Accelerating and enabling the design of complex diffractive gratings using physics-informed neural networks to evaluate scattering matrices

Eric Prehn[a] and Peter Jung[a]

[a]Institute of Space Research, German Aerospace Center (DLR), Berlin, Germany

ABSTRACT

Diffractive gratings in waveguides are utilized in various applications, including telecommunications, quantum photonics, optical sensing, display technology for augmented reality devices and space applications. As demand for diffractive gratings with multiple desirable properties increases in waveguide design, more complex grating structures need to be investigated. The design spaces of these complex gratings increase exponentially with the number of physical grating parameters (such as grating depth, duty cycle, layer thicknesses, blaze angles etc.). The effect of these gratings on incident light is often stored in electromagnetic scattering/Jones matrices that are calculated using rigorous coupled-wave analysis (RCWA). Optimising gratings within a waveguide involves storing, evaluating and often interpolation of these scattering matrices, resulting in a computational bottleneck for higher-dimensional grating parameter spaces. To address this challenge, we introduce a highly efficient, parallelisable, and lightweight approach to evaluating scattering matrices using a physics-informed neural network (PINN). This PINN is trained on a data set of scattering matrices and importantly is differentiable, can be GPU-accelerated and penalizes non-physical outputs. In this paper, we provide an optimisation example that demonstrates an exponential speedup in simulation time for high-dimensional grating parameter spaces compared to conventional methods. For inverse design of optical devices involving complex diffractive gratings, this approach opens up a new regime of computationally feasible optimisations.

Keywords: deep learning, physics-informed neural network, rigorous-coupled wave analysis, inverse design, diffractive gratings, waveguide design

1. INTRODUCTION

The finite-difference time domain method, finite element method and rigorous coupled-wave analysis (RCWA)[1–3] are full-wave electromagnetic (EM) simulation algorithms for modeling diffraction. This work focuses on RCWA and how optical engineers might be able to efficiently use RCWA for complex diffractive gratings. We emphasize that the primary goal of this work is to demonstrate our idea. An optical engineer faced with the task of simulating and optimising diffractive gratings, needs to simultaneously consider the accuracy and computational cost or complexity of their chosen method. Maintaining sufficient accuracy and limiting computational cost becomes an exponentially challenging task as the size of the design space increases. For example, optimising a grating with two physical geometric parameters (e.g. duty cycle and grating height), involves traversing a two-dimensional parameter space. However, as the number of grating parameters grows, the challenge intensifies significantly due to the curse of dimensionality. Not only does one need to simulate the physics of a grating in this high-dimensional space (potentially also multiple times for gradient calculations), one needs to store all of the EM results and be able to access and/or interpolate between them quickly. Consequently, advancing towards higher-dimensional and more intricate gratings calls for novel, more efficient methodologies.

At the same time, machine learning (ML) methods, specifically neural networks (NNs), typically benefit from larger datasets. In turn, if we train a NN to predict RCWA results, a large dataset is desirable (higher-dimensional spaces require larger datasets). Once trained, the NN can be plugged into auto-differentiable optical

Further author information:
Eric Prehn: E-mail: eric.prehn@dlr.de

Novel Optical Systems, Methods, and Applications XXVIII, edited by Cornelius F. Hahlweg, Joseph R. Mulley,
Proc. of SPIE Vol. 13595, 135950M · 2025 Published by SPIE
0277-786X · doi: 10.1117/12.3063950

simulations. Importantly, for higher-dimensional parameter spaces, the NN does not suffer from exponentially increasing time and memory costs at inference.

It is equally important to have accurate EM simulations. Ideally a surrogate model, whether it is a simple linear interpolator or a NN, should adhere to the underlying physics. In the worst cases, surrogate models can produce non-physical results for gratings, such as RCWA scattering matrices that have an increase in light intensity. A pioneering approach to embed physics into NNs is through physics-informed neural networks (PINNs),[4,5] with MaxwellNet[6] serving as an example for EM. In this work PINNs are developed with loss function terms that punish non-physical outputs. A PINN is trained on a data set of RCWA results (Jones or scattering matrices) and due to its inference speed, enables and accelerates optimisation of complex gratings.

As part of this paper, a toy optimisation of a multi-objective grating is performed. Trained PINNs are used within the simulation or optimisation, where each PINN outputs a Jones matrix for a different diffraction order. The PINNs easily integrate into the differentiable simulation, allowing for fast inverse-design. Our method demonstrates an exponential speedup in simulation time for high-dimensional grating parameter spaces compared to conventional methods. For inverse design of optical devices involving complex diffractive gratings, this approach opens up a new regime of computationally feasible optimisations. This has broad applications, as diffractive gratings play an important role in waveguide technology,[7] as well as high-efficiency solar cells,[8] ultra-thin metalenses and displays,[9] optical metrology for semiconductor fabrication,[10] X-ray diffraction for material analysis[11] and optical computation.[12]

2. BACKGROUND AND THEORY

RCWA solves PDEs in Fourier space beginning with Faraday's law and Ampére's law of Maxwell's equations:

$$\nabla \times E = -j\omega\mu_0 H, \quad \nabla \times E = j\omega\varepsilon_r E. \tag{1}$$

The electric field E and magnetic field H are complex-valued vector fields over $\mathbf{R}^3$, j denotes the imaginary unit, μ_0 and ε_0 are vacuum permeability and permittivity, respectively. The electromagnetic wave has angular frequency ω and ε_r denotes relative permittivity. For the diffraction problem, a grating is defined by a stack of layers which have identical periods in x and y directions.[13] RCWA begins by determining the eigenmodes of each layer via a Fourier expansion of the material distribution. These eigenmodes, combined with electromagnetic boundary conditions, are used to calculate the input-output relations and coupling coefficients for each layer. The overall input-output relations and coupling coefficients of the full structure are then derived by applying the Redheffer star product.[14] These results are often stored in scattering matrices or Jones matrices. For a more detailed description of the RCWA method, the reader is referred to.[1–3,15] Several excellent open-source software options are available for optical engineers. The classical options are Reticolo[13] and S4,[16] which are prominent in the field.

2.1 Scattering/Jones Matrices

For an RCWA calculation, the wavelength λ and k-vector of the incident light source are fixed. The solution is calculated for both transverse electric (TE) and transverse magnetic (TM) polarisations, and for linear combinations of them. All reflected and transmitted orders that exist, are stored in the global scattering matrix. The blocks of this matrix describe the phase and amplitude transformations between all possible incoming and outgoing waves (modes).

In most practical scenarios, interest is limited to a single incident wave (order) and a small number of its transmitted or reflected non-evanescent orders. The corresponding Jones matrices[17] characterise the response into each of these orders. To optimise a grating within an optical system, one aims to determine the optimal set of geometric parameters $p^* = \{p_1^*, \cdots, p_D^*\}$, which requires evaluation of the Jones matrices:

$$J^m(p_1, p_2, \cdots, p_D) = \begin{pmatrix} J_{EE} & J_{ME} \\ J_{EM} & J_{MM} \end{pmatrix}, \tag{2}$$

for each transmitted or reflected order indexed by m. Here D denotes the number of geometric grating parameters. For instance, when a TE-polarized ray propagates through an optical system and is diffracted by the grating into diffraction order s, the resulting transformation can be described by:

$$E^s_{\text{out}} = J^s(p_1, p_2, \cdots, p_D)E_{\text{in}} = \begin{pmatrix} J_{EE} & J_{ME} \\ J_{EM} & J_{MM} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} J_{EE} \\ J_{EM} \end{pmatrix}. \tag{3}$$

Note the entry J_{ME} denotes $J_{TM \longrightarrow TE}$, describing the transfer of phase and amplitude of TM into the TE polarisation when diffraction occurs. Hence, the diffraction efficiency of incident TE light into TM light for order s is equal to $|J^s_{EM}|^2$.

2.2 Jones Matrix Evaluation

As stated, to optimise a grating within an optical system, one seeks the optimal grating parameters:

$$p^* = (p_1^*, \cdots, p_D^*), \tag{4}$$

that maximize (or minimize) a given objective function C, which typically represents a performance metric of the system. The rays hitting a grating may also come from different directions. For each diffraction order m, where θ is the angle of incidence and ϕ is the azimuthal angle of the incoming ray, we aim to efficiently evaluate the function $J^m(p_1, p_2, \cdots, p_D, \theta, \phi)$ during an optical simulation. However, performing RCWA for every evaluation point becomes computationally prohibitive, especially when considering thousands of rays incident at various angles. To address this computational challenge, engineers approximate the mapping:

$$(p_1, p_2, \ldots, p_D, \theta, \phi) \mapsto J^m(p_1, p_2, \ldots, p_D, \theta, \phi). \tag{5}$$

A common method is to precompute RCWA results on a grid and use interpolation, but even linear interpolation evaluates slowly as D grows. Linear interpolation (also known as multi-linear interpolation) scales as $\mathcal{O}(2^D)$ in general, but there are certain speed-ups possible in some cases. This work introduces a PINN approach to approximate Jones matrices:

$$\hat{f}^m_{NN}(p_1, p_2, \cdots, p_D, \theta, \phi) = \hat{J}^m, \tag{6}$$

where $\hat{f}^m_{NN}$ is a PINN that has been trained on a precomputed RCWA dataset.

3. GRATING STRUCTURES AND DATASETS

For our example of a grating structure with $D = 2$, the duty cycle (the ratio of the grating ridge width to the period) and grating height are varied. The rest of the parameters (and materials) used to describe this grating remain fixed. RCWA results are generated for this grating structure on a grid. Note that in general the size (number of grid points) grows exponentially with increasing D. Details on the grating structures and corresponding RCWA datasets used for the optimisation examples are shown in Table 1 below. The datasets are generated using Reticolo[13] freeware. Visualisations of the corresponding gratings for $D = 2$ and $D = 3$ are shown in Fig. 1 and the added complexity of the grating structures with $D = 5$ can be clearly seen in Fig. 5.

4. PHYSICS-INFORMED NEURAL NETWORK APPROACH

A neural network is trained to predict each of the non-evanescent orders J^m. During training, each input to the NN is a grid point $(p_1, p_2, \ldots, p_D, \theta)$ and the corresponding target is the 2×2 matrix of complex values $J^m(p_1, p_2, \ldots, p_D, \theta)$. The loss function of the NN contains the mean squared errors (MSEs) of the predicted real and complex values, along with physics-informed loss terms. The physics-informed loss punishes when the NN predicts entries of J^m with absolute magnitude greater than one. This is due to the fact that entries of Jones matrices are physically restricted to lie within or on the unit complex circle, as they could increase intensity if they have absolute magnitude greater than one. It also encourages the diffraction efficiencies of the orders to sum up to one (conservation of energy for lossless gratings), and greatly punishes when the sum is greater than one. Each of these three loss terms are weighted by hyperparameters that are additionally tuned. Once trained the NNs are tested on a test set consisting of 1000 scattered points $\{p_{\text{test}}\}$. In this work, a model is kept and used for the optimisation examples once it achieved a MSE of less than 0.005 on the test set.

Table 1. For all datasets, the incoming light is green visible light with $\lambda = \lambda_0 = 0.532$ µm in free space. All gratings have a period in the x-direction of $\Lambda_x = 0.35$ µm. The azimuthal angle is fixed at $\phi = 10$. All angles are given in degrees and heights are in micrometers (µm). There are 30,000, 90,000 and 175,000 points in the $D = 2, 3, 5$ datasets, respectively.

$D = 2$	Duty Cycle	Grating Height	θ
Range	$[0.3, 0.8]$	$[0.6, 0.8]$	$[43.93, 49.96]$
Intervals	25	40	30

$D = 3$	Duty Cycle	Grating Height	Underlayer Height	θ
Range	$[0.3, 0.8]$	$[0.7, 0.8]$	$[0.1, 0.2]$	$[43.93, 49.96]$
Intervals	25	20	10	18

$D = 5$	Duty Cycle	Grating Height	Underlayer Height	Overlayer Height	Fill Height	θ
Range	$[0.3, 0.8]$	$[0.7, 0.8]$	$[0.1, 0.2]$	$[0.02, 0.4]$	$[0.1, 0.7]$	$[43.93, 49.96]$
Intervals	10	10	5	5	7	10

5. OPTIMISATION OF MULTI-OBJECTIVE GRATING

As a toy example we consider the optimisation of a multi-objective diffractive grating acting as an outcoupler in a waveguide. We seek to optimise the intensity of transmitted light within a region above the waveguide and minimize the intensity of reflected light within a region below the waveguide. We also care about the direction of light or field of view in this application. So in the transmitted region, we wish to have a uniform intensity distribution across the field of view. The simulation is performed along one dimension and a depiction of the simulation is shown in Fig. 2. The rays are launched from the bottom of the waveguide at uniformly randomly

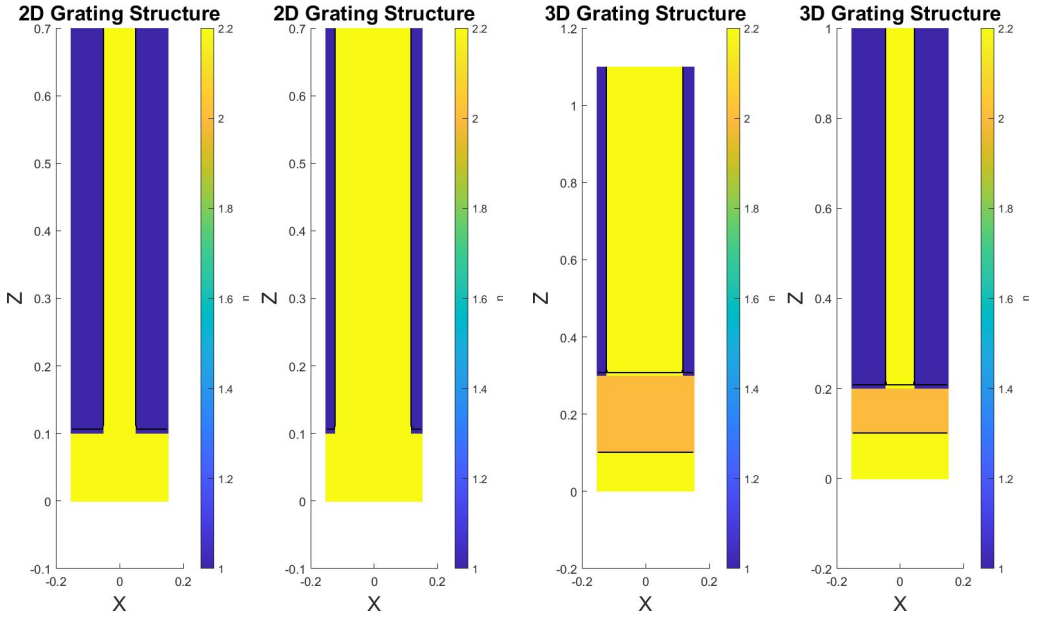

Figure 1. In this figure sample gratings from the two-dimensional and three-dimensional datasets are visualised. The relative permeability $\mu_r = 1$ for all layers and n denotes refractive index values. For the case with $D = 2$, the duty cycle and height of the grating are varied. For $D = 2$ we show a grating with a low duty cycle, while to its right we present the same grating structure with a high duty cycle. For $D = 3$ we add another layer (with $n = 2$) of variable height to the grating structure. All gratings are periodic in the x direction and uniform in the y direction and the images are rendered using Reticolo[13] freeware.

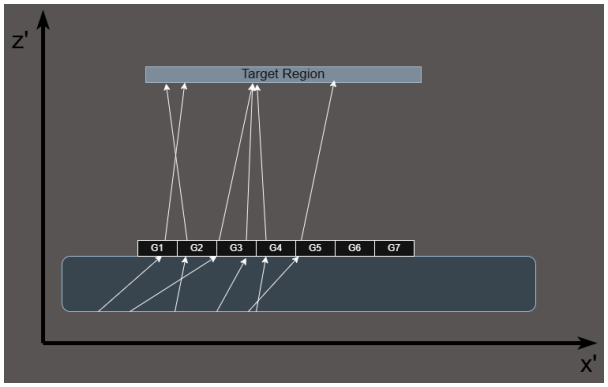

Figure 2. Illustration of toy optimisation example. The rays are launched randomly from the bottom of the waveguide, they hit different grating subregions G_i and diffract towards a target region (and a region below the waveguide not shown). Each grating region G_i consists of hundreds or thousands of identical diffraction gratings.

distributed angles and phases, with decreasing amplitude in the positive x-direction. They hit a grating region at the top of the waveguide. This grating region consists of several individual grating subregions, each denoted G_i. The optimisation involves finding the optimal geometric parameters p^i for each of the grating structures i located in subregions G_i. All of the geometric parameters in $(p_1^i, p_2^i, \cdots, p_D^i)$ except for the duty cycle must be the same across subregions.

The rays that exit the waveguide are those not trapped by total internal reflection (TIR) and come from non-evanescent diffraction orders. The interaction of these rays with the grating is modeled as in Equation (2). These rays are traced out of the waveguide and terminate at regions below or above the waveguide. The paths are traced during initialisation, and secondary bounces are ignored in the simulation. All rays reaching the target region are binned according to their angle of incidence $\Theta \in [-5, 5]$. This field of view (FOV) range of ten degrees is binned via eleven FOV bins, each denoted $\Theta_{f=1,\cdots,11}$ and the intensity contributions are summed for each bin. For the region below the waveguide, where undesired intensity arrives, the total contribution from all incoming angles is simply summed and denoted I_B. The objective function is defined as:

$$C = \sum_{f=1}^{11} I_{\Theta_f} - I_B - \sigma(I_\Theta), \tag{7}$$

where $\sigma(I_\Theta)$ is the standard deviation of the intensities across the FOV bins. Using jax[18] software the entire simulation is auto-differentiable. Both the linear interpolation method and PINN method for evaluation of the Jones matrices are seamlessly integrated into the optimisation using jax and equinox[19] software. The geometric parameters p^i for each of the grating subregions are optimised via backpropagation of the objective function using the Adam[20] optimiser.

6. RESULTS AND DISCUSSION

Optimisations were performed for each of the three grating structures (design spaces). Each optimisation was initiated with one thousand rays and was performed using both the PINN approach and linear interpolation. The exponential decrease in simulation time of the PINN approach, compared to traditional methods, as D increases, is verified and plotted in Fig. 4. This vastly different scaling behaviour outlines the benefit of using a PINN approach for complex gratings. In our toy example we optimise twenty gratings (twenty subregions) and an example optimisation trajectory is plotted in Fig. 3. This toy optimisation is a simple example of a basic, first-order approximation of a physical system. We emphasize that for more accurate modeling, more evaluations of Equation (3) will be required. Consequently, the advantages of a PINN approach will be even more pronounced.

The numerical experiments were conducted on a workstation with an Intel Core i7-1365U CPU @ 1800 MHz, 32 GB of RAM, and running Windows 11 Enterprise (64-bit). Optimisation trajectories of the geometric

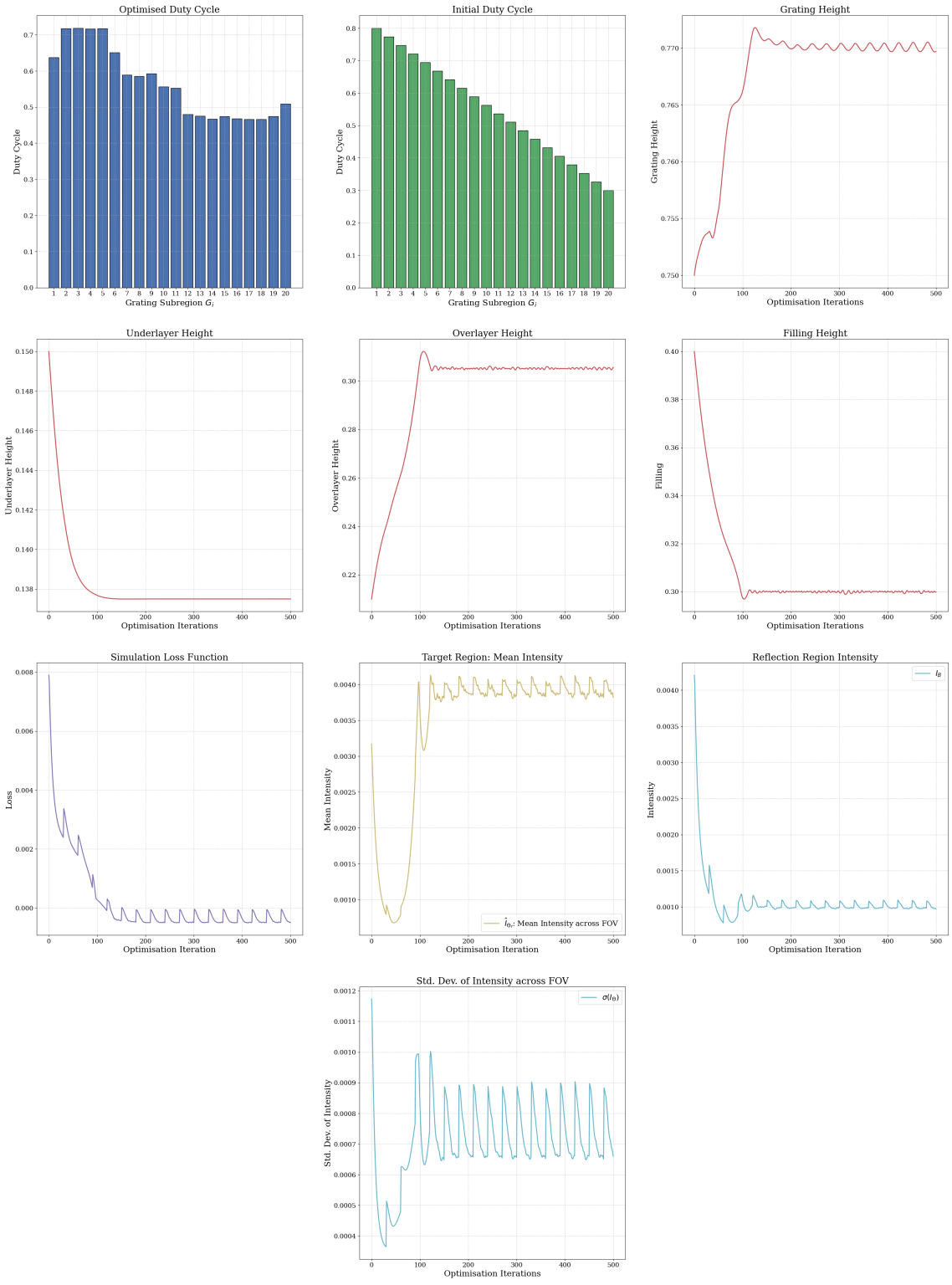

Figure 3. Example Optimisation trajectory for $D = 5$ grating structure using PINN approach. The loss function is simply set to be the negative of the objective function C and intensity values are normalised by the number of rays (1000 rays).

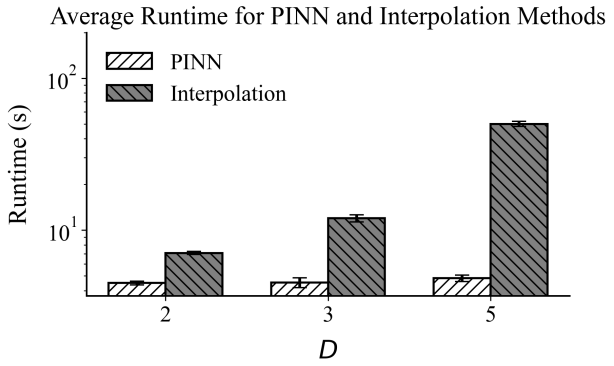

Figure 4. As D increases, the evaluation time for the linear interpolation method exhibits exponential growth, whereas the PINN method scales linearly. Error bars represent standard deviation over ten repetitions of each optimisation.

parameters depend on initial values of p. Of course, there is no guarantee of reaching the globally optimal p^* (finding the optimal set of gratings for our system). In practice, the optimisation needs to be repeated many times with different initial values of p. The exact trajectory taken is also sensitive to the evaluations of $J^m(p_1, p_2, \ldots, p_D, \theta, \phi)$. In order to find a grating that satisfies multiple objectives, careful weighting and balancing of the terms in the objective function C is required. However, this is not the subject of this paper. After completing the optimisation, the resulting set of geometric parameters can be used to compute the exact RCWA Jones matrices using software such as Reticolo. By replacing the approximate $\hat{J}^m$ with the exact Jones matrices, the performance of the final design can be verified.

7. CONCLUDING REMARKS

The use of PINNs for evaluating scattering or Jones matrices of complex gratings facilitates a new regime of computationally efficient grating design. The computational benefit of this method is clearly demonstrated through a toy optimisation example. We note that the curse of dimensionality still exists for design spaces with growing dimensionality and it is impossible to search the infinite design space of gratings. However, our method makes it feasible to have higher-dimensional design spaces in optical simulations. This capability is primarily driven by the rapid inference speed of NNs, on top of this the PINNs can be seamlessly integrated into auto-differentiable optical simulations for efficient inverse design.

ACKNOWLEDGMENTS

This project was made possible by the DLR Quantum Computing Initiative and the Federal Ministry for Economic Affairs and Climate Action; qci.dlr.de/projects/qcoptsens.

REFERENCES

[1] Moharam, M. G. and Gaylord, T. K., "Rigorous coupled-wave analysis of planar-grating diffraction," *J. Opt. Soc. Am.* **71**, 811–818 (7 1981).

[2] Moharam, M. G., Grann, E. B., Pommet, D. A., and Gaylord, T. K., "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," *J. Opt. Soc. Am. A* **12**, 1068–1076 (5 1995).

[3] Moharam, M. G., Pommet, D. A., Grann, E. B., and Gaylord, T. K., "Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach," *J. Opt. Soc. Am. A* **12**, 1077–1086 (5 1995).

[4] Raissi, M., Perdikaris, P., and Karniadakis, G. E., "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10561* (2017).

[5] Raissi, M., Perdikaris, P., and Karniadakis, G. E., "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.* **378**, 686–707 (2019).

[6] Lim, J. and Psaltis, D., "Maxwellnet: Physics-driven deep neural network training based on maxwell's equations," *Apl Photonics* **7** (2022).

[7] Amanti, F., Andrini, G., Armani, F., Barbato, F., Bellani, V., Bonaiuto, V., Cammarata, S., Campostrini, M., Dao, T. H., Matteis, F. D., Demontis, V., Donati, S., Giuseppe, G. D., Tchernij, S. D., Fontana, A., Forneris, J., Frontini, L., Gunnella, R., Iadanza, S., Kaplan, A. E., Lacava, C., Liberali, V., Martini, L., Marzioni, F., Morescalchi, L., Pedreschi, E., Piergentili, P., Prete, D., Rigato, V., Roncolato, C., Rossella, F., Salvato, M., Sargeni, F., Shojaii, J., Spinella, F., Stabile, A., Toncelli, A., and Vitali, V., "Integrated photonic passive building blocks on silicon-on-insulator platform," *Photonics* **11** (2024).

[8] Snaith, H. J., "Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells," *J. Phys. Chem. Lett.* **4**, 3623–3630 (2013).

[9] Zhang, L., Ding, J., Zheng, H., An, S., Lin, H., Zheng, B., Du, Q., Yin, G., Michon, J., Zhang, Y., et al., "Ultra-thin high-efficiency mid-infrared transmissive huygens meta-optics," *Nat. Commun.* **9**, 1481 (2018).

[10] Boef, A. J. D., "Optical metrology of semiconductor wafers in lithography," in [*International Conference on Optics in Precision Engineering and Nanotechnology (icOPEN2013)*], **8769**, 57–65 (2013).

[11] von Laue, M., "Concerning the detection of x-ray interferences," *Nobel lecture* **13** (1915).

[12] Silva, A., Monticone, F., Castaldi, G., Galdi, V., Alù, A., and Engheta, N., "Performing mathematical operations with metamaterials," *Science* **343**, 160–163 (2014).

[13] Hugonin, J. P. and Lalanne, P., "Reticolo software for grating analysis," *arXiv preprint arXiv:2101.00901* (2021).

[14] Kim, C. and Lee, B., "Torcwa: Gpu-accelerated fourier modal method and gradient-based optimization for metasurface design," *Comput. Phys. Commun.* **282**, 108552 (2023).

[15] Rumpf, R. C., "Improved formulation of scattering matrices for semi-analytical methods that is consistent with convention," *Prog. Electromagn. Res. B.* (2011).

[16] Liu, V. and Fan, S., "S4: A free electromagnetic solver for layered periodic structures," *Comput. Phys. Commun.* **183**, 2233–2244 (2012).

[17] Jones, R. C., "A new calculus for the treatment of optical systems I. description and discussion of the calculus," *J. Opt. Soc. Am.* **31**, 488–493 (7 1941).

[18] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q., "JAX: composable transformations of Python+NumPy programs," (2018).

[19] Kidger, P. and Garcia, C., "Equinox: neural networks in jax via callable pytrees and filtered transformations," *Differentiable Programming workshop at Neural Information Processing Systems 2021* (10 2021).

[20] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2017).

APPENDIX A. ADDITIONAL FIGURES

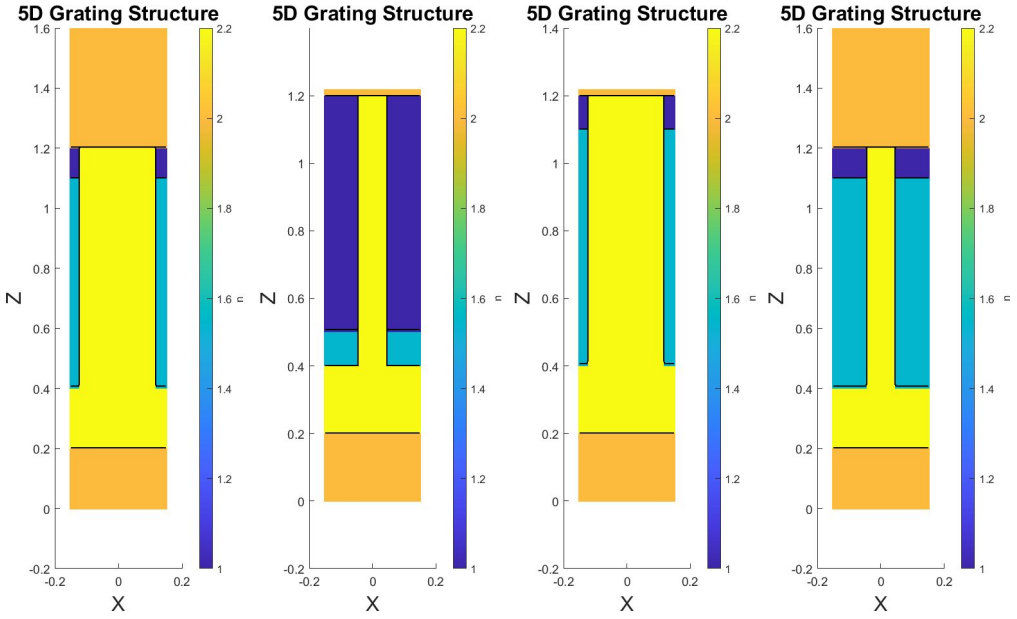

Figure 5. In this figure sample gratings from the five-dimensional dataset are visualised. The relative permeability $\mu_r = 1$ for all layers and n denotes refractive index values. In this case there are underlayer and overlayer heights that can be varied and the region of air ($n = 1$) in the grating can be partially filled by glass ($n = 1.5$) of variable height. All gratings are periodic in the x direction and uniform in the y direction and the images are rendered using Reticolo[13] freeware.