), p. 5026. ISSN: 1094-4087. DOI: 10.1364/OE.24.005026.

[110] A. He, X. Guo, K. Wang, Y. Zhang, and Y. Su. “Low Loss, Large Bandwidth Fiber-Chip Edge Couplers Based on Silicon-on-Insulator Platform”. In: *Journal of Lightwave Technology* 38.17 (Sept. 1, 2020), pp. 4780–4786. ISSN: 0733-8724, 1558-2213. DOI: 10.1109/JLT.2020.2995544.

[111] P.-I. Dietrich, R. J. Harris, M. Blaicher, M. K. Corrigan, T. J. Morris, W. Freude, A. Quirrenbach, and C. Koos. “Printed Freeform Lens Arrays on Multi-Core Fibers for Highly Efficient Coupling in Astrophotonic Systems”. In: *Optics Express* 25.15 (July 24, 2017), p. 18288. ISSN: 1094-4087. DOI: 10.1364/OE.25.018288.

[112] S. Singer, Y. Xu, S. T. Skacel, Y. Bao, H. Zwickel, P. Maier, L. Freter, P.-I. Dietrich, M. Kaschel, C. Menzel, S. Randel, W. Freude, and C. Koos. “3D-printed Facet-Attached Optical Elements for Beam Shaping in Optical Phased Arrays”. In: *Optics Express* 30.26 (Dec. 19, 2022), p. 46564. ISSN: 1094-4087. DOI: 10.1364/OE.456952.

[113] N. Lindenmann, G. Balthasar, D. Hillerkuss, R. Schmogrow, M. Jordan, J. Leuthold, W. Freude, and C. Koos. “Photonic Wire Bonding: A Novel Concept for Chip-Scale Interconnects”. In: *Optics Express* 20.16 (July 30, 2012), p. 17667. ISSN: 1094-4087. DOI: 10.1364/OE.20.017667.

[114] M. R. Billah, M. Blaicher, T. Hoose, P.-I. Dietrich, P. Marin-Palomo, N. Lindenmann, A. Nesic, A. Hofmann, U. Troppenz, M. Moehrle, S. Randel, W. Freude, and C. Koos. “Hybrid Integration of Silicon Photonics Circuits and InP Lasers by Photonic Wire Bonding”. In: *Optica* 5.7 (July 20, 2018), p. 876. ISSN: 2334-2536. DOI: 10.1364/OPTICA.5.000876.

[115] E. Luan, S. Yu, M. Salmani, M. S. Nezami, B. J. Shastri, L. Chrostowski, and A. Eshaghi. “Towards a High-Density Photonic Tensor Core Enabled by Intensity-Modulated Microrings and Photonic Wire Bonding”. In: *Scientific Reports* 13.1 (Jan. 23, 2023), p. 1260. ISSN: 2045-2322. DOI: 10.1038/s41598-023-27724-y.

[116] N. Lindenmann, S. Dottermusch, M. L. Goedecke, T. Hoose, M. R. Billah, T. P. Onanuga, A. Hofmann, W. Freude, and C. Koos. “Connecting Silicon Photonic Circuits to Multicore Fibers by Photonic Wire Bonding”. In: *Journal of Lightwave Technology* 33.4 (Feb. 15, 2015), pp. 755–760. ISSN: 0733-8724, 1558-2213. DOI: 10.1109/JLT.2014.2373051.

[117] X. Garcia-Santiago, S. Burger, C. Rockstuhl, and P.-I. Schneider. “Bayesian Optimization With Improved Scalability and Derivative Information for Efficient Design of Nanophotonic Structures”. In: *Journal of Lightwave Technology* 39.1 (Jan. 1, 2021), pp. 167–177. ISSN: 0733-8724, 1558-2213. DOI: 10.1109/JLT.2020.3023450.

[118] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization.* Jan. 29, 2017. arXiv: 1412.6980 [cs]. URL: http://arxiv.org/abs/1412.6980 (visited on 04/05/2024). preprint.

[119] Z. Zhang, X. Chen, Q. Cheng, A. Z. Khokhar, Z. Zhang, X. Yan, B. Huang, H. Chen, H. Liu, H. Li, D. J. Thomson, and G. T. Reed. “Two-Dimensional Apodized Grating Coupler for Polarization-Independent and Surface-Normal Optical Coupling”. In: *Journal of Lightwave Technology* (2020), pp. 1–1. ISSN: 0733-8724, 1558-2213. DOI: 10.1109/JLT.2020.2986043.

[120] A. M. Hammond, J. B. Slaby, M. J. Probst, and S. E. Ralph. "Multi-Layer Inverse Design of Vertical Grating Couplers for High-Density, Commercial Foundry Interconnects". In: *Optics Express* 30.17 (Aug. 15, 2022), p. 31058. ISSN: 1094-4087. DOI: 10.1364/OE.466015.

[121] P. Sun, T. Van Vaerenbergh, S. Hooten, and R. Beausoleil. "Adjoint Optimization of Polarization-Splitting Grating Couplers". In: *Optics Express* 31.3 (Jan. 30, 2023), p. 4884. ISSN: 1094-4087. DOI: 10.1364/OE.477532.

[122] C. Alonso-Ramos, L. Zavargo-Peche, A. Ortega-Moñux, R. Halir, I. Molina-Fernández, and P. Cheben. "Polarization-Independent Grating Coupler for Micrometric Silicon Rib Waveguides". In: *Optics Letters* 37.17 (Sept. 1, 2012), p. 3663. ISSN: 0146-9592, 1539-4794. DOI: 10.1364/OL.37.003663.

[123] J. H. Song, F. E. Doany, A. K. Medhin, N. Dupuis, B. G. Lee, and F. R. Libsch. "Polarization-Independent Nonuniform Grating Couplers on Silicon-on-Insulator". In: *Optics Letters* 40.17 (Sept. 1, 2015), p. 3941. ISSN: 0146-9592, 1539-4794. DOI: 10.1364/OL.40.003941.

[124] B. Zhang, M. Schiller, K. Al Qubaisi, D. Onural, A. Khilo, M. J. Naughton, and M. A. Popović. "Polarization-Insensitive 1D Grating Coupler Based on a Zero-Birefringence Subwavelength Corelet Waveguide". In: *Optics Letters* 47.13 (July 1, 2022), p. 3167. ISSN: 0146-9592, 1539-4794. DOI: 10.1364/OL.459306.

[125] T. W. Hughes, M. Minkov, V. Liu, Z. Yu, and S. Fan. "Full Wave Simulation and Optimization of Large Area Metalens". In: *OSA Optical Design and Fabrication 2021 (Flat Optics, Freeform, IODC, OFT)*. Flat Optics: Components to Systems. Washington, DC: Optica Publishing Group, 2021, FTh3C.5. ISBN: 978-1-943580-88-0. DOI: 10.1364/FLATOPTICS.2021.FTh3C.5.

[126] H. Gehring, M. Blaicher, W. Hartmann, P. Varytis, K. Busch, M. Wegener, and W. H. P. Pernice. "Low-Loss Fiber-to-Chip Couplers with Ultrawide Optical Bandwidth". In: *APL Photonics* 4.1 (Jan. 1, 2019), p. 010801. ISSN: 2378-0967. DOI: 10.1063/1.5064401.

[127] A. Bozzola, L. Carroll, D. Gerace, I. Cristiani, and L. C. Andreani. "Optimising Apodized Grating Couplers in a Pure SOI Platform to −0.5dB Coupling Efficiency". In: *Optics Express* 23.12 (June 15, 2015), p. 16289. ISSN: 1094-4087. DOI: 10.1364/OE.23.016289.

[128] D. Benedikovic, C. Alonso-Ramos, S. Guerber, X. Le Roux, P. Cheben, C. Dupré, B. Szelag, D. Fowler, É. Cassan, D. Marris-Morini, C. Baudot, F. Boeuf, and L. Vivien. "Sub-Decibel Silicon Grating Couplers Based on L-shaped Waveguides and Engineered Subwavelength Metamaterials". In: *Optics Express* 27.18 (Sept. 2, 2019), p. 26239. ISSN: 1094-4087. DOI: 10.1364/OE.27.026239.

[129] J. S. Jensen and O. Sigmund. "Topology Optimization of Photonic Crystal Structures: A High-Bandwidth Low-Loss T-junction Waveguide". In: *Journal of the Optical Society of America B* 22.6 (June 1, 2005), p. 1191. ISSN: 0740-3224, 1520-8540. DOI: 10.1364/JOSAB.22.001191.

[130] D. Sell, J. Yang, S. Doshay, R. Yang, and J. A. Fan. “Large-Angle, Multifunctional Metagratings Based on Freeform Multimode Geometries”. In: *Nano Letters* 17.6 (June 14, 2017), pp. 3752–3757. ISSN: 1530-6984, 1530-6992. DOI: 10.1021/acs.nanolett.7b01082.

[131] T. Hughes, G. Veronis, K. P. Wootton, R. Joel England, and S. Fan. “Method for Computationally Efficient Design of Dielectric Laser Accelerator Structures”. In: *Optics Express* 25.13 (June 26, 2017), p. 15414. ISSN: 1094-4087. DOI: 10.1364/OE.25.015414.

[132] Z. Xie, T. Lei, H. Qiu, Z. Zhang, H. Wang, and X. Yuan. “Broadband On-Chip Photonic Spin Hall Element via Inverse Design”. In: *Photonics Research* 8.2 (Feb. 1, 2020), p. 121. ISSN: 2327-9125. DOI: 10.1364/PRJ.8.000121.

[133] Y. Deng and J. G. Korvink. “Topology Optimization for Three-Dimensional Electromagnetic Waves Using an Edge Element-Based Finite-Element Method”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 472.2189 (May 2016), p. 20150835. ISSN: 1364-5021, 1471-2946. DOI: 10.1098/rspa.2015.0835.

[134] C. Barner-Kowollik, M. Bastmeyer, E. Blasco, G. Delaittre, P. Müller, B. Richter, and M. Wegener. “3D Laser Micro- and Nanoprinting: Challenges for Chemistry”. In: *Angewandte Chemie International Edition* 56.50 (Dec. 11, 2017), pp. 15828–15845. ISSN: 1433-7851, 1521-3773. DOI: 10.1002/anie.201704695.

[135] V. Hahn, P. Kiefer, T. Frenzel, J. Qu, E. Blasco, C. Barner-Kowollik, and M. Wegener. “Rapid Assembly of Small Materials Building Blocks (Voxels) into Large Functional 3D Metamaterials”. In: *Advanced Functional Materials* 30.26 (June 2020), p. 1907795. ISSN: 1616-301X, 1616-3028. DOI: 10.1002/adfm.201907795.

[136] M. P. Bendsøe. *Optimization of Structural Topology, Shape, and Material.* Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. ISBN: 978-3-662-03115-5. DOI: 10.1007/978-3-662-03115-5.

[137] G. Rozvany. “Aims, Scope, Methods, History and Unified Terminology of Computer-Aided Topology Optimization in Structural Mechanics”. In: *Structural and Multidisciplinary Optimization* 21.2 (Apr. 2001), pp. 90–108. ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s001580050174.

[138] M. P. Bendsøe and O. Sigmund. *Topology Optimization.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. ISBN: 978-3-662-05086-6. DOI: 10.1007/978-3-662-05086-6.

[139] O. Sigmund. “Morphology-Based Black and White Filters for Topology Optimization”. In: *Structural and Multidisciplinary Optimization* 33.4-5 (Feb. 26, 2007), pp. 401–424. ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s00158-006-0087-x.

[140] *fdfdpy.* Fan Group. URL: https://github.com/fancompute/fdfdpy (visited on 04/05/2024).

[141] C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch. “Adjoint Shape Optimization Applied to Electromagnetic Design”. In: *Optics Express* 21.18 (Sept. 9, 2013), p. 21693. ISSN: 1094-4087. DOI: 10.1364/OE.21.021693.

[142] O. Sigmund. “A 99 Line Topology Optimization Code Written in Matlab”. In: *Structural and Multidisciplinary Optimization* 21.2 (Apr. 2001), pp. 120–127. ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s001580050176.

[143] D. C. Liu and J. Nocedal. “On the Limited Memory BFGS Method for Large Scale Optimization”. In: *Mathematical Programming* 45.1-3 (Aug. 1989), pp. 503–528. ISSN: 0025-5610, 1436-4646. DOI: 10.1007/BF01589116.

[144] P.-I. Dietrich, M. Blaicher, I. Reuter, M. Billah, T. Hoose, A. Hofmann, C. Caer, R. Dangel, B. Offrein, U. Troppenz, M. Moehrle, W. Freude, and C. Koos. “In Situ 3D Nanoprinting of Free-Form Coupling Elements for Hybrid Photonic Integration”. In: *Nature Photonics* 12.4 (Apr. 2018), pp. 241–247. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/s41566-018-0133-4.

[145] T. Gissibl, S. Thiele, A. Herkommer, and H. Giessen. “Sub-Micrometre Accurate Free-Form Optics by Three-Dimensional Printing on Single-Mode Fibres”. In: *Nature Communications* 7.1 (June 24, 2016), p. 11763. ISSN: 2041-1723. DOI: 10.1038/ncomms11763.

[146] T. Gissibl, S. Thiele, A. Herkommer, and H. Giessen. “Two-Photon Direct Laser Writing of Ultracompact Multi-Lens Objectives”. In: *Nature Photonics* 10.8 (Aug. 2016), pp. 554–560. ISSN: 1749-4885, 1749-4893. DOI: 10.1038/nphoton.2016.121.

[147] S. Thiele, K. Arzenbacher, T. Gissibl, H. Giessen, and A. M. Herkommer. “3D-printed Eagle Eye: Compound Microlens System for Foveated Imaging”. In: *Science Advances* 3.2 (Feb. 3, 2017), e1602655. ISSN: 2375-2548. DOI: 10.1126/sciadv.1602655.

[148] D. A. B. Miller. “All Linear Optical Devices Are Mode Converters”. In: *Optics Express* 20.21 (Oct. 8, 2012), p. 23985. ISSN: 1094-4087. DOI: 10.1364/OE.20.023985.

[149] O. Sigmund. “Manufacturing Tolerant Topology Optimization”. In: *Acta Mechanica Sinica* 25.2 (Apr. 2009), pp. 227–239. ISSN: 0567-7718, 1614-3116. DOI: 10.1007/s10409-009-0240-z.

[150] J. Liu and Y. Ma. “A Survey of Manufacturing Oriented Topology Optimization Methods”. In: *Advances in Engineering Software* 100 (Oct. 2016), pp. 161–175. ISSN: 09659978. DOI: 10.1016/j.advengsoft.2016.07.017.

[151] Q. Li, W. Chen, S. Liu, and L. Tong. “Structural Topology Optimization Considering Connectivity Constraint”. In: *Structural and Multidisciplinary Optimization* 54.4 (Oct. 2016), pp. 971–984. ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s00158-016-1459-5.

[152] Yong-Jiu Zhao, Ke-Li Wu, and K.-K. Cheng. “A Compact 2-D Full-Wave Finite-Difference Frequency-Domain Method for General Guided Wave Structures”. In: *IEEE Transactions on Microwave Theory and Techniques* 50.7 (July 2002), pp. 1844–1848. ISSN: 0018-9480. DOI: 10.1109/TMTT.2002.800447.

[153] T. W. Hughes, I. A. D. Williamson, M. Minkov, and S. Fan. “Forward-Mode Differentiation of Maxwell’s Equations”. In: *ACS Photonics* 6.11 (Nov. 20, 2019), pp. 3010–3016. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.9b01238.

[154] A. Vaccari, A. Cala' Lesina, L. Cristoforetti, and R. Pontalti. “Parallel Implementation of a 3D Subgridding FDTD Algorithm for Large Simulations”. In: *Progress In Electromagnetics Research* 120 (2011), pp. 263–292. ISSN: 1559-8985. DOI: 10.2528/PIER11063004.

[155] P.-I. Schneider, X. Garcia Santiago, V. Soltwisch, M. Hammerschmidt, S. Burger, and C. Rockstuhl. “Benchmarking Five Global Optimization Approaches for Nano-optical Shape Optimization and Parameter Reconstruction”. In: *ACS Photonics* 6.11 (Nov. 20, 2019), pp. 2726–2733. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.9b00706.

[156] T. W. Hughes, M. Minkov, V. Liu, Z. Yu, and S. Fan. “A Perspective on the Pathway toward Full Wave Simulation of Large Area Metalenses”. In: *Applied Physics Letters* 119.15 (Oct. 11, 2021), p. 150502. ISSN: 0003-6951, 1077-3118. DOI: 10.1063/5.0071245.

[157] H.-C. Lin, Z. Wang, and C. W. Hsu. “Fast Multi-Source Nanophotonic Simulations Using Augmented Partial Factorization”. In: *Nature Computational Science* 2.12 (Dec. 15, 2022), pp. 815–822. ISSN: 2662-8457. DOI: 10.1038/s43588-022-00370-6.

[158] L. Li. “New Formulation of the Fourier Modal Method for Crossed Surface-Relief Gratings”. In: *Journal of the Optical Society of America A* 14.10 (Oct. 1, 1997), p. 2758. ISSN: 1084-7529, 1520-8532. DOI: 10.1364/JOSAA.14.002758.

[159] V. Liu and S. Fan. “S4 : A Free Electromagnetic Solver for Layered Periodic Structures”. In: *Computer Physics Communications* 183.10 (Oct. 2012), pp. 2233–2244. ISSN: 00104655. DOI: 10.1016/j.cpc.2012.04.026.

[160] M. Minkov, I. A. D. Williamson, L. C. Andreani, D. Gerace, B. Lou, A. Y. Song, T. W. Hughes, and S. Fan. “Inverse Design of Photonic Crystals through Automatic Differentiation”. In: *ACS Photonics* 7.7 (July 15, 2020), pp. 1729–1741. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.0c00327.

[161] D. Beutel, A. Groner, C. Rockstuhl, and I. Fernandez-Corbaton. “Efficient Simulation of Biperiodic, Layered Structures Based on the T-matrix Method”. In: *Journal of the Optical Society of America B* 38.6 (June 1, 2021), p. 1782. ISSN: 0740-3224, 1520-8540. DOI: 10.1364/JOSAB.419645.

[162] P. R. Wiecha, A. Arbouet, C. Girard, and O. L. Muskens. “Deep Learning in Nano-Photonics: Inverse Design and Beyond”. In: *Photonics Research* 9.5 (May 1, 2021), B182. ISSN: 2327-9125. DOI: 10.1364/PRJ.415960.

[163] R. Pestourie, Y. Mroueh, C. Rackauckas, P. Das, and S. G. Johnson. “Physics-Enhanced Deep Surrogates for Partial Differential Equations”. In: *Nature Machine Intelligence* 5.12 (Dec. 4, 2023), pp. 1458–1465. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00761-y. arXiv: 2111.05841 [physics].

[164] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. “Learning Nonlinear Operators via DeepONet Based on the Universal Approximation Theorem of Operators”. In: *Nature Machine Intelligence* 3.3 (Mar. 18, 2021), pp. 218–229. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00302-5.

[165] X. Chen, R. Chen, Q. Wan, R. Xu, and J. Liu. “An Improved Data-Free Surrogate Model for Solving Partial Differential Equations Using Deep Neural Networks”. In: *Scientific Reports* 11.1 (Sept. 30, 2021), p. 19507. ISSN: 2045-2322. DOI: 10.1038/s41598-021-99037-x.

[166] K. Hornik, M. Stinchcombe, and H. White. “Multilayer Feedforward Networks Are Universal Approximators”. In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90020-8.

[167] J. Baxter, J. Desautels, A. C. Lesina, P. Berini, and L. Ramunno. “Deep Learning for Engineering Optical Scattering from Plasmonic Nanostructures”. In: *OSA Optical Design and Fabrication 2021 (Flat Optics, Freeform, IODC, OFT)*. Flat Optics: Components to Systems. Washington, DC: Optica Publishing Group, 2021, JW2D.4. ISBN: 978-1-943580-88-0. DOI: 10.1364/FLATOPTICS.2021.JW2D.4.

[168] S. Krasikov, A. Tranter, A. Bogdanov, Y. Kivshar, Nonlinear Physics Center, Research School of Physics, The Australian National University, Canberra ACT 2601, Australia, School of Physics and Engineering, ITMO University, St. Petersburg 197101, Russia, and Centre for Quantum Computation and Communication Technology, Department of Quantum Science, Research School of Physics, The Australian National University, Canberra, ACT 2601, Australia. “Intelligent Metaphotonics Empowered by Machine Learning”. In: *Opto-Electronic Advances* 5.3 (2022), pp. 210147–210147. ISSN: 2096-4579. DOI: 10.29026/oea.2022.210147.

[169] S. An, C. Fowler, B. Zheng, M. Y. Shalaginov, H. Tang, H. Li, L. Zhou, J. Ding, A. M. Agarwal, C. Rivero-Baleine, K. A. Richardson, T. Gu, J. Hu, and H. Zhang. “A Deep Learning Approach for Objective-Driven All-Dielectric Metasurface Design”. In: *ACS Photonics* 6.12 (Dec. 18, 2019), pp. 3196–3207. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.9b00966.

[170] J. Jiang and J. A. Fan. “Simulator-Based Training of Generative Neural Networks for the Inverse Design of Metasurfaces”. In: *Nanophotonics* 9.5 (May 26, 2020), pp. 1059–1069. ISSN: 2192-8614, 2192-8606. DOI: 10.1515/nanoph-2019-0330.

[171] P. R. Wiecha and O. L. Muskens. “Deep Learning Meets Nanophotonics: A Generalized Accurate Predictor for Near Fields and Far Fields of Arbitrary 3D Nanostructures”. In: *Nano Letters* 20.1 (Jan. 8, 2020), pp. 329–338. ISSN: 1530-6984, 1530-6992. DOI: 10.1021/acs.nanolett.9b03971.

[172] R. Pestourie, Y. Mroueh, T. V. Nguyen, P. Das, and S. G. Johnson. “Active Learning of Deep Surrogates for PDEs: Application to Metasurface Design”. In: *npj Computational Materials* 6.1 (Oct. 29, 2020), p. 164. ISSN: 2057-3960. DOI: 10.1038/s41524-020-00431-2.

[173] L. Lu, R. Pestourie, W. Yao, Z. Wang, F. Verdugo, and S. G. Johnson. “Physics-Informed Neural Networks with Hard Constraints for Inverse Design”. In: *SIAM Journal on Scientific Computing* 43.6 (Jan. 2021), B1105–B1132. ISSN: 1064-8275, 1095-7197. DOI: 10.1137/21M1397908.

[174] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli. “Scientific Machine Learning Through Physics–Informed Neural Networks: Where We Are and What’s Next”. In: *Journal of Scientific Computing* 92.3 (Sept. 2022), p. 88. ISSN: 0885-7474, 1573-7691. DOI: `10.1007/s10915-022-01939-z`.

[175] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. *Neural Operator: Graph Kernel Network for Partial Differential Equations.* Mar. 6, 2020. arXiv: `2003.03485 [cs, math, stat]`. URL: `http://arxiv.org/abs/2003.03485` (visited on 04/05/2024). preprint.

[176] L. Lu, R. Pestourie, S. G. Johnson, and G. Romano. “Multifidelity Deep Neural Operators for Efficient Learning of Partial Differential Equations with Application to Fast Inverse Design of Nanoscale Heat Transport”. In: *Physical Review Research* 4.2 (June 13, 2022), p. 023210. ISSN: 2643-1564. DOI: `10.1103/PhysRevResearch.4.023210`.

[177] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. *Fourier Neural Operator for Parametric Partial Differential Equations.* May 16, 2021. arXiv: `2010.08895 [cs, math]`. URL: `http://arxiv.org/abs/2010.08895` (visited on 04/05/2024). preprint.

[178] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. “Handwritten Digit Recognition with a Back-Propagation Network”. In: *Advances in Neural Information Processing Systems.* Ed. by D. Touretzky. Vol. 2. Morgan-Kaufmann, 1989.

[179] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems.* Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc., 2012.

[180] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Vol. 9351. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: `10.1007/978-3-319-24574-4\_28`.

[181] N. Borhani, E. Kakkava, C. Moser, and D. Psaltis. “Learning to See through Multimode Fibers”. In: *Optica* 5.8 (Aug. 20, 2018), p. 960. ISSN: 2334-2536. DOI: `10.1364/OPTICA.5.000960`.

[182] Z. Li, H. Zheng, N. Kovachki, D. Jin, H. Chen, B. Liu, K. Azizzadenesheli, and A. Anandkumar. “Physics-Informed Neural Operator for Learning Partial Differential Equations”. In: *ACM / IMS Journal of Data Science* (Feb. 21, 2024), p. 3648506. ISSN: 2831-3194. DOI: `10.1145/3648506`.

[183] L. Lu, X. Meng, S. Cai, Z. Mao, S. Goswami, Z. Zhang, and G. E. Karniadakis. “A Comprehensive and Fair Comparison of Two Neural Operators (with Practical Extensions) Based on FAIR Data”. In: *Computer Methods in Applied Mechanics and Engineering* 393 (Apr. 2022), p. 114778. ISSN: 00457825. DOI: `10.1016/j.cma.2022.114778`.

[184] S. Goswami, M. Yin, Y. Yu, and G. E. Karniadakis. “A Physics-Informed Variational DeepONet for Predicting Crack Path in Quasi-Brittle Materials”. In: *Computer Methods in Applied Mechanics and Engineering* 391 (Mar. 2022), p. 114587. ISSN: 00457825. DOI: 10.1016/j.cma.2022.114587.

[185] S. Goswami, A. Bora, Y. Yu, and G. E. Karniadakis. “Physics-Informed Deep Neural Operator Networks”. In: *Machine Learning in Modeling and Simulation*. Ed. by T. Rabczuk and K.-J. Bathe. Cham: Springer International Publishing, 2023, pp. 219–254. ISBN: 978-3-031-36644-4. DOI: 10.1007/978-3-031-36644-4_6.

[186] V. S. Fanaskov and I. V. Oseledets. “Spectral Neural Operators”. In: *Doklady Mathematics* 108.S2 (Dec. 2023), S226–S232. ISSN: 1064-5624, 1531-8362. DOI: 10.1134/S1064562423701107.

[187] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. *Neural Operator: Learning Maps Between Function Spaces*. Apr. 7, 2023. arXiv: 2108.08481 [cs, math]. URL: http://arxiv.org/abs/2108.08481 (visited on 04/05/2024). preprint.

[188] F. Bartolucci, E. de Bézenac, B. Raonić, R. Molinaro, S. Mishra, and R. Alaifari. *Representation Equivalent Neural Operators: A Framework for Alias-free Operator Learning*. Nov. 2, 2023. arXiv: 2305.19913 [cs, eess]. URL: http://arxiv.org/abs/2305.19913 (visited on 04/05/2024). preprint.

[189] N. B. Kovachki, S. Lanthaler, and A. M. Stuart. *Operator Learning: Algorithms and Analysis*. Feb. 23, 2024. arXiv: 2402.15715 [cs, math]. URL: http://arxiv.org/abs/2402.15715 (visited on 04/05/2024). preprint.

[190] S. Ioffe and C. Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 7–9, 2015, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[191] D. Hendrycks and K. Gimpel. *Gaussian Error Linear Units (GELUs)*. June 5, 2023. arXiv: 1606.08415 [cs]. URL: http://arxiv.org/abs/1606.08415 (visited on 04/05/2024). preprint.

[192] D. E. Rumelhart, J. L. McClelland, and AU. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press, 1986. ISBN: 978-0-262-29140-8. DOI: 10.7551/mitpress/5236.001.0001.

[193] J. Schmidhuber. “Deep Learning in Neural Networks: An Overview”. In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. ISSN: 08936080. DOI: 10.1016/j.neunet.2014.09.003.

[194] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. Dec. 10, 2022. arXiv: 1312.6114 [cs, stat]. URL: http://arxiv.org/abs/1312.6114 (visited on 04/05/2024). preprint.

[195] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[196] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. Jan. 4, 2019. arXiv: `1711.05101 [cs, math]`. URL: `http://arxiv.org/abs/1711.05101` (visited on 04/05/2024). preprint.

[197] L. N. Smith and N. Topin. "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates". In: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*. Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. Ed. by T. Pham. Baltimore, United States: SPIE, May 10, 2019, p. 36. DOI: `10.1117/12.2520589`.

[198] FairScale authors. *FairScale: A General Purpose Modular PyTorch Library for High Performance and Large Scale Training*. URL: `https://github.com/facebookresearch/fairscale` (visited on 04/05/2024).

[199] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[200] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. *Self-Normalizing Neural Networks*. Sept. 7, 2017. arXiv: `1706.02515 [cs, stat]`. URL: `http://arxiv.org/abs/1706.02515` (visited on 04/05/2024). preprint.

[201] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. "Deconvolutional Networks". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA: IEEE, June 2010, pp. 2528–2535. ISBN: 978-1-4244-6984-0. DOI: `10.1109/CVPR.2010.5539957`.

[202] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. Van Langen, and R. A. Kievit. "Raincloud Plots: A Multi-Platform Tool for Robust Data Visualization". In: *Wellcome Open Research* 4 (Jan. 21, 2021), p. 63. ISSN: 2398-502X. DOI: `10.12688/wellcomeopenres.15191.2`.

[203] D. W. Scott. "On Optimal and Data-Based Histograms". In: *Biometrika* 66.3 (1979), pp. 605–610. ISSN: 0006-3444, 1464-3510. DOI: `10.1093/biomet/66.3.605`.

[204] M. Raissi, P. Perdikaris, and G. Karniadakis. "Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations". In: *Journal of Computational Physics* 378 (Feb. 2019), pp. 686–707. ISSN: 00219991. DOI: `10.1016/j.jcp.2018.10.045`.

[205] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. "Physics-Informed Machine Learning". In: *Nature Reviews Physics* 3.6 (May 24, 2021), pp. 422–440. ISSN: 2522-5820. DOI: `10.1038/s42254-021-00314-5`.

[206] S. Rabanser, O. Shchur, and S. Günnemann. *Introduction to Tensor Decompositions and Their Applications in Machine Learning*. Nov. 29, 2017. arXiv: `1711.10781 [cs, stat]`. URL: `http://arxiv.org/abs/1711.10781` (visited on 04/05/2024). preprint.

[207] J. Kossaifi, N. Kovachki, K. Azizzadenesheli, and A. Anandkumar. *Multi-Grid Tensorized Fourier Neural Operator for High-Resolution PDEs.* Sept. 29, 2023. arXiv: 2310.00120 [cs]. URL: http://arxiv.org/abs/2310.00120 (visited on 04/05/2024). preprint.

[208] N. Z. Zhao, S. Boutami, and S. Fan. "Accelerating Adjoint Variable Method Based Photonic Optimization with Schur Complement Domain Decomposition". In: *Optics Express* 27.15 (July 22, 2019), p. 20711. ISSN: 1094-4087. DOI: 10.1364/OE.27.020711.

[209] M. Chen, J. Jiang, and J. A. Fan. "Algorithm-Driven Paradigms for Freeform Optical Engineering". In: *ACS Photonics* 9.9 (Sept. 21, 2022), pp. 2860–2871. ISSN: 2330-4022, 2330-4022. DOI: 10.1021/acsphotonics.2c00612.

[210] K. Svanberg. "The Method of Moving Asymptotes–a New Method for Structural Optimization". In: *International Journal for Numerical Methods in Engineering* 24.2 (Feb. 1987), pp. 359–373. ISSN: 0029-5981, 1097-0207. DOI: 10.1002/nme.1620240207.

[211] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. "On the Expressive Power of Deep Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning.* Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 6–11, 2017, pp. 2847–2854. URL: https://proceedings.mlr.press/v70/raghu17a.html.

[212] J. Jiang, R. Lupoiu, E. W. Wang, D. Sell, J. Paul Hugonin, P. Lalanne, and J. A. Fan. "MetaNet: A New Paradigm for Data Sharing in Photonics Research". In: *Optics Express* 28.9 (Apr. 27, 2020), p. 13670. ISSN: 1094-4087. DOI: 10.1364/OE.388378.

[213] R. V. Woldseth, N. Aage, J. A. Bærentzen, and O. Sigmund. "On the Use of Artificial Neural Networks in Topology Optimisation". In: *Structural and Multidisciplinary Optimization* 65.10 (Oct. 2022), p. 294. ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s00158-022-03347-1.

[214] S. Cai, Z. Wang, L. Lu, T. A. Zaki, and G. E. Karniadakis. "DeepM&Mnet: Inferring the Electroconvection Multiphysics Fields Based on Operator Approximation by Neural Networks". In: *Journal of Computational Physics* 436 (July 2021), p. 110296. ISSN: 00219991. DOI: 10.1016/j.jcp.2021.110296.

[215] S. Wang, H. Wang, and P. Perdikaris. "Learning the Solution Operator of Parametric Partial Differential Equations with Physics-Informed DeepONets". In: *Science Advances* 7.40 (Oct. 2021), eabi8605. ISSN: 2375-2548. DOI: 10.1126/sciadv.abi8605.