Supplementary Information for

Evolution without iteration: Collapsing discrete and continuous photonic design spaces for differentiable global search

Haosong Huang[1], Zengyang Gao[3], Mingshuo Wang[4], Ye Guo[5], Jingang Zhang[2*], Ce Shang[4*], Yunfeng Nie[3*]

[1*]School of Aerospace Science and Technology, Xidian University, Xi'an, 710071, China.
[2]School of Future Technology, University of Chinese Academy of Sciences, Beijing, 100039, China.
[3]Brussels Photonics, Department of Applied Physics and Photonics, Vrije Universiteit Brussel, Brussels, 1050, Belgium .
[4]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100094, China .
[5]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 100039, China.

*Corresponding author(s). E-mail(s): zhangjg@ucas.ac.cn; shangce@aircas.ac.cn; Yunfeng.Nie@vub.be;

Contents

Supplementary Note 1. Comparison with existing inverse design methods **3**

Supplementary Note 2. Metasurface experiment configuration and surrogate modeling **4**
2.1 Physical configuration and dataset generation 4
2.2 FDTD simulation protocols 5
2.3 Physics-aware feature expansion and dual-stream surrogate architecture 7
2.3.1 Surrogate modeling: raw input vectorization strategy 7
2.3.2 Baseline: ResDNN and limitations 7
2.3.3 λ-ResNet: multi-dimensional advantages of point-wise wavelength prediction 8
2.3.4 Architecture implementation: Dual-stream net 8
2.4 Meta-Candidate Generation via Sparse Feature Clustering 9

Supplementary Note 3. Thin-film design space and candidate generation **11**
3.1 Material library and physical constraints 11
3.2 Diversity-driven hierarchical generation 11
3.3 Thin-film pruning process tracking 13

Supplementary Note 4. Computational complexity and scalability analysis **14**
4.1 Benchmarking metrics and definitions 14
4.2 Empirical Performance Evaluation 14
4.2.1 Impact of forward model complexity on scalability 14
4.2.2 System latency and dynamic pruning 15
4.2.3 Analysis of scalability trends 15

Supplementary Note 5. Effects of OptoNAS hyperparameters **16**
5.1 Pruning controller logic and convergence dynamics 16

Supplementary Note 6. Network training and implementation details **17**
6.1 Continuous variable constraints 17
6.2 VRAM memory optimization 17
6.3 Finetuning stage (stationary optimization) 18
6.4 Details of metasurface network training and results 18
6.5 Details of thin-film network training and results 19

Supplementary Note 1. Comparison with existing inverse design methods

Photonic inverse design has become an important strategy for metasurface optimization by identifying physical structures that satisfy target optical or system-level requirements. With advances in fabrication, electromagnetic modelling, and computation, a broad range of optimization methods has been developed for metasurface design. At a general level, existing methods can be grouped into gradient-based approaches for continuous variables, heuristic search methods for discrete or combinatorial variables, and end-to-end approaches that jointly optimize the optical frontend together with the backend computational procedure. As summarized in Supplementary Table S1, these representative methods are compared according to the degree of freedoms of continuous and discrete parameters, whether the optimization pipeline is differentiable, and whether the optimization is carried out at the component level or at the end-to-end system level. The design level distinguishes methods that optimize individual meta-atoms or isolated optical components for optical responses from those that optimize the full optical–computational system for final task performance.

Supplementary Table S1: Comparison of OptoNAS and representative metasurface optimization strategies.

Reference	Methodology	Continuous DoF	Discrete DoF	Differentiability	Design level	Application
[S1]	Genetic algorithm (GA) with Green Dyadic Method	2	1	No	Component	High scattering efficiency
[S2]	GA with finite-difference time-domain (FDTD)	1	1	No	Component	Infrared (IR)-absorbing visible (VIS)-transparent filter
[S3]	Topology optimization with FDTD	2	N/A	Yes	Component	High focusing efficiency metalens
[S4]	Adjoint optimization with Chebyshev-interpolation surrogate model	2	N/A	Yes	Component	High RGB focusing efficiency metalens
[S5]	Particle swarm optimization (PSO) with meta-atom library	2	N/A	No	Component	High efficiency polychromatic imaging
[S6]	GA with FDTD	1	N/A	No	Component	RGB color router
[S7]	Hybrid particle swarm optimization-genetic algorithm (PSO-GA) with full-wave simulation	3	1	No	Component	Spectro-polarimetric demultiplexing
[S8]	Neural network-based surrogate optimization	1	N/A	Yes	System	Large field of view RGB imaging
[S9]	Adjoint end-to-end optimization with the method of moving asymptotes (MMA) and meta-atom library	2	N/A	Yes	System	Wavelength–polarization-multiplexed holography
[S10]	End-to-end mixture probability sampling network (MPSN) with pretrained network and mixture density network (MDN)	4	N/A	Yes	System	Structural color inverse design
OptoNAS	Differentiable search in supernet architecture with dynamic pruning	8	5	Yes	System	Computational hyperspectral imaging

DoF denotes the degree of freedom.

A first class of methods performs optimization over continuous structural parameters, where differentiable optimization schemes are coupled with electromagnetic solvers or surrogate models that evaluate the optical response of candidate structures. For example, topology optimization with finite-difference time-domain (FDTD) simulation has been used to refine meta-atom geometries for high focusing efficiency metalens design[S3]. Adjoint optimization with a Chebyshev-interpolation surrogate model was further used to optimize meta-atom dimensions for RGB focusing[S4]. These methods are representative of component-level optimization based on continuous structural refinement. However, because they rely

on differentiability with respect to continuous design variables, they are generally less suited to directly handling discrete structural choices, such as material selection in metasurfaces.

A second class of methods relies on heuristic search when the design space includes discrete candidates or heterogeneous structural combinations. In these methods, the optimization algorithm is used to explore candidate structures, while the corresponding optical responses are obtained from electromagnetic solvers or precomputed libraries for fitness evaluation. For example, genetic-algorithm-based optimization was combined with the Green Dyadic Method to optimize meta-atom geometry and arrangement for polarization-dependent scattering control[S1], while related GA-based schemes using finite-difference time-domain (FDTD) evaluation were applied to visible/infrared (VIS/IR) filtering and RGB color routing[S2, 6]. Particle swarm optimization (PSO) using a precomputed meta-atom library has also been used for polychromatic imaging design[S5], and hybrid PSO–GA optimization with full-wave simulation was introduced to enlarge the jointly optimized structural space for spectro-polarimetric demultiplexing[S7]. Although these heuristic search methods can incorporate discrete variables, the lack of gradient information makes local refinement less efficient and can limit optimization precision in complex design spaces. Moreover, both classes of methods are mainly performed at the level of the metasurface component itself, rather than through joint optimization with the computational backend at the system level.

More recent efforts have shifted from component-level design to end-to-end system optimization, where the metasurface is optimized together with the computational backend. In this setting, the target is no longer only a local optical response of a meta-atom or device, but the final performance of the complete imaging or reconstruction pipeline. For example, a neural-network-based surrogate model was used to optimize the radius of meta-atoms for large-field-of-view RGB imaging by shaping the system point spread function in a task-oriented manner[S8]. Adjoint end-to-end optimization has also been combined with a meta-atom library to optimize metasurface parameters for wavelength- and polarization-multiplexed holography, using the method of moving asymptotes (MMA) as the optimizer[S9]. An end-to-end mixture probability sampling network (MPSN) was also reported for structural color inverse design, where a pretrained network is integrated with a mixture density network (MDN) to generate Gaussian mixture distributions of structural parameters, from which multiple structural candidates are sampled and evaluated jointly during training[S10]. This strategy improves the exploration of non-unique solution spaces and enables end-to-end optimization over multiple candidates within a predefined structural family. However, despite these advances, the optimized variable space in existing end-to-end approaches generally remains limited to a small number of differentiable parameters or predefined candidate settings within a single structural family.

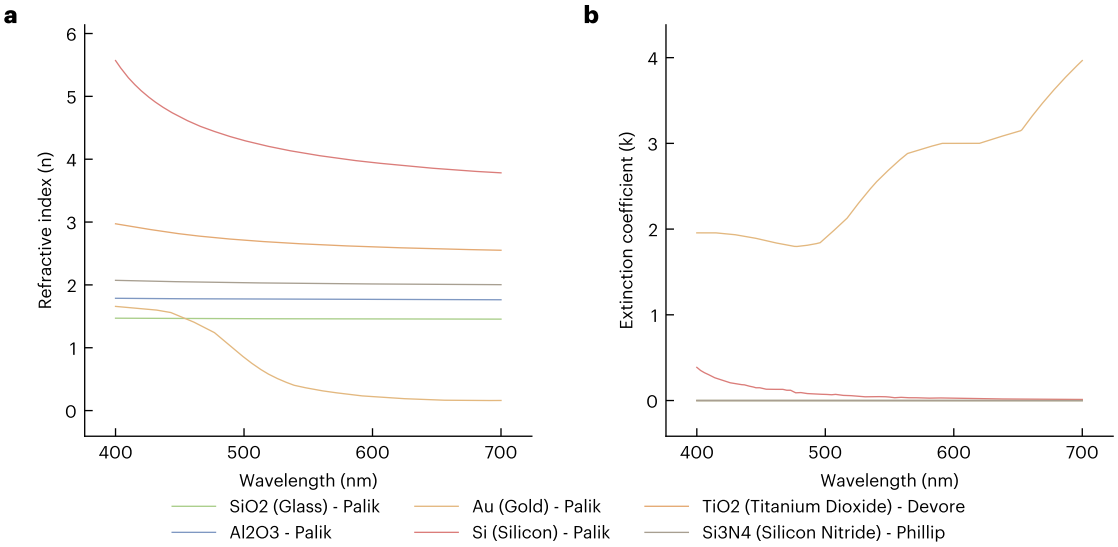
In contrast, OptoNAS is designed for end-to-end optimization in a higher-dimensional metasurface design space that includes both continuous and discrete parameters. Rather than restricting the search to one predefined structural family, it supports heterogeneous candidate structures within a unified supernet and jointly optimizes geometric parameters, structural types, and material-related choices under system-level objectives. In addition, by maintaining all active candidates in the differentiable optimization graph throughout training, OptoNAS enables end-to-end optimization over multiple candidates simultaneously, which improves the exploration of the search space and increases the chances of identifying an optimal system-level solution. In this sense, OptoNAS connects structural diversity at the component level with task-driven optimization at the system level, while maintaining differentiable training through supernet relaxation and dynamic pruning.

Supplementary Note 2. Metasurface experiment configuration and surrogate modeling

This section provides a detailed breakdown of the physical experimental setup, the dataset generation protocols, and the surrogate modeling framework used to accelerate the inverse design process.

Supplementary Note 2.1 Physical configuration and dataset generation

Unit cell architecture and fabrication constraints. The metasurface architecture is composed of sub-wavelength unit cells arranged with a fixed periodicity of $P = 400\,\mathrm{nm}$ along both lateral axes (x and y). This sub-wavelength periodicity was selected to suppress higher-order diffraction modes within the target visible spectrum ($\lambda \in [400, 700]\,\mathrm{nm}$), ensuring the device operates strictly in the zeroth-order transmission regime. To guarantee fabrication feasibility via standard lithography and etching processes, we impose a uniform height constraint: within any single design realization, all meta-atoms share a constant etch depth (h_{atom}). The structural topology of each meta-atom is selected from a set of four



Supplementary Figure S1: **Optical dispersion properties of the metasurface materials. a**, Refractive index (n), and **b**, extinction coefficient (k), plotted across the visible spectrum (400–700 nm). Dispersion data for the six candidate materials are sourced from the Lumerical FDTD 2025 R1 software database.

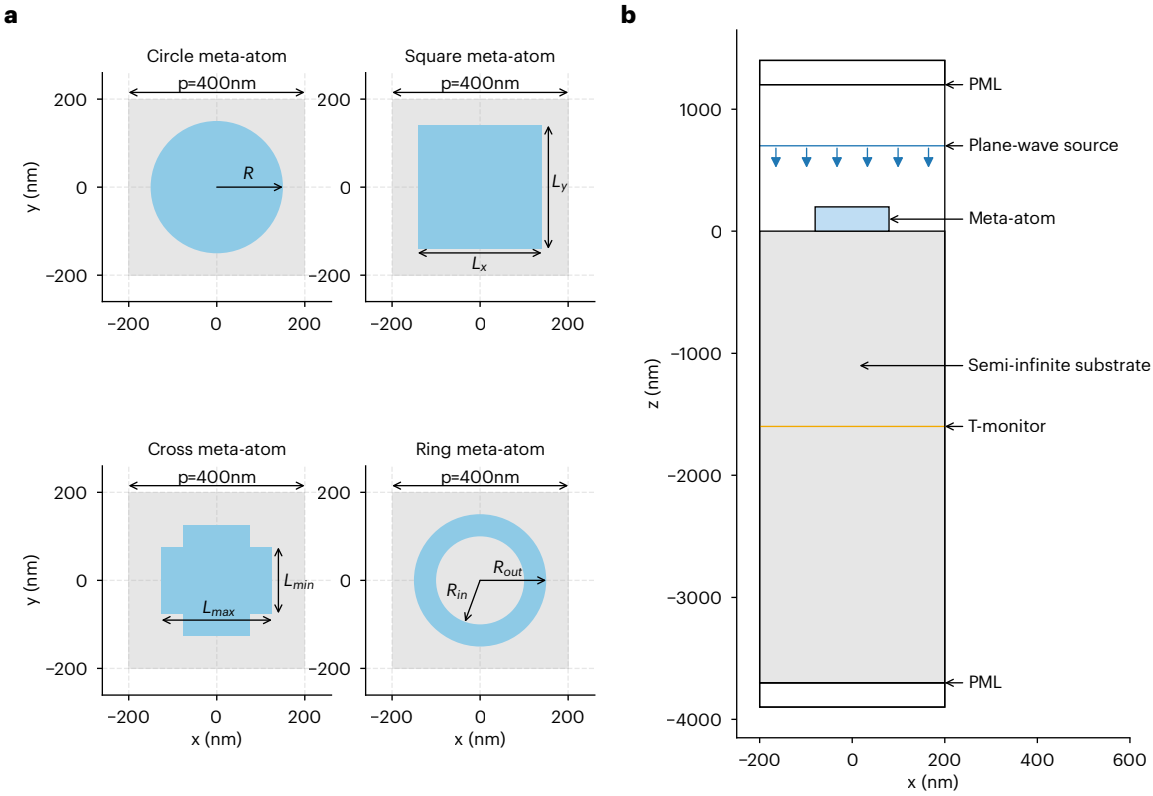
geometric primitives—cylinder, rectangular cuboid, ring, and cross—chosen to provide a versatile range of polarization sensitivities and spectral resonance characteristics.

Material selection. To encompass distinct optical regimes, we curated a material library comprising both plasmonic metals and high-index dielectrics. Refractive index and extinction coefficient data for these materials were acquired from Lumerical FDTD 2025 R1 software database. For the substrate, aluminum oxide (Al_2O_3) and silicon dioxide (SiO_2) were selected due to their low refractive indices and high optical transparency across the visible spectrum. The meta-atom library includes gold (Au), chosen to exploit localized surface plasmon resonances (LSPR) for broadband modulation, as well as high-index dielectrics including silicon (Si), titanium dioxide (TiO_2), and silicon nitride (Si_3N_4). These dielectric candidates were selected to support Mie-type magnetic and electric dipole resonances, offering high spectral selectivity with significantly lower absorption losses than their metallic counterparts. The underlying refractive index and extinction coefficient values of materials are drawn in Supplementary Figure S1.

Parameter space discretization. To construct a high-fidelity training dataset for the surrogate model, we performed a deterministic grid search over the valid design space, yielding a total of 9,240 unique physical configurations. The continuous parameters were discretized with step sizes chosen to balance spectral diversity against the resolution limits of standard nanofabrication. As illustrated in Supplementary Figure S2(a), four distinct meta-atom structures were used, each defined by its own set of optimization parameters. The geometric parameterization is as follows. The meta-atom height h_{atom} is swept from 100 to 300 nm in increments of $\Delta h = 20$ nm to modulate phase accumulation and resonance depth. For circular nanopillars, the radius R varies from 50 to 140 nm with a step size of 15 nm. Rectangular cuboids are parameterized by the lateral dimensions L_x and L_y, each varied independently from 100 to 280 nm in steps of 30 nm to enable polarization-dependent responses. Ring structures are defined by an outer radius R_{out} and an inner radius R_{in}; R_{out} is swept from 50 to 140 nm in 15 nm steps, while for each fixed R_{out}, R_{in} is dynamically sampled starting at 30 nm and increased in 30 nm increments subject to the constraint $R_{\text{in}} \leq R_{\text{out}} - 20$ nm to ensure mechanical stability and a physically realizable wall thickness. Cross-shaped structures are characterized by a major axis $L_{\max}$ and a minor axis $L_{\min}$;$L_{\max}$ ranges from 100 nm to 280 nm in 30 nm steps, and $L_{\min}$ is subsequently sampled in 20 nm increments from 100 nm up to the length of the major axis, with $L_{\min} \leq L_{\max}$.

Supplementary Note 2.2 FDTD simulation protocols

The ground-truth spectral responses for each meta-atom configuration were computed using Ansys Lumerical FDTD[11]. For each configuration, spectral responses were calculated under normal-incidence illumination, ensuring consistent comparison of transmission characteristics across diverse geometries and material combinations.



Supplementary Figure S2: **Meta-atom geometry**. (a) Four different planar meta-atom geometries and their corresponding seven topology-specific continuous learnable parameters. (b) Cross-sectional view of the simulation setup with a normally incident plane-wave source and a semi-infinite substrate.

Simulation domain and region layout. The FDTD simulation model was constructed with a carefully designed spatial layout to ensure accurate near-field evolution and to suppress numerical artifacts arising from finite computational boundaries. The simulation domain spanned a single lattice period of 400 nm × 400 nm in the lateral (x/y) directions, while being systematically defined along the z direction. As shown in Fig. S2(b), along the z axis, the computational domain comprised the air region, the functional substrate, a homogeneous host medium, and perfectly matched layer (PML) absorbing boundaries. Structural symmetry was exploited to further reduce the computational domain, as detailed below. On the air side, the meta-atom structures were positioned above the plane $z = 0$, with their maximum heights determined by the specific geometry and ranging from 60 to 300 nm. A broadband plane-wave source was injected along the $-z$ direction and placed above the meta-atoms. To suppress non-physical source–structure coupling and near-field truncation, the source was separated from the tallest meta-atom by approximately half of the maximum operating wavelength, while an additional free-space buffer of approximately one maximum wavelength was retained between the source and the upper PML boundary. This configuration ensured that the incident field interacted with the structure under near-ideal plane-wave conditions. On the substrate side, the functional substrate thickness was fixed at 100 nm. Beneath the substrate, a semi-infinite substrate layer was introduced. The transmitted field was recorded using a two-dimensional z-normal frequency-domain monitor positioned 1.5 µm. At this depth, evanescent and near-field components were sufficiently attenuated such that the recorded signal was dominated by propagating modes. To further suppress numerical absorption artifacts, an additional propagation buffer with a thickness of approximately 2–3 times the maximum operating wavelength was inserted between the transmission monitor and the lower PML boundary. The T-monitor tracks the transmittance response of the meta-atom in $\lambda \in [400, 700]$nm range with 301 sample points under equal 1nm wavelength spacing.

Boundary conditions and symmetry reduction. All four classes of meta-atom structures investigated in this work exhibited C_4 rotational symmetry. Accordingly, numerical simulations were performed on only a single quadrant of the unit cell, with appropriate symmetry boundary conditions applied to reconstruct the response of an infinitely periodic array. Specifically, anti-symmetric boundary conditions were applied on both boundaries along the x direction, while symmetric boundary conditions were applied

on both boundaries along the y direction. This boundary configuration preserved the geometric symmetry and the correctness of the optical response, while effectively reducing the computational domain to one quarter of the full unit cell, thereby significantly lowering the computational cost.

Mesh refinement and conformal interface treatment. All FDTD simulations employed the Conformal Variant 1 mesh refinement scheme provided by the Lumerical FDTD solver to enhance the geometric fidelity of curved surfaces and material interfaces. This approach introduced conformal corrections at grid cells intersecting material boundaries while preserving the underlying Cartesian Yee-grid topology, enabling a more accurate numerical representation of the true geometric interfaces. In contrast to conventional staircase meshing, where each grid cell is approximated as a single material, and interfaces are forced to align with the grid, the conformal mesh applied localized effective-material corrections at interfaces, thereby significantly reducing discretization-induced geometric errors at a given spatial resolution without requiring excessive global mesh refinement. The structures considered here—including circular, rectangular, ring, and cross geometries—generally exhibited subwavelength features, sharp corners, and high-curvature boundaries, and their optical responses were highly sensitive to the local electromagnetic field distributions near material interfaces. Implementing Conformal Variant 1 enabled a more precise characterization of these critical regions, thereby improving the coupling accuracy between near-field distributions and propagating modes and enhancing the numerical consistency of transmission spectra across varying geometric parameters. For these reasons, this mesh treatment was adopted as the default strategy for all simulations reported in this work.

Supplementary Note 2.3 Physics-aware feature expansion and dual-stream surrogate architecture

To achieve high-fidelity, low-latency spectral prediction in end-to-end inverse design, we propose a novel surrogate model architecture termed the dual-stream net. This section details the data input strategy and the design evolution logic of this architecture: first defining the vectorization scheme for heterogeneous parameters, subsequently establishing a point-wise wavelength prediction strategy to overcome the physical limitations of the baseline model, and finally resolving the computational bottlenecks introduced by this strategy through a decoupled architecture. For each meta-atom configuration, surrogate models generate predictions at 31 uniform wavelength intervals within the visible spectrum($\lambda \in [400, 700]$nm) to ensure consistency with the experimental hyperspectral dataset.

Supplementary Note 2.3.1 Surrogate modeling: raw input vectorization strategy

A central challenge in constructing a surrogate model within this design space is addressing the heterogeneity of input parameters, which comprise both discrete material categorical variables and continuous geometric structure parameters. To establish an efficient and unified data processing pipeline, we define a standardized raw vectorization scheme that maps any valid physical configuration into a compact, fixed-dimensional feature vector $\mathbf{v}_{raw} \in \mathbb{R}^{10}$.

This vector $\mathbf{v}_{raw}$ is concatenated from integer-encoded categorical identifiers and normalized continuous geometric parameters:

$$\mathbf{v}_{raw} = [m_{sub}, m_{atom}, h_{atom}, \mathbf{p}_{geo}] \tag{1}$$

$$\mathbf{p}_{geo} = [L_x, L_y, L_{min}, L_{max}, R, R_{in}, R_{out}] \tag{2}$$

To maximize storage efficiency, material configurations employ label encoding instead of sparse one-hot encoding. The substrate material index $m_{sub} \in \{0, 1\}$ corresponds to the set $\{\mathrm{SiO_2}, \mathrm{Al_2O_3}\}$, while the meta-atom material index $m_{atom} \in \{0, 1, 2, 3\}$ corresponds to the set $\{\mathrm{Au}, \mathrm{Si}, \mathrm{TiO_2}, \mathrm{Si_3N_4}\}$.

Here, h_{atom} denotes the height of the meta-atom. Geometric parameters are encoded within an allocated space $\mathbf{p}_{geo} \in \mathbb{R}^7$, where each dimension strictly corresponds to a specific shape attribute (e.g., cylinder radius, cuboid edge lengths). For parameter slots undefined by the current topology, we implement a zero-padding mechanism. For example, a meta-atom with a circular shape will have non-zero values for the circle radius R and zero values for all other six p_{geo} parameters. This standardized 10-dimensional format constitutes the foundational interface for the data loader.

Supplementary Note 2.3.2 Baseline: ResDNN and limitations

For comparative analysis, we first examine the baseline model (ResDNN). This model adopts an "early concatenation" and "full-spectrum output" strategy, where geometric parameters and material IDs (one-hot encoded) are concatenated at the input stage. Subsequently, a deep backbone network comprising 10 residual blocks (width of 128) directly regresses the 31-dimensional full-spectrum vector in a single

pass. Despite its capacity for deep feature extraction, the model suffers from two fundamental physical limitations when processing complex metasurface spectra. First, one-hot encoding severs the inherent physical correlations between materials and fails to natively represent the dispersion characteristics of the refractive index across varying wavelengths, forcing the network to expend a substantial number of parameters fitting an unstructured mapping rather than learning the underlying physical laws. Second, deep neural networks universally exhibit spectral bias, tending to preferentially fit low-frequency, smooth components. In the direct full-spectrum output mode, even with increased network depth (e.g., 10 residual blocks), the model struggles to acutely capture the sharp Fano resonances in metasurface transmission spectra, resulting in the smoothing or complete loss of high-frequency resonance peaks.

Supplementary Note 2.3.3 λ-ResNet: multi-dimensional advantages of point-wise wavelength prediction

To overcome the aforementioned challenges, we discard the paradigm of one-shot full-spectrum regression in favor of a point-wise wavelength prediction strategy. The introduction of this strategy is not merely intended to resolve a single fitting issue, but is rather rooted in profound considerations of physical principles and computational efficiency, aiming to construct a computational paradigm with the following dual advantages:

Capture of high-frequency features. The primary motivation for introducing a single-point wavelength λ is to leverage Fourier feature mapping:

$$\gamma(\lambda) = [\sin(2\pi \mathbf{B}\lambda), \cos(2\pi \mathbf{B}\lambda)] \tag{3}$$

By mapping the wavelength into a high-dimensional frequency space, we compel the network to focus on the high-frequency components of the spectrum. Combined with the raw wavelength λ_{raw} (utilized to eliminate periodic degeneracy), this strategy enables the model to acutely capture intricate resonance structures within the spectra.

Precise injection of physical ground truth. A critical design objective of this strategy is to eliminate the loss of physical information caused by discrete encoding. Through point-wise processing, we successfully construct an accurate dynamic lookup mechanism:

$$[n(\lambda), k(\lambda)] = \text{Lookup}(\text{MaterialID}, \lambda) \tag{4}$$

This mechanism permits the corresponding λ-ResNet model to directly access the exact optical constants (refractive index n and extinction coefficient k) at the queried frequency point, rather than relying on averaged values or static IDs. This rigorously ensures consistency between the input data and underlying physical laws, fundamentally resolving the representation deficiency of dispersion information.

Supplementary Note 2.3.4 Architecture implementation: Dual-stream net

Although point-wise wavelength prediction provides significant physical and algorithmic advantages, a naive "early concatenation" of geometric parameters with each wavelength point would lead to repeated computations of geometric features, resulting in a severe VRAM bottleneck. To this end, we devised the dual-stream net, implementing an asymmetric decoupled architecture to actualize this strategy efficiently.

Asymmetric dual-stream decoupling. To eliminate computational redundancy, we bifurcate the network into two functionally orthogonal computational streams. The first geometry stream employs a deep ResNet (6 layers, width 256) to process geometric parameters. Given that the geometric structure is invariant with respect to wavelength, this branch executes only once per sample, and therefore systematically decouples the most resource-intensive feature extraction process from spectral resolution. The second physics stream utilizes a lightweight MLP (2 layers, width 128) to process point-wise physical inputs. This branch is exclusively dedicated to modeling localized light-matter interactions at specific wavelengths. The geometric features are then passed through a Feature-wise Linear Modulation (FiLM) layer to generate affine transformation parameters, which subsequently modulate the feature distribution of the physics stream.

Hybrid skip connection. To preclude the attenuation of original geometric information induced by deep modulation, we introduce a hybrid skip connection at the network's terminus. We apply broadcasting to the unmodulated raw geometric embeddings and concatenate them directly prior to the final output layer. This architectural feature establishes a direct gradient backpropagation pathway, ensuring that the underlying geometric structural parameters bypass complex non-linear modulation layers to participate directly in the final prediction. This effectively guarantees training stability and the strict fidelity of geometric information.

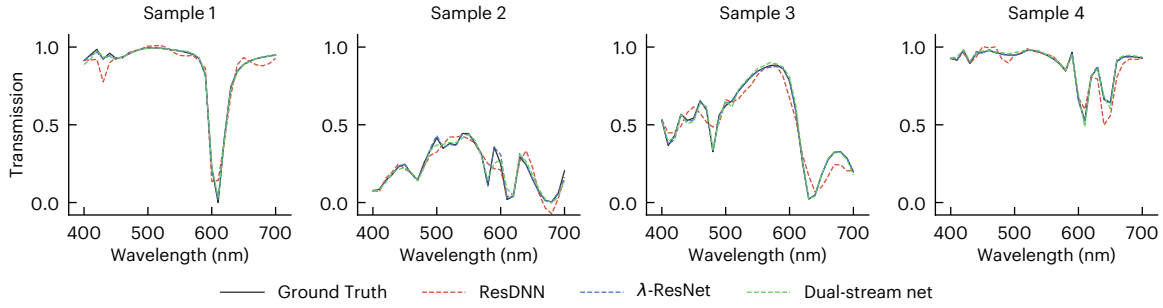
Computational overhead and accuracy comparison. Benefiting from the paradigm of feature sharing and the decoupled architecture, the dual-stream net maintains exceptionally high computational efficiency while delivering high-fidelity physical predictions. Supplementary Table S2 details the resource consumption during training alongside the overall Mean Squared Error (MSE) evaluated on the test set. Notably, the dual-stream network significantly enhances prediction accuracy with negligible computational overhead, despite its point-wise query mechanism. Such an architectural design drastically reduces the peak training memory footprint compared to the standard λ-ResNet approach. By decoupling the geometric processing from the wavelength-specific physical queries, the network avoids generating and storing highly redundant intermediate activation maps across the spectral dimension. Consequently, this reduced spatial complexity heavily shrinks the size of the computational graph, requiring far fewer intermediate states to be cached and calculated during backpropagation. Crucially, minimizing these cached activations prevents the surrogate from exhausting GPU VRAM during the critical backpropagation step, ensuring that end-to-end gradient flow to the physical parameters remains computationally tractable. Ultimately, the dual-stream framework achieves the high-precision modeling capabilities of a heavy, physics-informed network, but with a streamlined memory and time footprint that makes scalable, end-to-end inverse design computationally tractable.

Supplementary Table S2: **Computational overhead and accuracy comparison between ResDNN, λ-ResNet and Dual-stream net**

Model	Model VRAM	End-to-end VRAM	Calculation time	MSE
ResDNN (Baseline)	27.3 MB	1588.28 MB	0.78 ms	3.524e-03
λ-ResNet	150.6 MB	8986.11 MB	2.81 ms	8.661e-04
Dual-stream net (Ours)	92.1 MB	3260.41 MB	1.63 ms	8.643e-04

Note: End-to-end VRAM is tested under metasurface hyperspectral imaging experiments.

To illustrate the precision of the point-wise wavelength prediction strategy in capturing high-frequency features, Supplementary Figure S3 visualizes prediction results derived from randomly sampled instances in the test set. Compared to the baseline model (ResDNN), Dual-stream net fits sharp resonance peaks and intricate waveform details with markedly higher precision, thereby successfully circumventing the spectral bias intrinsic to conventional direct full-spectrum regression approaches.



Supplementary Figure S3: **Surrogate model prediction comparison.** Transmission spectra of four representative metasurface samples (400–700 nm). The exact numerical simulations are plotted alongside the corresponding predictions from the ResDNN, λ-ResNet, and Dual-stream net models to evaluate spectral reconstruction accuracy.

Supplementary Note 2.4 Meta-Candidate Generation via Sparse Feature Clustering

Feature space definition and metric quantification. To ensure the candidate pool exhibits maximal diversity across critical performance regimes, we developed a two-dimensional feature space that characterizes the spectral behavior of each meta-atom configuration. Let $\mathbf{S} \in \mathbb{R}^{N \times L}$ denote the spectral response of a super-pixel containing N meta-atoms across L wavelength points, and let $\bar{S}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} S_i(\lambda)$ be the spatially averaged super-pixel response. We define the feature vector $\mathbf{f} = [\eta, \rho]$ based on the following orthogonal metrics:

Optical Efficiency (η): This metric represents the integrated photon throughput of the device, ensuring the inclusion of candidates that balance resonance quality with signal strength. It is defined as the mean transmission intensity over the operational bandwidth:

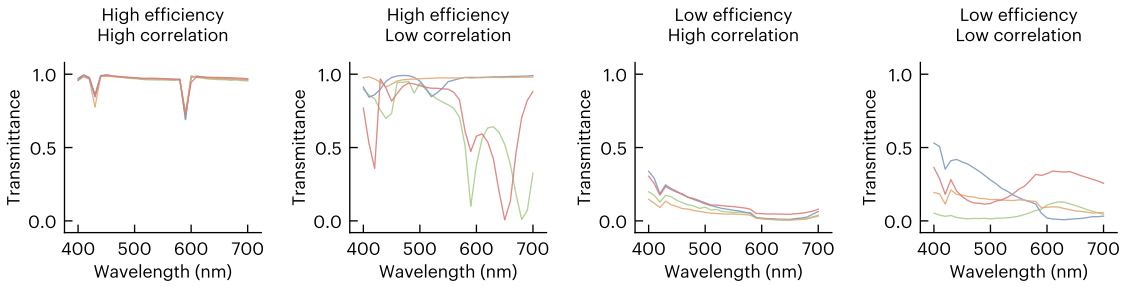
$$\eta = \frac{1}{L} \sum_{\lambda} \bar{S}(\lambda). \tag{5}$$

Internal correlation (ρ): For multi-atom super-pixels ($N > 1$), maximizing the orthogonality between constituent elements is essential for compressive sensing. We quantify redundancy by computing the mean off-diagonal elements of the Pearson correlation matrix between all meta-atom pairs (i, j):

$$\rho = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\mathrm{cov}(S_i, S_j)}{\sigma_{S_i} \sigma_{S_j}}, \tag{6}$$

where $\mathrm{cov}(\cdot)$ denotes covariance and σ_S denotes spectral standard deviation. A lower ρ implies higher informational diversity within the unit cell, prioritizing structures with distinct, non-overlapping spectral features.

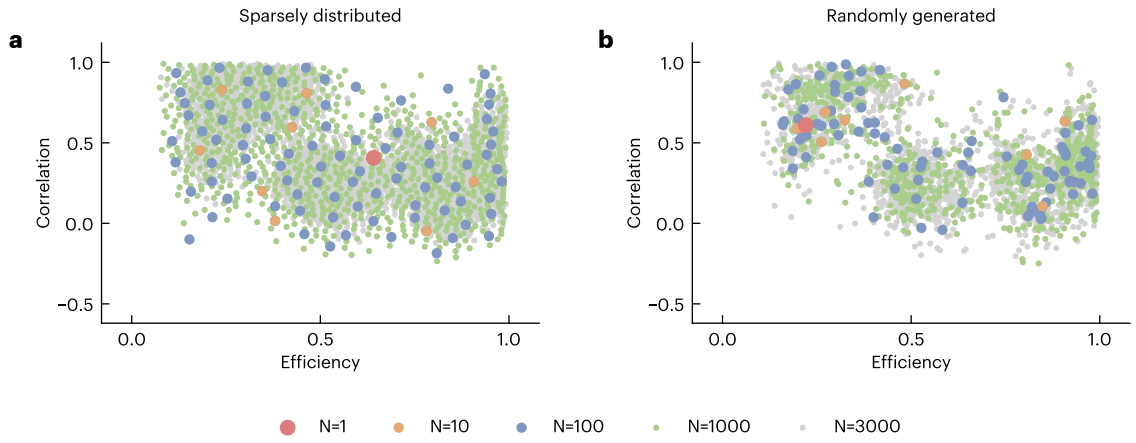
To illustrate the physical significance of these dimensions, representative spectral curves for the four corner cases of the feature space are presented in Supplementary Figure S4.



Supplementary Figure S4: **Representative spectra across the feature space.** Transmittance profiles illustrating the four extreme cases within the efficiency (η) and correlation (ρ) landscape. High efficiency corresponds to maximized overall signal throughput, whereas minimized correlation ensures distinct, non-overlapping spectral responses among the constituent meta-atoms.

Hierarchical Sampling Strategy. The metasurface design space is combinatorially vast. Parameterized by three superpixel scales ($2\times2, 3\times3, 4\times4$), four meta-atoms materials, two substrate materials, and four distinct meta-atom shapes, the structural library encompasses $(4^{16}+4^9+4^4)*4*2 = 34,361,837,568$ discrete configurations. Despite this immense structural diversity, the corresponding optical responses exhibit profound spectral redundancy, with many distinct structural combinations yielding highly overlapping transmittance profiles. With this over billions number of discrete possibilities, their transmittance coverage overlap significantly. To resolve this degeneracy and represent the global search space efficiently, a sparse, maximally informative subset of discrete candidates needs to be identified.

To avoid redundancy and ensure a uniform coverage of this feature space, we employ a hierarchical clustering strategy rather than simple random sampling. First, a massive pool of $N_{\mathrm{pool}} = 10^5$ random candidates is generated and mapped into the normalized feature space. We then apply a nested K-means clustering algorithm to iteratively select representative subsets of diminishing sizes $N_{set} \in \{3000, 1000, 100, 10, 1\}$. Starting from the largest subset, the feature space is partitioned into k clusters. The candidate closest to the centroid of each cluster is selected as the representative archetype. Crucially, this selection process is nested: the subset of size $N = 1000$ is strictly drawn from the survivors of the $N = 3000$ pool, and the single candidate ($N = 1$) corresponds to the geometric centroid of the most selective subset. This hierarchical reduction ensures that smaller, more computationally tractable datasets maintain the same topological diversity and broad feature-space coverage as the larger parent pools, effectively creating a coreset of the design space that maximizes information gain during the neural architecture search. As Supplementary Figure S5 shows, our sparse generation of candidates lead to more diverse and evenly spaced candidates than randomly generated candidates.



Supplementary Figure S5: **Comparison of candidate generation strategies.** Distributions of metasurface subsets ($N = 1$ to 3000) using **a**, sparsely distributed and **b**, random generation methods. The sparsely distributed approach ensures broad, uniform coverage of the efficiency–correlation landscape, whereas random generation results in significant clustering and reduced diversity.

Supplementary Note 3. Thin-film design space and candidate generation

This section delineates the construction of the thin-film optimization landscape, detailing the material constraints and the diversity-driven generation strategy developed to mitigate the mode-collapse phenomena inherent in uniform random sampling.

Supplementary Note 3.1 Material library and physical constraints

Material selection. The spectral encoder is implemented as an alternating multilayer stack of dielectric thin films. To maximize the achievable spectral diversity, we curated a material library spanning distinct refractive index regimes. The candidates, characterized by their complex refractive indices ($\tilde{n} = n + ik$) sourced from standard empirical datasets, include low-index materials ($n \approx 1.3-1.5$): magnesium fluoride (MgF_2) and silicon dioxide (SiO_2); medium-index materials ($n \approx 1.6-1.7$): aluminum oxide (Al_2O_3) and magnesium oxide (MgO); high-index materials ($n \approx 2.0-4.0$): gallium nitride (GaN), niobium pentoxide (Nb_2O_5), silicon nitride (Si_3N_4), titanium pentoxide (Ti_3O_5), zirconium dioxide (ZrO_2), rutile titanium dioxide (TiO_2), and silicon (Si). These materials were selected based on their low absorption in the visible spectrum ($\lambda \in [400, 700]$ nm) and compatibility with standard physical vapor deposition (PVD) processes. The underlying refractive index and extinction coefficient values of materials are drawn in Supplementary Figure S6.

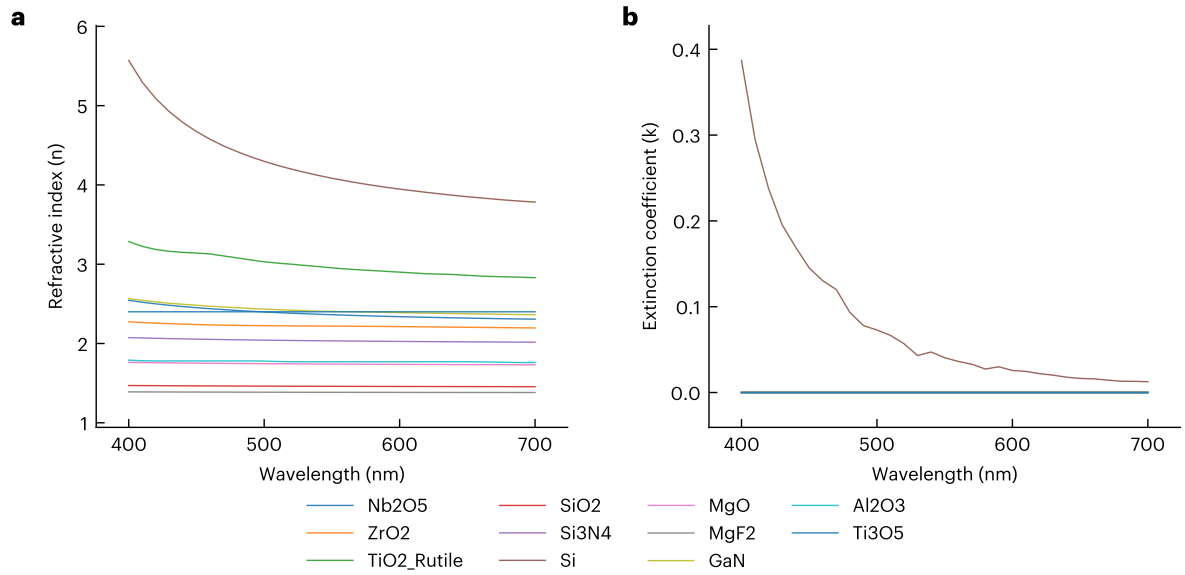
Structural constraints. To define the search space, the total layer count N_{layer} is sampled from the discrete set $\{5, 7, \ldots, 25\}$. This range allows the optimization to balance spectral complexity (associated with high N_{layer}) against fabrication cost. The physical thickness of each individual layer is constrained to the interval $t_i \in [30, 150]$ nm, ensuring feasible deposition rates while providing sufficient optical path length for phase accumulation and interference effects.

Supplementary Note 3.2 Diversity-driven hierarchical generation

A significant bottleneck in constructing a static candidate pool for neural architecture search is spectral redundancy. As illustrated in Supplementary Figure S7, stochastic generation processes tend to oversample averaged, low-contrast spectral shapes, while severely undersampling the complex, high-frequency resonance features located at the boundaries of the design space. To address this, we implemented a hierarchical generation process governed by a discriminative 2D feature space.

Feature space construction. Standard metric spaces (e.g., Euclidean distance) often fail to capture perceptually distinct spectral features; for example, distinguishing between a spectral peak and a valley of equal magnitude, or separating resonance shifts against a flat background. To resolve this, we introduced a linear reference baseline, $\mathbf{T}_{\text{ref}}(\lambda)$, which increases linearly from 0.1 at 400 nm to 1.0 at 700 nm. We then constructed a 2D embedding space using two complementary metrics.

Signed L1 deviance (polarity and magnitude sensitivity):



Supplementary Figure S6: **Optical dispersion properties of the thin-film materials. a**, Refractive index (n), and **b**, extinction coefficient (k), plotted across the visible spectrum (400–700 nm). Dispersion data for the candidate materials are sourced from the RefractiveIndex.INFO and KLA Filmetrics databases.

$$d_{\text{signed}}(\mathbf{T}, \mathbf{T}_{\text{ref}}) = \frac{1}{L} \sum_{\lambda} (\mathbf{T}(\lambda) - \mathbf{T}_{\text{ref}}(\lambda)). \tag{7}$$

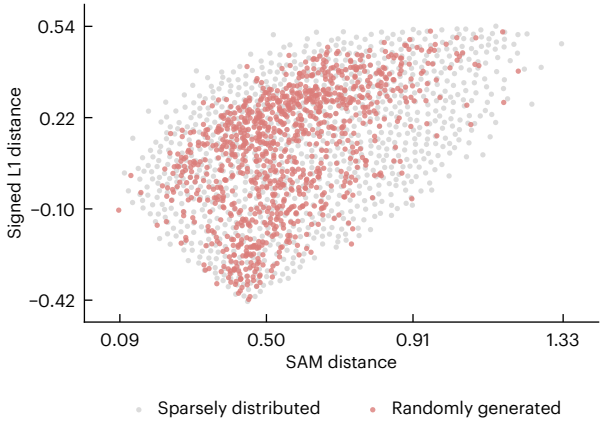
Unlike absolute distance metrics, this signed formulation strictly distinguishes between additive features (peaks, $d > 0$) and subtractive features (valleys, $d < 0$) relative to the baseline. This effectively separates spectral valleys from peaks along the vertical embedding axis.

Biased spectral angle mapper (spectral position sensitivity):

$$\theta_{\text{SAM}}(\mathbf{T}, \mathbf{T}_{\text{ref}}) = \arccos\left(\frac{\mathbf{T} \cdot \mathbf{T}_{\text{ref}}}{\|\mathbf{T}\|_2 \|\mathbf{T}_{\text{ref}}\|_2}\right). \tag{8}$$

The utilization of the ramped baseline $\mathbf{T}_{\text{ref}}$ introduces a spectral bias that breaks symmetry. A perturbation at a short wavelength (where $\mathbf{T}_{\text{ref}}$ is small) alters the projection angle differently than an identical perturbation at a long wavelength. This renders the metric sensitive to the spectral position of resonances, effectively separating blue-shifted features from red-shifted features along the horizontal embedding axis.

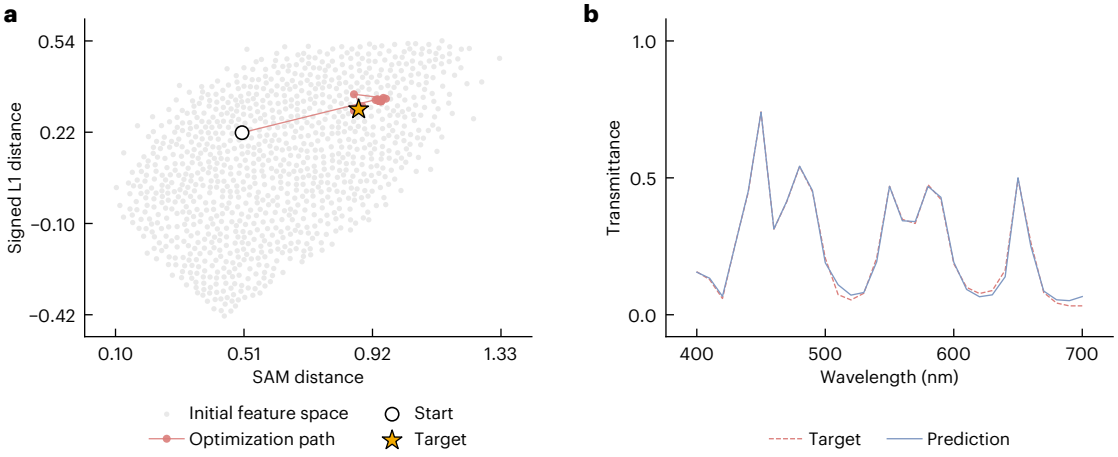
Hierarchical subset selection. We generated an initial universal pool of $N = 200,000$ candidates and projected them into this 2D feature space. To facilitate rigorous ablation studies, we constructed a nested sequence of candidate pools $S_{10} \subset S_{100} \subset \cdots \subset S_{200k}$ via a hybrid pruning strategy. For large subsets ($N > 2000$), we employed batched K-Means clustering to identify representative centroids, ensuring a uniform density coverage of the spectral manifold. For small subsets ($N \leq 2000$), we applied a greedy farthest point sampling (FPS) algorithm, iteratively selecting candidates that maximize the feature distance from the current set. This strategy ensures that even highly constrained search spaces retain maximum spectral variance and include extreme boundary candidates (e.g., high-frequency notch filters).



Supplementary Figure S7: **Candidate distribution in the optical feature space.** Scatter plot comparing 1,000 randomly generated thin-film candidates (red) against 1,000 sparsely distributed candidates (gray). The sparse sampling method achieves a significantly wider and more uniform coverage across the SAM and Signed L1 distance metrics.

Supplementary Note 3.3 Thin-film pruning process tracking

Throughout the thin-film inverse design optimization process, the composite optical property consistently approximates the target spectrum. Tracking this aggregated response therefore yields limited insight into the underlying search dynamics. To provide a transparent view of the search mechanism, we retrospectively track the trajectory of the finally selected structural candidate across the SAM/L1 feature space. As depicted in the Supplementary Figure S8, gradient guidance successfully drives this candidate from its initial configuration toward the target point. The optimization path exhibits periods of temporary stabilization at various intermediate locations. During these phases, the candidate actively participates in the composite property calculation alongside other active structures, sharing the representational burden of the target response. Distinct shifts in the trajectory occur when least-value pruning is executed. This pruning mechanism removes suboptimal topologies and forces the surviving candidate to adapt, learning new physical patterns to compensate for the reduced supernet capacity. Ultimately, this dynamic progression converges precisely at the target point. The accompanying transmittance curves confirm this successful convergence, demonstrating exceptional spectral agreement between the final isolated candidate and the designated target.



Supplementary Figure S8: **Structural optimization path for thin-film inverse design. a**, Track of the optimization trajectory from the initial configuration through the broad initial design space toward the target point in SAM/L1 feature space. **b**, Transmittance spectral comparison, showing the excellent agreement between the final optimized candidate and the designated target.

Supplementary Note 4. Computational complexity and scalability analysis

The computational tractability of large-scale inverse design is fundamentally constrained by the scaling of search algorithms[S12]. The total cost of optimization can be expressed as:

$$T = \frac{\mathcal{O}(N) \times \mathcal{O}(C)}{S},$$

where $\mathcal{O}(C)$ represents the average expense incurred per individual evaluation, $\mathcal{O}(N)$ denotes the algorithm's time complexity relative to the evaluation count, and $S(S \geq 1)$ stands for the speedup or acceleration achieved through the use of distributed and parallel computing methods. In the end-to-end differentiable inverse design task, evaluating a design space of N candidates necessitates training N distinct reconstruction networks to convergence, leading to a significant increase of $O(C)$ and therefore the optimization cost.

The OptoNAS framework circumvents this bottleneck by structurally decoupling the backend training cost from the search space size. By formulating a unified differentiable supernet, the heavy computational weights of the reconstruction backend ($\boldsymbol{\omega}$) are shared across all discrete physical candidates. Consequently, the marginal computational cost of expanding the search space is restricted to the lightweight forward modeling of the optical frontend. This effective amortization reduces the backend training complexity from $\mathcal{O}(N)$ to approximately $\mathcal{O}(1) + \mathcal{O}(\epsilon N)$, where ϵ represents the negligible cost of the lightweight surrogate relative to the heavy backend. Furthermore, our OptoNAS architecture has independent evaluation branching across different candidates, creating natural parallelism across GPUs, potentially further increasing S and reducing time cost.

Supplementary Note 4.1 Benchmarking metrics and definitions

To rigorously quantify these efficiency gains, we introduce two distinct acceleration metrics designed to isolate algorithmic performance from hardware scaling. The theoretical speedup (S_{theo}) quantifies the algorithmic efficiency of the supernet by normalizing the temporal throughput gain against the hardware resource cost (VRAM usage). It is defined as the ratio of the projected serial training time to the actual batched time, scaled by the inverse memory overhead:

$$S_{\text{theo}}(N) = \underbrace{\frac{N \times T_{\text{base}}}{T_{\text{supernet}}(N)}}_{\text{Time Acceleration}} \times \left(\underbrace{\frac{M_{\text{supernet}}(N)}{M_{\text{base}}}}_{\text{Memory Overhead}} \right)^{-1}, \tag{9}$$

where T_{base} and M_{base} represent the time per iteration and peak GPU memory usage for the baseline ($N = 1$), while $T_{\text{supernet}}(N)$ and $M_{\text{supernet}}(N)$ represent the corresponding metrics for the supernet size N. By incorporating the memory ratio ($M_{\text{base}}/M_{\text{supernet}}$), this metric strictly evaluates whether the architecture achieves acceleration through efficient weight amortization (sub-linear memory growth) rather than simply consuming proportional hardware resources. The effective speedup (S_{eff}), in contrast, accounts for the full system-level overhead, including data loading I/O, Python interpreter latency, and PyTorch compilation updates. It is derived from the total wall-clock GPU hours of the experimental execution. S_{eff} provides a conservative, real-world estimate of the performance gain achievable in a production environment.

Supplementary Note 4.2 Empirical Performance Evaluation

We conducted a systematic benchmark analysis using a workstation equipped with a single NVIDIA RTX 4090 GPU (24 GB VRAM). Detailed experimental configurations are provided in Supplementary Note 6. To assess scalability across varying levels of forward modeling complexity, the evaluation spanned two distinct domains: the metasurface domain, which necessitates a complex deep neural surrogate to map physical configurations to spectra, and the thin-film domain, which utilizes the analytical transfer matrix method. These domains impose different computational overheads on the optimization loop.

Supplementary Note 4.2.1 Impact of forward model complexity on scalability

As demonstrated in Supplementary Table S2, the metasurface experiment relies on a deep surrogate network to approximate Maxwell's equations; this introduces a steep gradient for resource consumption due to the retention of large computational graphs for backpropagation. Empirically, doubling the VRAM

usage in the metasurface regime only accommodates an increase in search space from $N = 1$ to $N = 3,000$. Conversely, the thin-film domain, utilizing the lightweight analytical TMM, scales from $N = 1$ to $N = 200,000$ within the same memory constraint—a capacity difference of approximately 66 times. This comparison demonstrates that the scalability of the framework is governed by the architectural complexity and learnable parameter count of the forward model, as the GPU memory bottleneck is dictated by the size of the gradient graph required for end-to-end differentiation.

Supplementary Note 4.2.2 System latency and dynamic pruning

The theoretical speedup (S_{theo}) projections are calculated based on the conservative assumption that the full candidate pool N remains active throughout the optimization process. However, the OptoNAS pruning controller dynamically contracts the search space over time. Theoretically, this should yield a super-linear reduction in cumulative training time as candidates are discarded.

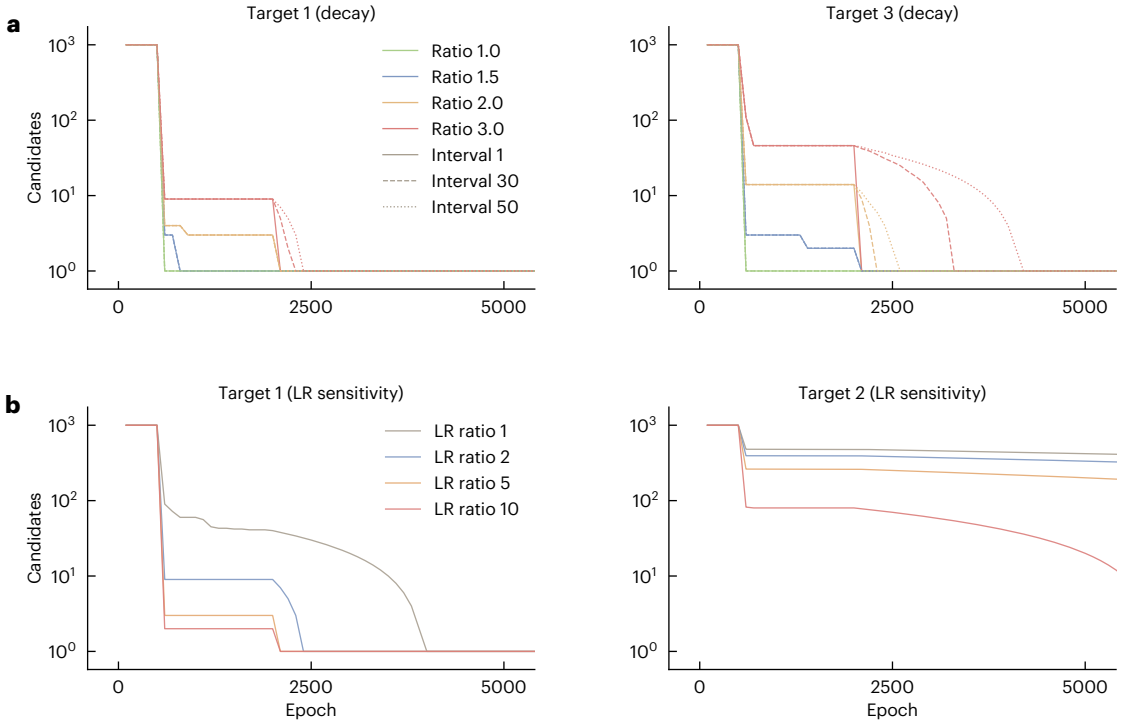
In practice, this algorithmic advantage is counterbalanced by non-ideal system constraints, including CPU data loading latency, RAM bandwidth bottlenecks, and Python interpreter overheads. As evidenced in Supplementary Table S3, the effective speedup (S_{eff}) closely tracks the theoretical speedup across all scales. This convergence implies that the computational time reclaimed by dynamic pruning is diminished by residual system overheads. Consequently, while the current results validate the framework's high efficiency, they also indicate that further engineering optimizations—specifically targeting data throughput and system latency—could unmask the full benefits of pruning, potentially driving the effective speedup significantly beyond the current theoretical baseline.

Supplementary Note 4.2.3 Analysis of scalability trends

As illustrated in Supplementary Table S3, the framework exhibits two critical performance characteristics that confirm its suitability for large-scale inverse design. First, a computational complexity gap is evident: the logarithmic-scale comparison reveals a dramatic divergence between the theoretical linear cost of standard independent search and the sub-linear trajectory of OptoNAS. In the high-throughput thin-film regime, increasing the candidate pool from a single design to $10,000$ incurs a temporal overhead of merely 43% (1.62s $\rightarrow$ 2.32s). At the extreme scale of $N = 200,000$, OptoNAS achieves a theoretical speedup of nearly $10,000\times$ compared to the projected baseline. This confirms that the framework successfully amortizes the backend training cost, effectively decoupling the computational budget from the size of the design space. Second, linear memory scaling is observed: the tensorized supernet formulation maps the architectural search into dense, batched matrix operations that are inherently suited for the SIMD architecture of modern GPUs. The memory profile exhibits a highly efficient linear gradient rather than an exponential explosion. For the thin-film model, the system stores approximately $16,000$ additional candidates per gigabyte of VRAM. This high memory density ensures that massive search spaces fit comfortably within the memory envelope of standard consumer-grade GPUs.

Supplementary Table S3: **Comparative computational scaling analysis.** This table contrasts resource utilization and acceleration metrics between the computationally dense Meta-surface experiment and the high-throughput Thin-film experiment. The divergence in maximum search space (N) and speedup magnitudes is driven by the intrinsic complexity of the surrogate models: the Meta-surface model requires heavy spatial tensor processing, whereas the Thin-film model operates on lightweight vector embeddings, allowing for greater scalability within the same VRAM constraints.

	Meta-Surface Experiment					**Thin-Film Experiment**		
	Theoretical (100 batches)			Real (Effective)		Theoretical (30 batches)		
Search Space (N)	**GPU VRAM (MB)**	**Time** (s)	**Theor. Speedup(×)**	**Eff. GPU Hours**	**Eff. Speedup(×)**	**GPU VRAM(MB)**	**Time** (s)	**Theor. Speedup(×)**
1 (Baseline)	9,256	28.38	1.00	17.41	1.00	10,394	1.62	1.0
10	9,302	38.71	9.53	18.26	9.53	10,394	1.84	8.8
100	9,498	40.27	94.36	18.45	94.36	10,398	2.17	74.6
1,000	14,158	44.31	418.73	21.91	794.61	10,444	2.19	737.8
3,000	**22,746**	**65.50**	**528.95**	**24.44**	**2,137.07**	–	–	–
10,000	–	–	–	–	–	10,978	2.32	6,611.29
100,000	–	–	–	–	–	**16,532**	**8.31**	**12,256.63**
200,000	–	–	–	–	–	22,858	15.05	9,789.33



Supplementary Figure S9: **Parametric control of candidate pool pruning dynamics. a**, Active candidate trajectories for Targets 1 and 3 under varying soft pruning ratios and selection intervals. **b**, Optimization stability and decay rate benchmarked against different learning rate (LR) ratios for Targets 1 and 2.

Supplementary Note 5. Effects of OptoNAS hyperparameters

Just as traditional deep learning relies on the tuning of optimizers and learning rates to navigate the loss landscape effectively, the OptoNAS framework is governed by a set of specific hyperparameters that modulate the behavior of the pruning controller. These parameters determine the trajectory of the search process, mediating the critical trade-off between exploration (retaining a diverse pool of physical topologies) and exploitation (converging rapidly to a single solution).

In this section, we analyze the sensitivity of the framework to three governing variables: the architecture learning rate multiplier (μ_α), the pruning interval (N_{interval}), and the threshold ratio (r_{thresh}). We discuss how manipulating these variables systematically alters the candidate decay rate—the velocity at which the active search space N contracts from the initial universal pool to the final design.

Supplementary Note 5.1 Pruning controller logic and convergence dynamics

The pruning controller intervenes periodically in the differentiable supernet training to filter the design space. As detailed in the framework workflow, this process relies on two distinct elimination mechanisms. Threshold elimination is a probabilistic filter where candidates are removed if their normalized probability weight falls below a dynamic threshold defined by the population statistics. Least value elimination, in contrast, serves as a deterministic approach that forcibly removes the lowest-ranked candidate if the population size remains static for an extended duration, ensuring the search does not stagnate.

To understand the dynamics of this controller, we visualize the reduction of active candidates over time under different hyperparameter regimes (Supplementary Figure S9).

Architecture learning rate multiplier (μ_α). The structural parameters $\boldsymbol{\alpha}$ govern the probability distribution over the candidate pool. These parameters often require a distinct learning pace compared to the physical frontend parameters $\boldsymbol{\theta}$ or the backend reconstruction weights $\boldsymbol{\omega}$. We introduce the multiplier μ_α to scale the structural learning rate relative to the base learning rate: $\eta_\alpha = \mu_\alpha \cdot \eta_\theta$.

As illustrated in Supplementary Figure S9a, μ_α acts as the primary accelerator for the search phase. Increasing μ_α amplifies the magnitude of probability updates, resulting in a steeper decay curve, as the optimizer can rapidly distinguish and isolate dominant candidates. However, excessive acceleration risks premature convergence, potentially locking the search into a local optimum before the physical parameters have fully adapted. Decreasing μ_α, on the contrary, slows the differentiation between candidates. A lower

multiplier causes the structural weights to evolve at a similar pace to the physical parameters, leading to a more prolonged exploration phase where the candidate pool remains large for more epochs.

Pruning interval (N_{interval}). This parameter defines the frequency (in epochs) at which the controller evaluates the candidate pool for potential elimination. It effectively controls the granularity of the pruning process (see Supplementary Figure S9a). Short intervals result in frequent checks, producing a smooth, continuous decay profile (solid lines). While this minimizes the computational waste of training rejected candidates, it requires the physical parameters to adapt rapidly between pruning events. Long intervals, in contrast, introduce a staircase-like decay behavior (dotted lines). This provides a longer adaptation period for the physical parameters of surviving candidates to stabilize after the probability distribution is normalized. However, this stability comes at the cost of total training time, as the system retains non-competitive candidates for longer durations.

Threshold ratio (r_{thresh}). The threshold ratio modulates the aggressiveness of the threshold elimination strategy. A candidate is retained only if its probability exceeds a fraction of the uniform average, scaled by this ratio: $w_k \geq \frac{1}{|N_{\text{active}}|} \cdot \frac{1}{r_{\text{thresh}}}$. This parameter effectively sets the safety margin for survival (see Supplementary Figure S9b). Increasing r_{thresh} lowers the probability bar required for survival, making the controller more conservative and retaining a larger number of candidates for a longer duration. As the ratio increases, the decay curve flattens, indicating that the system relies more heavily on the slower least-value elimination mechanism to finalize the search. This improves robustness against noise but extends the computational budget. Decreasing r_{thresh}, on the contrary, raises the bar for survival. A lower ratio imposes more stringent population filtering, rapidly contracting the search space. While this accelerates convergence, it increases the likelihood of erroneous rejection, in which high-potential designs are prematurely discarded due to stochastic fluctuations in their loss values.

Supplementary Note 6. Network training and implementation details

Supplementary Note 6.1 Continuous variable constraints

To ensure that the optimized physical parameters remain within the fabrication limits of the lithography and deposition processes, we impose a soft differentiability constraint on the search space. The unbounded latent variables $\mathbf{p}_i$ utilized by the optimizer are mapped to the physical domain $\boldsymbol{\theta}_{c,i}$ via a sigmoid activation function:

$$\boldsymbol{\theta}_{c,i} = v_{\min} + (v_{\max} - v_{\min}) \cdot \sigma(\mathbf{p}_i), \tag{10}$$

$$\sigma(\mathbf{p}_i) = \frac{1}{1 + e^{-\mathbf{p}_i}} \tag{11}$$

where $v_{\min}$ and $v_{\max}$ denote the lower and upper bounds of the parameter (e.g., layer thickness or hole radius), and $\sigma(\cdot)$ is the standard logistic sigmoid function. This mapping guarantees that the physical parameters strictly respect the defined boundaries throughout the gradient descent process. However, the saturation regions of the sigmoid function can lead to vanishing gradients if the latent variables drift too far from the origin. To mitigate this and stabilize the optimization dynamics, particularly during the initial phase of the search, we apply global gradient clipping. This prevents the latent parameters from growing excessively large, thereby maintaining the sensitivity of the mapping function and ensuring that the physical parameters can effectively traverse the design space without being trapped near the boundaries.

Supplementary Note 6.2 VRAM memory optimization

The computational throughput of the OptoNAS framework critically depends on efficient utilization of GPU memory bandwidth and tensor core parallelism. To maximize evaluation speed, we preload the discrete candidate parameter θ_d into the GPU VRAM, rather than streaming batches from the host memory. This allows the forward model—whether the neural surrogate or the analytical Transfer Matrix Method—to be executed as a single fused kernel operation across the candidate dimension. Furthermore, we leverage Just-In-Time (JIT) compilation techniques, specifically `torch.compile`, to fuse pointwise operations and minimize Python interpreter overhead. Empirical benchmarks demonstrate that this static graph compilation achieves an approximately threefold reduction in wall-clock time per iteration compared to eager execution, which is essential for scaling the search to hundreds of thousands of candidates.

Supplementary Note 6.3 Finetuning stage (stationary optimization)

To guarantee optimal convergence, we decouple the search termination from the final network training phase. Although dynamic pruning typically isolates a single candidate ($N_{active} = 1$) at the middle of the optimization process, we continue joint optimization until a pre-defined finetuning epoch $T_{finetune}$. This buffer period allows the continuous parameters of the surviving candidate to converge to a local optimum. At $T_{finetune}$, the entire optical frontend—including discrete topology and physical parameters—is frozen. The subsequent training phase is then dedicated exclusively to optimizing the reconstruction network $\boldsymbol{\omega}$. This strategy isolates backend training from residual parameter fluctuations, enabling the network to fully adapt to the fixed, stationary physics of the final optical design.

Supplementary Note 6.4 Details of metasurface network training and results

The training protocol for the metasurface experiment is designed to handle the heterogeneity of the feature space, which includes both categorical material indices and continuous geometric parameters. The optimization is driven by the AdamW optimizer with a base learning rate of 1×10^{-3} and a weight decay of 1×10^{-4} to prevent overfitting in the reconstruction backend. A crucial hyperparameter in this regime is the architecture learning rate multiplier μ_α, which we set to 5.0. This amplification ensures that the structural probability distribution evolves more rapidly than the physical parameters, allowing the pruning controller to identify and discard sub-optimal topologies efficiently prior to geometric refinement. The surrogate model, a pre-trained dual-stream net, remains frozen during the search phase; however, we backpropagate gradients through it to update the continuous latent inputs. Due to the memory-intensive nature of the deep surrogate, the batch size for the spectral reconstruction target is set to 64, while the candidate pool is managed dynamically to fit within the 24 GB VRAM envelope.

To rigorously assess the reconstruction quality of the hyperspectral imaging system, the framework is evaluated across three complementary metrics quantifying spatial fidelity, spectral shape alignment, and absolute magnitude error.

The Peak Signal-to-Noise Ratio (PSNR) quantifies the spatial fidelity of the reconstructed hyperspectral datacube. It evaluates the ratio between the maximum possible power of the true image signal and the power of the reconstruction error. The metric is defined as:

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}} \right) \tag{12}$$

where MAX_I represents the maximum valid pixel intensity of the image and MSE denotes the mean squared error between the predicted and ground-truth spatial dimensions. A higher value indicates superior spatial reconstruction accuracy.

The Spectral Angle Mapper (SAM) evaluates the spectral shape fidelity by treating the spectrum at a single spatial pixel as a high-dimensional vector. It calculates the angular deviation between the reconstructed spectral vector $\hat{\mathbf{y}}$ and the ground-truth spectral vector $\mathbf{y}$ across all wavelength bands. The formula is expressed as:

$$\mathrm{SAM} = \arccos \left(\frac{\hat{\mathbf{y}} \cdot \mathbf{y}}{\|\hat{\mathbf{y}}\|_2 \|\mathbf{y}\|_2} \right) \tag{13}$$

Because this calculation normalizes the vector magnitudes, it is highly sensitive to the spectral profile while remaining invariant to overall intensity fluctuations. A lower angular value measured in radians signifies a more precise alignment with the true material spectral signature.

The L1 Loss provides a direct measure of the absolute magnitude discrepancy. It calculates the average absolute pixel-wise difference between the predicted hyperspectral volume $\hat{\mathbf{Y}}$ and the target volume $\mathbf{Y}$

Supplementary Table S4: **OptoNAS performance scaling.** Comparison of reconstruction quality metrics (PSNR, SAM, L1) as the number of candidates increases. The data demonstrates a monotonic improvement in performance (higher PSNR, lower error) with larger search spaces.

Metric	**1**	**10**	**100**	**1000**	**3000**
PSNR (dB) ↑	40.22(3.97)	40.73(4.64)	41.44(4.84)	43.18(4.62)	**43.66(4.68)**
SAM (rad) ↓	3.89(1.40)	3.37(1.86)	3.23(1.45)	2.68(1.60)	**2.64(1.52)**
L1 Loss ↓	0.0067(0.0040)	0.0063(0.0043)	0.0063(0.0056)	0.0048(0.0036)	**0.0045(0.0031)**

*Note: **Bold** text indicates headers and metric names. ↑ indicates higher is better; ↓ indicates lower is better.*

across all spatial and spectral coordinates:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}_i - Y_i| \tag{14}$$

where N is the total number of elements in the datacube, a lower L1 loss reflects a reduction in the overall magnitude of the reconstruction error.

Supplementary Table S4 details the quantitative performance scaling of the OptoNAS framework. The data track the end-to-end reconstruction quality as the size of the discrete structural candidate pool expands from a single baseline structure to a comprehensive pool of 3,000 candidates. The results demonstrate a strict monotonic improvement across all three evaluation metrics as the search space grows.

Supplementary Note 6.5 Details of thin-film network training and results

In the thin-film regime, the computational efficiency of the analytical TMM forward model enables an accelerated training schedule. We initialize the search with an extensive candidate pool of $K = 200,000$, utilizing the full vectorization capabilities of the GPU. The layer thicknesses are initialized uniformly within the valid range of $[30, 150]$ nm via the inverse sigmoid mapping. Given the convex nature of the thin-film spectral loss landscape relative to the metasurface domain, we employ a standard learning rate of 1×10^{-3} without requiring architectural acceleration ($\mu_\alpha = 1.0$). The pruning controller is configured with an interval of 10 epochs and a threshold ratio of 0.8, facilitating a steady, non-stochastic reduction of the design space. Following the isolation of the optimal stack configuration, we perform a dedicated stationary optimization phase of 50 epochs. During this stage, the layer thicknesses are locked, and the reconstruction backend is fine-tuned to minimize the mean squared error, ensuring the final spectral response is accurately characterized.

The initial thin-film inverse design experiment utilized 50 random structural configurations and their corresponding target spectra to establish distinct candidate sets. During optimization, 8 trials involving the 1,000-candidate pool failed to converge stably due to strict training time constraints and hyperparameter sensitivity. Consequently, the statistical analysis is based exclusively on the remaining 42 successful experiments. As Supplementary Table S5 shows, expanding the active search space from 1 to 1,000 candidates significantly reduces both the mean prediction error and its variance, demonstrating substantially improved optimization stability and spectral fidelity.

Supplementary Table S5: **Prediction error and stability across different candidate pool sizes.** The table presents the mean squared error and its standard deviation evaluated across the test groups as a function of the initial candidate count.

Candidates	Mean MSE	Std MSE
1	4.794e-02	8.109e-02
10	2.979e-03	3.607e-03
100	2.277e-03	3.861e-03
1000	**1.794e-03**	**1.815e-03**

In the second experiment, 50 target spectra exhibiting an average transmittance greater than 0.2 were selected from the OptoGPT test dataset. The pre-trained OptoGPT model serves as the generative baseline for comparison. Because OptoGPT deterministically yields the highest probability output with minimal variance, we utilized its direct results for evaluation. Conversely, to account for the stochastic variance inherent to initial pool sampling, OptoNAS was evaluated across 3 independent trials using distinct sets of 1,000 candidates. The trial yielding the maximum reconstruction fidelity was selected for the comparative analysis. As detailed in Supplementary Table S6, OptoNAS achieves a substantially lower mean prediction error and variance relative to the OptoGPT baseline, highlighting its superior spectral fidelity and robustness.

Supplementary Table S6: **Quantitative performance comparison between OptoNAS and OptoGPT.** The table summarizes the overall mean squared error and its standard deviation evaluated across the test targets.

Method	Mean MSE	Std MSE
OptoGPT	3.239e-03	4.471e-03
OptoNAS (Ours)	**1.565e-03**	**1.969e-3**

References

S1 Wiecha, P. R. *et al.* Evolutionary multi-objective optimization of colour pixels based on dielectric nanoantennas. *Nat. Nanotechnol.* **12**, 163–169 (2017).

S2 Jafar-Zanjani, S., Inampudi, S. & Mosallaei, H. Adaptive genetic algorithm for optical metasurfaces design. *Sci. Rep.* **8**, 11040 (2018).

S3 Phan, T. *et al.* High-efficiency, large-area, topology-optimized metasurfaces. *Light Sci. Appl.* **8**, 48 (2019).

S4 Li, Z. *et al.* Inverse design enables large-scale high-performance meta-optics reshaping virtual reality. *Nat. Commun.* **13**, 2409 (2022).

S5 Ou, X. *et al.* Tunable polarization-multiplexed achromatic dielectric metalens. *Nano Lett.* **22**, 10049–10056 (2022).

S6 Zou, X. *et al.* Pixel-level bayer-type colour router based on metasurfaces. *Nat. Commun.* **13**, 3288 (2022).

S7 Chen, J. *et al.* Polychromatic full-polarization control in mid-infrared light. *Light Sci. Appl.* **12**, 105 (2023).

S8 Tseng, E. *et al.* Neural nano-optics for high-quality thin lens imaging. *Nat. Commun.* **12**, 6493 (2021).

S9 Yin, Y. *et al.* Multi-dimensional multiplexed metasurface holography by inverse design. *Adv. Mater.* **36**, 2312303 (2024).

S10 Wei, Z. *et al.* Ultraprecision, high-capacity, and wide-gamut structural colors enabled by a mixture probability sampling network. *Light Sci. Appl.* **15**, 164 (2026).

S11 Ansys Inc. Ansys Lumerical FDTD: 3D electromagnetic simulator. https://www.ansys.com/products/photonics/fdtd (2026).

S12 Li, J.-Y., Zhan, Z.-H. & Zhang, J. Evolutionary computation for expensive optimization: A survey. *Mach. Intell. Res.* **19**, 3–23 (2022).