Enhancing Adjoint Optimization-based Photonics Inverse Design with Explainable Machine Learning

Christopher Yeung¹,², David Ho¹, Benjamin Pham¹, Katherine T. Fountaine², and Aaswath P. Raman¹,*

¹Department of Materials Science and Engineering, University of California, Los Angeles, CA 90095, USA
²NG Next, Northrop Grumman Corporation, Redondo Beach, CA 90278, USA
*Corresponding Author: aaswath@ucla.edu

ABSTRACT: A fundamental challenge in the design of photonic devices and electromagnetic structures more generally is the optimization of their overall architecture to achieve a desired response. To this end, topology or shape optimizers based on the adjoint variables method have been widely adopted due to their high computational efficiency and ability to create complex freeform geometries. However, the functional understanding of such freeform structures remains black box. Moreover, unless a design space of high-performance devices is known in advance, such gradient-based optimizers can get trapped in local minima, limiting performance achievable through this inverse design process. To elucidate the relationships between device performance and nanoscale structuring, while mitigating the effects of local minima trapping, we present an inverse design framework that combines adjoint optimization, automated machine learning (AutoML), and explainable artificial intelligence (XAI). Integrated with a numerical electromagnetics simulation method, our framework reveals structural contributions towards a figure-of-merit (FOM) of interest, then leverages this information to minimize the FOM further than that obtained through adjoint optimization alone, overcoming local minima. We demonstrate our framework in the context of waveguide design and achieve between 43% and 74% increases in device performance relative to state-of-the-art adjoint optimization-based inverse design across a range of telecom wavelengths. Results of this work thus highlight machine learning strategies that can substantially extend and improve the capabilities of a conventional, optimization-based inverse design algorithm while revealing deeper insights into the algorithm’s designs.

KEYWORDS: nanophotonics, deep learning, explainability, adjoint optimization, automated machine learning
Effectively optimizing nanophotonic structures is key to their use in a broad range of optical applications. For example, photonic integrated circuits, metasurfaces, and guided-wave systems can be geometrically manipulated at subwavelength scales to deliver a wide range of functionalities [1][2][3][4][5]. However, a large design space must be rapidly explored in order to optimize the geometry for a particular application. To effectively navigate such a design space, gradient-based optimization algorithms such as the adjoint variables method have been widely adopted to design non-intuitive or irregularly-shaped electromagnetic structures that are highly efficient at accomplishing a particular goal. By calculating the shape derivatives at all points in space using only two electromagnetic simulations per iteration [6], adjoint optimizations are orders of magnitude more computationally efficient than alternative optimization methods and capable of achieving state-of-the-art performance [7][8][9][6].

Although adjoint optimization-based methods have been successfully applied to a variety of photonic systems [10][11][12][13], the method’s reliance on gradient-based information means that the method is local in nature, and therefore bounded by the corresponding limitations. Specifically, since the design space for electromagnetic structures is predominantly non-convex, adjoint optimizations are susceptible to getting stuck in local minima [14]. Thus, unless a region of high-performance devices is known in advance, multiple optimization runs are needed (typically by using random starting points) to arrive at a single optimization target [15]. To overcome these limitations, recent efforts have combined machine learning (ML) with adjoint optimization to address some of the latter’s limitations. For example, population-based inverse design was demonstrated using global topology-optimization networks, or GLOnets [16], which integrate the adjoint method directly into the training process. Alternative strategies also include the integration of ML and adjoint optimization as a two-step process, where the ML-component performs an initial global-search approximation, then the optimization improves design performance further [16][17][18]. Although both approaches can improve upon the algorithm’s performance, the underlying issue of local minima trapping remains unaddressed, since the integration and use of a gradient-based optimizer inherently indicates that the issue is still present. In this regard, metaheuristic techniques such as simulated annealing have been proposed to escape local minima in the search process [19][20]. This method may directly address the issue of local minima trapping through neighbor-based exploration, but its applications in
photonic shape and topology optimization have been severely limited due to relatively low computational efficiency (on the order of 1,000 iterations) [21][22].

To comprehensively address the issue of local minima trapping, we seek to identify the root of the problem and ask: what caused the algorithm to get trapped in the first place? In the context of the optical structures being optimized, arriving at certain geometric elements and their resulting electromagnetic response must be responsible for (or contribute to) guiding the optimization towards suboptimal results. To discover the geometric features responsible for local minima trapping, and to then overcome them, here we employ an explainable artificial intelligence (XAI) based approach. XAI serves as a promising candidate for addressing local minima trapping due to its ability to reveal a model’s decision-making process as well as the contributing factors thereof [23][24][25]. For example, XAI can reveal the spatial regions of a nanophotonic structure that contribute to the presence or lack of an absorption peak [26]. Thus, to explain the causes of local minima trapping in gradient-based adjoint optimization, and subsequently use this information to prevent the optimization from converging onto suboptimal states, we present an XAI-based framework that utilizes the relationships between device efficiency and nanoscale structuring to increase optimization performance.

We demonstrate our optimization framework in the context of Y-splitter waveguide design (Figure 1a), where the objective is to optimize the shape of the silicon-oxide interface to maximize power transmission efficiency from an input port to two output ports of the same width. We represent the objective function as a decreasing figure-of-merit (FOM) which ranges from 1 to 0, where 0 represents ideal performance. With this definition in place, adjoint shape optimizations are performed on an initial Y-splitter design to minimize the FOM (Figure 1b) across a range of target wavelengths in the telecommunications range (1.3 to 1.8 µm). The design and FOM information from the optimizations are used in conjunction with neural architecture search to automate the training of an ML model (AutoML). The model learns the relationships between device structure and performance by accurately predicting the FOM and target wavelength of an input design (Figure 1c). We then use a suite of XAI algorithms, SHapley Additive exPlanations or SHAP (a post hoc explanation technique based on game theory [27]), on the model to extract the structure-performance relationships as “feature explanation” heatmaps (Figure 1d). By interrogating our trained ML model, the explanations here inform the structural features that contribute to the FOM of interest. Using this information, we devise a
boundary extraction algorithm that takes the explanations and makes guided design changes (Figure 1e). These design changes then provide a new starting point for the local adjoint optimization method, which allows the method to reach lower FOMs than before (at multiple target wavelengths).

**Figure 1.** (a) Nanophotonic device optimization: silicon-on-insulator Y-junction splitter for telecom applications. (b) Multiple adjoint optimization runs are applied to the Y-splitter design at various target wavelengths. (c) Results of the optimizations are used as training data in automated machine learning (AutoML) to train a neural network, where the inputs are images and the outputs are device figure-of-merits (FOM) and target wavelengths. (d) Explainable AI algorithms are used on the neural network to capture feature explanations, (e) which are used to optimize device performance further by allowing the algorithm to escape its local minima.
Results and Discussion

Adjoint Optimization and Convolutional Neural Network Training

We first developed our training dataset by performing multiple adjoint shape optimization runs on a starting Y-splitter design (Figure 2a). We applied a widely-adopted implementation of the adjoint method that is integrated with a commercial finite-difference time-domain (FDTD) solver [28]. The 2D cross-sections of the Y-splitter designs are represented as black and white images, where the black and white pixels represent the permittivity of silicon and SiO$_2$, respectively. In our configuration of the adjoint method, the optimizable region is the area between the input and output ports, while the port sizes remain fixed. The optimizable geometry within this region is defined using the level set method and cubic spline interpolations [28][29]. Each optimization run was performed using different operating wavelengths as optimization objectives (1.3 to 1.8 µm in 0.1 µm steps) to produce a collection of device designs with gradual performance improvements. Performance improvement is indicated by a decreasing FOM (as design iteration increases), until a plateau is reached. As shown in Figure 2a, each design iteration consists of a forward and adjoint (i.e., time-reversed) simulation, which calculates the shape derivative over the entire optimizable region and modifies the geometry (per iteration) in proportion to the FOM gradient [6]. At the final iteration “N” (which may vary across each optimization run), device geometry is tailored to achieve maximum attainable performance with respect to the sought target. In our application of the adjoint method, the FOM represents the power coupling of guided modes, and is defined as:

$$FOM = \frac{1}{P(\lambda)} \int |T_0(\lambda) - T(\lambda)| d\lambda,$$

(1)

where $\lambda$ is the evaluated wavelength, $T$ is the actual power transmission through the output ports, $T_0$ is the ideal power transmission, and $P$ is the source power (in Watts). Thus, the FOM is the difference between the input and output transmission normalized by the power injected by the source, which results in maximum performance at 0. Figure 2b shows the results of each optimization run, where the collective FOM information and corresponding designs (at each iteration; not including the starting design) are used as training data for deep learning.
From the optimized structures shown in Figure 2, it can be observed that unique geometries are obtained for each target wavelength, which in turn yield a range of FOM values. Therefore, the FOM and target wavelength are coupled with one another, and both are dependent on the waveguide structure. Thus, to ensure that our model simultaneously learns both of these properties, which in turn captures more information regarding the structure than models trained on the properties individually, we designed a single neural network that takes the Y-splitter geometries as inputs (here, 128×64 pixel images, or 2.5×1.25 µm² domains) and outputs both FOM and target wavelength. In the particular design space we explored, 192 input and output pairs were generated for the neural network. For ease of training, target wavelengths were converted into categorical labels, where a position-specific output node value of 1 represents the wavelength of a specific design, while the other positional nodes equal 0. For example, a target wavelength \( T_i \) of 1.3 µm is represented as \( T_{1.3}=[1,0,0,0,0,0] \), 1.4 µm is \( T_{1.4}=[0,1,0,0,0,0] \), and this pattern is repeated up to 1.8 µm. Alternatively, \( \arg \max(T_i) \), or the index of the maximum value along the vector, represents the target wavelength. Combined with a floating point value (ranging from 0 to 1) to serve as the FOM, we devise a training data structure that is amenable to both classification and regression-based tasks. With this input-output relationship defined, as well as a 90:10 training-validation data split, we use neural architecture search (AutoKeras [30]) with image blocks to automate the deep learning process by testing different model variants across multiple trials. We observe that the optimal architecture was identified after 17 trials, which has a validation loss of \( 8.6 \times 10^{-5} \). The final training and validation losses of each trial are presented in Figure S1a, and the evolution of the convolutional neural network (CNN) architecture from the first trial to the last is shown in Figure S1b. The optimized CNN possesses four convolutional blocks (with 512, 64, 32, and 32 filters, respectively) followed by a dense layer. Each block contains Leaky ReLu, batch normalization, and max pooling layers. Training progression of the optimized architecture is shown in Figure 2c, where a strong convergence between the training and validation losses can be observed.
Figure 2. (a) Training data generation. Adjoint optimization runs are performed across target wavelengths ranging from 1.3 µm to 1.8 µm to produce high-performance devices in the telecom window. (b) FOM vs. design iterations across each optimization run. (c) Training and validation losses for the AutoML-optimized neural network shows high training accuracy and model convergence.

Structure Explanation and Re-Optimization

After training our machine learning model, we next sought to explain the relationship between the overall shape and FOM such that this information can be leveraged to potentially further optimize the devices, and overcome any local minima the adjoint method may have arrived at. To verify that the model properly learned the structure-FOM relationship, we passed the final design iterations (of each target wavelength) into the trained model and compared the ground truth outputs to the model’s predictions. The comparison is shown in Figure 3a, where we observe a strong match (over 90% accuracy) between the predictions (blue points) and ground truths (orange points). Inset images are model inputs. From this result, we can infer that the model accurately learned the key features on the optimized structures which contribute to the target wavelength-specific FOM values. Therefore, we can utilize XAI to reveal the structure-performance relationships of each device. Specifically, we employed an explanation strategy for photonics design, using SHAP, to highlight the device feature contributions to their
respective FOM [26]. These feature contribution heatmaps (represented as SHAP values, $\Phi(x,y)$, ranging from -1 to 1) are illustrated in Figure 3b, where the blue and red pixels indicate positive and negative contributions towards the FOM, respectively. We then leverage the information captured by the XAI algorithm, and manipulate the structure accordingly, to assess its effect on device performance.

**Figure 3.** (a) Comparisons between model predictions and ground truths, for FOM (regression) and target wavelength (classification) values, show that the model accurately learned the relationship between device structure and performance. Inset images show the model inputs, which are adjoint-optimized devices. (b) SHAP explanation heatmaps of the optimized devices reveal the structural features that contribute positively (blue) or negatively (red) towards optimal device performance.

To determine how to practically use the SHAP values (represented as red/blue heatmap pixels), we first note that high concentrations of blue pixels are located throughout a majority of each structure, while the center of the structures and select portions of the outer boundaries contain large regions of red pixels. In this regard, since the training data solely consists of geometries with varying degrees of shape changes at the SOI boundary, and no geometry change is introduced within the structure (i.e., no material subtractions or white pixels are inside the main island of black pixels), we focus our analysis on the SHAP values located at the structure
boundary rather than the center. Following the aforementioned principles of positive and negative contributions, we define the red regions along the structure boundaries as negative contributions towards device performance that should be removed from the design. Accordingly, we devised a boundary extraction algorithm to systematically adjust the shape of the adjoint-optimized devices using the explanation heatmaps. The algorithm, conceptually illustrated in Figure 4a, consists of an initial filtering procedure, which identifies the red-to-blue transition points along the structure boundary. In this procedure, a binarization function is applied to the SHAP values that converts the structure into existing and non-existing elements (shown in Figure 4a as white and black pixels, respectively). Thresholds for binarization ($\rho(x,y)$) are given by the following step function:

$$
\rho(x, y) = \begin{cases} 
1 & \text{for } \Phi(x, y) \leq 0, \\
0 & \text{for } \Phi(x, y) > 0,
\end{cases}
$$

(2)

where $\rho(x,y)=1$ and $\rho(x,y)=0$ indicates existing and non-existing elements, respectively. A median filter is applied to the binarization to reduce noise. Next, we “draw” a boundary around the existing elements by capturing an array of points $\eta(x,y)=[X_i;Y_i]$, in which $X_i=[x_1,x_2,...,x_i]$ and $Y_i=[y_1,y_2,...,y_i]$ are vectors of length $i$. $X_i$ is an evenly spaced set of x-coordinate values from the left to the right of the image. Since $i$ ultimately determines the resolution of the shape, to ensure that the optimizable geometry is within feasible fabrication range, we set its value to 20 points (matching the interval used in the initial optimization runs). This interval equates to 100 nm spacing along the x-axis, which is well within CMOS lithography resolutions. To find the y-coordinate values in $Y_i$, we apply Algorithm 1 (detailed in the Supporting Information).
Figure 4. (a) Schematic representation of our explanation-optimization algorithm and workflow, consisting of explanation, filtering, and boundary extraction steps. (b) Comparison between the adjoint-optimized and explanation-optimized geometries.

Using Algorithm 1 we raster the image (from top to bottom) across all values in $X_i$ and find the points where existing elements are found (indicated by $P(x,y)=1$), then mark these points for $Y_i$. For quality purposes, we add the $\alpha$ hyperparameter to enhance robustness by reducing sharp changes in the structure as a result of filtering or noise from the explanations. We apply this workflow to each wavelength-specific adjoint-optimized structure from the previous step and present the new “explanation-optimized” results in Figure 4b. As an example of our method’s application, for the 1.3 $\mu$m target design, we note that the explanation algorithm deemed the large vertical spike at the left as a negative contribution. After applying our explanation-based boundary extraction process, it can be observed that this spike was removed.
To assess whether the explanation (or SHAP value) based modifications to the optimized structures (e.g., the spike removal) yield meaningful or effective contributions, we simulated the explanation-optimized designs and used them as new starting points for a second stage of adjoint optimization runs. In Figure 5, we show the FOM evolutions over the entire optimization cycle. The red arrows indicate the end of the first optimization stage and the beginning of the second explanation-based re-optimization stage. Further observation reveals that in the second stage of the 1.3 μm target design, removing the vertical spike immediately reduced the FOM from 0.139 to 0.101 (a 27% improvement) at the first iteration, while the end of the optimization resulted in a final FOM of 0.038 (a 73% improvement compared to the end of the first stage). In some of the other examples (e.g., 1.5-1.8 μm), we note that the first step of the second stage did not always result in an immediate FOM reduction, particularly when the FOM value was already exceedingly low (<0.075). However, across all the optimization targets, the end of every second-stage optimization consistently resulted in a lower final FOM than the first-stage FOM (by 43% on average). Moreover, an increase in FOM followed by a further global reduction is indicative of an objective function that was previously stuck in a local minimum [14]. Thus, we demonstrate that our explanation-based re-optimization technique is capable of enhancing the performance of the adjoint optimization algorithm by allowing the FOM to escape its local minima at various optimization targets and performance ranges. We note that this procedure only used two optimization runs per target: one for feature contribution learning and the other for local minima escape or global FOM reduction. As previously mentioned, alternative methods at identifying lower minima typically involve repeated optimizations at random starting points, which can scale well-beyond two optimization runs per target.

To show that the achieved performance enhancements were not simply obtained through arbitrary perturbations to the optimized structure, we performed an additional “random change” analysis where we randomly modified the first-stage structures, repeated the second stage of optimizations, and compared the results. This comparison is presented in Figure S2, where five random modifications (defined in the Supporting Information) were made to each structure. Across the 30 tests performed on the six optimized designs, all of the randomly modified structures possess higher FOM values (i.e., lower performance) than the explainability-optimized devices, while only two results fall within 25% of the latter’s final device performance. Additionally, 28 tests yield higher final FOM values than the initial optimized design. Therefore,
not only are the random changes ineffective in terms of escaping the local minima, but they can also inadvertently push the optimization into a worse state than where the optimization started at. As such, we demonstrate that our XAI-based approach is not stochastic in nature, but can deterministically tune a structure in order to maximize performance.

![Figure 5. Two-stage optimization of SOI waveguide designs. Red arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage, where the FOM values are improved by 43%, on average, across all target wavelengths.](image)

Lastly, we evaluated the generalizability of the proposed framework by applying it to other contemporary nanophotonics design challenges. We note that prior studies have successfully applied XAI to alternative nanophotonic structures, such as metasurfaces, and demonstrated performance enhancements in the form of spectral property tuning [26] (though no optimization algorithm integration was employed). Thus, to differentiate this work further and to maximize implementation time, we focused this generalizability analysis on material alternatives. In this regard, over the past few years, Si₃N₄ has emerged as a promising alternative to silicon in photonic systems. Compared to silicon, Si₃N₄ has low propagation losses and does
not exhibit two-photon absorption in the telecommunications range [31][32][33] (among other pros and cons [34]). Accordingly, we performed the same two-stage optimization study on a Si$_3$N$_4$ waveguide (for the same Y-splitter starting geometry) and found that the second-round optimization was also able to surpass the results of the first at every target wavelength (Figure S3). Across all the test cases, an average FOM improvement of 74% was achieved. Subsequently, we show that the presented approach is generally applicable to numerous applications of adjoint optimization for electromagnetic design, including those with different constituent materials, structures, and optimization targets.

Conclusions

In summary, we present an inverse design framework that extends the capabilities of gradient-based shape or topology optimization algorithms for photonic inverse design, while elucidating the relationships between device performance and nanoscale structuring. Our framework combines adjoint optimization, AutoML, and XAI to enhance device performance beyond that obtainable through the optimization algorithm alone. We applied our method to SOI waveguide design and showed that the optimization algorithm initially reaches a performance plateau (i.e., local minima). After utilizing XAI to reveal the device’s structural contributions towards a designated FOM (where 0 represents ideal performance), we leveraged this information (in conjunction with a boundary extraction algorithm) to push the optimization out of its local minima and reduce the FOM further. Across a range of performance-plateaued devices, optimized for various wavelengths within the 1.3 μm to 1.8 μm telecom window, our method was able to improve device performance by an average of 43%. The entire procedure only requires two optimization runs per optimization target, which is potentially more computationally efficient than alternative approaches, particularly those that rely on multiple optimization runs and random starting points. Additionally, generalizability tests performed on Si$_3$N$_4$ waveguides showed an average of 74% device performance improvement. Thus, we demonstrate that our XAI-based approach provides an automated and systematic solution for an electromagnetic optimization algorithm to escape local minima and achieve greater device performance. Looking ahead, integrating conventional optimization and data-driven machine
learning will likely prove a fruitful direction for inverse design and physics discovery in photonic systems.

Supporting Information

AutoML training results, details of the boundary extraction algorithm, random change comparison, and generalizability analysis.

Acknowledgements

This work was supported by the Sloan Research Fellowship from the Alfred P. Sloan Foundation. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. P. Sandborn, N. Quack, N. Hoghooghi, J. B. Chou, J. Ferrara, S. Gambini, B. Behroozpour, L. Zhu, B. Boser, C. Chang-Hasnain, and M. C. Wu, "Linear frequency chirp generation employing optoelectronic feedback loop and integrated silicon photonics," CLEO: 2013, 1-2 (2013).
2. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, “Adjoint shape optimization applied to electromagnetic design,” Opt. Express 21, 21693-21701 (2013).
3. I. A. Krutova, M. Yu. Saygin, I. V. Dyakonov, and S. P. Kulik, “Optimized low-loss integrated photonics silicon-nitride Y-branch splitter,” AIP Conference Proceedings 2241, 020027 (2020).
4. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, “Deep learning with coherent nanophotonic circuits,” Nature Photonics 11, 441 (2017).
5. N. C. Harris, J. Carolan, D. Bunandar, M. Prabhu, M. Hochberg, T. Baehr-Jones, M. L. Fanto, A. M. Smith, C. C. Tison, P. M. Alsing, and D. Englund, “Linear programmable nanophotonic processors,” Optica 5, 1623 (2018).
6. O.D. Miller, “Photonic Design: From Fundamental Solar Cell Physics to Computational Inverse Design,” PhD dissertation, University of California, Berkeley (2012).
7. W. Li, F. Meng, Y. Chen, Y. Li, and X. Huang, “Topology optimization of photonic and phononic crystals and metamaterials: a review,” Adv. Theory Simul. 2, 1900017 (2019).
8. S. D. Campbell, D. Sell, R. P. Jenkins, E. B. Whiting, J. A. Fan, and D. H. Werner, "Review of numerical optimization techniques for meta-device design," Opt. Mater. Express 9, 1842-1863 (2019).
9. S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez, “Inverse design in nanophotonics,” Nature Photon 12, 659–670 (2018).
10. Sell, D., Yang, J., Doshay, S., Yang, R. & Fan, J. A. Large-angle, multifunctional metagratings based on freeform multimode geometries. Nano Lett. 17, 3752–3757 (2017).
11. R. Matzen, J. S. Jensen, and O. Sigmund, “Topology optimization for transient response of photonic crystal structures,” J. Opt. Soc. Am. B 27, 2040-2050 (2010).
12. Y. Deng, & J. G. Korvink, “Topology optimization for three-dimensional electromagnetic waves using an edge element-based finite-element method,” Proc. R. Soc. A 472, 20150835 (2016).
13. B. Shen, P. Wang, R. Polson, and R. Menon, “Integrated metamaterials for efficient and compact free-space-to-waveguide coupling,’ Opt. Express 22, 27175–27182 (2014).
14. J. Jiang, M. Chen, and J. A. Fan, “Deep neural networks for the evaluation and design of photonic devices,” Nat Rev Mater 6, 679–700 (2021).
15. J. A. Fan, “Freeform metasurface design based on topology optimization,” MRS Bulletin 45, 196–201 (2020).
16. J. Jiang, and J. A. Fan, “Multiobjective and categorical global optimization of photonic structures based on ResNet generative neural networks,” Nano 10, 361-369 (2021).
17. C. Yeung, R. Tsai, B. Pham, B. King, Y. Kawagoe, D. Ho, J. Liang, M. W. Knight and A. P. Raman, “Global Inverse Design across Multiple Photonic Structure Classes Using Generative Deep Learning,” Adv. Optical Mater. 2100548, (2021).
18. W. Ma, Z. Liu, Z. A. Kudyshnev, A. Boltasseva, W. Cai, and Y. Liu, “Deep learning for the design of photonic structures,” Nat. Photonics 15, 77–90 (2021).
19. J. Nalep, “Smart Delivery Systems: Solving Complex Vehicle Routing Problems,” Elsevier (2020).
20. W. J. Kim and J. D. O'Brien, "Optimization of a two-dimensional photonic-crystal waveguide branch by simulated annealing and the finite-element method," J. Opt. Soc. Am. B 21, 289-295 (2004).
21. C. Zhao and J. Zhang, "Binary plasmonics: launching surface plasmon polaritons to a desired pattern," Opt. Lett. 34, 2417-2419 (2009).
22. Z. Fayyaz, N. Mohammadian, F. Salimi, A. Fatima, M. R. R. Tabar, and M. R. N. Avanaki, "Simulated annealing optimization in wavefront shaping controlled transmission," Appl. Opt. 57, 6233-6242 (2018).
23. G. Vilone, and L. Longo, “Explainable Artificial Intelligence: a Systematic Review,” arXiv (2020).
24. S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. Low, S. Newman, J. Kim, and S. Lee, “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” Nat Biomed Eng 2, 749–760 (2018).
25. A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Info. Fus. 58 (2020).
26. C. Yeung, J. Tsai, B. King, Y. Kawagoe, D. Ho, M. W. Knight, and A. P. Raman, “Elucidating the Behavior of Nanophotonic Structures through Explainable Machine Learning Algorithms,” ACS Photo. 7, 8, 2309–2318 (2020).
27. S. Lundberg, and S. Lee, “A Unified Approach to Interpreting Model Predictions,” arXiv (2017).
28. C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, "Adjoint shape optimization applied to electromagnetic design," Opt. Express 21, 21693-21701 (2013).
29. S. Osher and J. A. Sethian, “Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations,” J. Comput. Phys. 79(1), 12–49 (1988).
30. H. Jin, Q. Song, and X. Hu, “Auto-Keras: An Efficient Neural Architecture Search System,” arXiv (2018).
31. Y. Zhang, C. Husko, S. Lefrancois, I. H. Rey, T. F. Krauss, J. Schröder, and B. J. Eggleton, "Non-degenerate two-photon absorption in silicon waveguides: analytical and experimental study," Opt. Express 23, 17101-17110 (2015).
32. I. A. Krutova, M. Y. Saygin, I. V. Dyakonov, and S. P. Kulik, “Optimized low-loss integrated photonics silicon-nitride Y-branch splitter,” AIP Conf. Proc. 2241, 020027 (2020).
33. H. E. Dirani, L. Youssef, C. Petit-Etienne, S. Kerdiles, P. Grosse, C. Monat, E. Pargon, and C. Sciancalepore, "Ultralow-loss tightly confining Si3N4 waveguides and high-Q microresonators," Opt. Express 27, 30726-30740 (2019).
34. R. Baets, A. Z. Subramanian, S. Clemmen, B. Kuyken, P. Bienstman, N. Le Thomas, G. Roelkens, D. Van Thourhout, P. Helin, and S. Severi, "Silicon Photonics: silicon nitride versus silicon-on-insulator," Opt. Fiber Comm. Conf. (2016).
Supporting Information

AutoML Training and Execution

In this work, we use neural architecture search (AutoKeras) with image blocks to automate the deep learning process by testing different model variants across multiple trials. We observe that the optimal architecture was identified after 17 trials, which has a validation loss of 8.6×10⁻⁵. The final training and validation losses of each trial are presented in Figure S1a, and the evolution of the convolutional neural network (CNN) architecture from the first trial to the last is shown in Figure S1b. The first trial initially begins with a two-block architecture and a validation loss of 3.4×10⁻². At the final trial, the optimized CNN possesses four convolutional blocks (with 512, 64, 32, and 32 filters, respectively) followed by a dense layer. Each block contains Leaky ReLu, batch normalization, and max pooling layers.

![Figure S1](image)

**Figure S1.** (a) Training and validation loss progression across the AutoML trials. (b) Schematic of the model architecture changes between the first and final trials.

Boundary Extraction Algorithm

As described in the main text, we devised a boundary extraction algorithm to systematically adjust the shape of the adjoint-optimized devices using the explanation heatmaps. This algorithm is represented as:

**Algorithm 1: Boundary Extraction**

**Input:** Filtered image array \( (P) \) of \( \rho(x,y) \) values at every image pixel.
for $x \in X_i$, do
  for $y = 1:L$, where $l$ is the top of the image and $L$ is the height or bottom of the image, do
    if $P(x,y) = 1$ and $y_i - y_{i-1} < \alpha$, where $\alpha$ is a hyperparameter for enhancing robustness, then
      $Y_i.append(y_i)$
    else
      $Y_i.append(y_{i-1})$
  Return: $Y_i$

As shown in Algorithm 1, we raster the image (from top to bottom) across all values in $X_i$ and find the points where existing elements are found (indicated by $P(x,y)=1$), then mark these points for $Y_i$. For quality purposes, we add the $\alpha$ hyperparameter (with a default value of 5 points or 100 nm in device length) to enhance robustness by reducing sharp changes in the structure as a result of filtering or noise from the explanations.

Random Change Comparison

To show that the achieved performance enhancements were not simply obtained through arbitrary perturbations to the optimized structure, we performed an additional “random change” analysis where we randomly modified the first-stage structures, repeated the second stage of optimizations, and compared the results. This comparison is presented in Figure S2, where five random modifications were made to each structure. Random modifications were introduced by adding 300 nm to $y_n$ and $y_{n+1}$ in the optimizable parameter vector $Y_i$, where $n$ is varied from 1 to $i-1$ (where $i$ is the length of the vector). Across the 30 tests performed on the six optimized designs, all of the randomly modified structures possess higher FOM values (i.e., lower performance) than the explainability-optimized devices, while only two results fall within 25% of the latter’s final device performance. Additionally, 28 tests yield higher final FOM values than the initial optimized design. Therefore, not only are the random changes ineffective in terms of escaping the local minima, but they can also inadvertently push the optimization into a worse state than where the optimization started at.
Generalizability Assessment

We evaluate the generalizability of the proposed framework by applying it to other contemporary nanophotonics design challenges. Accordingly, we performed the same two-stage optimization study on a Si₃N₄ waveguide (for the same Y-splitter starting geometry). Similar to the results presented in the main text (Figure 5), here we observe that the second stage (red-boxed region) of each target wavelength-specific optimization produces a lower overall FOM than the final FOM obtained through adjoint optimization alone (indicated by the red arrows). In particular, as shown in Figure S3, we observe a minimum of 31% FOM improvement (for the 1.6 µm design) and a maximum of 90% improvement (for the 1.3 µm design), for an average of 74% improvement across all the test cases. Subsequently, we show that the presented approach is generally applicable to numerous applications of adjoint optimization for
electromagnetic design, including those with different constituent materials, structures, and optimization targets.

**Figure S3.** Two-stage optimization of Si₃N₄ waveguide designs. Red arrows indicate the end of the first adjoint optimization stage and the beginning of the second explanation-based re-optimization stage. Inset images capture a closer look at the second stages, where the FOM values are improved by 74%, on average, across all target wavelengths.