Title

Investigating the Behavior of Nanophotonic Structures using Explainable Convolutional Neural Network

Author

Tsai, Ju-Ming

Publication Date

2020

UNIVERSITY OF CALIFORNIA

Los Angeles

Investigating the Behavior of Nanophotonic Structures using Explainable Convolutional Neural Network

A thesis submitted in partial satisfaction

of the requirements for the degree of Master of Science

in Materials Science and Engineering

by

Ju-Ming Tsai

2020

ABSTRACT OF THE THESIS

Investigating the Behavior of Nanophotonic Structures using Explainable Convolutional Neural Network

by

Ju-Ming Tsai

Master of Science in Materials Science and Engineering

University of California, Los Angeles, 2020

Professor Aaswath Pattabhi Raman, Chair

Reaching the true potential of nanophotonic devices requires the broadband control of spectral and angular selectivity in the absorption and emission of electromagnetic waves. To this end, previously investigated design methods for nanophotonic structures and have encompassed both conventional forward and inverse optimization approaches as well as nascent machine learning (ML) strategies. While far more computationally efficient than optimization processes, ML-based methods that are capable of generating complex

nanophotonic structures are still ‘black boxes’ that lack explanations for their predictions. In that regard, we demonstrate that well-established deep learning architectures such as convolutional neural networks (CNN), which are highly proficient at forward design, can be explained to derive unique design insights by extracting the underlying physical relationships learned by network. To illustrate this capability, we trained a CNN model with 10,000 images of selective mid-infrared thermal emitters and their corresponding absorption spectra. The trained CNN predicted the spectra of new and unknown designs with over 95% accuracy. After training the CNN, we applied the Shapley Additive Explanations (SHAP) algorithm to the model to determine features that made positive or negative contributions towards specific spectral points, thereby informing which features to create or eliminate in order to meet a target spectrum. Using this strategy, we show that a starting electromagnetic metasurface design can be selectively manipulated to create target spectral properties. Our results reveal that the physical relationships between structure and spectra can be obtained, and new designs can be achieved, by exposing the valuable information hidden within a neural network.

The thesis of Ju-Ming Tsai is approved.

Jaime Marian

Amartya Sankar Banerjee

Aaswath Pattabhi Raman, Chair

University of California, Los Angeles

2020

Table of Contents

Table of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Professor Aaswath Raman for his guidance and support in the project and this thesis. I would like to thank Professor Jaime Marian and Professor Amartya Banerjee for being the committee on my thesis. My research partner, Christopher Yeung, was instrumental in discussing the path of my research. I am extremely grateful for this. I thank all the colleagues in Professor Raman's lab for the help. I thank all the support from my friends in UCLA.

Last but not least, I am remarkably appreciative to my parents and my family members, for being constantly supportive. I would not be able to finish the degree without their support.

Chapter 1

Introduction

Photonic devices play an important role in many areas of our daily life and it has enabled a wide range of transformative technologies such as such as photonic integrated circuits for ultra-high speed optical communication[1, 2, 3] and metasurfaces that precisely control the propagation of electromagnetic waves (*e.g.* waveguides, directional output couplers, and thermal emitters)[4, 5, 6]. How the electromagnetic wave interacts with the materials at each different specific wavelength enables the different applications for the nanophotonic devices.

Different interactions such as scattered, refracted, filtered between light and the photonic device can be achieved by using metallic and dielectric 2D nanostructures and 3D nanoarchitectures. Even some fascinating new interactions occur that are impossible with natural materials and in conventional geometries. The control over light has enabled a variety of applications, including optical computing, medical technologies and many other novel applications[7].

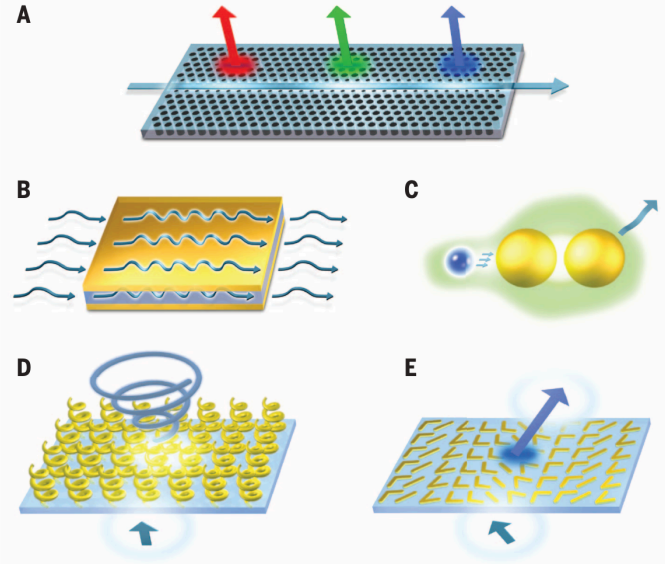

Figure 1. Photonic device applications (A) 2D photonic crystal waveguide coupled to resonant cavities serves as a wavelength division multiplexer. (B) Metal-insulator-metal surface plasmon polariton waveguide strongly confines light and shrinks the wavelength. (C) Plasmonic dimer nanoantenna coupled to an optical emitter creates directional emission of light. (D) Metasurface composed of chiral antennas offers selectivity to circularly polarized light. (E) Metasurface composed of graded plasmonic or dielectric antenna geometries enables wavelength-dependent control over the reflection and refraction of the optical wave front[7].

Figure 1 shows some well-established applications for the photonic usage. Among one of them is the metal-insulator-metal (MIM) surface plasmon polariton (SPP) waveguide. The size and density of optical devices employing conventional dielectric optical waveguide and photonic crystals will in principle be limited by the diffraction limit of light, where the smallest diameter of the beam propagating in dielectric is of the order of λ_0/n[8, 9]. The MIM structure serves as the plasmonic waveguide to squeeze the SPP field into the dielectric core and the wavelength along the direction of propagation can be shortened significantly[10]. MIM-SPPs can therefore be guided in waveguides with very small transverse dimensions[11] and allow the realization of nanocavities with extremely small mode volumes[12, 13].

The finite-difference time-domain (FDTD) method is a powerful computational technique

which is widely used to calculate the optical properties of nanostructures[14, 15, 16]. The major advantage of the FDTD method compared to other methods is its ability to provide a full spectrum in a single simulation by propagating a short pulse in the time-domain[17]. Thus to describe realistic materials, the frequency dependent dielectric functions of the constituent materials need to be modeled in an analytical form, then transformed into the time-domain.

However, forward design is limited by trial and error guided by domain knowledge and human intuition. It is also strongly limited to the degree of the insight to the problem. To get a nontrivial design that may be overseen by human intuition, a deterministic optimization algorithm is needed to search for the possible solution. The inverse design solves problem by optimizing the design parameters with certain constraint. It is a computational algorithm that enables automatically design optimal structure consisting of dielectric or metal materials in the system[18].

In optical device optimization, various optimization methods have been used, such as genetic algorithm, or topology optimization. The topology optimization method was originally developed for structural optimization problems, but has recently been extended to some other design problems[19]. For example, broadband photonic crystal waveguide[20] or even 3D nanophotonic devices structure[21] can be optimized. However, these optimization algorithms usually optimize for some specific metrics, and they rarely directly achieve the most suitable structure parameters for a complete transmission spectrum in a wide wavelength range[22].

In recent years, machine-learning based method has shown the power on the exploration and optimization of complex problems. For example, the quantum many-body problem could be solved by utilizing ANNs[23]. Moreover, the light scattering of multilayer nanoparticles with different thicknesses can be simulated with a trained ANNs[24]. ANNs could solve the spectrum prediction and inverse design problems more quickly than the numerical simulation method[24, 25].

However, among all the machine learning methods which have been utilized on the photonic material problems, the complex physical relationship between input parameters and output optical responses have kept a mystery. In this thesis, we tried to open the 'black box' of the machine learning method by implementing an external explanation model and acquired more insight of the machine learning model for the inverse design problem.

Chapter 2

Background

Nanophotonic devices have many applications in different fields. There is tremendous potential for further advancements in these applications, but they are critically limited by existing design processes.

To address the problems of designing nanophotonic device, people have come up with several different design methods for it. For example, the most common approach to photonic device design is via numerical simulations based on physical laws (*e.g.*, Maxwell's Equations) to optimally design nanophotonic structures with specific, target functionality. This design technique, also known as "forward design"[26].

2.1 Electromagnetic Simulation

Fundamental problem for the simulation of the photonic devices is solving the Maxwell's equations in both time and space simultaneously.

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0} \quad \text{——————} \quad (2.1)$$

$$\nabla \cdot \mathbf{H} = 0 \quad \text{——————} \quad (2.2)$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad \text{——————} \quad (2.3)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \quad \text{——————} \quad (2.4)^{27}$$

Where $\boldsymbol{E}$ is the electric field vector, $\boldsymbol{H}$ is the magnetic field vector, μ is the magnetic permeability, ε is the electric permittivity, $\boldsymbol{J}$ is the current density.

The Maxwell's equations is composed of four different partial equation that describes different relationship between electric field, magnetic field, charge density and current density in the system. Despite the great linear and first-order mathematical properties of the equations, the analytical form of the solutions for the Maxwell's equations does not exist. The numerical method can only be applied to approximate distribution of fields in time and space in the system.

The Finite-difference Time-domain (FDTD) method has been widely used as a forward simulation method to solve the electromagnetic wave propagation numerically[28]. FDTD simulation was first invented by K. S. Yee et al and it discretizes the Maxwell equation in space domains using Yee's discretization cells[29]. The Maxwell's curl equations are then calculated in isotropic, linear, non-dispersive media. The approximation of the simulation eventually matches the continuous equation as the grid and the time steps is fine enough.

The FDTD uses the central difference approximation to the Maxwell's equations. It solves the equation and update the result of both the electric and magnetic fields at each time step and discretized space point in the defined system using leapfrog method. Moreover, while analyzing the electromagnetic response from the material structure, an absorbing boundary condition or perfectly matched layer, which suppress redundant reflection, needs to be applied

if the boundary is unbounded.

2.1.1 Discretized Cell in Cartesian Coordination System

In 3D cartesian coordination system, the total system can be divided into cubic or rectangular cells with Δx, Δy, Δz in length, as shown in figure 2.

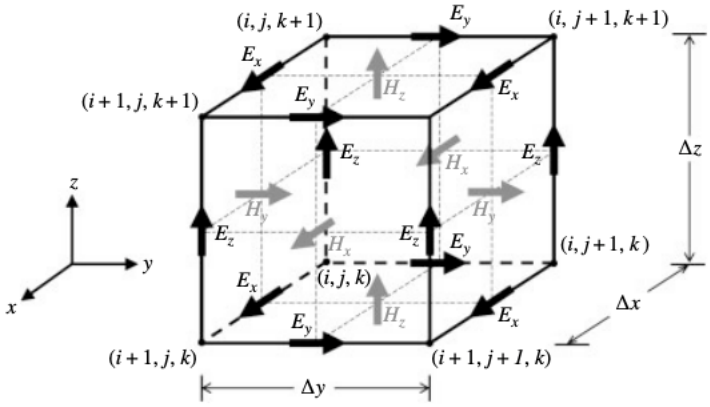

Figure 2. Relative placement of electric field and magnetic fields components in the cell.

The electric field components are placed at the center of the slides of the cells: $\boldsymbol{E_x}$ components are placed at the midpoints of sides aligned in the x direction, $\boldsymbol{E_y}$ components are placed at the midpoints of sides aligned in the y direction, $\boldsymbol{E_z}$ components are placed at the midpoints of sides aligned in the x direction. The magnetic field components are placed at the center of the faces of the cells and components are normal to the faces. $\boldsymbol{H_x}$ is placed at the center on y-z planes, $\boldsymbol{H_y}$ are placed at the center on x-z planes, $\boldsymbol{H_z}$ are placed at the center on x-y planes.

2.1.2 Leapfrog Method

Leapfrog method is used for updating the electric and magnetic fields in each cells

for every time step in FDTD simulation. The electric field components are calculated at time steps nΔt, where n is an integer and Δt is the time increment. On the other hand, the magnetic field components are computed at half integer time steps $\left(n+\frac{1}{2}\right)\Delta t$.

Time-update equations for electric field components $\boldsymbol{E_x}$, $\boldsymbol{E_y}$, $\boldsymbol{E_z}$ are derived from Ampere's law as Eq. (2.4), and the magnetic field components $\boldsymbol{H_x}, \boldsymbol{H_y}, \boldsymbol{H_z}$ are derived from Faraday's law, as shown in Eq. (2.3).

$$\nabla \times \boldsymbol{H}^{n-\frac{1}{2}} = \varepsilon \frac{\partial \boldsymbol{E}^{n-\frac{1}{2}}}{\partial t} + \boldsymbol{J}^{\boldsymbol{n-\frac{1}{2}}} = \varepsilon \frac{\partial \boldsymbol{E}^{n-\frac{1}{2}}}{\partial t} + \boldsymbol{\sigma E}^{\boldsymbol{n-\frac{1}{2}}} \quad \text{—— (2.5)}$$

Where σ is the electric conductivity. Eq. (2.5) shows how the magnetic field components are computed at time step $(n-\frac{1}{2})\Delta t$. $\varepsilon \frac{\partial \boldsymbol{E}}{\partial t}$ is the displacement induced by the electric field vector. Moreover, it is showing that the conduction current with time variation of electric field induces magnetic field in right-hand curl direction.

$$\nabla \times \boldsymbol{H}^{n-\frac{1}{2}} = \varepsilon \frac{\partial \boldsymbol{E}^{n-\frac{1}{2}}}{\partial t} + \sigma \boldsymbol{E}^{\boldsymbol{n-\frac{1}{2}}} \approx \varepsilon \frac{\boldsymbol{E}^{n}-\boldsymbol{E}^{n-1}}{\Delta t} + \sigma \frac{\boldsymbol{E}^{n}+\boldsymbol{E}^{n-1}}{2} \quad \text{—— (2.6)}$$

Approximating $\boldsymbol{E}^{\boldsymbol{n-\frac{1}{2}}}$ term by using its average value $\frac{\boldsymbol{E}^{n}+\boldsymbol{E}^{n-1}}{2}$, we can get Eq. (2.6). If we rearrange the term in Eq. (2.6), the update equation for the electric field vector at the time step number n $\boldsymbol{E}^{\boldsymbol{n}}$ can be obtained from a time-step previous value $\boldsymbol{E}^{\boldsymbol{n-1}}$ and the half time-step previous magnetic field curl value $\nabla \times \boldsymbol{H}^{n-\frac{1}{2}}$, as shown below:

$$\boldsymbol{E}^{\boldsymbol{n}} = \left(\frac{1-\frac{\sigma \Delta t}{2\varepsilon}}{1+\frac{\sigma \Delta t}{2\varepsilon}}\right)\boldsymbol{E}^{\boldsymbol{n-1}} + \left(\frac{\frac{\Delta t}{\varepsilon}}{1+\frac{\sigma \Delta t}{2\varepsilon}}\right)\nabla \times \boldsymbol{H}^{n-\frac{1}{2}} \quad \text{—— (2.7)}$$

The update equation $\boldsymbol{E}_{\boldsymbol{x}}^{\boldsymbol{n}}$ at location (i+1/2, j, k) (shown in figure 3) can be calculated from equation (2.7) and is expressed as follow in Eq. (2.8).

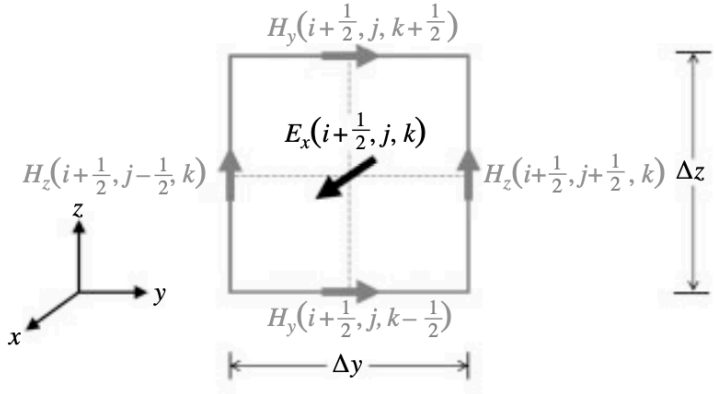

Figure 3. Electric field component in the x direction $\boldsymbol{E_x^n}$ at a location (i + 1/2, j, k) and the circulating magnetic field components closest to it.

$$\boldsymbol{E_x^n}\left(i+\frac{1}{2},j,k\right)=\frac{1-\frac{\sigma\left(i+\frac{1}{2},j,k\right)\Delta t}{2\varepsilon\left(i+\frac{1}{2},j,k\right)}}{1+\frac{\sigma\left(i+\frac{1}{2},j,k\right)\Delta t}{2\varepsilon\left(i+\frac{1}{2},j,k\right)}}\boldsymbol{E_x^{n-1}}\left(i+\frac{1}{2},j,k\right)+\frac{\frac{\Delta t}{\varepsilon\left(i+\frac{1}{2},j,k\right)}}{1+\frac{\sigma\left(i+\frac{1}{2},j,k\right)\Delta t}{2\varepsilon\left(i+\frac{1}{2},j,k\right)}}\left[\frac{\partial \boldsymbol{H_z^{n-\frac{1}{2}}}\left(\boldsymbol{i+\frac{1}{2},j,k}\right)}{\partial y}-\frac{\partial \boldsymbol{H_y^{n-\frac{1}{2}}}\left(\boldsymbol{i+\frac{1}{2},j,k}\right)}{\partial z}\right]$$

$$=\frac{1-\frac{\sigma\left(i+\frac{1}{2},j,k\right)\Delta t}{2\varepsilon\left(i+\frac{1}{2},j,k\right)}}{1+\frac{\sigma\left(i+\frac{1}{2},j,k\right)\Delta t}{2\varepsilon\left(i+\frac{1}{2},j,k\right)}}\boldsymbol{E_x^{n-1}}\left(i+\frac{1}{2},j,k\right)+\frac{\frac{\Delta t}{\varepsilon\left(i+\frac{1}{2},j,k\right)}}{1+\frac{\sigma\left(i+\frac{1}{2},j,k\right)\Delta t}{2\varepsilon\left(i+\frac{1}{2},j,k\right)}}\frac{1}{\Delta z\Delta y}\left[\boldsymbol{H_z^{n-\frac{1}{2}}}\left(i+\frac{1}{2},\ j+\frac{1}{2},\ k\right)\Delta z-\right.$$

$$\left.\boldsymbol{H_z^{n-\frac{1}{2}}}\left(i+\frac{1}{2},\ j-\frac{1}{2},\ k\right)\Delta z-\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i+\frac{1}{2},\ j,\ k+\frac{1}{2}\right)\Delta y+\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i+\frac{1}{2},\ j,\ k-\frac{1}{2}\right)\Delta y\right]$$

(2.8)

The spatial derivative terms are approximated by the central finite differences $\frac{\boldsymbol{H_z^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j+\frac{1}{2},k\right)-\boldsymbol{H_z^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j-\frac{1}{2},k\right)}{\Delta y}$ and $\frac{\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)+\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j,k-\frac{1}{2}\right)}{\Delta z}$. $\boldsymbol{E_y^n}$ and $\boldsymbol{E_z^n}$ terms are derived and updated in the same manner as follow:

$$\boldsymbol{E_y^n}\left(i,j+\frac{1}{2},k\right)=\frac{1-\frac{\sigma\left(i,j+\frac{1}{2},k\right)\Delta t}{2\varepsilon\left(i,j+\frac{1}{2},k\right)}}{1+\frac{\sigma\left(i,j+\frac{1}{2},k\right)\Delta t}{2\varepsilon\left(i,j+\frac{1}{2},k\right)}}\boldsymbol{E_y^{n-1}}\left(i,j+\frac{1}{2},k\right)+\frac{\frac{\Delta t}{\varepsilon\left(i,j+\frac{1}{2},k\right)}}{1+\frac{\sigma\left(i,j+\frac{1}{2},k\right)\Delta t}{2\varepsilon\left(i,j+\frac{1}{2},k\right)}}\frac{1}{\Delta z\Delta x}\left[\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)\Delta x-\right.$$

$$\left.\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i,j+\frac{1}{2},k-\frac{1}{2}\right)\Delta x-\boldsymbol{H_z^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j+\frac{1}{2},k\right)\Delta z+\boldsymbol{H_z^{n-\frac{1}{2}}}\left(i-\frac{1}{2},j+\frac{1}{2},k\right)\Delta y\right]$$

$$\boldsymbol{E_z^n}\left(i,j,k+\frac{1}{2}\right)=\frac{1-\frac{\sigma\left(i,j,k+\frac{1}{2}\right)\Delta t}{2\varepsilon\left(i,j,k+\frac{1}{2}\right)}}{1+\frac{\sigma\left(i,j,k+\frac{1}{2}\right)\Delta t}{2\varepsilon\left(i,j,k+\frac{1}{2}\right)}}\boldsymbol{E_z^{n-1}}\left(i,j,k+\frac{1}{2}\right)+\frac{\frac{\Delta t}{\varepsilon\left(i,j,k+\frac{1}{2}\right)}}{1+\frac{\sigma\left(i,j,k+\frac{1}{2}\right)\Delta t}{2\varepsilon\left(i,j,k+\frac{1}{2}\right)}}\frac{1}{\Delta x\Delta y}\left[\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)\boldsymbol{\Delta y}-\right.$$

$$\left.\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i-\frac{1}{2},j,k+\frac{1}{2}\right)\Delta y-\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)\Delta x+\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i,j-\frac{1}{2},k+\frac{1}{2}\right)\Delta x\right]$$

(2.9)

For Faraday's law, it states the time variation of magnetic field induces the electric field in the negative direction of right hand curl. The equation is given as follow.

$$\nabla \times \boldsymbol{E^n} = -\mu \frac{\partial \boldsymbol{H^n}}{\partial t} \quad \text{—— (2.10)}$$

If the time variation term is approximated by the central finite difference, it can be showed as follow:

$$\mu \frac{\partial \boldsymbol{H^n}}{\partial t} \approx \mu \frac{\boldsymbol{H^{n+\frac{1}{2}}} - \boldsymbol{H^{n-\frac{1}{2}}}}{\Delta t} \approx -\nabla \times \boldsymbol{E^n} \quad \text{—— (2.11)}$$

$$\boldsymbol{H^{n+\frac{1}{2}}} = \boldsymbol{H^{n-\frac{1}{2}}} - \frac{\Delta t}{\mu} \nabla \times \boldsymbol{E^n} \quad \text{—— (2.12)}$$

We rearrange Eq. (2.11) and get Eq. (2.12). The update equation for the magnetic field at time step n+1/2 is obtained from one-step previous value $\boldsymbol{H^{n-\frac{1}{2}}}$ and half-step previous electric field curl value $\nabla \times \boldsymbol{E^n}$.

From Eq. (2.12), the update equation for $\boldsymbol{H_x^{n+\frac{1}{2}}}$ at a location (i, j + 1/2, k + 1/2) (see Figure 4), for example, is expressed as follows:

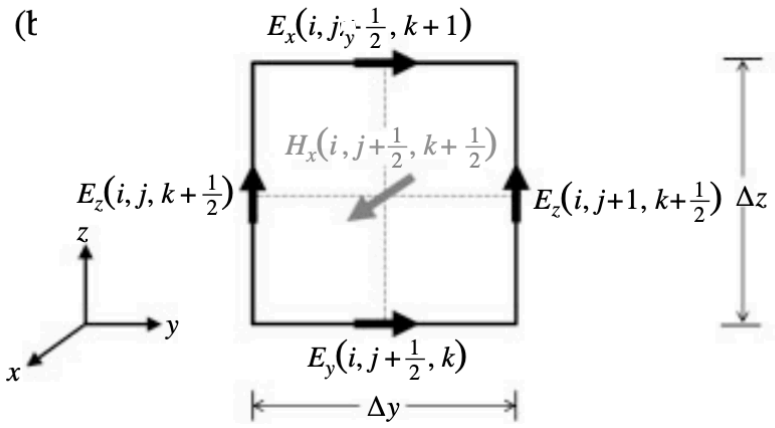

Figure 4. Magnetic field component in the x direction $\boldsymbol{H_x^{n+\frac{1}{2}}}$ at a location (i, j + 1/2, k + 1/2) and the circulating electric field components closest to it.

$$\boldsymbol{H_x^{n+\frac{1}{2}}}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)=\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)-\frac{\Delta t}{\mu\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}\left[\frac{\partial \boldsymbol{E_z^n}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}{\partial y}-\frac{\partial \boldsymbol{E_y^n}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}{\partial z}\right]$$

$$=\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)-\frac{\Delta t}{\mu\left(i,j+\frac{1}{2},k+\frac{1}{2}\right)}\frac{1}{\Delta y\Delta z}\left[\boldsymbol{E_z^n}\left(i,j+1,k+\frac{1}{2}\right)\Delta z-\boldsymbol{E_z^n}\left(i,j,k+\frac{1}{2}\right)\Delta z-\boldsymbol{E_y^n}\left(i,j+\frac{1}{2},k+1\right)\Delta y+\boldsymbol{E_y^n}\left(i,j+\frac{1}{2},k\right)\Delta y\right]$$

—— (2.13)

In Eq. (2.13), spatial derivatives are approximated by the central finite differences. $\boldsymbol{H_y^{n+\frac{1}{2}}}$ and $\boldsymbol{H_z^{n+\frac{1}{2}}}$ are updated in the same manner.

$$\boldsymbol{H_y^{n+\frac{1}{2}}}\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)=\boldsymbol{H_y^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)-\frac{\Delta t}{\mu\left(i+\frac{1}{2},j,k+\frac{1}{2}\right)}\frac{1}{\Delta z\Delta x}\left[\boldsymbol{E_x^n}\left(i+\frac{1}{2},j,k+1\right)\Delta x-\boldsymbol{E_x^n}\left(i+\frac{1}{2},j,k\right)\Delta x-\boldsymbol{E_z^n}\left(i+1,j,k+\frac{1}{2}\right)\Delta z+\boldsymbol{E_z^n}\left(i,j,k+\frac{1}{2}\right)\Delta z\right]$$

$$\boldsymbol{H_z^{n+\frac{1}{2}}}\left(i+\frac{1}{2},j+\frac{1}{2},k\right)=\boldsymbol{H_x^{n-\frac{1}{2}}}\left(i+\frac{1}{2},j+\frac{1}{2},k\right)-\frac{\Delta t}{\mu\left(i+\frac{1}{2},j+\frac{1}{2},k\right)}\frac{1}{\Delta x\Delta y}\left[\boldsymbol{E_y^n}\left(i+1,j+\frac{1}{2},k\right)\Delta y+\boldsymbol{E_y^n}\left(i,j+\frac{1}{2},k\right)\Delta y-\boldsymbol{E_x^n}\left(i+\frac{1}{2},j+1,k\right)\Delta x-\boldsymbol{E_x^n}\left(i+\frac{1}{2},j,k\right)\Delta x\right]$$

(2.14)

The transient electric and magnetic fields inside the defined system space can be simulated by updating $\boldsymbol{E_x^n}$, $\boldsymbol{E_y^n}$, $\boldsymbol{E_z^n}$, $\boldsymbol{H_x^{n+\frac{1}{2}}}$, $\boldsymbol{H_y^{n+\frac{1}{2}}}$, and $\boldsymbol{H_z^{n+\frac{1}{2}}}$ at ever point in the discretized cell. For the FDTD solution to be stable, the time increment Δt needs to be set to fulfill the Courant stability condition (Courant et al. 1928), given as follows:

$$\Delta t\leq\frac{1}{c\sqrt{\left(\frac{1}{\Delta x}\right)^2+\left(\frac{1}{\Delta y}\right)^2+\left(\frac{1}{\Delta z}\right)^2}}\quad\text{—— (2.13)}$$

Where c is the light speed.

2.1.3 Perfectly Matched Layer (PML)

One inconvenience in the FDTD simulation is that the Maxwell's equations have to be solved in a discretized domain space which the size is needed to be restrained. Nevertheless, the boundless theoretical space can be solved if the special boundary condition is applied to absorb the outgoing of the electromagnetic wave[30]. With the use of the perfectly matched layer (PML), the reflection factor of a plane wave striking a vacuum-layer interface is null at any frequency and incident angle. The general frame of PML is illustrated in figure 5. It is a widely used technique after invented and has addressed one of the big problem in the FDTD simulation.

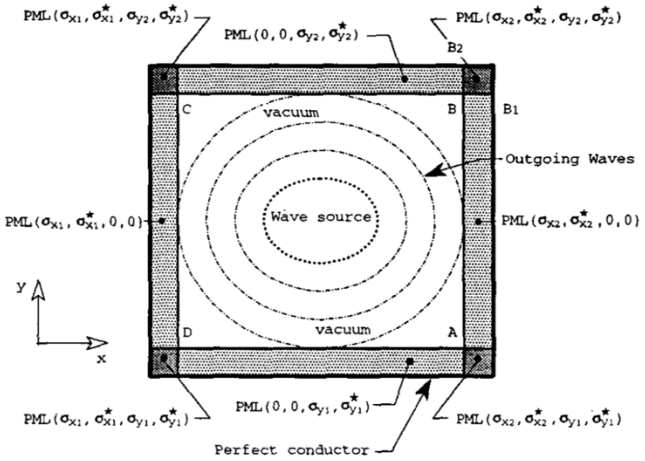

Figure 5. The PML technique[30]

Forward design is well established in many fields, but their solution space is restricted by human intuition, and requires computationally expensive trial-and-error process to obtain target results. As the complexity of such devices rise, so has the doubt that a library of traditional 'templates structures' and intuition will be able to meet the demand for high performance and highly tailored functionality[31]. To address the limitations of forward design,

"inverse design" methods were developed, in which a nanophotonic device structure can be generated through specification of a predefined target[31].

2.2 Inverse Design

Among various inverse design methodologies such as topology[32,33,34] and adjoint-based optimization[35,36] have been utilized to design complex structures which selectively interact with light on the nanometer-scale.

2.2.1 Topology Optimization

The topology optimization was developed for the mechanical and structural problem originally[37] and is a highly important weight optimization tool in industry for machine parts, cars and airplanes[35]. In inverse design method for nanophotonic devices, topology optimization can produce optimized geometry without any constraint on the geometric space for the design. The idea in topology optimization is discretizing the entire domain volume into pixels, each being a design variable that represents the material property. The total number of variables can thus be very large for a complex design task, and the structures are not restricted to any certain class of geometries[26].

An iterative topology optimization is based on repeated finite element analyses followed by gradient computation and updates by the deterministic mathematical optimization procedure. It first contains the numerical modeling, for example a numerical integration algorithm in time domain with defined boundary condition. It is followed by

the design parametrization, which the design is parametrized locally that the material properties at every spatial point is a design variable to be optimized. This facilitates the basic feature of the topology optimization which is the unlimited feature freedom in the domain volume. The iterative optimization process is then conducted by using different mathematical techniques.

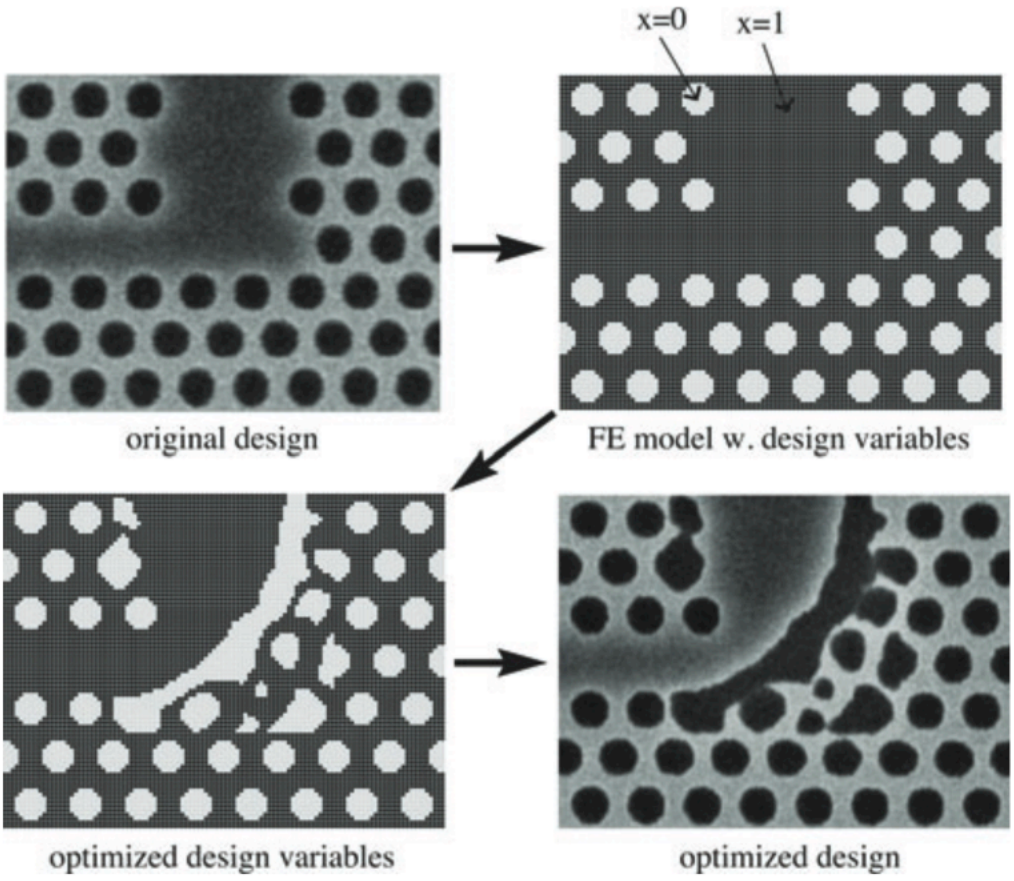

Figure 6. Parametrization of the topology design problem.[35]

Figure 6 is an example of topology optimization inverse design with a PHC Z-bend[38].

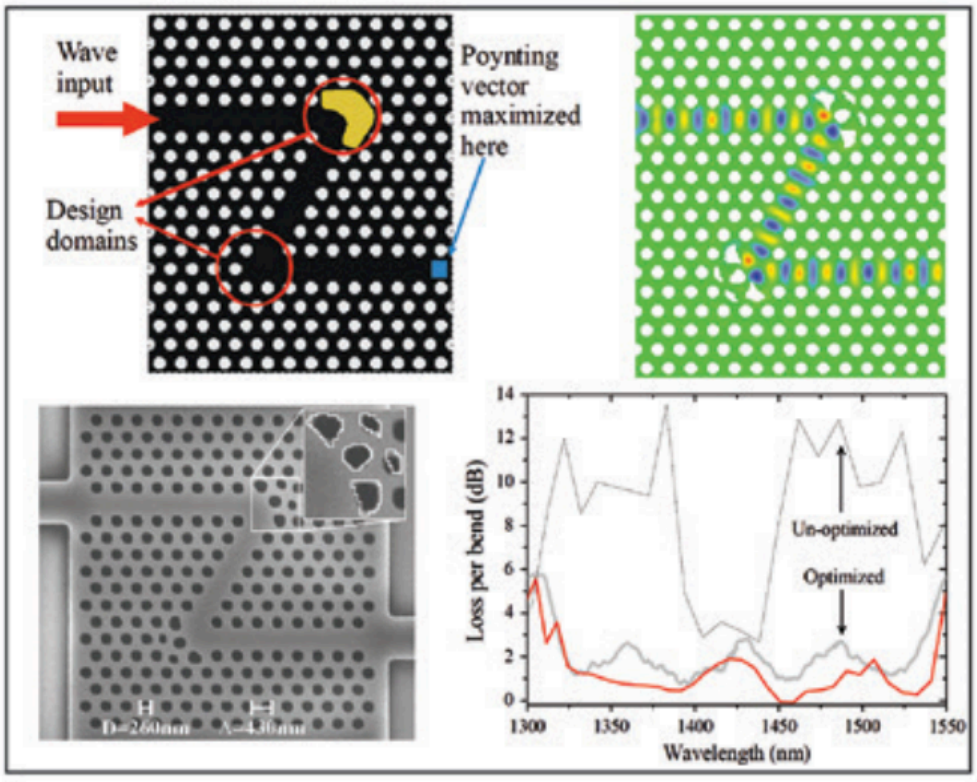

Figure 7. A photonic crystal waveguide Z-bend showing exceptional transmission. Panels in zigzag order: Schematic of optimization, simulated wave propagation, SEM image of the fabricated Z-bend, and comparison of bend losses between optimized (thick grey/red for measurement/ simulation) and unoptimized (thin black) structures.[38]

The sharp bend in a waveguide is notorious for causing significant bending loss and its poor transmission. In conventional optimization methods, the problem can only be solved by adjusting the hole size and the lattice distribution in the whole bending area if the method is not free from the geometric constraints. Topology optimization is shown to have a higher efficiency and lower holes to be adjusted during the optimization.

2.2.2 Adjoint-based Optimization

In recent years, some significant results have been made in designing nanophotonic devices by using adjoint-based optimization which allows the gradient of an objective

function to be computed with respect to an arbitrarily large number of degrees of freedom using only two simulations[39]. The adjoint variable method and the gradient optimization concept, however, are general and can be implemented with other simulation methods, such as the harmonic balance method[40, 41] for nonlinear circuit and laser simulations[42].

One example for the inverse design for electromagnetic design is by using adjoint method to calculate the shape derivatives at all points in domain space[43]. The method enables the process to compute only two electromagnetic simulations per iteration. In figure 8, the gradient of Figure-of-Merit can be computed with single simulation by following equation.

$$\frac{\Delta FoM}{\Delta \varepsilon_r} = Re\left[\left(\varepsilon_0 \Delta V \overline{\overline{G^{EP}}}(x, x_0)\overline{E^{old}(x_0)}\right) \cdot E^{old}(x)\right] \equiv Re[E^{adj}(x) \cdot E^{old}(x)] \quad -(2.14)$$

Where $E^{old}(x)$ is the value of the electric field at a given point before any change and $E^{adj}(x)$ is the new adjoint electric field. $E^{adj}(x)$ is defined as follow:

$$E^{adj}(x) = \left(\varepsilon_0 \Delta V \overline{\overline{G^{EP}}}(x, x_0)\overline{E^{old}(x_0)}\right) \quad -(2.15)$$

which is the electrical field induced at x from an electric dipole at x_0 driven with amplitude $\varepsilon_0 \Delta V \overline{E^{old}(x_0)}$, as illustrated in figure 8.

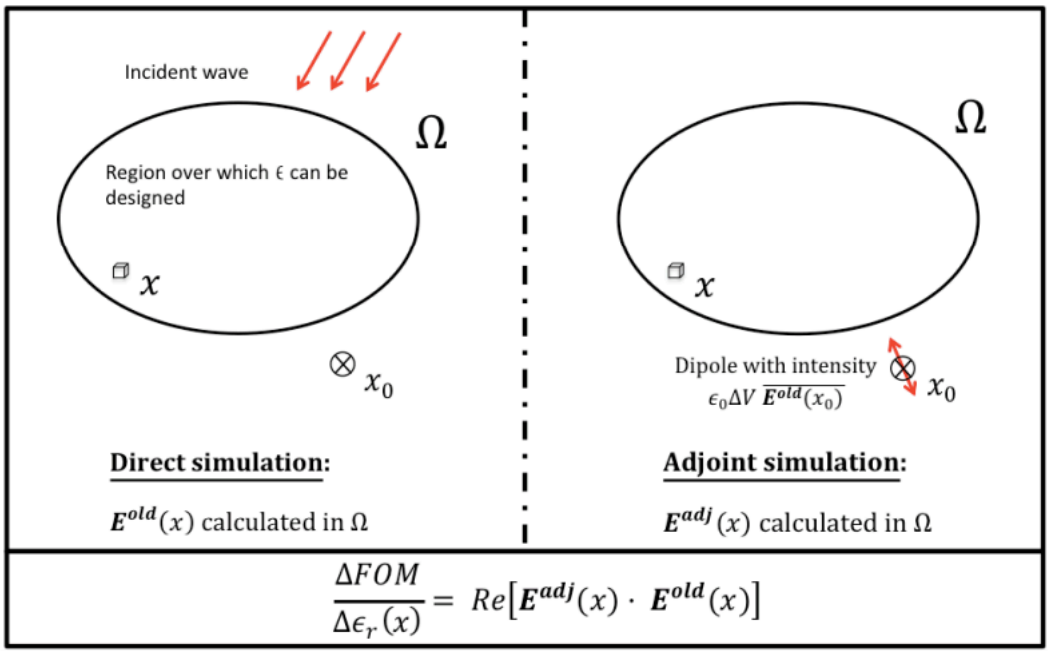

Figure 8. Adjoint method schematic: two simulations are needed for every iteration; the direct and the adjoint simulation. Sources for each simulation are drawn in red.[43]

Given the formula above, only one forward simulation which is used to calculate the Figure-of-Merit in all optimization schemes, and one adjoint simulation is needed to calculate the shape derivative over the entire design region, for arbitrarily degrees of freedom. As the gradient of the Figure-of-Merit is obtained, the change in geometry can be introduced proportional to the gradient. After iterations, this can lead to the optimum of the design.

Inverse design algorithms allow for the discovery of unintuitive and intricate structures, but can exhibit nonexistent or unstable results[44]. Furthermore, these methods are constrained by long runtimes, and can potentially miss globally optimal designs[26,31]. Additionally, the aforementioned optimization methods do not attempt to explain the underlying relationship

between the physical structure and its spectral response. As a result, the physical insights of nanophotonic device behavior remain a mystery with inverse design.

2.3 Introduction to Machine Learning Method

In recent years, machine learning (ML) techniques have emerged as promising strategies for the inverse design of nanoscale metamaterials, which overcome some of the shortcomings of previous inverse design paradigms. For example, tandem neural networks have been used to design multilayer thin films based on target transmission spectra[45]. For spatially-complex geometries, Generative Adversarial Networks (GANs) were developed to produce images of novel structures, given an input of desired spectral properties. GANs were used to design diffractive metagratings[46], sub-wavelength antenna[47], and two-dimensional metamaterials[48] with minimal computational time and cost. It has been demonstrated that a GAN is capable of generating devices with highly tailored spectral properties, even when the complexity of the device far exceeds the scope of human intuition.

The basic idea of the machine learning is given the dataset with input and output, make the machine to map the underlying relationship that cannot be easily understood by human beings between the input and output by optimizing the desired loss function. How the machine is 'learning' the relationship is by using the 'weight' of every neuron. The most basic model for machine learning problem is called: Multilayer Perceptron. It is composed of the input layer,

multilayer with arbitrary numbers of neurons in each layer, and followed by the output layer. Figure 9 is showing the structure of the multilayer perceptron schematically. In this example, there are three input features, two hidden layers with 5 and 4 neurons each layer and two output predictions.

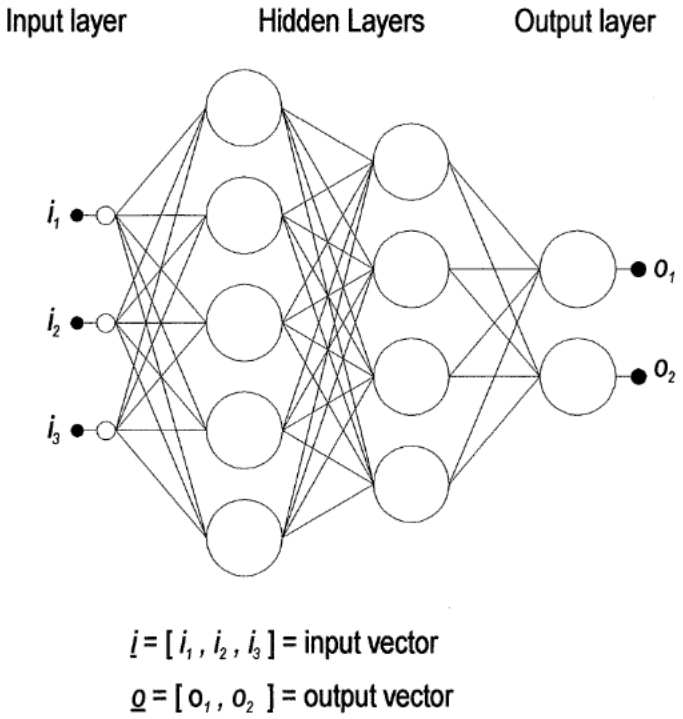

Figure 9. A multilayer perceptron with two hidden layers.[49]

Consider the network given in figure 9, how the node is calculated is by the following formula:

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(w^T x + b) \quad -(2.16)$$

Where w denotes the vector of weights, x is the vector of input (output from previous layer), b is the vector of bias and phi is the activation function.

There are several kinds of activation functions for the neural network, for example

sigmoid function, tanh function and also Rectified Liner Unit (ReLU) function has been widely used in neural network. After the output values have been computed from all the weights and bias in each node, it will be used in the loss function to calculate the 'total loss' for this node setup. Based on the loss function, the model will update their weights biases from the 'back propagation' rule. One training iteration is ended after the parameters are updated.

The multilayer perceptron has been applied in many different fields of study such as materials properties prediction, environmental carbon dioxide emission or even for biology purposes[50, 51]. However, this kind of machine learning model is not suitable for more complex materials design if "image recognition" technique is needed.

2.4 Convolutional Neural Network (CNN)

In the case of image recognition problem, the network will be fed with size-normalized images. There are problems with using traditional network on this kind of problem[52]. First, images are large with several hundreds of variables (pixels). A fully-connected layer would already have hundreds of neurons in the first layer, and thousands of weights in the whole network. However, the main deficiency for the unstructured nets is they have no built-in invariance with respect to translations or local distortions of the inputs.

In a CNN, some degree of shift, scale, and distortion invariance are allowed due to the architectural ideas: local receptive fields, shared weights, and spatial subsampling. Also, the

problem of a oversizing network is solved because of the shared weight properties.

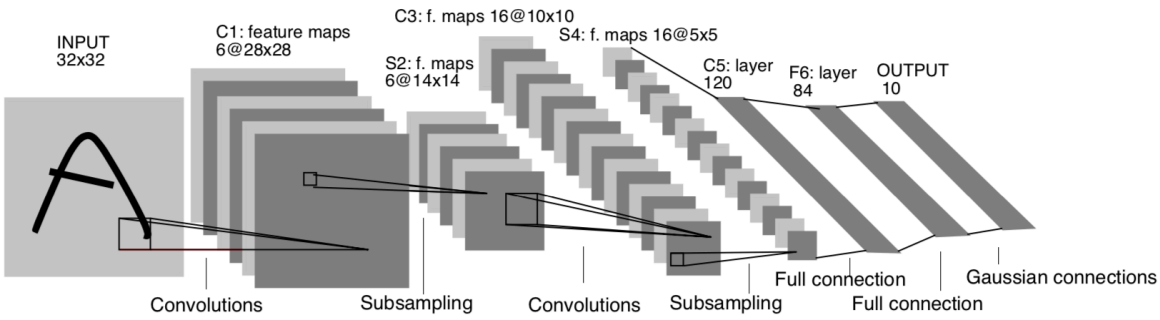

Figure 10. Architecture for a convolutional neural network (LeNet-5)[52].

A basic architecture of a CNN is showed in figure 10. There are three main kinds of layer inside the convolutional neural network: convolutional layers, pooling layers, fully connected layers[53].

2.4.1 Convolutional Layers

Convolutional layers work as feature extractors and they learn the feature representation of the images. Each neuron in the layers has a receptive field and the inputs are convolved with a learned weight to attain a new feature map. The convolved results are sent to a non-linear activation function afterward. Different feature maps in the same convolutional layers have the different weights so that several features can be captured at the same location. A formal notation of the output feature Y_k can be showed as:

$$Y_k = f(w_k * x) \quad -(3.2)$$

Where the input images is x, the weights of convolutional filter related to *k*th is w_k; the multiplication sign refers to 2D convolution operator, and f(·) is the non-linear activation function.

2.4.2 Pooling Layers

The purpose of pooling layers is to decrease the spatial resolution of the feature maps to reach the spatial invariance to input distortion. There are two common pooling layers: average pooling and max pooling. Average pooling layer process the input values in the small neighbor by averaging them and pass the value to the next layer, whereas the max pooling layer takes the maximum value within the receptive field and pass it to the next layer. The examples for two pooling layer are showed in figure 11.

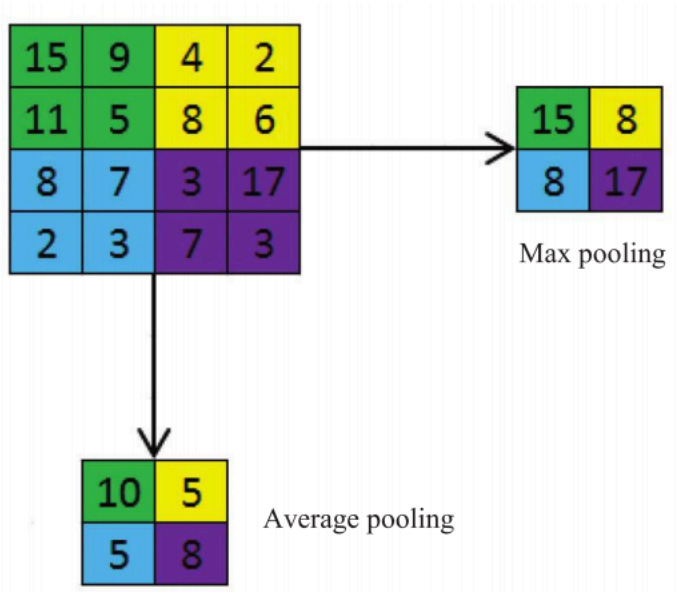

Figure 11. Average and Max pooling.

2.4.3 Fully Connected Layer

After processed by several layers of convolution and pooling layers, abstract feature representations are extracted into a stack of information. The followed up fully connected layers interpret these representations and it serves as the function of high-level reasoning [53]. Depends on the machine learning problem, different activation function can be used on

the fully connected layer. In our case, we are using regression layer.

Chapter 3

Convolutional-Neural-Network-Based Forward Design

3.1 Introduction

Convolutional Neural Networks (CNNs) are deep neural networks widely used for image analysis and classification[54]. We show that CNNs can be effectively used for forward design and, by explaining design features that are critical to the CNN's predictions, we can further our understanding of what the CNN has learned. With this newfound understanding, physical relationships can be extracted, and the same CNNs can be used to strategically transform designs and achieve new targets. As presented in figure 12a, our approach uses the Deep SHapley Additive exPlanations (SHAP) framework, which combines the DeepLIFT and SHAP methods[55], to explain a CNN's predictions by calculating 'contribution' values for each feature in the input image. We then leveraged these explanations to identify changes to the original structure that would produce a new target output spectrum. With the Deep SHAP explainer, we can obtain a heatmap plot for each wavelength, where the red pixels represent positive contributions towards the model's prediction, and the blue pixels represent negative

contributions. By utilizing explanations at specific wavelengths, we can make informed design decisions on how to obtain absorption peaks at desired resonance wavelengths.

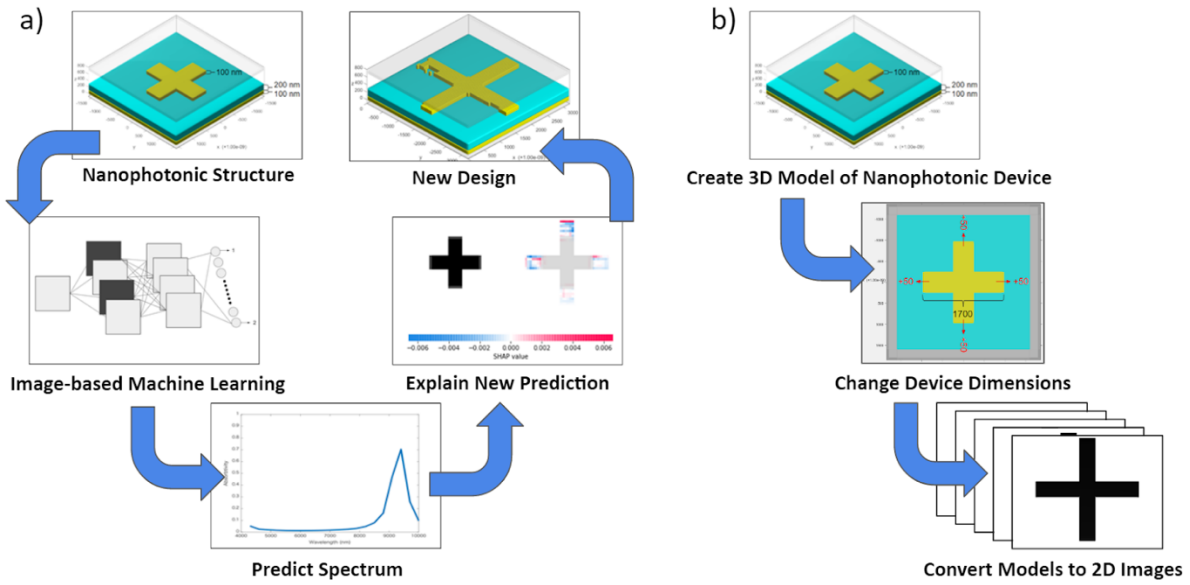

Figure 12. Utilizing a CNN's explanations for information extraction and design transformation.
(a) Schematic of the explanation and optimization approach to elucidating the underlying physics learned by a CNN. A CNN is trained on nanophotonic metal-insulator-metal structures and their corresponding absorption spectra in the mid-infrared regime. The relationships between structural features and absorption peaks are exposed with the SHAP algorithm, then used to construct new designs with new target resonant wavelengths. (b) The process of converting 3D device models into 2D representations for the training data set.

3.2 Training Data Generation

To train a CNN for the forward design of nanophotonic structures, we performed finite-difference time domain (FDTD) simulations of metal-insulator-metal metamaterials in Lumerical, operating at mid-infrared wavelengths, to generate images of 10,000 unique devices and their corresponding absorption spectra. Figure 12b illustrates the process of creating training data for the CNN. Full three-dimensional models of the mid-infrared resonators were

built in Lumerical, as shown in figure 13, and their dimensions were progressively adjusted for design variation. The models were then converted into two-dimensional images, and each of the 10,000 images were associated with an 800-point vector of absorption values (ranging from 0 to 1) across fixed wavelengths (4 μm to 12 μm). The simulated devices, previously demonstrated in literature to possess selective thermal emissivity over a large bandwidthta[56], consist of a 100 nm gold bottom layer, a 200 nm Al_2O_3 dielectric middle layer, and a 100 nm gold resonator top layer with various dimensions (within a 3.2 μm × 3.2μm unit cell). Periodic boundary conditions were applied along the x- and y-planes. Each image was resized to 40 × 40 pixels and converted to grey-scale for ease of training.

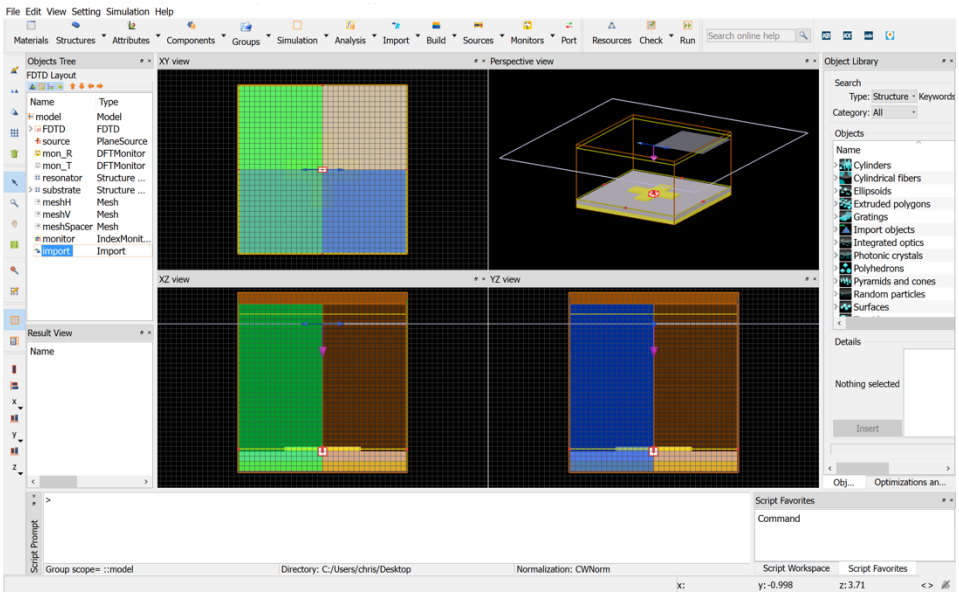

Figure 13. Simulation setup of Lumerical.

3.3 Training of CNN

After generating the training data, we trained multiple CNN architectures, with 10% of

the training dataset used for validation, to determine the optimum hyperparameters. Table 1 presents each of the trained models along with their validation root-mean-square error (RMSE) and training time. Model 1 served as the starting point, which consisted of three convolutional layer-stacks, each proceeding with a batch normalization layer, rectified linear unit (ReLU) activation layer, and average pooling layer (except the final stack). Each convolutional layer used 3 × 3 filters, numbering in 8, 16, and 32 in each subsequent layer. The pooling layer used 2 × 2 windows with a stride of 2. By testing incremental changes to the model (Model 2-8), we determined that a four-stack architecture with leaky ReLU layers trained with the adaptive moment estimation (Adam) algorithm yielded the lowest error (Model 9).

The CNN was implemented using TensorFlow and Keras, and trained on one Intel Core i5-8600T CPU for 300 epochs. In addition to the model information presented in Table 1, the CNN was trained with a learning rate of 0.001, beta1 of 0.9, beta2 of 0.999, and test dataset of 10%. The DeepExplainer module from the SHAP Python library was used to explain the predictions of the CNN. Image analysis and conversion was performed in Python.

Model 1			**Model 2**			**Model 3**		
Layers	Param.	Options	Layers	Param.	Options	Layers	Param.	Options
conv2d	3x3,8	sgdm	conv2d	3x3,16	sgdm	conv2d	3x3,16	sgdm
ReLU		256	ReLU		256	leakyReLU		256
avgPool	2x2, 2	minibatch	avgPool	2x2, 2	minibatch	avgPool	2x2, 2	minibatch
conv2d	3x3,16	100 epochs	conv2d	3x3,32	100 epochs	conv2d	3x3,32	100 epochs
ReLU			ReLU			leakyReLU		
avgPool	2x2, 2		avgPool	2x2, 2		avgPool	2x2, 2	

conv2d ReLU	3x3,32		conv2d ReLU	3x3,64		conv2d leakyReLU	3x3,64	
RMSE	0.15313		RMSE	0.10648		RMSE	0.11762	
Time	63 min		Time	167 min		Time	218 min	
Model 4			**Model 5**			**Model 6**		
Layers	Param.	Options	Layers	Param.	Options	Layers	Param.	Options
conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32 2x2, 2 3x3,64 2x2, 2 3x3,128	sgdm 256 minibatch 100 epochs	conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	adam 256 minibatch 100 epochs	conv2d ReLU maxPool conv2d ReLU maxPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 256 minibatch 100 epochs
RMSE	0.13289		RMSE	0.11497		RMSE	0.16737	
Time	87 min		Time	77 min		Time	58 min	
Model 7			**Model 8**			**Model 9**		
Layers	Param.	Options	Layers	Param.	Options	Layers	Param.	Options
conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 256 minibatch 300 epochs	conv2d ReLU avgPool conv2d ReLU avgPool conv2d ReLU	3x3,8 2x2, 2 3x3,16 2x2, 2 3x3,32	sgdm 128 minibatch 100 epochs	conv2d leakyReLU avgPool conv2d leakyReLU avgPool conv2d leakyReLU avgPool	3x3,16 2x2, 2 3x3,32 2x2, 2 3x3,64 2x2, 2	adam 128 minibatch 300 epochs

						conv2d leakyReLU	3x3,128	
RMSE	0.097562		RMSE	0.14086		RMSE	0.07709	
Time	229 min		Time	42 min		Time	340 min	

Table 1 CNN hyperparameter optimization. Table of trained CNN architectures and corresponding RMSE values.

Figure 14 illustrates the predictions of the CNN when six new and unknown images were used as inputs. On average, each prediction was generated in 0.270 ± 0.043 seconds ($n = 10$), while each simulation took approximately 15 minutes. FDTD simulations were performed on the new images, after converting them into the top layer of the metal-insulator-metal structure. The simulated absorption spectra were then compared to the CNN predictions, as shown in figure 14, where the CNN predictions exhibit a high degree of accuracy in comparison to their corresponding ground truths. Although there are minor perturbations at the regions of the spectra which approach 0 absorption, the wavelength and amplitude of the predicted resonance peaks are aligned with the simulated peaks (with over 95% accuracy). The results here demonstrate that the CNN can successfully perform forward design-based tasks with high accuracy and with orders of magnitude faster than conventional numerical simulation.

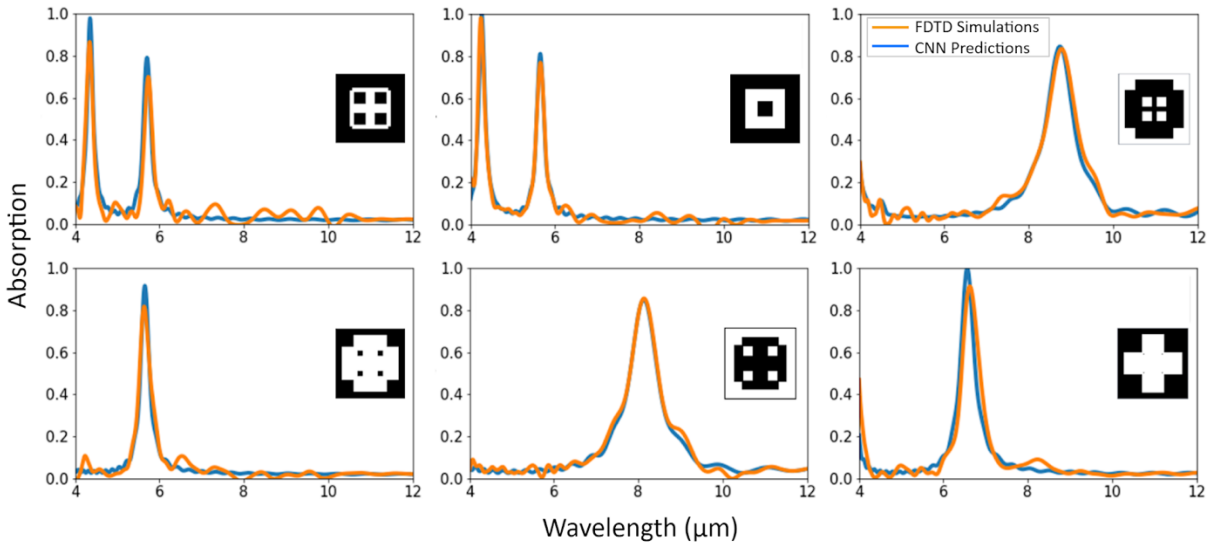

Figure 14. Simulating nanophotonic material response with CNNs. CNN-predicted absorption spectra *vs.* ground truth simulations of six new nanophotonic structures (shown in the inset images).

Chapter 4

Inverse Design by Shapley Additive Explanation (SHAP)

4.1 Introduction

Despite recent interest in ML-based inverse design, ML-based forward design remains largely unexplored. ML-based forward design offers the same advantages as the inverse design counterpart, and can similarly calculate light-matter interactions with orders of magnitude less computation complexity and time than numerical simulations. However, regardless of the ML-design direction, the internal decision models built by neural networks are not well-understood; their contents are widely regarded as a 'black box'[57,58], and thus alike traditional inverse optimization methods. This 'black box' challenge emerges from the fact that neural networks, and supervised ML algorithms in general, 'learn' by optimizing hundreds of thousands to millions of internal variables (*e.g.* weights and biases) that fit the training data[59]. Consequently, it is challenging to explain *why* a neural network makes one prediction over another. This lack of explainability is a key limitation in both traditional inverse design and ML-based methods.

To this end, we aim to uncover what a neural network has learned regarding the underlying physical principles which govern specific nanophotonic structures and their properties.

In this work, we leverage recent advances in explainable ML to demonstrate that the data-driven relationships learned by a neural network can be revealed to derive physical insights between material structure and optical response. We present a forward design and explanation methodology for nanophotonic devices which combines convolutional neural networks (CNNs) and explainability algorithms, to design nanophotonic structures that can meet a target spectrum, while elucidating the underlying physics of design features that contribute to specific electromagnetic behavior. The presented strategy demystifies some of the relationships that enable accurate deep-learning model predictions, while unveiling the potential limits of the model itself.

4.2 CNN Explainability with Shapley Additive Explainations

The high accuracy of the CNN's predictions indicate that the network has, to some extent, learned the physical relationships between the class of nanophotonic structures we explored and its absorption spectra. Normally, this information is embedded within thousands of internal weights and parameters. To draw useful conclusions and design principles from the network's internal model, we utilized SHAP. This empirical method calculates a 'SHAP value' for each pixel that represents how important that pixel is to the CNN's overall prediction. With the

SHAP values, we can explain a nanophotonic device feature's contribution to various resonant wavelengths. SHAP values are calculated through the following equation:

$$\emptyset_i(f,x) = \sum_{z\prime \subseteq x\prime} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z'\backslash i)] \quad -(4.1)$$

Where $\emptyset_i$ is the SHAP value, x' are simplified inputs that mapped binary values into the original input space (x), M is the number of simplified input features, z' is a subset of non-zero indices in x', $f_x(z')$ is a model trained with the feature present, and $f_x(z'^{\backslash}i)$ is a model trained with the feature withheld[55]. The SHAP model captures the effect of withholding a feature, then iterates the computation across all possible subsets ($z' \subseteq x'$).

We validated this approach by comparing the SHAP explanations against a standard antenna-based analytical relationship between the length of resonant arms in the metal-dielectric-metal metamaterials we developed and the resonant wavelengths:

$$\lambda = (2n_{eff})\, L+C \quad -(4.2)$$

Here, λ is the resonant wavelength, n_{eff} is the effective index of the transverse electric (TE) mode, L is the length of the resonator, and C is a correction phase term[60,61,62]. To explain this relationship explicitly, we performed SHAP explanations on the CNN model trained on 10,000 images. As shown in figure 15a, SHAP explanation heatmaps were captured at 6.0, 6.4, 6.8, 7.2, 7.6, 8.0, 8.4, and 8.8 μm with single-reference backgrounds (described in the Methods section), while the base image possessed a Lorentzian peak absorption at 5.2 μm and arm lengths of 1.4 μm. These explanations reveal the features, or lack thereof, that the CNN deems

critical towards achieving resonance at the designated wavelengths. Specifically, as the resonant wavelength increases, the explanations show regions of blue (negative contributions) which gradually migrate from the center of the image to the edges, indicating that starting from the base image, the antenna arm lengths must become longer in order to achieve resonance at larger wavelengths. Inversely, Fig. 15b shows that for a base image with longer initial antenna arm lengths (2.9 μm), the antenna arms must become shorter in order to achieve resonance at smaller wavelengths. This behavior is evident from the regions of blue pixels converging towards the center of the image as the resonant wavelength decreases. Both cross-arm tests suggest that the CNN has deduced the relationship between antenna length and resonant wavelength defined in Eqn. (4.2). With SHAP, we were able to analytically confirm this deduction and infer the CNN's decision process for a class of nanophotonic structures.

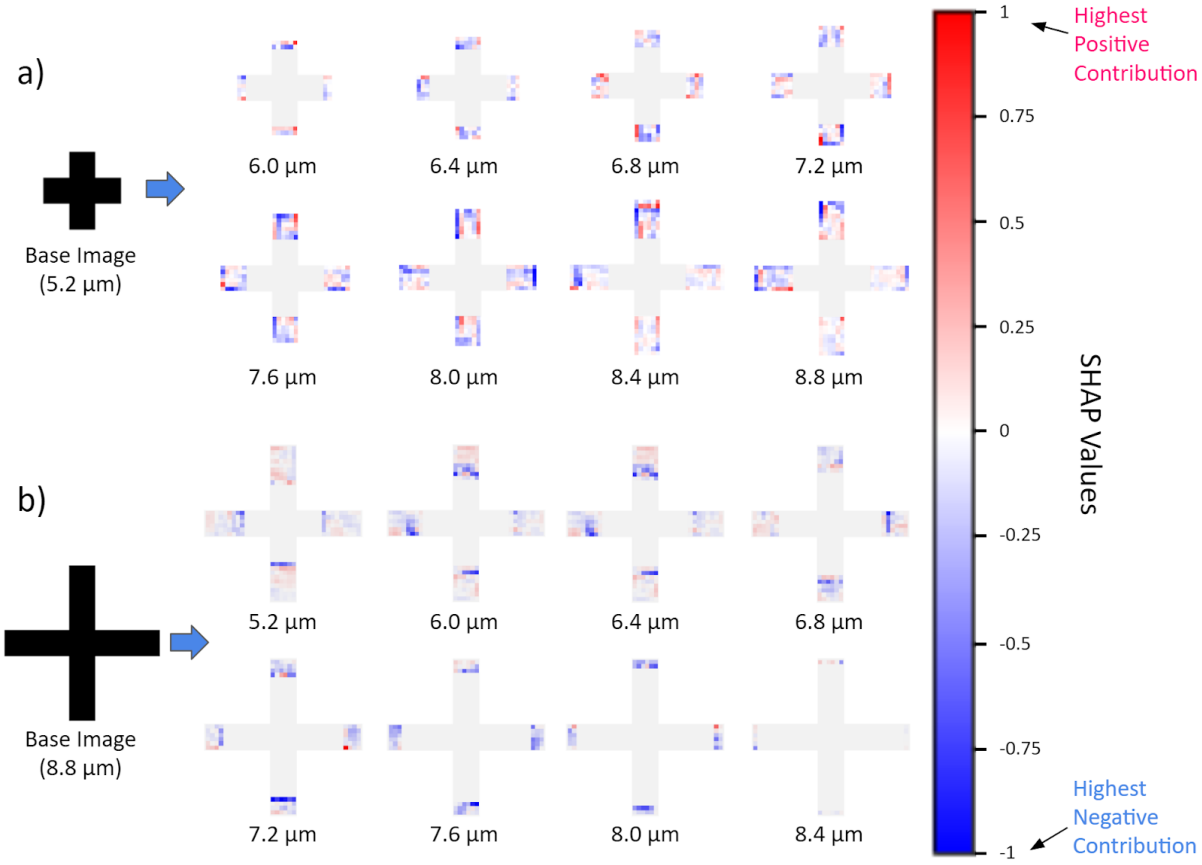

Figure 15. SHAP explanation heatmaps. SHAP explanations for a (a) 'short-arm' cross (1.4 µm lengths) at increasing resonant wavelengths and a (b) 'long-arm' cross (2.9 µm lengths) at decreasing resonant wavelengths, revealing the CNN learned that the cross-arms must increase to achieve resonance at longer wavelengths and vice versa.

However, we observe varying degrees of red and blue pixels throughout the explanation heatmaps. For example, on the 8.8 µm explanation with the 5.2 µm base image, there are higher-intensity blue pixels on the top and left arms of the cross, indicating that the CNN weighs each arm differently in determining the resonant wavelength, when in reality, all of the arms are equally important to achieving resonance at the designated wavelength. In addition, the magnitude of the blue pixels are greater towards the edges of the structure, while the remaining areas have red pixels scattered throughout. Both results can be attributed to the filters developed by the CNN during training, which dictates the features that contribute the least or

the most to a resonant wavelength. CNNs extract information from images by applying a hierarchy of filters to the input image[63]. The filters are optimized such that the error is minimized when comparing the CNN's output to the target output. CNNs have a tendency to develop edge detection filters, since non-edge patterns (*e.g.*, a patch of black pixels) do not typically provide sufficient information to differentiate discrete objects[54]. Therefore, our CNN was tasked with creating the minimum set of filters that captures the most important features and distinctions (*i.e.*, the cross-arm edges) required to correlate the images to their respective absorption spectra. Naturally, this determines the range of the CNN's feature recognition capabilities and the extent of which it can generalize (or accurately predict new and unknown images), which may be limited to an unknown degree. However, we can address the model uncertainty by using the SHAP explanations to observe the most prominent sections of the structure contributing to resonance as well as the sections that were disregarded; two pieces of information that provide insight on what the CNN did and did not learn, respectively. Thus, in addition to uncovering the predominant physical relationships learned by the CNN, the presented CNN-explanation approach is also effective at determining the limitations and risks associated with a trained ML model by enabling users to further understand how the CNN behaves.

4.3 Using SHAP Explanations for Targeted Design Transformation

To confirm the physical relationship learned by the CNN, and to assess the possibility of using the SHAP explanations for designing nanophotonic devices with targeted functionalities, we used the SHAP value heatmaps from the previous section to transform the base image such that it met the previously identified resonance wavelengths. These transformed designs were then compared with the corresponding FDTD simulated background images (as shown in figure 16a) to ensure that the CNN learned the relationship between cross-arm length and resonance wavelength. Transformation was performed by converting all the blue pixels (and the first 1% of pixels greater than 0 to account for noise), in the images shown in figure 15a, to black pixels on the base image. Figure 16b and 16c show the spectra of the transformed structures and the original FDTD simulated structures, respectively. On figure 16d, the resonant wavelengths at peak absorption and antenna arm lengths of both sets of structures are plotted (with linear fits of R=0.998). Here, the FDTD structures possess an n_{eff} of 1.13 and C of 2.21. Similarly, the transformed structures display an n_{eff} of 1.15 and C of 2.10, yielding an n_{eff} error of 1.8% and a C error of 4.9%. The comparison between the SHAP-generated and the FDTD simulated structures demonstrates that the information extracted by the CNN aligns strongly with the physical relationship established in Eqn. (4.2), and that the extracted data can be effectively utilized to make reasonably accurate targeted design transformations.

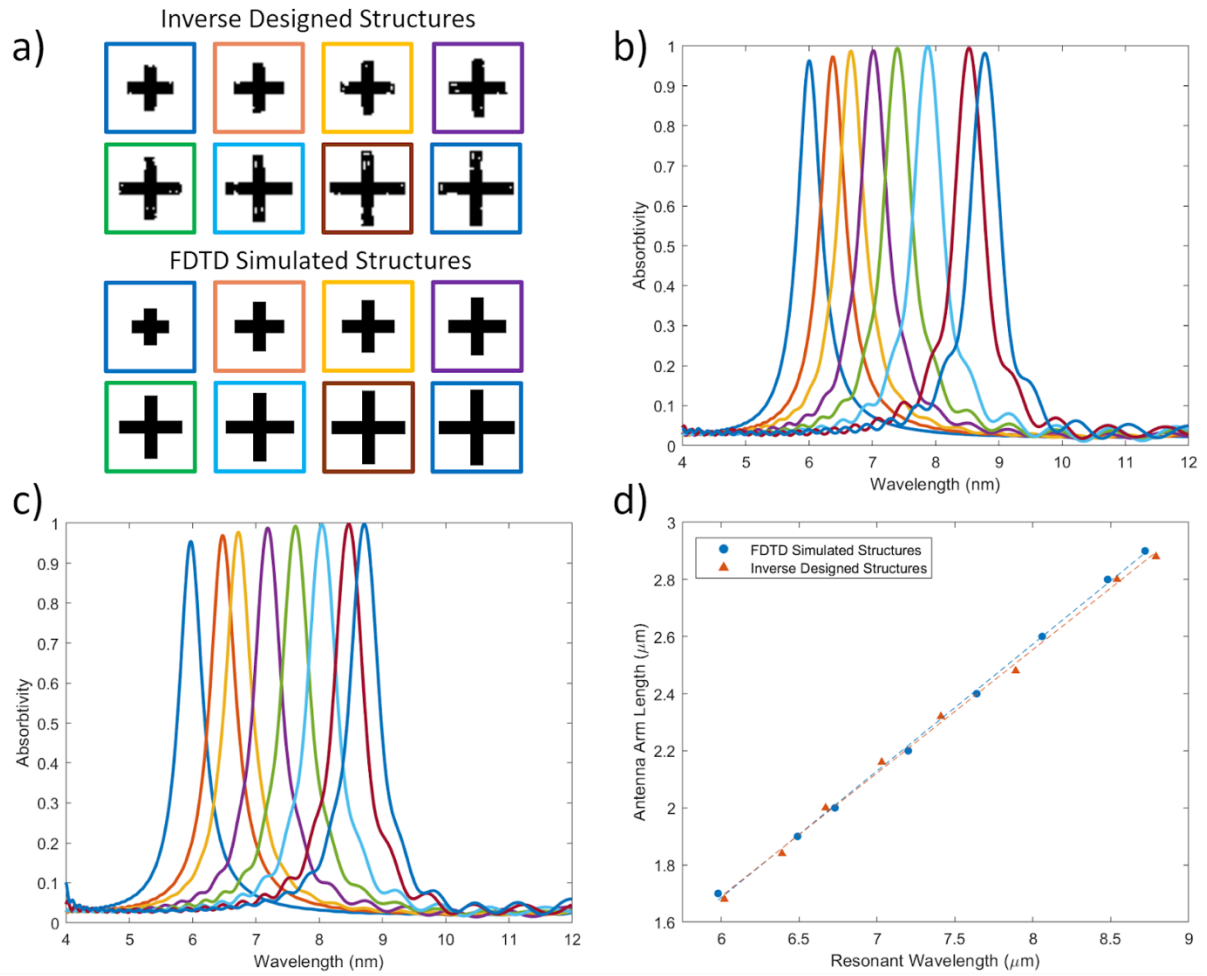

Figure 16. Inverse design with a forward-trained CNN and SHAP. (a) Images of the inverse designed structures and FDTD simulated structures. The absorption spectra for the corresponding (b) inverse-designed and (c) simulated structures. Image colors correspond to the plot colors. (d) Comparison of the physical relationship between antenna arm length and resonant wavelength for the two sets of structures. The antenna arm lengths in the inverse designed structures exhibited resonant wavelengths with an nefferror of 1.8% and a C error of 4.9%, in comparison to the ground truths (linear fit of plots shown with R^2=0.998).

4.4 Deriving Physical Insights by Explaining Complex Spectral Patterns

To demonstrate the applicability of our explanation and design transformation method to increasingly complex structures and target spectra, we performed additional tests on the same CNN. Figure 17 presents a series of test cases, where explanations of a dual absorption peak structure (L-shaped) and a single-peak structure (I-shaped) were captured at the peak wavelengths of each structure (marked in figure 17a). The SHAP explanation heatmaps at the designated wavelengths are shown in figure 17b, where the I-shaped image was used as the background for the L-shaped image and vice versa. The complete distribution of SHAP values from each heatmap are plotted and quantified in figure 17c and 17d for the I-shaped resonator and L-shaped resonator, respectively. The inset bar graphs present the average SHAP values across each explanation. From these plots, we observe that at the peak/target wavelengths of the background image, the explanation of the base image at those wavelengths (indicated by the red-dashed boxes in figure 17b) yield higher-magnitude and more negative SHAP values (blue pixels) than the explanations at non-peak wavelengths. Thus, the results here reveal that the CNN learned the relationship between the two structures. Specifically, the inclusion of the horizontal-bar on the I-shaped structure renders two absorption peaks at 5.5 μm and 8.2 μm while the removal of the bar on the L-shaped structure renders a single peak at 6.3 μm.

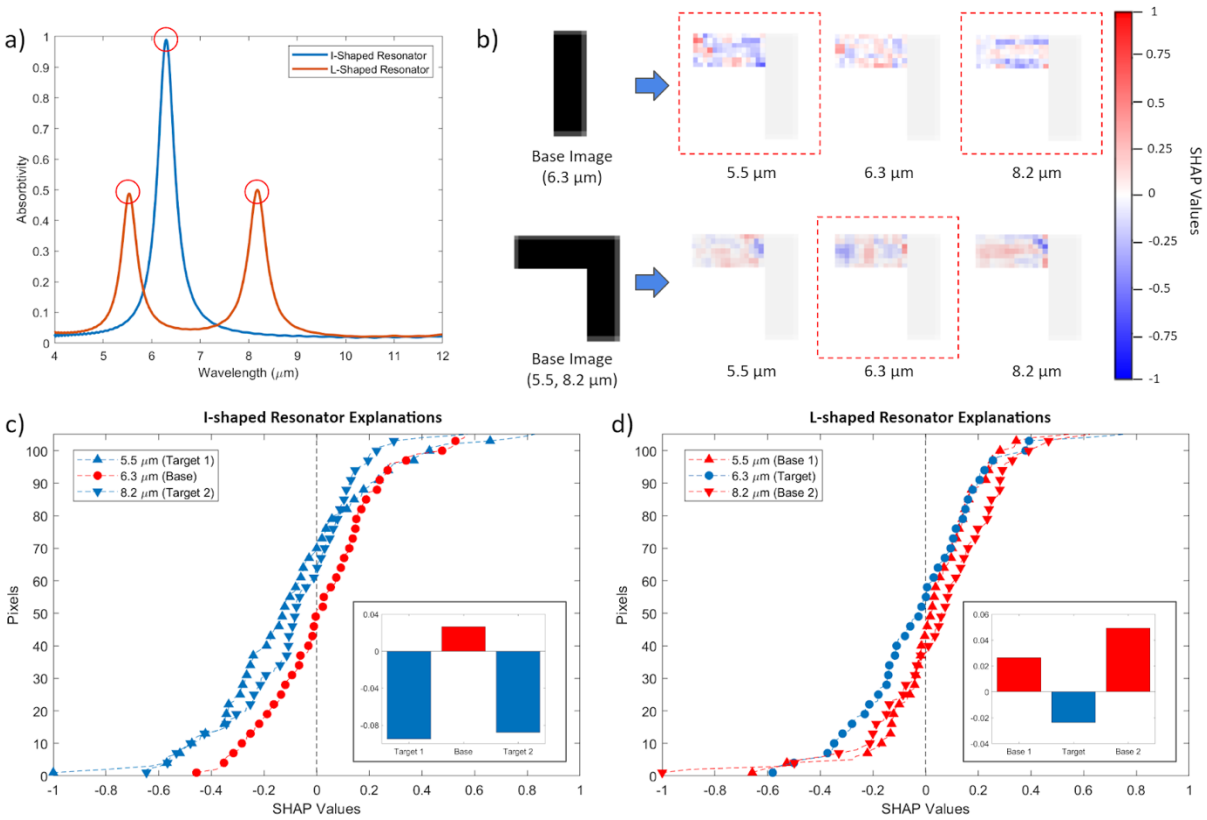

Figure 17. Explanations of a dual-peak structure and a single-peak structure at various wavelengths. (a) Absorption spectra of a single-peak I-shaped resonator and a dual-peak L-shaped resonator. Red circles indicate the resonant wavelengths. (b) SHAP explanations of the resonators at the previously identified resonant wavelengths. Red dashed boxes indicate the explanations for obtaining new target resonant wavelengths of the opposing shape. Distribution of SHAP values across the explanation pixel-maps for the (c) I-shaped resonator and the (d) L-shaped resonator. Inset bar graphs represent the average SHAP values of each explanation, where the negative SHAP values (blue pixels) are dominant on all target explanations.

Moreover, the explanations are able to inform different areas of the structure which contribute to different resonances. For example, for the dual-peak L-shaped structure, the explanation at each peak (5.5 μm and 8.2 μm) illustrates different red-pixel dominant regions (features contributing to resonance at these wavelengths). This phenomenon bears resemblance to the fact that the electric field concentrations in these structures are not uniform at different resonant wavelengths (shown in figure 18). Similar to the electric field of the L-shaped

structure at 8.2 µm, SHAP informs us that roughly the entire horizontal-arm length evenly contributes to resonance, while at 5.2 µm, the center of the arm contributes to an unevenly distributed electric field pattern that is responsible for resonance. However, the precise connections between the SHAP explanations and the electric fields are unclear and requires further investigation.

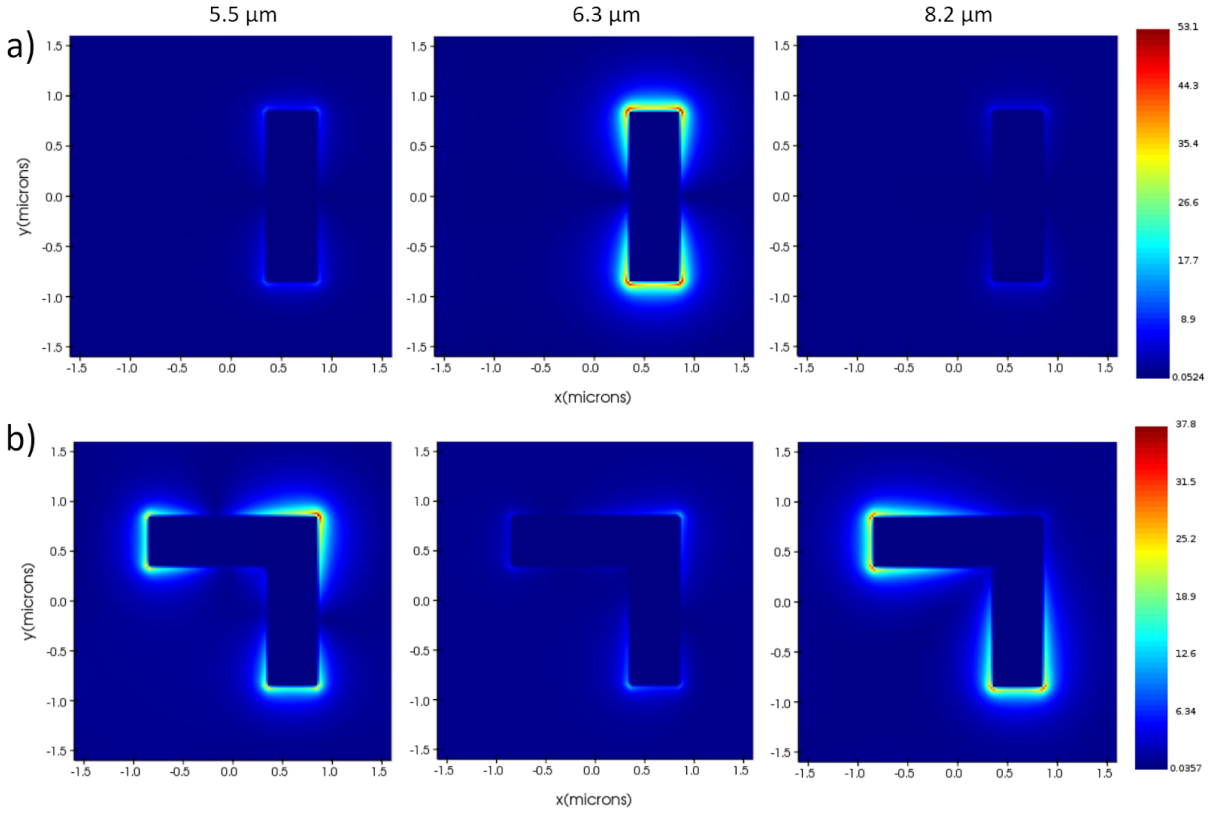

Figure 18. Electric field simulations of MIM resonators at various resonant wavelengths. The electric field profile of (a) an L-shaped resonator and (b) an I-shaped resonator at 5.5 µm, 6.3 µm, and 8.2 µm.

Following the analysis of the SHAP explanations for the dual and single-peak structures, we performed the same design transformation study from the previous section to assess its applicability to complex, multi-peak transformations. In figure 19a, the L-shaped device was transformed by utilizing the explanation generated at 6.3 µm, then all of the blue pixels to the

opposite state on the original image. The resulting structure exhibited a single absorption peak of approximately 0.9 at 5.4 μm. Using the same approach, we attempted the reverse scenario of generating a dual-peak structure from a single-peak structure (figure 19b). We leveraged the explanation from one of the dual-peak wavelengths (as either wavelength provided no noticeable difference) and applied it to the design transformation process. The transformed structure possessed an absorption peak of approximately 0.6 at 4.8 μm and 0.48 at 6.9 μm.

Through our design transformation studies, we demonstrate that complex spectral targets can be met by converting the pixels identified by the SHAP heatmaps. In the first case, by focusing the image conversion process on the explanations of a single target wavelength, we converted a dual-peak structure into a single-peak structure. In the second case, the single-peak structure was converted into a dual-peak structure by using the SHAP values of two target wavelengths. However, the resonant wavelengths of the transformed structures deviate from the target wavelengths by approximately 8.9% (relative to the evaluated wavelength range), indicating that there are limitations associated with the precision of the presented design transformation approach. Such limitations may be attributed to one of the most prominent shortcomings of SHAP: its inability to account for feature dependence[64, 65], which may have inhibited the identification of key structural features required for resonance. Despite the discrepancy between the target and resulting resonant wavelengths, the general patterns identified by the transformed designed structures still offer significant insights into the critical

features which contribute to resonance; a crucial element which was unobtainable in previous ML studies pertaining to nanophotonic devices.

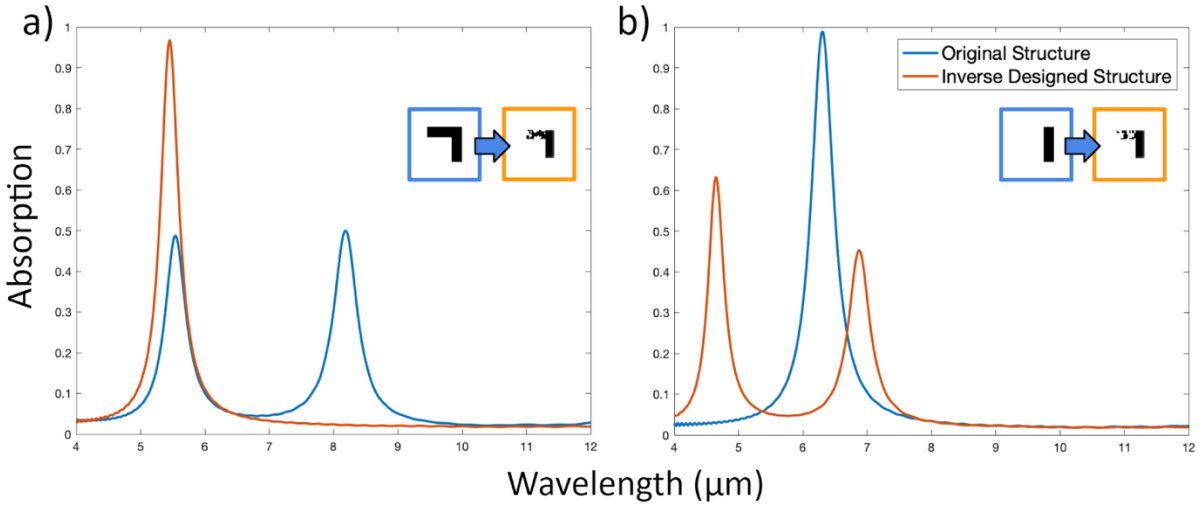

Figure 19: Targeted design transformation for complex spectral responses. Transformation of a (a) dual-peak structure to a single-peak structure and a (b) single-peak structure to a dual-peak structure by utilizing the SHAP values at targeted resonant wavelengths for image conversion.

Chapter 5

Conclusion and Future Work

Artificial neural networks can predict the optical and thermal properties of nanophotonic structures with astounding precision. However, neural networks are traditionally classified as 'black boxes'; the contents of which are esoteric or even incomprehensible. Thus, the general understanding of the physical relationships learned by the ML model are severely limited. To address this shortcoming, we present a visual, scalable, and universal approach to unraveling the mysteries hidden within a ML model trained on the electromagnetic response of infrared metamaterials. The presented method leverages an explanation algorithm (Shapley Additive Explanations, or SHAP) to rank the contributions of individual features on an image towards each of the network's predictions. To demonstrate model explainability, first, we trained a CNN on 10,000 images of nanophotonic structures and their absorption spectra. The trained CNN predicted the spectra of new and unknown structures with over 95% accuracy, and orders of magnitude faster (~0.3 seconds) than conventional simulation (~15 minutes). Then, by generating SHAP explanations at designated wavelengths, we determined the structural components (or lack thereof) of a base design which contributed positively or negatively towards resonant behavior at these wavelengths. We further demonstrated that by examining the SHAP explanations, both qualitative and quantitative relationships between structure and

spectra can be obtained (*i.e.*, resonator arm length vs resonant wavelength), and that the explanations themselves can be used to make targeted design transformations. Additionally, the explanations also revealed what the CNN did *not* learn, thus exposing potential limitations and risks associated with the trained model. As a result, the presented explanation and design transformation method shows that the patterns and principles encoded within the ML model can be extracted to derive valuable insights into nanophotonic devices physics, thereby enabling potentially new discoveries in the understanding of electromagnetic wave-matter interactions and other extendable applications. Future studies will encompass alternative explanation techniques and the explanation of additional device-property relationships.

Chapter 6

References

1. Chen, R. *et al.* Nanophotonic integrated circuits from nanoresonators grown on silicon. *Nat. Commun.* **5**, 1–10 (2014).

2. Pelton, M. Modified spontaneous emission in nanophotonic structures. *Nature Photonics* vol. 9 427–435 (2015).

3. Odebo Länk, N., Verre, R., Johansson, P. & Käll, M. Large-Scale Silicon Nanophotonic Metasurfaces with Polarization Independent Near-Perfect Absorption. *Nano Lett.* **17**, 3054–3060 (2017).

4. Pralle, M. U. *et al.* Photonic crystal enhanced narrow-band infrared emitters. *Appl. Phys. Lett.* **81**, 4685–4687 (2003).

5. Lin, S. Y., Moreno, J. & Fleming, J. G. Three-dimensional photonic-crystal emitter for thermal photovoltaic power generation. *Appl. Phys. Lett.* **83**, 380–382 (2003).

6. Laroche, M., Carminati, R. & Greffet, J. J. Coherent thermal antenna using a photonic crystal slab. *Phys. Rev. Lett.* **96**, 123903 (2006).

7. Koenderink, A. F., Alù, A. & Polman, A. Nanophotonics: Shrinking light-based technology. *Science* vol. 348 516–521 (2015).

8. Tanaka, K., Tanaka, M. & Sugiyama, T. Simulation of practical nanometric optical circuits based on surface plasmon polariton gap waveguides. *Opt. Express* **13**, 256 (2005).

9. Takahara, J., Yamagishi, S., Taki, H., Morimoto, A. & Kobayashi, T. Guiding of a one-dimensional optical beam with nanometer diameter. *Opt. Lett.* **22**, 475 (1997).

10. Verhagen, E., Dionne, J. A., Kuipers, L., Atwater, H. A. & Polman, A. Near-field visualization of strongly confined surface plasmon polaritons in metal-insulator-metal waveguides. *Nano Lett.* **8**, 2925–2929 (2008).

11. Dionne, J. A., Lezec, H. J. & Atwater, H. A. Highly confined photon transport in subwavelength metallic slot waveguides. *Nano Lett.* **6**, 1928–1932 (2006).

12. Miyazaki, H. T. & Kurokawa, Y. Squeezing visible light waves into a 3-nm-thick and 55-nm-long plasmon cavity. *Phys. Rev. Lett.* **96**, 097401 (2006).

13. Maier, S. A. Effective mode volume of nanoscale plasmon cavities. *Opt. Quantum Electron.* **38**, 257–267 (2006).

14. Dmitriev, A., Pakizeh, T., Käll, M. & Sutherland, D. S. Gold-sitica-gold nanosandwiches: Tunable bimodal plasmonic resonators. *Small* **3**, 294–299 (2007).

15. Oubre, C. & Nordlander, P. Optical properties of metallodielectric nanostructures calculated using the finite difference time domain method. *J. Phys. Chem. B* **108**, 17740–17747 (2004).

16. Foteinopoulou, S., Vigneron, J. P. & Vandenbem, C. Optical near-field excitations on plasmonic nanoparticle-based structures. *Opt. Express* **15**, 4253 (2007).

17. Hao, F. & Nordlander, P. Efficient dielectric function for FDTD simulation of the optical properties of silver and gold nanoparticles. *Chem. Phys. Lett.* **446**, 115–118 (2007).

18. Bhargava, S. & Yablonovitch, E. Inverse design of optical antennas for sub-wavelength energy delivery. in *Conference on Lasers and Electro-Optics Europe - Technical Digest* vols 2014-January CM2F.2 (Institute of Electrical and Electronics Engineers Inc., 2014).

19. Tsuji, Y., Hirayama, K., Nomura, T., Sato, K. & Nishiwaki, S. Design of optical circuit devices based on topology optimization. *IEEE Photonics Technol. Lett.* **18**, 850–852 (2006).

20. Frandsen, L. H. *et al.* Broadband photonic crystal waveguide 60□ bend obtained utilizing topology optimization. *Opt. Express* **12**, 5916 (2004).

21. Elesin, Y., Lazarov, B. S., Jensen, J. S. & Sigmund, O. Time domain topology optimization of 3D nanophotonic devices. *Photonics Nanostructures - Fundam. Appl.* **12**, 23–33 (2014).

22. Zhang, T. *et al.* Efficient spectrum prediction and inverse design for plasmonic waveguide systems based on artificial neural networks. *Photonics Res.* **7**, 368 (2019).

23. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science (80-.).* **355**, 602–606 (2017).

24. Peurifoy, J. *et al.* Nanophotonic particle simulation and inverse design using artificial neural networks. *Sci. Adv.* **4**, eaar4206 (2018).

25. Kojima, K., Wang, B., Kamilov, U., Koike-Akino, T. & Parsons, K. Acceleration of FDTD-based inverse design using a neural network approach. in *Optics InfoBase Conference Papers* vol. Part F52-IPRSN 2017 ITu1A.4 (OSA - The Optical Society, 2017).

26. Yao, K., Unni, R. & Zheng, Y. Intelligent nanophotonics: Merging photonics and artificial intelligence at the nanoscale. *Nanophotonics* vol. 8 339–366 (2019).

27. A. Taflove, S. C. H. Computational electrodynamics: the finite-difference time-method. *Sencond Ed. Artech house Publ. Norwood.* **ISBN**, 1006 (2000).

28. Pflaum, C. & Rahimi, Z. A finite difference frequency domain (FDFD) method for materials with negative permittivity. in *Proceedings of the 2009 International Conference on Electromagnetics in Advanced Applications, ICEAA '09* 799–802 (2009). doi:10.1109/ICEAA.2009.5297314.

29. Yee, K. S. Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media. *IEEE Transactions on Antennas and Propagation* vol. 14 302–307 (1966).

30. Berenger, J. P. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.* **114**, 185–200 (1994).

31. Molesky, S. *et al.* Inverse design in nanophotonics. *Nature Photonics* vol. 12 659–670 (2018).

32. Lin, Z., Groever, B., Capasso, F., Rodriguez, A. W. & Lončar, M. Topology-Optimized Multilayered Metaoptics. *Phys. Rev. Appl.* **9**, 044030 (2018).

33. Sell, D., Yang, J., Doshay, S. & Fan, J. A. Periodic Dielectric Metasurfaces with High-Efficiency, Multiwavelength Functionalities. *Adv. Opt. Mater.* **5**, 1700645 (2017).

34. Yang, J. & Fan, J. A. Topology-optimized metasurfaces: impact of initial geometric layout. *Opt. Lett.* **42**, 3161 (2017).

35. Jensen, J. S. & Sigmund, O. Topology optimization for nano-photonics. *Laser and Photonics Reviews* vol. 5 308–321 (2011).

36. Xiao, T. P. *et al.* Diffractive Spectral-Splitting Optical Element Designed by Adjoint-Based Electromagnetic Optimization and Fabricated by Femtosecond 3D Direct Laser Writing. *ACS Photonics* **3**, 886–894 (2016).

37. Bendsøe, M. P. & Kikuchi, N. Generating optimal topologies in structural design using a homogenization method. *Comput. Methods Appl. Mech. Eng.* **71**, 197–224 (1988).

38. Borel, P. I. *et al.* Topology optimization and fabrication of photonic crystal structures. *Opt. Express* **12**, 1996 (2004).

39. Hughes, T. W., Minkov, M., Williamson, I. A. D. & Fan, S. Adjoint Method and Inverse Design for Nonlinear Nanophotonic Devices. *ACS Photonics* **5**, 4781–4787 (2018).

40. Troyanovsky, B., Yu, Z. & Dutton, R. W. Physics-based simulation of nonlinear distortion in semiconductor devices using the harmonic balance method. *Comput. Methods Appl. Mech. Eng.* **181**, 467–482 (2000).

41. Kundert, K. S. & Sangiovanni-Vincentelli, A. Simulation of Nonlinear Circuits in the Frequency Domain. *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **5**, 521–535 (1986).

42. Esterhazy, S. *et al.* Scalable numerical approach for the steady-state ab initio laser theory. *Phys. Rev. A - At. Mol. Opt. Phys.* **90**, 023816 (2014).

43. Lalau-Keraly, C. M., Bhargava, S., Miller, O. D. & Yablonovitch, E. Adjoint shape optimization applied to electromagnetic design. *Opt. Express* **21**, 21693 (2013).

44. BURGER, M., OSHER, S. J. & YABLONOVITCH, E. Inverse Problem Techniques for the Design of Photonic Crystals. *IEICE Trans. Electron.* **E87-C**, 258–265 (2004).

45. Liu, D., Tan, Y., Khoram, E. & Yu, Z. Training Deep Neural Networks for the Inverse Design of Nanophotonic Structures. *ACS Photonics* **5**, 1365–1369 (2018).

46. Jiang, J. *et al.* Free-form diffractive metagrating design based on generative adversarial networks. *ACS Nano* **13**, 8872–8878 (2019).

47. So, S. & Rho, J. Designing nanophotonic structures using conditional deep convolutional generative adversarial networks. *Nanophotonics* **8**, 1255–1261 (2019).

48. Liu, Z., Zhu, D., Rodrigues, S. P., Lee, K. T. & Cai, W. Generative Model for the Inverse Design of Metasurfaces. *Nano Lett.* **18**, 6570–6576 (2018).

49. Gardner, M. W. & Dorling, S. R. Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998).

50. Acheampong, A. O. & Boateng, E. B. Modelling carbon emission intensity: Application of artificial neural network. *J. Clean. Prod.* **225**, 833–856 (2019).

51. Gentiluomo, L. *et al.* Application of interpretable artificial neural networks to early monoclonal antibodies development. *Eur. J. Pharm. Biopharm.* **141**, 81–89 (2019).

52. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2323 (1998).

53. Rawat, W. & Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation* vol. 29 2352–2449 (2017).

54. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep*

Convolutional Neural Networks. http://code.google.com/p/cuda-convnet/ (2012).

55. Lundberg, S. M., Allen, P. G. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*. https://github.com/slundberg/shap (2017).

56. Liu, X. *et al.* Taming the blackbody with infrared metamaterials as selective thermal emitters. *Phys. Rev. Lett.* **107**, 045901 (2011).

57. Olden, J. D. & Jackson, D. A. Illuminating the 'black box': A randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Modell.* **154**, 135–150 (2002).

58. Qiu, F. & Jensen, J. R. Opening the black box of neural networks for remote sensing image classification. *Int. J. Remote Sens.* **25**, 1749–1768 (2004).

59. Han, S., Pool, J., Tran, J. & Dally, W. J. *Learning both Weights and Connections for Efficient Neural Networks*. (2015).

60. Wada, K. & Kimerling, L. C. *Photonics and Electronics with Germanium*. *Photonics and electronics with germanium* (Wiley-VCH Verlag GmbH & Co. KGaA, 2015). doi:10.1002/9783527650200.

61. Omeis, F. *et al.* Metal-insulator-metal antennas in the far-infrared range based on highly doped InAsSb. *Appl. Phys. Lett.* **111**, 121108 (2017).

62. Chen, K., Adato, R. & Altug, H. Dual-band perfect absorber for multispectral plasmon-enhanced infrared spectroscopy. *ACS Nano* **6**, 7998–8006 (2012).

63. Brachmann, A. & Redies, C. Using Convolutional Neural Network Filters to Measure Left-Right Mirror Symmetry in Images. *Symmetry (Basel).* **8**, 144 (2016).

64. Molnar, C. 5.10 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning. *Christophm.github.io* (2020). at <https://christophm.github.io/interpretable-ml-book/shap.html#kernelshap>

65. S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv e-prints, art. arXiv:1802.03888, Feb 2018.