NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Computation cost reduction in 3D shape optimization of nanophotonic components

Masnad, Md Mahadi; Salsabil, Nishat; Xu, Dan-Xia; Liboiron-Ladouceur, Odile; Grinberg, Yuri

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.
For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1088/2040-8986/ad3a76
Journal of Optics, 26, 5, 2024-04-11

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=bc27788c-c965-4f6b-95fe-68de4a427fbf
https://publications-cnrc.canada.ca/fra/voir/objet/?id=bc27788c-c965-4f6b-95fe-68de4a427fbf

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

OPEN ACCESS
IOP Publishing
Journal of Optics
J. Opt. **26** (2024) 055702 (10pp)
https://doi.org/10.1088/2040-8986/ad3a76

Computation cost reduction in 3D shape optimization of nanophotonic components

Md Mahadi Masnad[1,*], Nishat Salsabil[2], Dan-Xia Xu[3], Odile Liboiron-Ladouceur[1] and Yuri Grinberg[4]

[1] Electrical and Computer Engineering Department, McGill University, Montreal, Canada
[2] Materials and Metallurgical Engineering Department, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[3] Advanced Electronics and Photonics Research Center, National Research Council Canada, Ottawa, Canada
[4] Digital Technologies Research Center, National Research Council Canada, Ottawa, Canada

E-mail: md.masnad@mail.mcgill.ca

Received 11 September 2023, revised 23 January 2024
Accepted for publication 4 April 2024
Published 11 April 2024

Abstract

Inverse design methodologies effectively optimize many design parameters of a photonic device with respect to a primary objective, uncovering locally optimal designs in a typically non-convex parameter space. Often, a variety of secondary objectives (performance metrics) also need to be considered before fabrication takes place. Hence, a large collection of optimized designs is useful, as their performance on secondary objectives often varies. For certain classes of components such as shape-optimized devices, the most efficient optimization approach is to begin with 2D optimization from random parameter initialization and then follow up with 3D re-optimization. Nevertheless, the latter stage is substantially time- and resource-intensive. Thus, obtaining a desired collection of optimized designs through repeated 3D optimizations is a computational challenge. To address this issue, a machine learning-based regression model is proposed to reduce the computation cost involved in the 3D optimization stage. The regression model correlates the 2D and 3D optimized structural parameters based on a small dataset. Using the predicted design parameters from this model as the initial condition for 3D optimization, the same optima are reached faster. The effectiveness of this approach is demonstrated in the shape optimization-based inverse design of TE_0-TE_1 mode converters, an important component in mode-division multiplexing applications. The final optimized designs are identical in both approaches, but leveraging a machine learning-based regression model offers a 35% reduction in computation load for the 3D optimization step. The approach provides a more effective means for sampling larger numbers of 3D optimized designs.

Keywords: inverse design, shape optimization, supervised learning, regression, mode converter

* Author to whom any correspondence should be addressed.

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

© 2024 The Author(s). Published by IOP Publishing Ltd

1. Introduction

When designing a photonic integrated device, besides the primary functionality, other performance attributes may be critical for specific applications. Often, trade-offs are necessary among various performance metrics. For example, the optimization of broadband grating couplers often comes at the cost of reduced peak coupling efficiency [1]. Similarly, a topology-optimized wavelength multiplexer, designed to be more tolerant to fabrication imperfections, may result in diminished optical transmission per channel [2]. For high-speed data transmission, minimizing crosstalk at the operating wavelength is of particular importance. Therefore, having a collection of optimized designs that exhibit superior performance for different metrics is useful. In conventional design approaches for simple geometries, the design parameters can often be directly correlated to specific performance attributes, for instance, longer photodetectors tend to have higher responsivity [3], longer tapered directional couplers tend to be broadband [4], etc. In other words, the effects of a handful of design parameters can often be easily mapped to the performance attributes of the photonic device. As the functionality of the photonic devices expands [5–7], and more complicated geometries are optimized in limited footprints [8], advanced optimization methods such as particle swarm optimization [9], genetic algorithm [10], topology optimization [11] and shape optimization [12] in inverse design are deployed. These methods optimize device performance by tuning many design parameters, making it challenging to map specific design parameters to performance attributes. In gradient descent-based optimization, the objective function can be defined such that it optimizes several performance attributes simultaneously. However, it may take many trial-and-error cycles to find the proper objective function that provides the best trade-off among all the requirements. Therefore, an extensive search of the top-performing designs based on the main performance metric is useful. This collection allows the designer to understand such trade-offs and make educated choices regarding the devices that are most suitable for fabrication.

Recently, in several investigations, a computational mapping between design parameters and device performance has been established. Principal component analysis was implemented to identify lower-dimensional design subspaces where top-performing designs reside [1, 13–15]. This analytical process requires the acquisition of an initial collection of (locally) optimized designs. However, the task of locating these designs through random initialization in high-dimensional design spaces proves to be computationally expensive. In the context of topology optimization for nanophotonic components, a method based on time-reversal techniques has been introduced to derive a physics-aware initial topological structure, rather than a random initialization, which facilitates faster convergence [16]. A similar idea of predicting the electromagnetic field profile, leading to a favorable choice of the initial design structure demonstrates 47.3% computation cost reduction in design optimization through finite-difference frequency-domain (FDFD) simulations [17]. However, this method is less likely to yield diverse initial structures, which is essential in our work for having a collection of diverse final designs. Among other endeavors to accelerate the inverse design convergence are the adaptive mesh scheme, achieving 1.79-fold acceleration [18], the boundary integral method [19], and the deployment of deep learning in solving various inverse design problems [20, 21].

While the optimization of 2D nanophotonic components is relatively fast, the accuracy of 2D models may not always be preserved when the designs are re-simulated in 3D finite-difference time-domain (FDTD) simulations. Therefore, it is imperative to further optimize and assess their performance in 3D simulations before selecting designs for fabrication, which is a notably time-consuming process. In this study, a machine learning regression model is proposed to learn the mapping from 2D to 3D optimized designs, facilitating a more efficient collection of 3D optimized structures for subsequent analysis. Instead of using 2D optimized parameters as the starting conditions for a 3D optimization, we employ parameters predicted by the regression model, as illustrated in figure 1. A flow chart of the optimization process is shown in figure 1(a). The idea is depicted in figure 1(b) which shows a hypothetical design space (for illustration purpose only) defined by two fictitious design parameters. The process of obtaining a 2D optimized design Y from a random shape X is a rapid and shared procedure in both approaches. In this work, we find one iteration in 3D FDTD simulation is approximately 18 times slower than one iteration in 2D optimization. Thus, acquiring the final design Z from Y becomes computationally intensive. We deploy the regression model to obtain a predicted structure Z', which is located closer to the final design Z, thereby expediting the 3D optimization step.

Figure 1(c) shows the convergence of the optimization and the change in the figure of merit (FOM) across successive iterations. The initial 2D optimization phase is common to both approaches, exhibiting a typical progression within the first hour. Notably, during the subsequent 3D optimization stage, the predicted structure requires fewer iterations compared to the 2D optimized structure to achieve the final 3D optimized design. In the design parameter space, the predicted structure resides in closer proximity to the ultimate design than its 2D-optimized counterpart. Consequently, this is reflected in a higher FOM value at the onset of the 3D optimization step when compared to the 2D optimized design. It is important to note that for a given design area, i.e. the spacing between the input/output (I/O) waveguides, once the n-number of design parameters are chosen, an n-dimensional non-convex design parameter space is already well-defined, where the FOM values at the local optima are fixed. The proposed optimization approach finds these optimal designs more efficiently than the conventional approach. Although various regression methods could be employed, our investigation reveals that a simple linear regression suffices to achieve a notable computational efficiency.

The simulations are executed utilizing the Ansys Lumerical MODE and FDTD solvers [22]. The optimization procedures are conducted through the utilization of Lumopt, an

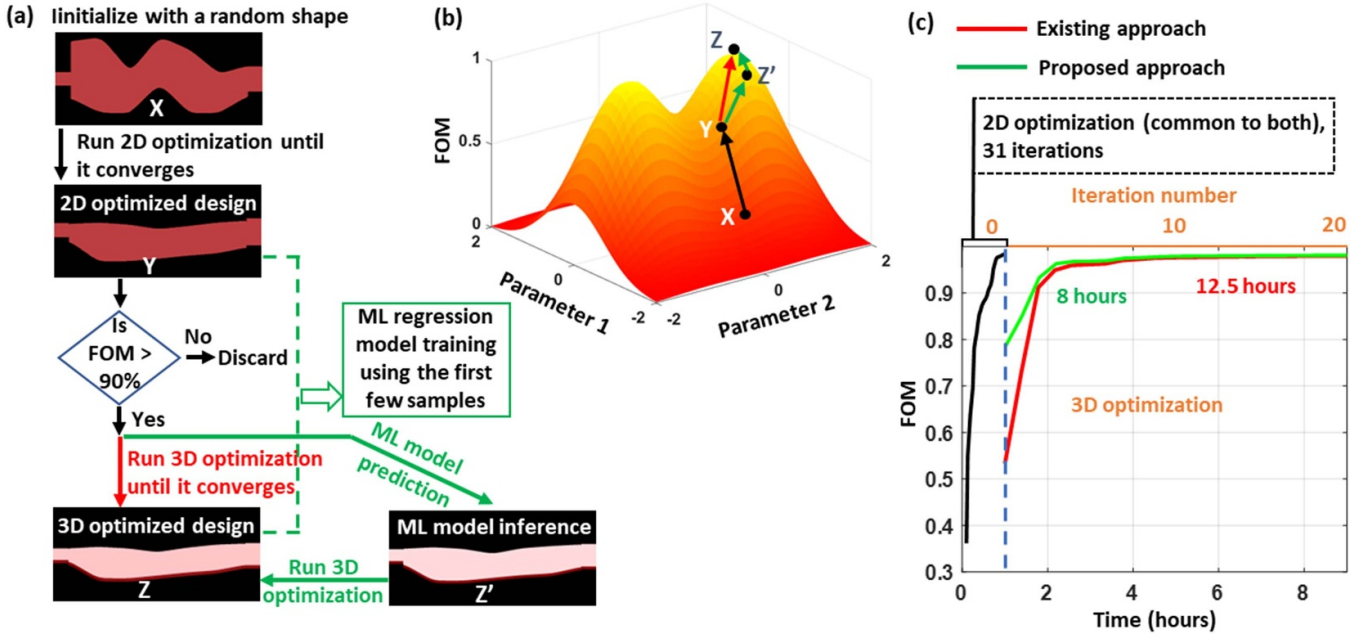

Figure 1. A schematic block diagram showing the different steps of the existing approach and the proposed approach. (a) A flow chart of the optimization process. (b) Obtaining an (local) optimum Y starting from a random initialization X in a hypothetical design space with two design parameters (not the real design space; for illustration purpose only). After the 2D optimization, the existing approach directly optimizes shape Y to reach the optimum shape Z in 3D optimization, while the proposed approach uses a regression model to predict Z through shape Z', located closer to the final design leading to a reduction in the number of iterations needed in the 3D optimization step by 35% on average. (c) Convergence of the optimization (existing and proposed approaches) and the change in the FOM with successive iterations. 2D optimization is common to both approaches and takes a small fraction of the optimization time.

open-source package, in conjunction with the scipy package—a Python library dedicated to scientific and technical computing. We apply the proposed methodology to the design of TE_0-TE_1 mode converters (MCs) as a case study. Leveraging the regression model results in a 35% reduction in the re-optimization duration without compromising the optimized performance. This saves approximately 4.5 h of computation time per design sample, particularly when executed on a personal computer equipped with an Intel® Core™ i7-3770 processor and 32 GB of RAM. The cumulative time savings become particularly significant when applied across a large number of design samples. Consequently, the proposed approach significantly enhances the utilization of computational resources in the context of design optimization, ultimately leading to expedited convergence toward 3D-optimized designs.

2. Methodology

To mitigate computational expenses in 3D FDTD optimization, a linear regression model serves as an intermediary design step, as depicted in figure 1. Initially, a set of 2D optimized designs encompassing a broad spectrum of local optima is generated through random parameter initializations within a predefined design area bounded by input and output waveguides. Guided by the primary design objective, well-performing design samples are selectively retained to construct a dataset for subsequent analysis. A random subset of these selected designs, large enough to effectively train the model for all design parameters, undergoes re-optimization using 3D FDTD simulations. The training set is comprised of these pairs of 2D and 3D optimized designs. It is noteworthy that the model is trained on design samples already exhibiting good FOM so that it can efficiently predict the designs with high FOM values. Designs with low FOM are excluded to prevent potential degradation of the model's predictive accuracy. Next, the model is trained using the design parameters of the 2D optimized design as the input features and the 3D re-optimized design as the output labels. During this phase, the model establishes correlations between these features and labels by determining the weight terms in equation (1). Assuming that the design is defined by n design parameters, we denote the ith parameter of a 2D optimized and 3D optimized design as p_{2D_i} and p_{3D_i}, respectively. We define $\boldsymbol{W}$ as the $n \times n$ weight matrix and $\boldsymbol{w}_0$ as a $n \times 1$ column bias vector, where

$$\begin{bmatrix} p_{3D_1} \\ \vdots \\ p_{3D_n} \end{bmatrix}_{train} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix} \begin{bmatrix} p_{2D_1} \\ \vdots \\ p_{2D_n} \end{bmatrix}_{train} + \begin{bmatrix} w_{10} \\ \vdots \\ w_{n0} \end{bmatrix}. \quad (1)$$

In a compact form, this equation can be rewritten as $\boldsymbol{p}_{3D_train} = \boldsymbol{W} \cdot \boldsymbol{p}_{2D_train} + \boldsymbol{w}_0$, where $\boldsymbol{p}_{2D_train} = [p_{2D_1}\ p_{2D_2} \cdots p_{2D_n}]^T_{train}$ and $\boldsymbol{p}_{3D_train} = [p_{3D_1}\ p_{3D_2} \cdots$

$p_{3D_n}]^T_{train}$. The bias vector is included to account for the solutions not passing through the origin of the linear coordinate system. The training set is used to train the regression model to obtain deterministic values of the weight terms w_{ij} and the bias terms w_{i0}. After that, this regression model is used to predict the design parameters $\boldsymbol{p}_{reg}$ of the 3D re-optimized counterpart of a new 2D optimized design (not included in the training set). These predicted design parameters end up closer to the final design parameters of the 3D optimized design. Thus, the predicted parameters are expected to make a better initialization of the 3D re-optimization step than the 2D optimized parameters. Let p_{reg_i} denote the ith predicted parameter by the regression model, then the predicted parameters are obtained using equation (2), where $\boldsymbol{W}$ and $\boldsymbol{w}_0$ are determined in the training phase

$$\boldsymbol{p}_{reg} = \boldsymbol{W} \cdot \boldsymbol{p}_{2D_new} + \boldsymbol{w}_0 \tag{2}$$

where $\boldsymbol{p}_{2D_new} = [p_{2D_1}\ p_{2D_2} \ldots p_{2D_n}]^T_{new}$ are the parameters of a new 2D optimized design, which is not included in the training set, and $\boldsymbol{p}_{reg} = [p_{reg_1}\ p_{reg_2} \ldots p_{reg_n}]^T$. We compare the time needed to converge to the final design with the time needed for convergence when the 2D optimized design is directly used for initialization. The resulting device performance is also compared.

3. Device

For the demonstration of the proposed optimization method, the TE_0 to TE_1 optical MC is selected. The MCs are designed for a silicon-on-insulator (SOI) platform with a silicon layer thickness of 220 nm. A 450 nm wide single-mode input waveguide is used to support the fundamental TE mode (TE_0), while the width of the output waveguide is set at 850 nm to accommodate the first-order TE mode (TE_1) while avoiding higher order modes and constraining modal crosstalk to the first two TE modes. The separation between the I/O waveguides is set to 3.6 μm, allowing for a variety of base designs. As depicted in figure 2(a), 14 boundary points of the design structure are taken as the design parameters for optimization using the gradient-descent-based algorithm, bound-constraint limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS-B) [23]. L-BFGS-B, an extension of the L-BFGS algorithm, is specifically tailored to handle optimization problems with bound constraints. It incorporates mechanisms to address scenarios where the step size becomes exceptionally small or large during the optimization process. This is a standard off-the-shelf general-purpose optimization tool that is well suited for problems where gradient information is available, such as the adjoint-based method. The time taken by a single iteration is fully dictated by the time it takes to run the forward and adjoint optical simulations. The design optimization iteration begins with an initial guess of the design parameter vector denoted as $\boldsymbol{x}_0$. At the kth iteration, the algorithm computes the gradient (∇) of the objective function, defined in this context as

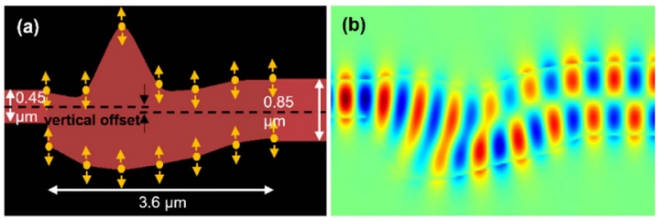

Figure 2. (a) Initialization of 2D optimization of the TE_0-TE_1 mode converter with arbitrary initial structure and a vertical offset between the I/O waveguides. The gradient-descent based algorithm moves the boundary points around in the design area to come up with a shape of the structure which maximizes the specified figure-of-merit. (b) The y-component of the electric field distribution in a TE_0-TE_1 mode converter showing the mode conversion mechanism.

$f(x_k) = 1 - \mathrm{FOM}(x_k)$, with respect to the current design parameters: $\nabla f(\boldsymbol{x}_k)$. The algorithm efficiently maintains a limited memory history of preceding iterations to approximate the inverse Hessian matrix, which is used to determine the search direction $\boldsymbol{s}_k$ [23]. A line search is subsequently performed along the search direction to determine the step size $\boldsymbol{\alpha}_k$ that minimizes the objective function along the search direction. The design parameter vector is then updated according to equation (3)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{\alpha}_k . \boldsymbol{s}_k. \tag{3}$$

The behavior of the objective function, including its smoothness, convexity, and the presence of local minima, can impact convergence. Nevertheless, the L-BFGS-B is guaranteed to converge to a local minimum, depending on the initial parameter values. The convergence criteria are either the change in successive objective function evaluations below some threshold or the magnitude of the gradient is below some threshold. A total of 200 points are obtained through linear interpolation to define the polygon representing the design structure. To further refine the boundary, a Gaussian smoothing filter is convolved with the interpolated points. The judicious selection of the step size, linear interpolation, and the application of the smoothing filter collectively mitigate the risk of structure boundaries self-intersecting and guide the optimization towards the nearest local minimum.

To increase the variability of the designs and the problem complexity, an additional design parameter is incorporated—a vertical offset between the input and output waveguides. A nonzero waveguide offset has been shown to be beneficial in achieving high-performance designs [24]. Experimentally validated, these MCs exhibit a maximum insertion loss of 0.7 dB with a maximum modal crosstalk of -19 dB in the C-band (1530–1565 nm). While the remainder of the work is based on simulated results, the experimental measurements of this class of MCs suggest that simulation performance is a good proxy for the true performance of the designs. The detailed design approach and performance analysis with experimental results of those MCs are reported in [25–27].

Despite the inherent variations in design structures and conversion efficiencies (CEs), the operational principle remains the same for all these MCs. The fundamental TE mode has a symmetrical field profile with the largest field amplitude at the center. The fundamental TE mode is characterized by a symmetrical field profile with the largest field amplitude at the center. In contrast, the TE_1 mode exhibits two lobes with a null point in the middle, accompanied by a phase difference of π between the lobes. To convert the TE_0 mode into TE_1, the spatial field amplitude and phase need to be rearranged. This transformation is achieved through imbalanced optical paths, introducing a delay to one segment of the wavefront. As evidenced by the simulated electric field (E_y) distribution within the design area of a TE_0-TE_1 MC in figure 2(b), the lower segment of the wavefront traverses a longer path, accumulating phase shifts and changes in field amplitude. This results in the transformation of the TE_0 mode into the TE_1 mode, characterized by two antisymmetric lobes in the wavefront.

4. Dataset

Initially in 2D optimization, the I/O waveguides are positioned on the same horizontal axis, i.e. zero vertical offset. We select the FOM as the average optical mode conversion (between TE_0 and TE_1) efficiency over 11 equally spaced wavelength points between 1.5 μm and 1.6 μm. Out of 64 random initializations (excluding identical final designs), 12 designs turn out to have a FOM above 90%. Further attempts at random shape initializations produced similar optimized designs to ones already discovered. Therefore, these 12 designs can be regarded as a reasonable representation of the subspace of good designs. The 12 base designs are shown in figure 3, highlighting the wide variety of high performing structures with zero vertical offset between the I/O waveguides. For a fixed distance of 3.6 μm between the I/O waveguides, the area of the silicon structure varies between 3.4 μm^2 and 5.5 μm^2.

For the subsequent step of the study, a two-step curation process is applied, and the FOM serves as a criterion for selection. First, the 12 good designs are retained, and the remaining designs having FOMs below 90% are eliminated. Additional copies of these 12 base designs are created by introducing vertical offsets ranging from 50 nm to 200 nm in both upward and downward directions. These additional copies can be considered 'variants' of the 12 base designs. For example, eight variants of design 1 (in figure 3) are created when eight copies of design 1 for eight vertical offset values are optimized through 2D simulations. If any of these variants yields an FOM below the threshold of 90%, that specific variant is discarded, and the remaining variants are retained for 3D optimization. This process is applied to derive variants of the remaining base designs. As a result, a total of 85 2D-optimized designs are obtained, each defined by 15 design parameters (14 boundary points and a vertical offset) and exhibiting FOMs surpassing the threshold of 90% in 2D simulations. The 85 2D optimized designs undergo a subsequent phase of optimization through 3D FDTD simulations. In the 3D simulations, 59 out of 85 designs exhibit FOMs exceeding 90%. Next, these 59 3D-optimized designs with FOMs above 90% are retained, and a dataset is constructed. The dataset comprises 59 pairs of 2D- and 3D-optimized design samples with 15 design parameters and four performance attributes: (1) average mode CE across the bandwidth (FOM), (2) modal crosstalk, changes in (3) FOM, and (4) crosstalk at 1.55 μm wavelength for $\pm$20 nm waveguide width variations (to quantify the device robustness to dimensional variations).

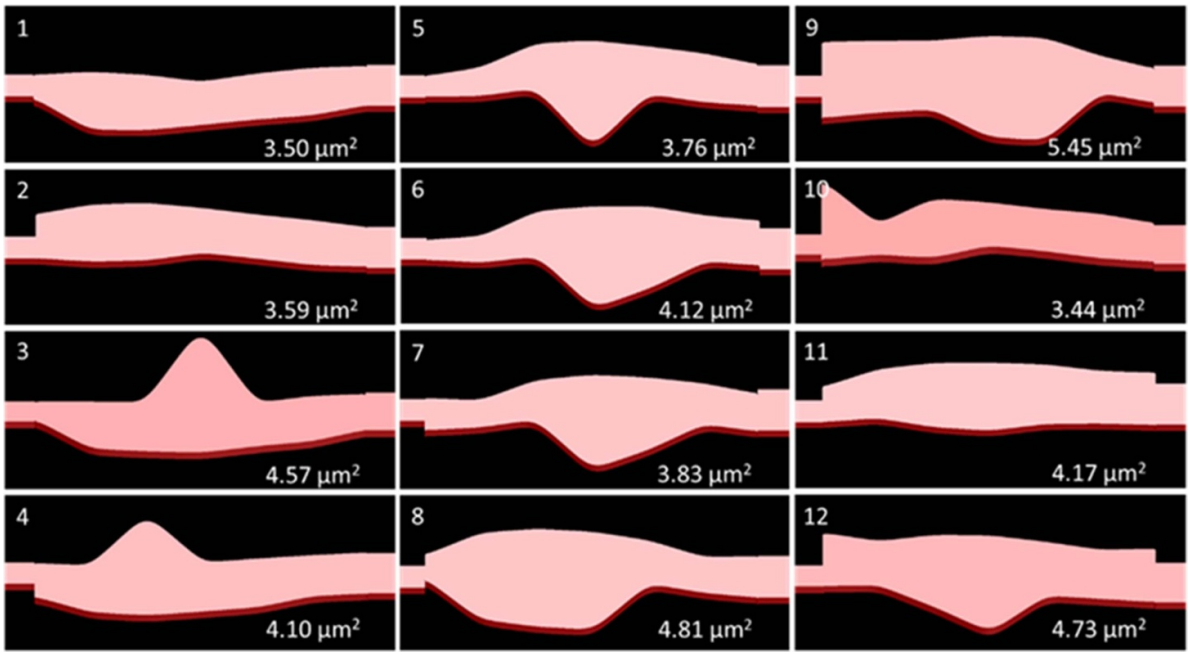

Figure 3. Tilted top view of the 12 base designs (optimized in 3D FDTD simulations), zero offset between the I/O waveguides. The area of the design structure is mentioned below each of the images.

5. Implementation of the regression model

In this section, we demonstrate an implementation of the linear regression model that results in a 35% reduction in computation costs in the 3D optimization. Given the parameters of a 2D optimized design, the regression model predicts the expected design parameters of 3D optimization. The dataset, consisting of 59 pairs of optimized designs, is partitioned into a training set comprising 24 randomly selected designs and a test set encompassing the remaining 35 designs. The training set, comprising 24 pairs of 2D and 3D optimized designs, is utilized to train the regression model, determining the values of the weight terms w_{ij} and the bias terms w_{i0}, as described in equation (1). Subsequently, this trained regression model is applied to predict the 3D optimized parameters of the 2D optimized designs within the test set. Then the predicted design parameters are obtained using equation (2) where $n = 15$.

In this approach, the predicted parameters are computed for all 35 samples in the test set. The differences between the 3D optimized parameters and the 2D optimized parameters ($|p_{2D_i} - p_{3D_i}|$) and the predicted parameters ($|p_{3D_i} - p_{reg_i}|$) are then compared for all test set samples. Figure 4 shows the box plots of these differences, inclusive of their percentiles and outlier points. The average difference (MEAN) between the 2D and final 3D optimized parameters (all parameters combined, 35 $\times$ 15 parameter tuples) is approximately 86 nm, with a standard deviation (SD) of about 87 nm and a standard error of the mean (SEM) of 14.7 nm. On the other hand, the average difference between the 3D optimized parameters and the regression predicted parameters is approximately 46.8 nm with a standard deviation of approximately 51.2 nm and a SEM of 8.7 nm. Evidently, the regression model is beneficial since the predicted structure (formed by the predicted parameters) is closer to the final 3D optimized structure than the 2D optimized design. Thus, it is expected that using the predicted structure obtained through the regression model as the starting point for 3D optimization will effectively reduce the computation cost by reducing the number of iterations needed for the convergence of the optimization. To validate this hypothesis, all 35 predicted designs are optimized in 3D FDTD simulations. Results demonstrate that 30 out of 35 predicted designs require fewer iterations to converge without compromising the final performance, substantiating the effectiveness of the regression model in reducing computation costs.

Table 1 summarizes some statistical data of the two optimization approaches. The average number of iterations needed until convergence starting from the 3D predicted designs is approximately 21 while it is 32 for the 2D optimized designs as the starting points, resulting in a 35% computation cost reduction. Additionally, the predicted structures show an average mode CE (FOM) of 76.2% in 3D FDTD simulation before optimization begins, while it is only 59% for the 2D optimized parameters when simulated in 3D FDTD. Within five iterations, the predicted structures reach an average FOM of 93.5% which is only 2.3% behind the average final FOM (95.7%),

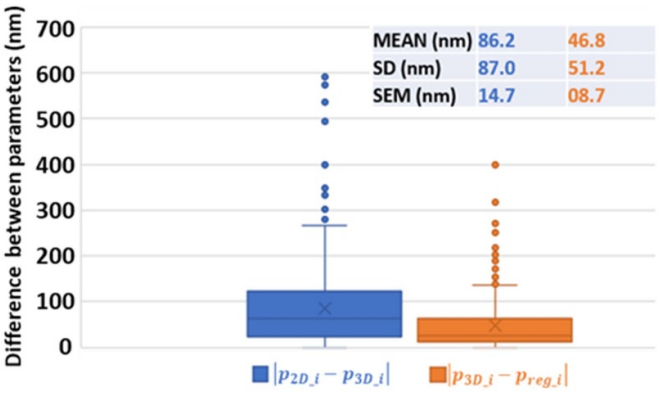

Figure 4. The box plots showing the distributions of $|p_{2D_i} - p_{3D_i}|$ (blue) and $|p_{3D_i} - p_{reg_i}|$ (orange).

whereas it would take approximately 9 iterations to reach the same FOM value if the optimization starts from the 2D optimized parameters. Thus, by predicting the design structures utilizing the regression model before the 3D optimization shows significant leverage over the conventional approach of performing 3D optimization starting from 2D optimized designs. This regression-based design prediction model can further leverage the computation cost reduction in dimensionality reduction-based design approach reported in [1, 13–15] since regression can help collect the initial set of optimized designs faster.

6. Utility of a collection of designs

Once we have a collection of good designs, exploring the dataset leads to different designs when a different performance attribute is prioritized. Figure 5(a) displays the average mode CE and the maximum modal crosstalk of the selected 59 designs across the wavelength range of 1.5–1.6 μm. As expected, a higher mode CE correlates with a lower modal crosstalk. However, when considering only the C-band, a small subset of designs with FOMs around 97% demonstrates lower modal crosstalk than the MC with the maximum CE, as depicted in figure 5(b). In applications where modal crosstalk is a critical performance metric, these MCs can be utilized with a minor sacrifice in CE.

The average values of the performance metrics over such a wide bandwidth of 100 nm do not often provide sufficient information since the modal crosstalk at the central wavelength can be significantly lower than the maximum value across the bandwidth. In some applications, the bandwidth of interest may be narrow, and only the crosstalk in that bandwidth matters. We pick two groups of designs, the variants of design 1, superior in the mode CE, and the variants of design 2, superior in modal crosstalk, (listed in figure 3) and analyze their performance in more details. The mode CEs of these two designs across the 100 nm bandwidth (denoted by solid lines) are plotted in figure 6(b). The average CE of some of the variants (with different vertical offsets) of these two designs

Table 1. Statistical comparison between the optimization from 2D designs and the optimization from the predicted designs using the regression model built on 24 pairs of 2D and 3D optimized designs.

	2D → 3D	Prediction → 3D
Number of samples in test set	35	35
Average number of iterations to converge to the final design	31.1	20.4
Standard deviation of the average number of iterations	13.5	8.4
Standard error of the mean (on iteration number)	2.3	1.4
Average FOM at the start/average final FOM	58.8%/95.7%	76.2%/95.7%
Standard deviation of FOM at the start	18.6%	22.0%
Standard error of the mean of FOM at the start	3.1%	3.7%
Average FOM at the 5th iteration	90.4%	93.5%
Standard deviation of FOM at 5th iteration	5.8%	5.5%
Standard error of the mean of FOM at 5th iteration	1%	0.9%

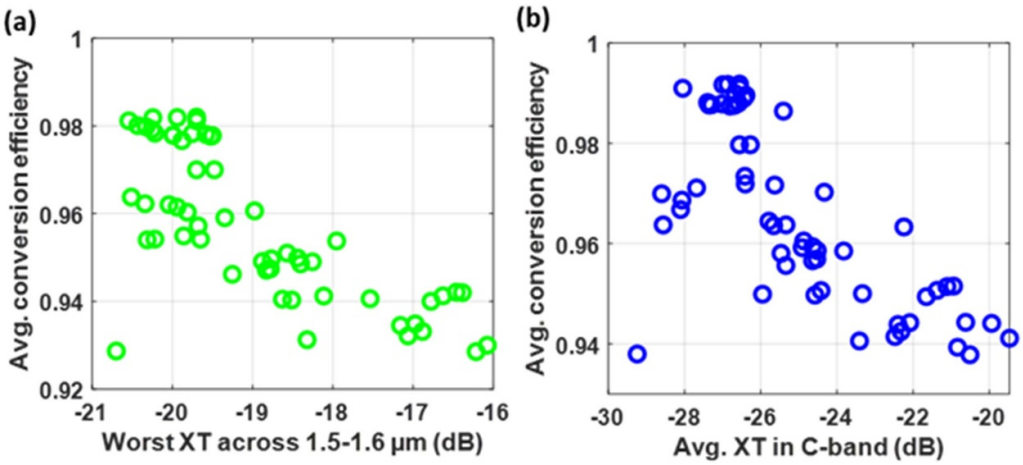

Figure 5. Average mode conversion efficiency of the selected 59 3D optimized MCs as a function of (a) the worst modal crosstalk across the bandwidth of 1.5–1.6 μm and (b) the average modal crosstalk in the C-band (1.530–1.565 μm).

(denoted by circles and squares) are also plotted in figure 6(b). As can be observed, introducing the vertical offset results in slight improvement in the CE (around 0.5%) for some design variants, which in turn leads to a reduction of the XT by several dBs (as only a fraction of the optical power couples to the XT mode). Therefore, some of the variants, with slightly higher CE, have appreciably lower modal crosstalk than the base design at different operating wavelengths, as depicted in figures 6(b) and (c). The preferred direction of the offset (positive or negative), for improved performance, depends on the specific geometry.

Another important performance attribute of photonic devices is robustness to fabrication process variations. Since inverse-designed devices generally have compact footprints containing many small features, common fabrication errors such as under-etch and over-etch cause notable impacts on the device performance. Shape-optimized MCs have been demonstrated to be relatively tolerant to dimensional variations such as 10 nm under/over etches [25]. In this section, we investigate the fabrication robustness in the dataset containing 12 base designs (no vertical offset between the I/O waveguides). The base design and its variants have visually similar geometric shapes. Therefore, the variants are left out of this analysis. It is likely that under/over-etch fabrication errors affect the optical performance of different designs to different extents. To quantify the effect of etch variations on their CE and crosstalk performances, additional copies with expanded and shrunk boundaries (10 nm on each side) of the 12 base designs are simulated to mimic the effect of under and over-etch, respectively. A similar study on the same class of MCs was conducted, and its experimental results are reported in [25, 27]. Furthermore, the impacts on device performance resulting from the smoothing of sharp edges/corners in the design structure during the fabrication process are discussed in [26–28]. In figure 7, the effects on the mode CE of the MCs are presented, simulated in 3D FDTD. The average CEs across the 1.5–1.6 μm wavelength range of the nominal designs, under-etched, and over-etched designs are plotted in figure 7 using 'square', '+', and '×', respectively. For comparing the CE, the average value is a good representation since the minimum and maximum values of CE across the bandwidth of 1.5–1.6 μm are within ±2% of the average value (figure 6(a)). Design 1 and design 11 are found to be the most robust among the samples with an approximate 1.2% reduction in the average CE, while design 8 is the most sensitive to the dimensional variations with an approximate 3.1% reduction.

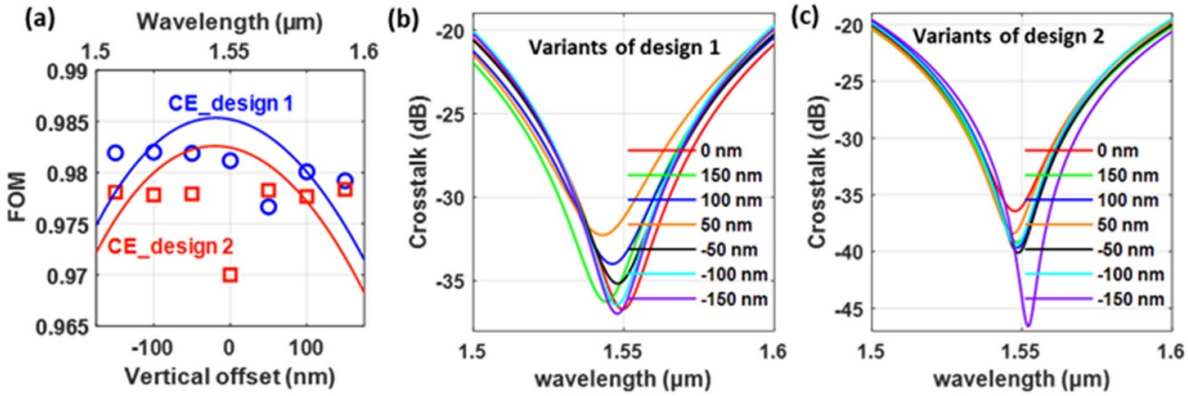

Figure 6. (a) Mode conversion efficiency (solid lines) of design 1 and design 2 as a function of optical wavelength (top horizontal axis. The data points denoted by circles (design 1) and squares (design 2) are the average mode conversion efficiencies (FOM) with respect to vertical offsets (bottom horizontal axis) of all the variants of the two 3D optimized base designs; (b) crosstalk of a few variants of design 1; (c) crosstalk of a few variants of design 2.

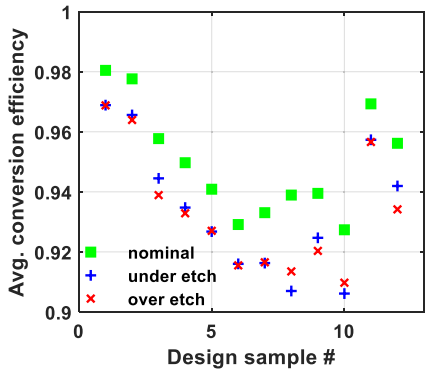

Figure 7. Average mode conversion efficiency (FOM) of the nominal, 10 nm under etched, and 10 nm over etched base designs.

Similarly, the effects of under and over etches on the crosstalk of these MCs are examined. The average crosstalk along with the minimum and maximum values across the bandwidth of 1.5–1.6 μm for the nominal designs, and the average crosstalk of under-etched, and over-etched designs are plotted in figure 8(a) in linear scale for clarity, using 'square', '+', and '×', respectively. Measuring the XT and its change in the logarithmic scale may falsely guide the designer since 3 dB increase from an arbitrary $-x$ dB accounts for twice the optical power corresponding to 3 dB increase from $-(x + 3)$ dB. Thus, measuring the XT in linear scale makes more sense for this kind of comparison. However, unlike mode CE, the modal crosstalk varies widely across the bandwidth as depicted in figure 8(a) with green vertical lines. Thus, the average value across the bandwidth is not a very good representation of the device robustness of crosstalk performance to the dimensional variations. Therefore, it is more practical to choose an operating wavelength and compare the changes to the XT for under/over etch. Figure 8(b) presents the effect on the XT performance of these 12 design samples at the center wavelength of 1.55 μm. On the linear scale, design sample 7 stands out to be the most robust with respect to under etch and over etch with an increase by approximately three folds, while design samples 1, 2, and 3 show lowest nominal XT as well as good robustness compared to other samples.

7. Discussion

In this work, the training set is kept smaller than the test set. The primary objective is to adeptly train the regression model using a modest-sized training set, achieving sufficiently accurate predictions of design parameters closely aligned with the final 3D-optimized parameters. We have demonstrated that a modest training set is sufficient for the regression. However, using a larger training set will certainly enhance the model's performance and lead to the predicted structure located closer to the final design, and hence, the optimization would converge even faster. To validate the concept, another study has been conducted with the training set and the test set comprising 36 and 23 design samples, respectively. As expected, the model's prediction accuracy increased, leading to better initializations for the design samples in the test set, which is manifested by the faster convergence of the 3D optimization, requiring an average of 17 iterations. This translates to an approximately 46% reduction in the cumulative computation cost per design sample, but the computation cost is reduced only for 23 design samples.

While a comprehensive comparison between the existing approach and the proposed approach is presented in this article, it is also noteworthy to compare them with a possible approach of directly starting with the 3D optimization. A single iteration through 2D simulations in Ansys Lumerical requires approximately 1.33 min, whereas each iteration through 3D FDTD simulations demands approximately 24 min on the computing system employed in this study. The time it takes to do a single iteration of L-BFGS-B is fully dictated by the time it takes to run the forward and adjoint optical simulations. In the case of direct 3D optimization with random initializations, convergence to the final design requires an average of 43 iterations per design sample, translating into

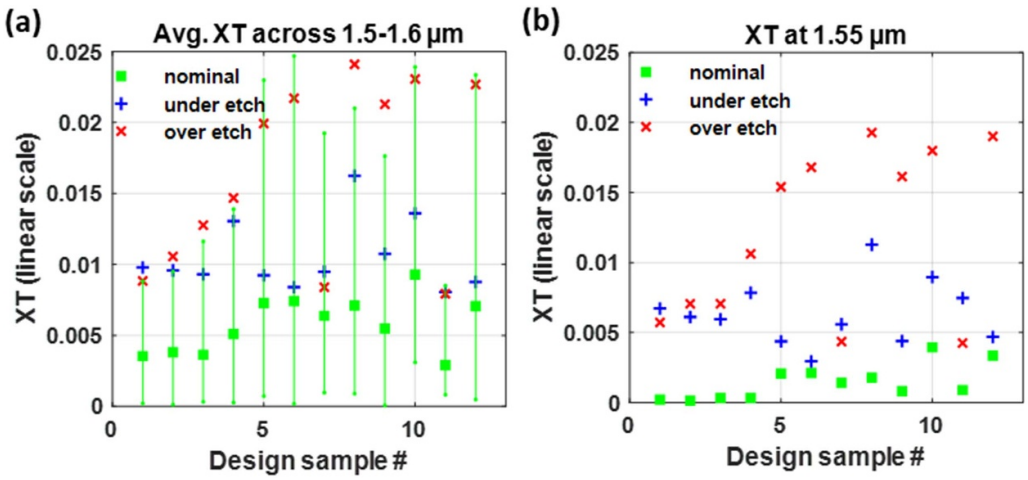

Figure 8. (a) Average XT along with the minimum and maximum values in linear scale across the bandwidth of 1.5–1.6 μm for the nominal designs, and average XT of the under etched and over etched designs, (b) crosstalk comparison among the nominal, under etched, over etched designs at 1.55 μm.

17.2 h. Consequently, a random initialization in 3D optimization extends the convergence time significantly. Moreover, the attainment of a local optimum with an acceptable FOM value is not guaranteed in a single attempt. To illustrate, in the context of the 2D optimization step undertaken in this study, out of 64 random initializations, only 12 designs surpassed the threshold FOM of 90%. This outcome suggests an approximately 20% likelihood of obtaining a good design with acceptable performance. In contrast, given that the 2D optimization of a design sample requires roughly an hour to complete, the designer, at the end of an hour, can readily decide whether to discard the design (if the FOM falls below a predetermined threshold) or proceed to the subsequent step (if the FOM exceeds the threshold).

8. Conclusion

For inverse-designed components using the shape optimization method, 2D optimized designs need substantial re-optimization in 3D to obtain fabrication-ready layout. We propose a regression method that correlates the 2D and 3D optimized structures using a small dataset. Using parameters predicted with this regression model as the initialization condition for 3D optimization, we demonstrate a significant reduction in the number of iterations required for convergence, saving 35% of the computation cost involved in 3D optimization used to design a TE_0-TE_1 MC when 24 pairs of 2D and 3D optimized designs are used to build the regression model. To the best of our knowledge, this is the first demonstration of utilizing a regression model to bridge the gap between the 2D and 3D models. The fact that linear regression model proves to be beneficial suggests that 3D re-optimization step has predictable patterns, which even simple machine learning techniques are able to pick up. Other more involved machine learning methods can potentially be used to discover more complex patterns and save on the 3D computations even further. This approach provides a range of design options of nanophotonic components with varied performance attributes in a more computationally efficient manner. It can also be useful for building a process design kit library containing multiple designs customized for different applications. It is also worth noting that this design methodology can be readily adopted for any other photonic integrated design using the shape optimization technique.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Conflict of interest

The authors declare that there are no conflicts of interest related to this article.

Funding

This work is supported by the AI4Design and HTSN (High Throughput and Secure Networks) Challenge Programs of National Research Council (NRC) Canada.

ORCID iD

Md Mahadi Masnad https://orcid.org/0000-0001-9385-7588

References

[1] Melati D, Grinberg Y, Dezfouli M K, Janz S, Cheben P, Schmid J H, Sánchez-Postigo A and Xu D-X 2019 Mapping the global design space of nanophotonic components using machine learning pattern recognition *Nat. Commun.* **10** 4775

[2] Zhang G, Xu D-X, Grinberg Y and Liboiron-Ladouceur O 2021 Topological inverse design of nanophotonic devices with energy constraint *Opt. Express* **29** 12681
[3] Fard M M, Cowan G and Liboiron-Ladouceur O 2016 Responsivity optimization of a high-speed germanium-on-silicon photodetector *Opt. Express* **24** 27738
[4] Ding Y, Xu J, Ros F D, Huang B, Ou H and Peucheret C 2013 On-chip two-mode division multiplexing using tapered directional coupler-based mode multiplexer and demultiplexer *Opt. Express* **21** 10376
[5] Subbaraman H, Xu X, Hosseini A, Zhang X, Zhang Y, Kwong D and Chen R T 2015 Recent advances in silicon-based passive and active optical interconnects *Opt. Express* **23** 2487
[6] Mirjalili S M, Abedi K and Mirjalili S 2013 Optical buffer performance enhancement using particle swarm optimization in ring-shape-hole photonic crystal waveguide *Optik* **124** 5989
[7] Thomson D *et al* 2016 Roadmap on silicon photonics *J. Opt.* **18** 073003
[8] Molesky S, Lin Z, Piggott A Y, Jin W, Vucković J and Rodriguez A W 2018 Inverse design in nanophotonics *Nat. Photon.* **12** 659
[9] Kennedy J and Eberhart R 1995 Particle swarm optimization *IEEE Proc. ICNN'95-Int. Conf. on Neural Networks* vol 4 p 1942
[10] Mirjalili S 2019 *Genetic Algorithm, Evolutionary Algorithms and Neural Networks* vol 43 (Springer)
[11] Jensen J S and Sigmund O 2004 Systematic design of photonic crystal structures using topology optimization: low-loss waveguide bends *Appl. Phys. Lett.* **84** 2022
[12] Lalau-Keraly C M, Bhargava S, Ganapati V, Scranton G and Yablonovitch E 2014 *Shape Optimization of Nanophotonic Devices Using the Adjoint Method Science and Innovations, STu2M-6* (Optica Publishing Group)
[13] Melati D *et al* 2020 Design of multi-parameter photonic devices using machine learning pattern recognition, integrated photonics platforms: fundamental research *Proc. SPIE* **11364** 29–35
[14] Dezfouli M K *et al* 2020 Perfectly vertical surface grating couplers using subwavelength engineering for increased feature sizes *Opt. Lett.* **45** 3701
[15] Torrijos-Morán L and García-Rupérez J 2021 Design of slow-light-enhanced bimodal interferometers using dimensionality reduction techniques *Opt. Express* **29** 33962
[16] Wang Z, Wang B-Z, Liu J-P and Wang R 2021 Method to obtain the initial value for the inverse design in nanophotonics based on a time-reversal technique *Opt. Lett.* **46** 2815
[17] Schulte L, Butz M, Becker M, Risse B and Schuck C 2023 Accelerating finite-difference frequency-domain simulations for inverse design problems in nanophotonics using deep learning *Opticaopen Preprints* (https://doi.org/10.1364/opticaopen.24147402.v1)
[18] Wang Q, Liu H, Cui X, Teng G, Zhao Y, Liu Z and Dai F 2021 Acceleration of inverse design process on adaptive mesh *Opt. Express* **29** 25064–83
[19] Garza E and Sideris C 2022 Fast inverse design of 3D nanophotonic devices using boundary integral methods *ACS Photonics* **10** 824–35
[20] Pan Z and Pan X MDPI 2023 Deep learning and adjoint method accelerated inverse design in photonics: a review *Photonics* **10** 852
[21] Wiecha P R, Arbouet A, Girard C and Muskens O L 2021 Deep learning in nano-photonics: inverse design and beyond *Photon. Res.* **9** B182–200
[22] Ansys lumerical photonics simulation & design software (available at: www.ansys.com/products/photonics) 6 January 2024
[23] Zhu C, Byrd R H, Lu P and Nocedal J 1997 Algorithm 778: l-BFGS-B: fortran subroutines for large-scale bound-constrained optimization *ACM Trans. Math. Softw.* **23** 550–60
[24] Masnad M M, Grinberg Y, Xu D and Liboiron-Ladouceur O 2021 Physics-guided inverse design for SiPh mode manipulation *OSA Photonics in Switching and Computing* p Tu4A.4
[25] Masnad M M, Zhang G, Xu D-X, Grinberg Y and Liboiron-Ladouceur O 2022 Fabrication error tolerant broadband mode converters and their working principles *Opt. Express* **30** 25817
[26] Masnad M M, Zhang G, Xu D-X, Grinberg Y and Liboiron-Ladouceur O 2022 Dimensional variation tolerant inverse designed broadband mode converter *Optical Fiber Communications Conf. and Exhibition (OFC)* (IEEE) pp 1–3
[27] Masnad M M 2022 Computationally efficient and fabrication error tolerant inverse-designed mode converters and mode-division (de) multiplexers *MSc Dissertation* (McGill University)
[28] Gostimirovic D, Masnad M M, Xu D-X, Grinberg Y and Liboiron-Ladouceur O 2022 Pre-fabrication performance verification of a topologically optimized mode demultiplexer using deep neural networks *European Conf. on Optical Communication (ECOC)* (IEEE) pp 1–4