Photonic-structure optimization using highly data-efficient deep learning: Application to nanofin and annular-groove phase masks

Nicolas Roy [1,2,3,*] Lorenzo König,[4] Olivier Absil [4] Charlotte Beauthier [5] Alexandre Mayer [1,2,3] and Michaël Lobet [6,2,3]

[1]*Namur Institute for Complex Systems, University of Namur, Rue de Bruxelles 61, 5000 Namur, Belgium*
[2]*Namur Institute of Structured Matter, University of Namur, Rue de Bruxelles 61, 5000 Namur, Belgium*
[3]*Department of Physics, University of Namur, Rue de Bruxelles 61, 5000 Namur, Belgium*
[4]*STAR Institute, University of Liège, Allée du Six Août 19c, 4000 Liège, Belgium*
[5]*Minamo Developpement Team, Cenaero, Avenue des Frères Wright 29, 6041 Gosselies, Belgium*
[6]*John A. Paulson School of Engineering and Applied Sciences, Harvard University, 9 Oxford Street, Cambridge, Massachusetts 02138, USA*

(Received 14 September 2023; accepted 18 December 2023; published 23 January 2024)

Metasurfaces offer a flexible framework for the manipulation of light properties in the realm of thin-film optics. Specifically, the polarization of light can be effectively controlled through the use of thin phase plates. This study aims to introduce a surrogate optimization framework for these devices. The framework is applied to develop two kinds of vortex phase masks (VPMs) tailored for application in astronomical high-contrast imaging. Computational intelligence techniques are exploited to optimize the geometric features of these devices. The large design space and computational limitations necessitate the use of surrogate models like partial least-squares kriging, radial basis functions, or neural networks. However, we demonstrate the inadequacy of these methods in modeling the performance of VPMs. To address the shortcomings of these methods, a data-efficient evolutionary optimization setup using a deep neural network as a highly accurate and efficient surrogate model is proposed. The optimization process in this study employs a robust particle swarm evolutionary optimization scheme, which operates on explicit geometric parameters of the photonic device. Through this approach, optimal designs are developed for two design candidates. In the most complex case, evolutionary optimization enables optimization of the design that would otherwise be impractical (requiring too many simulations). In both cases, the surrogate model improves the reliability and efficiency of the procedure, effectively reducing the required number of simulations by up to 75% compared to conventional optimization techniques.

DOI: 10.1103/PhysRevA.109.013514

I. INTRODUCTION

Metasurfaces offer a flexible framework for shaping the behavior of light in the realm of thin-film optics [1–3]. However, their utilization in some applications necessitates intricate material configurations, posing a significant design challenge. To tackle this obstacle, inverse design methods fueled by computational intelligence have attracted interest [4,5]. Among them, evolutionary optimization [6] has the power to leverage deep learning within a versatile surrogate optimization framework. Its performance, however, should be evaluated on a challenging case. The design of metasurfaces such as vortex phase masks, specifically, all-dielectric phase plates [7,8], tailored for use in coronagraphy, is certainly a relevant choice.

Coronagraphy is a powerful technique for imaging exoplanets. It enables the detection of faint planetary signals in a star's surrounding regions. One promising coronagraphic implementation is the annular-groove phase mask (AGPM), which employs a focal-plane phase mask comprised of a circular subwavelength dielectric grating [8–11]. The grating acts as a spatially variant half waveplate, which creates a helical phase ramp (i.e., an optical vortex) on the optical axis of the telescope, ending up creating a dark region in the field of view.

The performance of AGPMs has traditionally been analyzed using the rigorous coupled-wave analysis (RCWA) method [12]. However, the validity of this infinite, one-dimensional grating model reaches its limits as focus shifts from the outskirts to the center of the device [13]. Therefore, the use of three-dimensional (3D) electromagnetic solvers has become necessary for modeling of the center of AGPMs. These new tools, by allowing more freedom in the AGPM design, provide a challenging benchmark for our framework and new ways to improve the AGPMs.

Two approaches are devised for designing the center of the phase mask, as explained below. Each one leads to photonic devices that are difficult to optimize due to the complex interplay between numerous mask parameters and the optical system. The figure of merit of the optimization is the simulated efficiency of the devices in producing the optical vortex. Both designs present large design spaces that require a tremendous number of simulations to be properly sampled: Up to a trillion simulations are required to try only two values for each of the 38 design parameters in one of the two approaches. Therefore, we need an optimization procedure that is efficient in terms of the number of simulations required while exploring the vast search space.

*nicolas.roy@unamur.be

In this work we combine a particle swarm optimization (PSO) global optimization algorithm with a U-Net surrogate model in order to optimize the mask parameters. The model is improved throughout the optimization process using active learning [14,15]. This combination aims to achieve an automatic and efficient exploration of the phase mask design space. On one hand, U-Net, a deep learning architecture borrowed from image segmentation [16], is effective in providing quick (measured in milliseconds) and accurate predictions of field distributions based on the structure topology [17]. On the other hand, PSO is a well-known global optimization heuristic algorithm. Importantly, the methodology proposed here, while being applied to coronagraphy, is general enough to optimize any complex photonic device defined by 10–100 parameters with efficient exploitation of 200–1000 simulations. The code [18] and the data [19] required to reproduce our numerical experiments are available on Github.

The present work is organized as follows. Sections II A and II B describe the proposed U-Net surrogate modeling methodology for optimization. Sections II C and II D provide an overview of the vortex phase mask coronagraph parameters and simulations. Section II E describes the model enrichment process, which consists in the interaction between the optimizer, the simulation, and the U-Net. Section III A compares the implemented surrogate accuracy and efficiency with other existing models such as dense neural networks, radial basis functions, and partial least-squares Kriging. Section III B investigates the influence of the number of simulations used in the training of the surrogate model in order to identify a minimal data-set size. Section III C follows with the AGPM designs produced by the proposed process. Section IV summarizes the paper and discusses future research directions.

II. METHODS

A. Evolutionary optimization approach to metasurface design

Evolutionary computation served electromagnetic design for a long time [20]. Global optimizers such as PSO [21] or a genetic algorithm [22] allows us to find a solution to problems that otherwise would require an extremely high number of simulations [23]. However, for the design of metasurfaces, this approach can result in unreasonably high computational needs, as they most often require complex solvers such as the finite-difference time domain (FDTD). In fact, evolutionary global optimizers still require thousands of figure of merit evaluations in order to find an optimal combination for several tens of variables.

Surrogate models have been devised [24–26] to reduce the number of simulations required by partially relying on the quick predictions they provide for evaluation. Surrogate models also have other interesting properties [27]: A simpler model of a complex simulation will offer a smoother performance metric to the optimizer.

We chose this global evolutionary optimization approach over more popular inverse design methods [17,28,29] for several reasons. First, the surrogate optimization approach provides an interesting compromise between the fast convergence of adjoint solvers [29] and the global exploration of the design space allowed by global optimizers. Second, a metasurface design consists in a tight geometrical description of the device using bounded parameters, contrary to inverse design techniques that use a freeform density-based description [29]. Finally, the surrogate optimization scheme is straightforward to implement, consistent across different devices, and it enables massively parallel workloads for the full solver. Other types of parameters such as categorical variables can easily be implemented for materials and shapes.

B. Surrogate solver: U-Net

Various black-box methods can be substitutes (surrogate models) for simulations, including interpolation techniques such as radial basis functions, Kriging [30,31], or other classic machine learning methods such as regression trees [32,33]. In recent years, deep neural networks (DNNs) were often used as surrogate models [34,35], particularly in photonics [36]. These DNNs are described as universal approximators [37,38] for any continuous bounded function, making them a versatile tool used in many tasks. In fact, DNNs can take many forms, referred to as architectures, to adapt to the data of the problem at hand. In this work we will be using an architecture of DNN called the U-Net [16] implemented using PYTORCH [39]. The number of layers and the activation functions are kept identical to those in the seminal article [16], while the number of neurons saw around a fivefold decrease, following the parsimony principle [40] to avoid overfitting. This architecture, illustrated in Fig. 1, excels in image-translation tasks such as image segmentation, where the spatial topology of the image is preserved while the meaning of each pixel changes. In our case, the input is the relative dielectric permittivity and the output is a real function of the electromagnetic fields, namely, the leakage field. While the U-Net is able to handle 3D data [41], a 2D representation of both the structure and the polarization leakage field is numerically more efficient and sufficient to reproduce the physics of the whole system. Therefore, characteristic 2D slices of the dielectric and polarization leakage field are used as represented in Fig. 1. The slices go through a sequence of convolution operations until they reach a low-dimensional representation at the center (bottleneck) of the U-Net (Fig. 1). This compressed representation is then transformed back to match its original size at the output, using transposed convolution operators.

Up to this point, U-Net is very similar to a convolutional autoencoder (CAE) [42]. However, U-Net is completed by skip connections (shortcuts in Fig. 1), allowing low-level local features to pass through and merge with features emerging from the bottleneck, along the decoder. These shortcuts make U-Net effective in modeling the behavior of electromagnetic fields inside metasurfaces as they link more easily the fields to the local dielectric topology. This network architecture has been shown recently to perform well for a diffracting system [36].

C. Metasurfaces: Case of annular-groove phase masks

The optimization framework presented in this paper is applied to the design of the AGPM center. The AGPM is one way of implementing a vector vortex phase mask capable

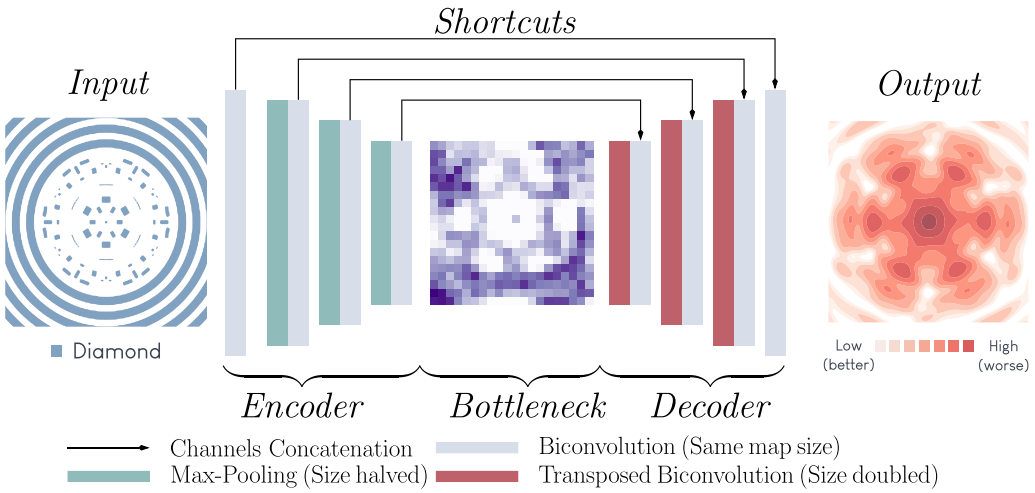

FIG. 1. Representation of the U-Net architecture. The network is comprised of an encoding and a decoding module. These are located at each side of a bottleneck. The input of the network is a slice of the dielectric structure (left), while the output is a conforming slice of the polarization leakage field (right). The shortcuts directly connecting layers along the encoder and decoder are represented by the black arrows.

of achieving contrasts of several orders of magnitude across a large bandwidth [10,43]. Several AGPMs have been successfully installed on the world's most advanced telescopes to date [44–47]. The AGPM uses the artificial birefringence of subwavelength gratings etched onto a diamond substrate to create the helical phase ramp characteristic of a vortex phase mask. The grating parameters are tuned to provide a π phase shift between orthogonal polarizations and create an achromatic half waveplate. The characteristic helical phase ramp is obtained by spatially varying the fast axis of the half waveplate across the mask, resulting in the circular groove pattern of the AGPM.

D. Physics solver: Finite-difference time-domain simulations

While the grating parameters of the AGPM are optimized using RCWA [12], which is well suited for describing infinite periodic gratings, at the center of the AGPM, the pattern is no longer periodic. The FDTD method [48] is used here to fully describe the behavior of the AGPM at its center. A circularly polarized plane wave is propagated through the AGPM. The half-waveplate character of the AGPM flips the helicity of the wave while imprinting the textbook helical phase ramp, leaving a small fraction unaffected, referred to as polarization leakage. The polarization leakage is numerically computed as the mean intensity of the circular polarization with the same handedness as the input in a slice 2.25 μm inside the substrate. This polarization leakage quantifies the amount of light that does not acquire the phase ramp due to the chromaticity of the design. The effects of the curved grating lines near the AGPM center can be described accurately and an optimal size can be estimated for the central pillar [13].

Here the AGPM center is inversely designed by using a more complex metasurface structure. Starting from the AGPM pattern, a region with a radius of five grating periods is optimized. These five periods correspond to the region in which the central leakage is localized (optimizable area in Fig. 2). A simple pattern of concentric grooves of varying line width and position is first considered to minimize the leakage term at the center of the mask. Figure 2(a) shows an example of a concentric groove pattern defined by its inner and outer radii for each groove. The design freedom is then drastically increased by using rectangular nanofins [2] placed in a hexagonal pixelization grid [Fig. 2(b)], leading to an optimization problem with hundreds of free parameters. Each nanofin in the pixelization grid has five free parameters: its position (two), size (two), and its tilt angle (one). In principle, for 91 blocks shown in Fig. 2(b) within the optimized region this results in 455 free parameters. However, for the case of the AGPM, the circular symmetry of the problem is exploited by using the symmetry of the hexagonal pixelization grid. Forcing the orientation of the blocks to be parallel or orthogonal to the annular grooves further reduces the number of free parameters: While keeping the vectorial nature of the mask, the number of independent parameters is reduced to 38. Figure 2(b) shows an example of a nanofin pattern, including the hexagonal pixelization grid overlaid as thin gray lines. The depth of the structures is fixed throughout the mask for both patterns and is optimized for the annular-groove pattern beyond the central region considered for optimization. Such parameters are available in the Supplemental Material (S3) [49].

E. Global optimization scheme

In this section, a surrogate optimization process [50] is presented. This active learning [14] process will iteratively feed the U-Net with new designs and leakage maps computed through the FDTD method to improve its accuracy. The choice of those designs is left to a PSO algorithm [51] that will probe designs with increasing performance. This choice is further refined by a selection strategy similar to that in [15]. The whole model building and optimization process is illustrated in Fig. 3.

The process starts by generating N_0 random designs [Fig. 3(a)]. This value is fixed following a study in an upcoming section ($N_0 = 50$). These initial designs enter the main loop and are evaluated using FDTD simulations, as described in Sec. II C [Fig. 3(b)]. Once this first batch of simulations

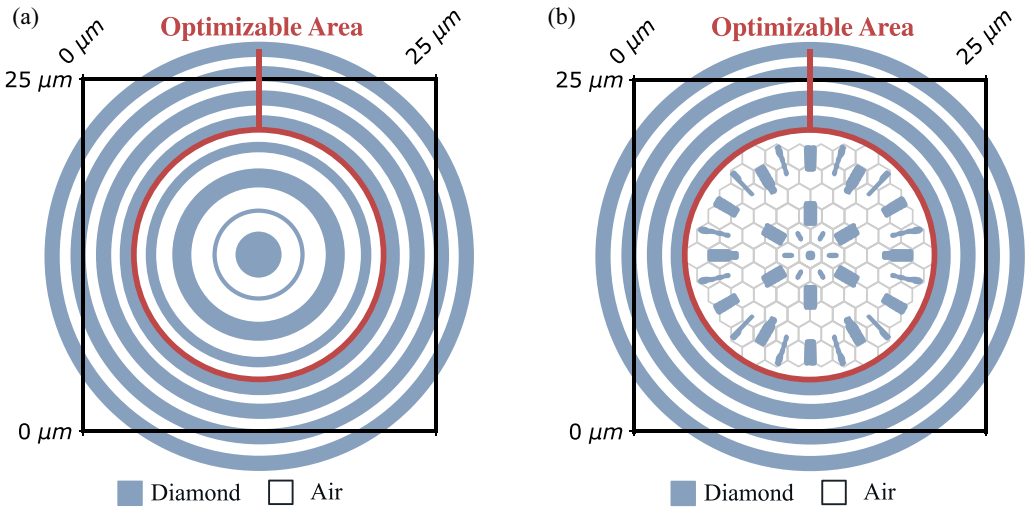

FIG. 2. Dielectric permittivity ϵ_r profile of (a) an annular-groove phase mask and (b) a nanofin phase mask. The AGPM pattern is maintained in the region beyond the optimizable area (red circle) and determines the depth of the full pattern. The hexagonal pixelization grid used to define the metasurface pattern is overlaid as thin gray lines in (b).

is performed, the first U-Net is trained on the initial set of simulations [Fig. 3(c)]. The U-Net will then be explored by multiple (n_{PSO}) parallel executions of the PSO algorithm [Fig. 3(d)]. This exploration leads to pairs consisting of the design that is probed by PSO and the corresponding leakage maps predicted by U-Net. Since PSO is a stochastic algorithm, each instance will lead to a different path and sequence of probed designs. These design-prediction pairs are then stored in a raw database for further analysis [Fig. 3(d)].

This raw database cannot be directly used to improve U-Net as its performance is only an estimation made by U-Net itself. As such, it may include significant errors, especially at the early stages of training. Furthermore, the data set is large and redundant. For all these reasons, a selection process [Fig. 3(e)] is required to pick just a few designs that are allowed to go back to the evaluation step [Fig. 3(b)]. Due to this selection, only the most relevant of these designs will be associated with correct

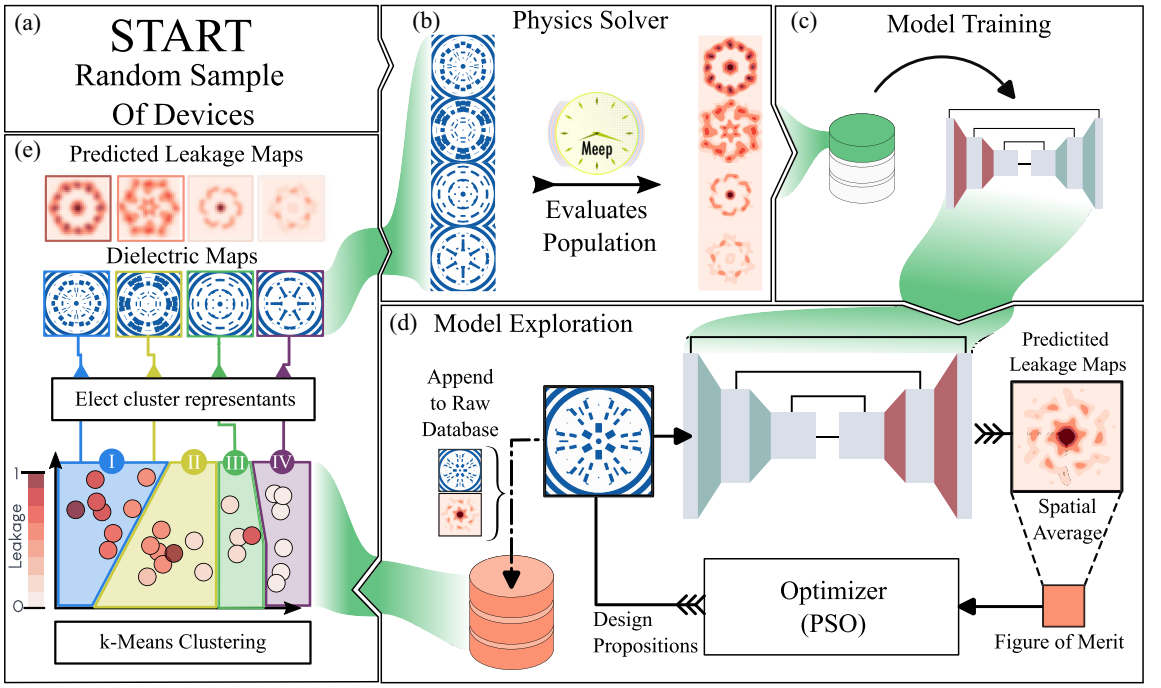

FIG. 3. Summary of the surrogate based optimization scheme. (a) To initiate the algorithm, a randomly generated selection of designs is formed. (b) The selected designs are evaluated in the expensive FDTD simulations. (c) The new designs are appended to the data set and the U-Net is updated. (d) A PSO algorithm searches for better designs using the U-Net model. (e) The designs proposed by PSO are filtered using k-means clustering and a design is elected in each cluster. The loop is closed by going back to step (b).

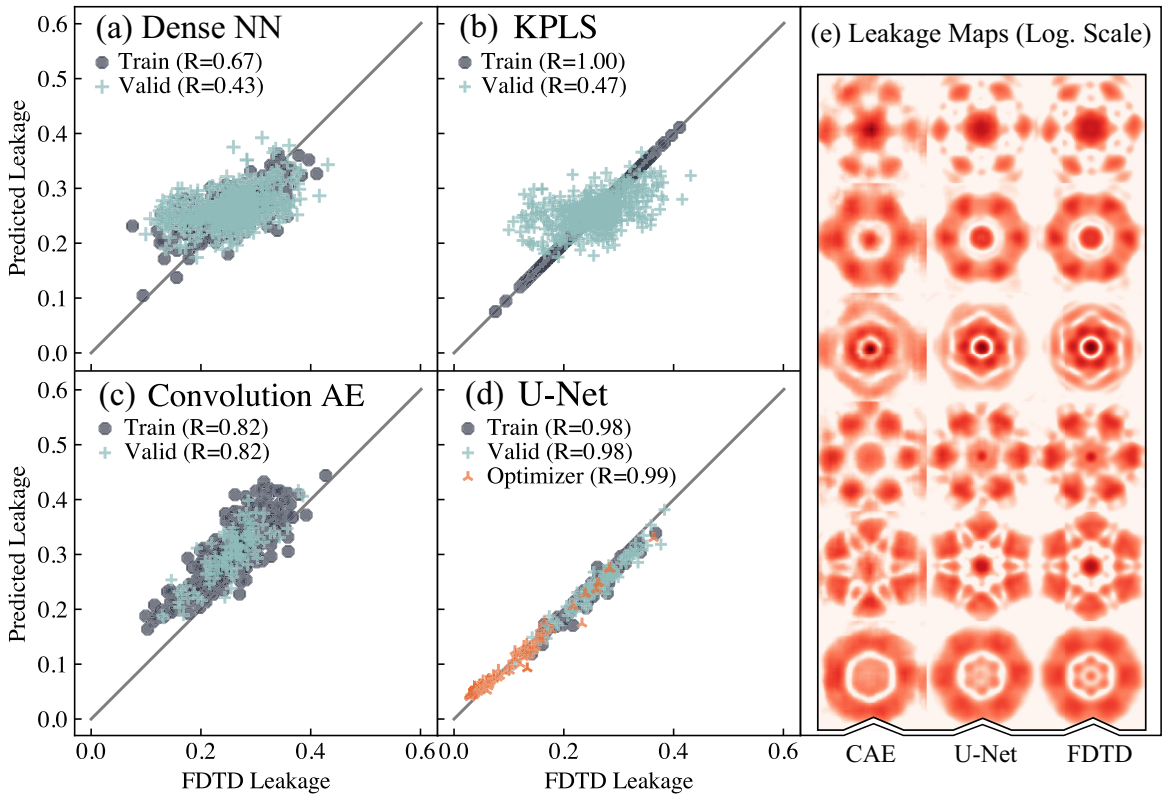

FIG. 4. Benchmark of U-Net against dense neural network, KPLS, and CAE models. Scatter plots of the predicted versus the FDTD mean leakage from various methods are shown: (a) DNN, (b) KPLS, (c) CAE, and (d) U-Net. The ground-truth leakage computed by the FDTD method is found on the abscissa while the ordinate represents the corresponding predicted values. The results for two groups of designs are shown and correspond to the training set and the validation set. Predictions made during optimization are also shown for the U-Net. The Pearson correlation coefficient R is displayed in the legend. (e) CAE, U-Net, and FDTD leakage maps are shown for six random designs.

leakage maps in the step in Fig. 3(b) using the FDTD solver.

Relevant designs are those that provide information about the search space (diversified) and achieve optimal optical performance. Failing to balance exploration (attempting less promising designs) and exploitation (refining the best ones) would lead to either early convergence on a suboptimal design or failure to converge altogether. To make the selection, the raw database is split in k clusters by applying k-means [52] ($k = 4$) on the topology of the designs as illustrated in Fig. 3(e). The approach works in synergy with PSO due to the observed tendency of particle swarms to form a sequence of design niches during the optimization process of the step in Fig. 3(c). Evidence for this statement is available in the Supplemental Material (S2) [49]. The clusters of k-means then allow us to sample these niches as they each correspond to a local optimum for the design coined archetype. Finally, a design is picked with a probability inversely proportional to its performance in each cluster. More details about the selection can be found in the Supplemental Material (S2) [49].

Once filtered, the few remaining designs take the same path as the initial random sample: They are evaluated accurately with the FDTD method in Fig. 3(b) and the pairs of dielectric and leakage maps are stored in the database [Fig. 3(c)]. The U-Net model then incorporates the increased data set in Fig. 3(c). This new model is finally used in Fig. 3(d), like the initial one, closing on the surrogate optimization loop.

Using $n_{\mathrm{PSO}} = 5$ parallel PSO instances, with one design sampled in each of $k = 4$ clusters, turns out to perform well, leading to $k \times n_{\mathrm{PSO}} = 20$ FDTD evaluations for each model update. From 15 to 20 model updates were found to be sufficient for the optimization to converge. These parameters were used in the optimization of nanofin and annular-groove phase masks.

III. RESULTS

A. Quality of the polarization leakage predictions

In the previous sections, U-Net was introduced as a promising architecture for surrogate modeling of complex metasurfaces. In this section, this statement is supported by comparing the U-Net approach to competing methods for the prediction of the spatially averaged polarization leakage.

Figure 4 shows, for the nanofin designs, the predicted spatial average of the polarization leakage $\bar{l}$ against its ground-truth value obtained through the FDTD method. Results for different groups of designs are shown in Figs. 4(a)–4(d): the training data set used for building each model and the validation data set that is unknown to each model. The optimizer data set represents all designs probed during an optimization session and is only shown in Fig. 4(d) as red ⅄'s. The training and validation sets were the same for all methods and contained 2500 designs each. While the training or validation set is randomly sampled, the optimizer set is biased towards better

designs not found in a random sampling, as can be seen in Fig. 4(d) by the lower values of $\bar{l}$.

The U-Net model showed consistent accuracy over the training, validation, and optimizer data sets with a Pearson correlation coefficient (PCC) [53] R of 0.98, 0.98 and 0.99, respectively. The CAE reached the second best performance with a correlation coefficient of 0.82. This is expected as the CAE and U-Net share the way they handle the inference: They match spatial distributions (i.e., maps) of the dielectric permittivity and leakage fields. Maps of the dielectric permittivity are directly built from the design parameters described in Sec. II C, while the produced leakage field maps are spatially averaged to obtain the mean leakage $\bar{l}$. The relationship between the dielectric and the leakage field spatial distributions proves to be well modeled as shown in Fig. 4(e).

The performance gap between the CAE and U-Net can be observed through the predicted leakage maps of Fig. 4(e). The only architectural difference brought by the U-Net is the presence of skip connections. Due to these, U-Net predicts finer local design features accurately, which is an expected property of U-Net [16]. Some nonphysical field artifacts are present in the CAE predictions, while they do not appear with U-Net.

Contrasting with CAE and U-Net approaches, DNNs and Kriging partial least-squares (KPLS) regression models [24,31] attempt to predict the mean leakage $\bar{l}$ directly starting from the geometric parameters of the dielectric structure. These approaches failed, barely reaching a moderate correlation ($R = 0.5$) with this large training data set. This poor result is mainly due to the complexity of the interaction between the geometric structure parameters and the mean leakage. The KPLS interpolation is particularly ineffective as the mean leakage reacts abruptly to many of the geometric parameters.

Our approach yields a robust deep learning model, capable of being repurposed for future predictions, even when applied to a similar but different devices. Specifically, the model trained on nanofins serves as a solid foundation for making predictions in the annular-groove case and vice versa. Moreover, this network is also valuable for tasks such as inverse design or Shapley additive explanation analysis, as introduced in the work of Lundberg and Lee [54].

The key takeaway message is the superior accuracy of U-Net, with $R \geqslant 0.95$. These results strongly justify the choice of U-Net as a surrogate for FDTD simulations in photonics, particularly when characteristic slices can be extracted to reduce the complexity and size of the model.

B. Influence of the data-set size

While the accuracy displayed by U-Net is compelling for optimization, its benefits have to be balanced against the number of simulations required to train it. In fact, machine learning models generally require thousands of simulations, with a great dependence on the problem at hand. It is difficult, if not impossible, to meet this expectation for numerically expensive simulations. This study aims to define the minimal number N_0 of simulations required to initiate a coarse U-Net model that brings the validation PCC R above 0.5. This particular value was selected based on observation of the field maps and corresponds to a moderate correlation in statistics.

The efficiency of the U-Net predictions is assessed in Fig. 5, where the model is trained with increasingly scarce training data. The validation set PCC R is plotted against a decreasing number of training set simulations. The training and validation sets are obtained by splitting a common source data set of 5000 simulations from randomly chosen designs.

Figure 5 shows that U-Net maintains an $R \geqslant 0.5$ for the validation set for as few as $N_0 = 100$ simulations in the training set. Even better results can be achieved through data augmentation. Data augmentation [55] helps to improve the robustness of machine learning models by exposing them to a greater variety of data, which enables them to better generalize to unseen data by coping with noise, variations, and biases. In our case, the data are augmented by cropping and rotating the designs and corresponding polarization leakage maps at random angles while training. U-Net with data augmentation is shown to achieve $R \geqslant 0.5$ with as few as $N_0 = 50$ simulations (Fig. 5, on the right). For large simulation data sets, on the other hand, data augmentation has a limited impact on the validation PCC (Fig. 5, on the left). This is expected as the available data become sufficient to fully train the network. In short, data augmentation proves to be a key tool to handle the small data-set sizes at the beginning of an optimization process.

C. Performance analysis of vortex-phase-mask designs

Figure 6 compiles the optimizations of the annular-groove (AG) and the nanofin (NF) vortex phase mask designs for the above-discussed surrogate optimization method and the direct PSO optimizer (directly using the FDTD solver instead of U-Net to evaluate the figure of merit $\bar{l}$). The four resulting processes are labeled as follows: AG-D and AG-S optimize the annular-groove design with direct and surrogate optimizers, respectively, while NF-D and NF-S optimize the nanofin pattern. Each marker in Fig. 6 corresponds to a full optimization process and shows the optimal performance computed using the FDTD solver with regard to the number of evaluations required. For efficient use of computational resources, optimizers were stopped when no progress happened during 40 evaluations using simulations. This condition corresponds roughly to two iterations for the surrogate optimizer (AG-S or NF-S) as well as for the direct optimizer (AG-D or NF-D).

Figure 6 shows how efficient our surrogate optimizer approach is when compared to direct optimization. In the NF-S case, surrogate optimization finds equivalent or better solutions in four times fewer evaluations than in the NF-D case. The surrogate method is also more reliable, with two-thirds of processes ending up with a polarization leakage below 0.04, while the direct optimizer only has one-fourth.

Another noteworthy aspect is the reduced leakage observed in annular-groove designs, suggesting that the concentric rings are superior in converting right-handed circular polarization and generating the anticipated helical phase ramp compared to the nanofins. This also appears in the inset nanofin designs, for two of them (A and C) mimic an annular groove. This systematic approach confirms that the highest geometrically induced anisotropy is achieved by the 1D grating, even at the center.

For the annular-groove optimization, results are similar on average to the ones reported in a previous work [13]

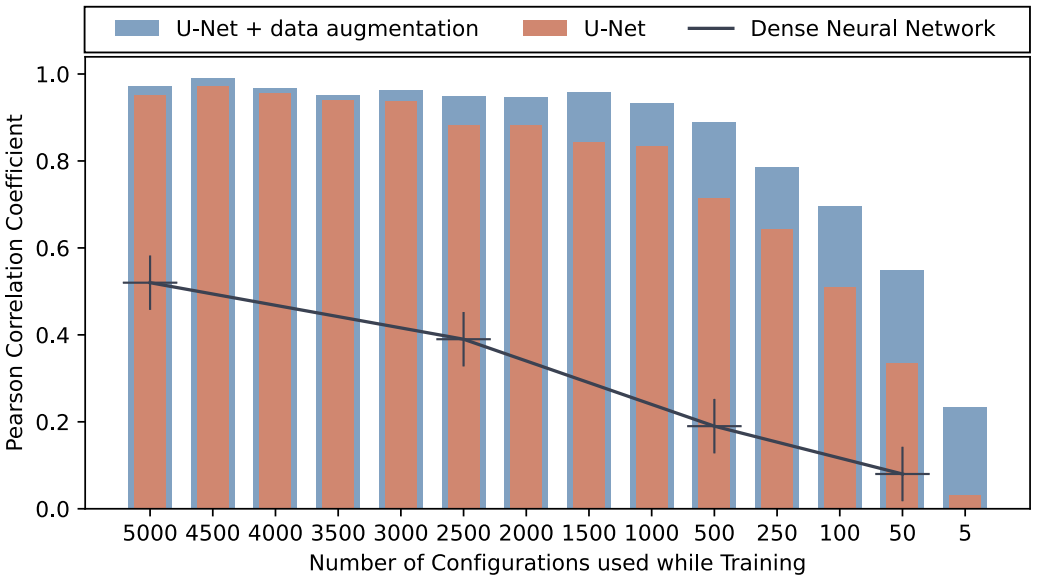

FIG. 5. Regression quality with varying data-set size. The PCC on the validation set is shown for different sizes of the training set. Five randomly seeded trainings were considered for each data-set size. The data augmented training (blue) is compared to a standard training (orange). The performance of the simpler dense neural network is charted in black for a benchmark.

($\bar{l} = 0.02$), where an annular-groove pattern (defined by two parameters) was devised by plotting the leakage value for 25 values of both parameters, amounting to 625 evaluations. Best designs of the direct optimizer (E) managed to obtain leakage $\bar{l}$ of 0.0143, which represents a 25% improvement with respect to previous work [13]. Meanwhile, the surrogate optimizer enabled a slightly lower (better) leakage (D) of 0.017 with a similar simulation budget. The superior efficiency of the direct optimizer in this case is expected since it operates within a smaller search space, consisting of only 10 free parameters, compared to the nanofin design, which involves 38 parameters. Moreover, designs D and E

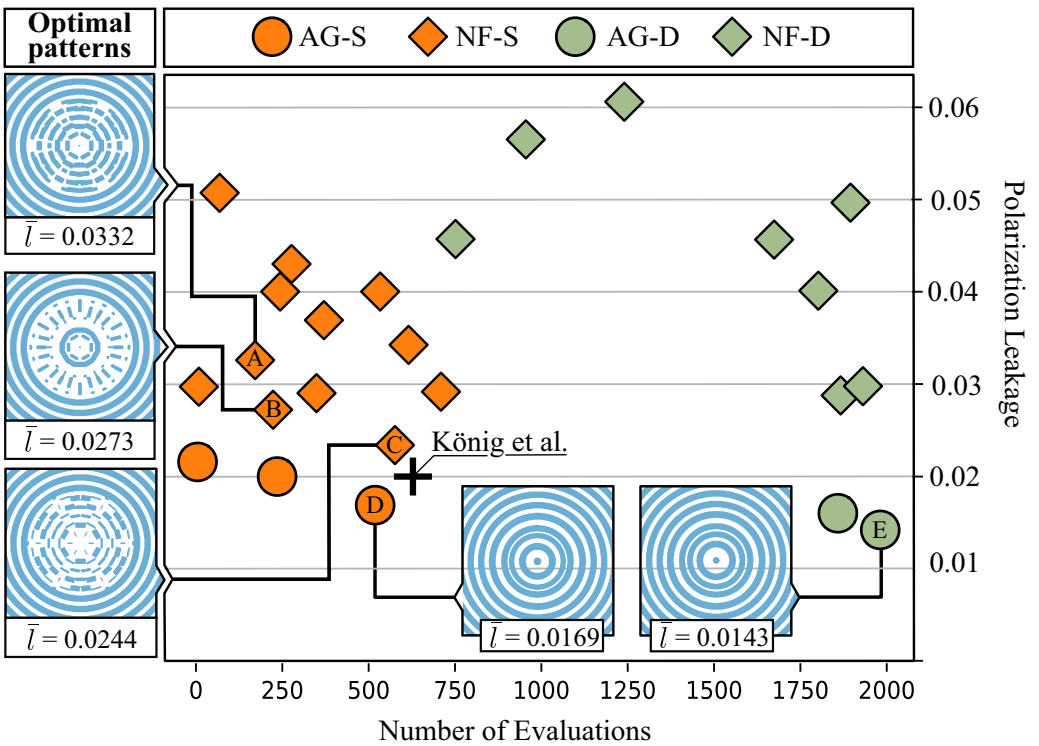

FIG. 6. Overview of the optimization performance. Optimal polarization leakage is plotted against the number of evaluations required in four optimizations schemes. An overview of some designs, including the best one for both the AG-S and NF-S processes, is provided on the left. Previously published results [13] are plotted as König *et al.*

are very practically similar, suggesting that the difference of performance originates from small variations in the design. Still, the surrogate optimizer found it with four times fewer evaluations.

While many produced designs tend to imitate an annular-groove pattern such as in Fig. 6, designs A and C, some of the best performing ones use a mix of agglomerated fins parallel to or orthogonal with the external annular groove. This is a sensible design as the birefringence can also be produced by a radial pattern.

IV. CONCLUSION

This work has demonstrated the potential of global optimization methods in a new era of photonic design lead by novel machine learning and adjoint optimization methods. Specifically, we have highlighted the role of evolutionary optimization (PSO), which enables a stable global exploration of the search space. The staggering need for evaluations of evolutionary algorithms was alleviated due to a U-Net surrogate solver.

An already efficient surrogate FDTD solver (U-Net) was made even more efficient due to data augmentation. This surrogate provided swift and accurate evaluations of designs for a PSO global evolutionary optimization algorithm. The U-Net worked by matching characteristic slices of the design and figure of merit instead of considering the whole simulation domain, resulting in a simpler and faster model. The resulting surrogate optimization framework enabled optimization of designs with four times fewer simulations than required by the evolutionary algorithm. The optimization scheme was also shown to be more reliable given the lower variance in the results obtained when compared to direct optimization of the simulation.

The surrogate-based optimization was applied on two types of devices: an annular-groove and a nanofin pattern. An optimum was identified in the former case, achieving performance that is either comparable to or surpasses previous attempts. This optimization scheme is versatile: It could integrate other types of design parameters such as categorical parameters for choosing materials and shapes aside from geometric lengths. Moreover, as this scheme makes no preliminary hypothesis on the underlying simulation, it can be applied to a wide range of photonic design problems such as gratings, multiplexers or demultiplexers, scattering problems, and multilayered photonic crystals.

ACKNOWLEDGMENTS

Computational resources were provided by the Consortium des Équipements de Calcul Intensif, funded by the Fonds de la Recherche Scientifique de Belgique (FRS-FNRS) under Grant No. 2.5020.11 and by the Walloon Region. This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 819155). A.M., M.L., and O.A. are Research Associates of "Fonds de la Recherche Scientifique de Belgique (FRS-FNRS)." The authors thank Prof. O. Deparis (University of Namur) for generously dedicating their time and expertise to meticulously review this manuscript. Their insightful feedback and attention to detail greatly enhanced the clarity and accuracy of our work.

[1] J. B. Mueller, N. A. Rubin, R. C. Devlin, B. Groever, and F. Capasso, Metasurface polarization optics: Independent phase control of arbitrary orthogonal states of polarization, Phys. Rev. Lett. **118**, 113901 (2017).

[2] W. T. Chen, A. Y. Zhu, and F. Capasso, Flat optics with dispersion-engineered metasurfaces, Nat. Rev. Mater. **5**, 604 (2020).

[3] W. T. Chen, A. Y. Zhu, V. Sanjeev, M. Khorasaninejad, Z. Shi, E. Lee, and F. Capasso, A broadband achromatic metalens for focusing and imaging in the visible, Nat. Nanotechnol. **13**, 220 (2018).

[4] S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez, Inverse design in nanophotonics, Nat. Photon. **12**, 659 (2018).

[5] Z. Lin, R. Pestourie, C. Roques-Carmes, Z. Li, F. Capasso, M. Soljačić, and S. G. Johnson, End-to-end metasurface inverse design for single-shot multi-channel imaging, Opt. Express **30**, 28358 (2022).

[6] Z. Li, R. Pestourie, Z. Lin, S. G. Johnson, and F. Capasso, Empowering metasurfaces with inverse design: Principles and applications, ACS Photon. **9**, 2178 (2022).

[7] A. Arbabi, Y. Horie, M. Bagheri, and A. Faraon, Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission, Nat. Nanotechnol. **10**, 937 (2015).

[8] Z. Bomzon, G. Biener, V. Kleiner, and E. Hasman, Space-variant Pancharatnam–Berry phase optical elements with computer-generated subwavelength gratings, Opt. Lett. **27**, 1141 (2002).

[9] K. Koshelev and Y. Kivshar, Dielectric resonant metaphotonics, ACS Photon. **8**, 102 (2021).

[10] D. Mawet, P. Riaud, O. Absil, and J. Surdej, Annular groove phase mask coronagraph, Astrophys. J. **633**, 1191 (2005).

[11] D. Mawet, P. Riaud, J. Surdej, and J. Baudrand, Subwavelength surface-relief gratings for stellar coronagraphy, Appl. Opt. **44**, 7313 (2005).

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

[13] L. König, O. Absil, M. Lobet, C. Delacroix, M. Karlsson, G. O. de Xivry, and J. Loicq, Optimal design of the annular groove phase mask central region, Opt. Express **30**, 27048 (2022).

[14] B. Settles, Active learning literature survey, University of Wisconsin-Madison Department of Computer Sciences 2009 Technical Report Nr. 1648, https://minds.wisconsin.edu/handle/1793/60660.

[15] Z. Bodó, Z. Minier, and L. Csató, in *Active Learning and Experimental Design Workshop, in conjunction with AISTATS 2010, Sardinia, 2010*, edited by I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov [JMLR 16, 127 (2011)].

[16] O. Ronneberger, P. Fischer, and T. Brox, in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, edited by N. Navab, J. Hornegger, W. Wells, and A. Frangi, Lecture Notes in Computer Science Vol. 9351 (Springer, Cham, 2015), pp. 234–241.

[17] M. Chen, R. Lupoiu, C. Mao, D.-H. Huang, J. Jiang, P. Lalanne, and J. A. Fan, in *High Contrast Metastructures XI*, edited by C. J. Chang-Hasnain, J. A. Fan, and W. Zhou, *SPIE Proc.* Vol. 12011 (SPIE, Bellingham, 2022), p. 120110C.

[18] https://github.com/Kaeryv/Keever.

[19] https://github.com/Kaeryv/PRA23Suppl.

[20] D. Weile and E. Michielssen, Genetic algorithm optimization applied to electromagnetics: A review, IEEE Trans. Antennas Propag. **45**, 343 (1997).

[21] R. Eberhart and J. Kennedy, *Proceedings of the IEEE International Conference on Neural Networks, Perth, 1995* (IEEE, Piscataway, 1995), Vol. 4, pp. 1942–1948.

[22] J. H. Holland, Genetic algorithms and the optimal allocation of trials, SIAM J. Comput. **2**, 88 (1973).

[23] A. Mayer, H. Bi, S. Griesse-Nascimento, B. Hackens, J. Loicq, E. Mazur, O. Deparis, and M. Lobet, Genetic-algorithm-aided ultra-broadband perfect absorbers using plasmonic metamaterials, Opt. Express **30**, 1167 (2022).

[24] M. A. Bouhlel, J. T. Hwang, N. Bartoli, R. Lafage, J. Morlier, and J. R. R. A. Martins, A python surrogate modeling framework with derivatives, Adv. Eng. Softw. **135**, 102662 (2019).

[25] H. Wang, Y. Jin, and J. Doherty, Committee-based active learning for surrogate-assisted particle swarm optimization of expensive problems, IEEE Trans. Cybern. **47**, 2664 (2017).

[26] A. I. Forrester and A. J. Keane, Recent advances in surrogate-based optimization, Prog. Aeronaut. Sci. **45**, 50 (2009).

[27] Y.-S. Ong, Z. Zhou, and D. Lim, in *Proceedings of the IEEE International Conference on Evolutionary Computation, Vancouver, 2006* (IEEE Computational Intelligence Society, Piscataway, 2006), pp. 2928–2935.

[28] Y. Augenstein, T. Repän, and C. Rockstuhl, Neural operator-based surrogate solver for free-form electromagnetic inverse design, ACS Photon. **10**, 1547 (2023).

[29] A. M. Hammond, A. Oskooi, M. Chen, Z. Lin, S. G. Johnson, and S. E. Ralph, High-performance hybrid time/frequency-domain topology optimization for large-scale photonics inverse design, Opt. Express **30**, 4467 (2022).

[30] M. J. D. Powell, in *Advances in Numerical Analysis: Wavelets, Subdivision Algorithms, and Radial Basis Functions*, edited by W. Light (Oxford University Press, Oxford, 1992), Vol. 11, Chap. 3, pp. 105–210.

[31] M. A. Bouhlel, N. Bartoli, A. Otsmane, and J. Morlier, Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction, Struct. Multidiscip. Optim. **53**, 935 (2016).

[32] R. O. L. Breiman, J. Friedman, and C. Stone, *Classification and Regression Trees* (Wadsworth, Belmont, 1988).

[33] T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016* (ACM, New York, 2016), pp. 785–794.

[34] G. Sun and S. Wang, A review of the artificial neural network surrogate modeling in aerodynamic design, Proc. Inst. Mech. Eng. G **233**, 5863 (2019).

[35] D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer, Machine learning–accelerated computational fluid dynamics, Proc. Natl. Acad. Sci. USA **118**, e2101784118 (2021).

[36] M. Chen, R. Lupoiu, C. Mao, D.-H. Huang, J. Jiang, P. Lalanne, and J. A. Fan, High speed simulation and freeform optimization of nanophotonic devices with physics-augmented deep learning, ACS Photon. **9**, 3110 (2022).

[37] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signal Syst. **2**, 303 (1989).

[38] Y. Lu and J. Lu, in *Advances in Neural Information Processing Systems 33*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran, Red Hook, 2020), pp. 3094–3105.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai *et al.*, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran, Red Hook, 2019), pp. 8024–8035.

[40] D. M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. **44**, 1 (2004).

[41] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*, edited by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells (Springer, Cham, 2016), pp. 424–432.

[42] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, in *Artificial Neural Networks and Machine Learning—ICANN 2011*, edited by T. Honkela, W. Duch, M. Girolami, and S. Kaski (Springer, Berlin, 2011), pp. 52–59.

[43] C. Delacroix, P. Forsberg, M. Karlsson, D. Mawet, O. Absil, C. Hanot, J. Surdej, and S. Habraken, Design, manufacturing, and performance analysis of mid-infrared achromatic half-wave plates with diamond subwavelength gratings, Appl. Opt. **51**, 5897 (2012).

[44] C. Delacroix, O. Absil, D. Mawet, C. Hanot, M. Karlsson, P. Forsberg, E. Pantin, J. Surdej, and S. Habraken, A diamond AGPM coronagraph for VISIR, Proc. SPIE **8446**, 84468K (2012).

[45] D. Mawet, O. Absil, C. Delacroix, J. H. Girard, J. Milli, J. O'Neal, P. Baudoz, A. Boccaletti, P. Bourget, V. Christiaens, P. Forsberg, F. Gonte, S. Habraken, C. Hanot, M. Karlsson, M. Kasper, J. L. Lizon, K. Muzic, R. Olivier, E. Peña *et al.*, L'-band AGPM vector vortex coronagraph's first light on VLT/NACO: Discovery of a late-type companion at two beamwidths from an F0V star, Astron. Astrophys. **552**, L13 (2013).

[46] D. Defrère, O. Absil, P. Hinz, J. Kuhn, D. Mawet, B. Mennesson, A. Skemer, K. Wallace, V. Bailey, E. Downey, C. Delacroix, O. Durney, P. Forsberg, C. Gomez Gonzales, S. Habraken, W. F. Hoffmann, M. Karlsson, M. Kenworthy, J. Leisenring, M. Montoya *et al.*, L'-band AGPM vector vortex coronagraph's first light on LBTI/LMIRCam, Proc. SPIE **9148**, 91483X (2014).

[47] E. Serabyn, E. Huby, K. Matthews, D. Mawet, O. Absil, B. Femenia, P. Wizinowich, M. Karlsson, M. Bottom, R. Campbell, B. Carlomagno, D. Defrère, C. Delacroix, P. Forsberg, C. Gomez Gonzalez, S. Habraken, A. Jolivet, K. Liewer, S. Lilley, P. Piron *et al.*, The W. M. Keck Observatory

infrared vortex coronagraph and a first image of HIP 79124 B, Astron. J. **153**, 43 (2017).

[48] A. F. Oskooi, D. Roundy, M. Ibanescu, P. Bermel, J. D. Joannopoulos, and S. G. Johnson, MEEP: A flexible free-software package for electromagnetic simulations by the FDTD method, Comput. Phys. Commun. **181**, 687 (2010).

[49] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevA.109.013514 for dielectric and leakage profiles for some of the best designs (S1), details for the reproducibility of the surrogate optimization procedure (S2), details for the reproducibility of the FDTD simulations (S3), and visualization of the evolution of U-Net accuracy during the surrogate optimization procedure (S4).

[50] H.-M. Gutmann, A radial basis function method for global optimization, J. Global Optim. **19**, 201 (2001).

[51] N. Roy, C. Beauthier, and A. Mayer, *Proceedings of the 2022 IEEE Congress on Evolutionary Computation, Padua, 2022* (IEEE, Piscataway, 2022), pp. 1–8.

[52] J. A. Hartigan and M. A. Wong, A K-means clustering algorithm, Appl. Stat. **28**, 100 (1979).

[53] *SPSS tutorials: Pearson correlation*, https://libguides.library.kent.edu/SPSS/PearsonCorr (Kent State University, Kent, 2023).

[54] S. M. Lundberg and S.-I. Lee, in *Advances in Neural Information Processing Systems 30, Long Beach, 2017*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran, Red Hook, 2017).

[55] C. Shorten and T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data **6**, 60 (2019).