

Supporting Information

Delayed-Gratification Deep Reinforcement Learning for Sample-Efficient Photonic Inverse Design

Honggang Li¹, Shujing Ren¹, Yongchang Feng¹, Yihang Chen², and Kunyuan Xu^{2,*}

¹Key Laboratory of Atomic and Subatomic Structure and Quantum Control (Ministry of Education), Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter, South China Normal University, Guangzhou 510006, China

²Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials, Guangdong-Hong Kong Joint Laboratory of Quantum Matter, South China Normal University, Guangzhou 510006, China

*Corresponding author: xuky@scnu.edu.cn.

This document provides the algorithm, ablation configurations, task-specific implementation settings, complete numerical results, and supplementary figures. Notation and abbreviations follow the main text.

S1 Algorithm and simulation-call accounting

Algorithm S1 gives the pseudocode for the DG-DRL interaction loop. The design representation, variable-wise policy, action mask, and reward are defined in Eqs. (1)–(4) of the main text. Updating the design vector and selecting an action do not involve simulations. The algorithm stops when the maximum number of simulations is reached. Training transitions include the intermediate steps with zero reward and the terminal step with the task reward.

Algorithm S1 DG-DRL interaction and training loop

Require: Design dimension N , variable domains, action constraints, simulator f , evaluation function F , network, replay buffer, maximum simulation count $C_{\max}$

Ensure: Evaluated candidate designs and trained network

```
1:  $C \leftarrow 0$ 
2: while  $C < C_{\max}$  do
3:   Initialize design  $\mathbf{S}_0$ 
4:   for  $t = 1, \dots, N$  do
5:     Set  $s_t = (\mathbf{S}_{t-1}, t)$  and identify the current variable
6:     Construct the valid-action mask for  $s_t$ 
7:     Select a valid action  $a_t$  and obtain  $\mathbf{S}_t$ 
8:     if  $t < N$  then
9:        $r_t \leftarrow 0$ 
10:    else
11:       $y_N \leftarrow f(\mathbf{S}_N)$ ;  $r_t \leftarrow F(y_N)$ ;  $C \leftarrow C + 1$ 
12:    end if
13:    Store the transition, including the successor state and terminal flag
14:    Update the network using the training procedure
15:  end for
16: end while
```

S2 Ablation configurations and results

The Task 3 ablation compares the four configurations in Table S1. The delayed-feedback local-action variant simulates a structure every eight actions, while DG-DRL runs a simulation only at the end of its 64-step episode. Both binary actions are valid in Task 3, so the action mask leaves the action set unchanged for all configurations.

All four variants use the same Double DQN implementation in Stable Baselines3, an MLP with layer widths [128, 256, 256, 128], batch size 128, learning rate 6×10^{-4} , and discount factor $\gamma = 1.0$. The ε -greedy exploration rate decreases linearly from 1.0 to 0.04 over the first 70% of training steps. Each configuration uses seeds 1–10 and completes 40,000 RCWA simulations. Performance is the best diffraction efficiency found within this simulation count. Table S2 reports the checkpoint statistics, and Fig. S1 shows the learning curves.

Table S1: Configuration of ablation variants on Task 3.

Variant	Action space	Reward mechanism	State space	Horizon
T-DRL	Micro-action $\{0, \dots, 63\}$ (flip one bit)	Stepwise $r_t = \Delta \text{EFF}$	64-dim current design	Variable (up to ~ 128 steps)
T-DRL+SDAS	SDAS $\{0, 1\}$ (keep/flip current cell)	Stepwise $r_t = \Delta \text{EFF}$	65-dim design + step index	Fixed $H = 64$
T-DRL+DRM	Micro-action $\{0, \dots, 63\}$ (flip one bit)	Semi-delayed reward every 8 steps	64-dim current design	Variable (up to ~ 1024 steps)
DG-DRL	SDAS $\{0, 1\}$ (keep/flip current cell)	Terminal-only reward at $t = 64$	65-dim design + step index	Fixed $H = 64$

Table S2: Ablation results on Task 3: mean $\pm$ standard deviation of diffraction efficiency (%) over 10 runs at matched RCWA simulation counts.

Method	10k simulations	20k simulations	30k simulations	40k simulations
T-DRL	48.36 ± 5.87	51.15 ± 4.28	53.93 ± 4.83	53.93 ± 4.83
T-DRL+DRM	48.58 ± 2.21	50.30 ± 3.47	51.31 ± 3.04	51.31 ± 3.04
T-DRL+SDAS	40.74 ± 2.83	41.05 ± 2.83	41.63 ± 3.26	42.41 ± 3.66
DG-DRL	50.62 ± 4.67	59.27 ± 3.77	69.96 ± 7.70	79.37 ± 4.39

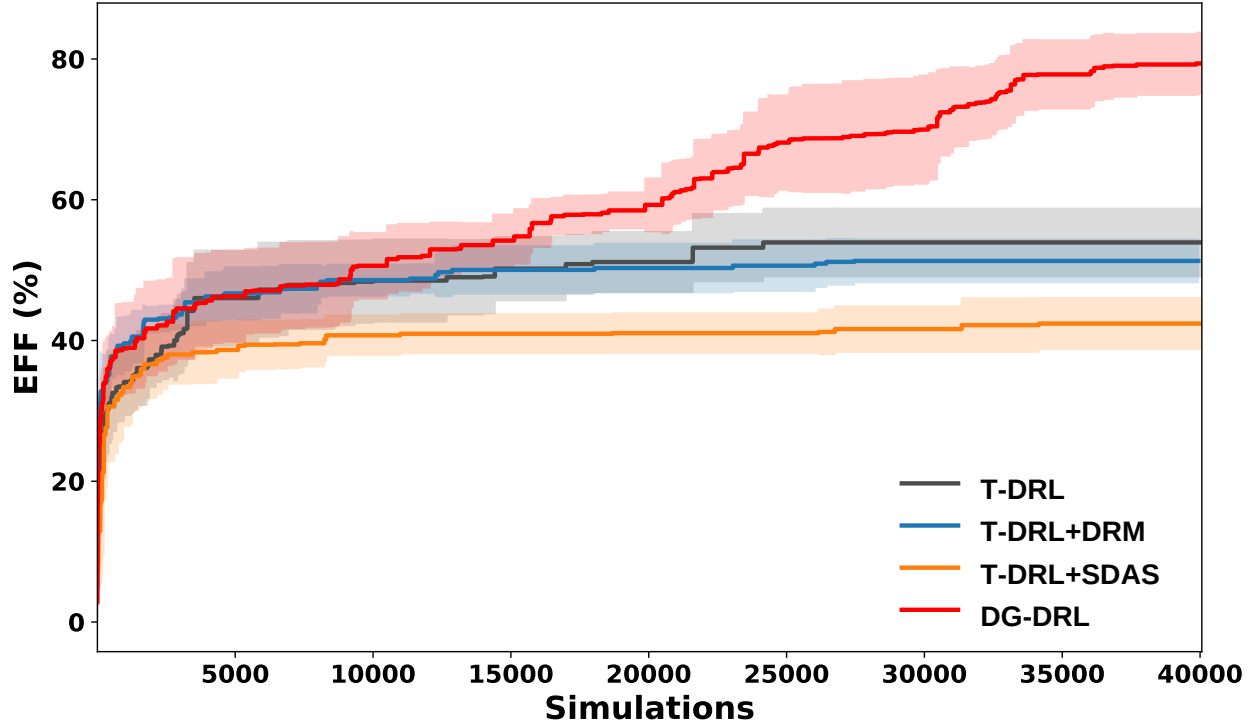


Figure S1: Learning curves of the four Task 3 ablation configurations over ten runs. Solid curves show the mean best-so-far diffraction efficiency, and shading shows one standard deviation. The horizontal axis counts completed RCWA simulations.

S3 Task-specific settings and extended data

S3.1 Task 1: High-transmission structural-color films

Parameterization and electromagnetic simulation

The six design parameters are selected by integer indices from the catalogs in Table S3. Disk and film materials use the same catalog and ordering, listed in Table S4.

Table S3: Discretized design parameters for the bilayer nanophotonic structure in Task 1.

Parameter	Range	Step size	Number of actions
Disk diameter	100–1000 nm	10 nm	91
Unit-cell width	100–1000 nm	10 nm	91
Disk thickness	50–1000 nm	10 nm	96
Film thickness	50–1000 nm	10 nm	96
Disk material	1–13	1	13
Film material	1–13	1	13

Table S4: Material catalog and corresponding IDs used in Task 1.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13
Material	SiO ₂	Al ₂ O ₃	ZnO	Si ₃ N ₄	ZnS	TiO ₂	MgF ₂	GaAs	InAs	Si	LiNbO ₃	TiN	Ge

Each design is simulated using the same Lumerical FDTD template over 300–1200 nm. The

computational region covers one square periodic cell laterally and has a length of $6.4\ \mu\text{m}$ along z . Periodic boundary conditions are used along x and y , with perfectly matched layers along z . The broadband source is above the structure, with injection axis z , backward propagation, and polarization angle 180° . The incidence angles in the template are $\theta = 0^\circ$ and $\phi = 0^\circ$.

Reflection and transmission monitors are placed at $z = +2.1\ \mu\text{m}$ and $z = -2.1\ \mu\text{m}$, respectively, with lateral spans of 1.2 times the unit-cell period. Both monitors use 901 frequency points over 300–1200 nm. The local mesh-override region extends from $z = -h_{\text{film}} - 5a/20$ to $z = h_{\text{disk}} + 5a/20$, where a is the unit-cell period. Its mesh spacing is $\Delta x = \Delta y = \Delta z = a/20$. Mesh accuracy is 4, simulation time is 10 ps, and the auto-shutoff threshold is 10^{-5} . The materials are implemented using the same local dispersive Lumerical entries for every method.

The visible-range transmittance is averaged over 380–780 nm. Color is calculated from the reflection spectrum under AM1.5 illumination, with black, corresponding to zero reflectance, as the reference for CIEDE2000. The task evaluation function is Eq. (5) of the main text. A DG-DRL episode contains six decisions, with one simulation after the sixth decision. T-DRL uses 12 actions, which increase or decrease one of the six catalog indices by one step within its bounds. Each T-DRL action is followed by an FDTD simulation, and the same task evaluation function is used to calculate the reward. Its network, optimizer, and training hyperparameters are matched to DG-DRL.

Extended results

A design is counted as valid when it lies in the given transmittance interval and passes the physical feasibility checks. Table S5 reports design counts and chromaticity convex-hull areas for the four intervals. Table S6 reports stage-wise mean rewards over three seeds, and Fig. S2 shows representative spectra.

Table S5: Task 1 extended results: valid-design counts and color-gamut areas, together with the corresponding DG-DRL/T-DRL ratios.

Metric	Method	60–70%	70–80%	80–90%	90–100%
Valid designs	DG-DRL	321	343	211	134
	T-DRL	195	103	74	55
	DG-DRL/T-DRL	1.65	3.33	2.85	2.44
Color-gamut area (convex hull)	DG-DRL	0.0665	0.0510	0.0754	0.0766
	T-DRL	0.0141	0.0091	0.0058	0.0080
	DG-DRL/T-DRL	4.72	5.60	13.00	9.58

Table S6: Task 1 extended results: phase-wise mean reward after 2,500 FDTD simulations. Values are mean $\pm$ standard deviation over three random seeds.

Method	1–500	501–1000	1001–1500	1501–2000	2001–2500
DG-DRL	84.43 ± 0.76	95.01 ± 1.92	107.13 ± 2.20	114.10 ± 1.43	122.97 ± 1.53
T-DRL	87.00 ± 3.05	83.39 ± 2.54	87.30 ± 6.50	85.06 ± 4.27	86.73 ± 1.81
CEM	83.20 ± 1.05	98.50 ± 1.16	99.76 ± 0.92	100.38 ± 1.72	103.59 ± 1.67
PSO	86.56 ± 4.51	105.65 ± 5.22	107.13 ± 2.90	109.28 ± 4.44	115.93 ± 2.38

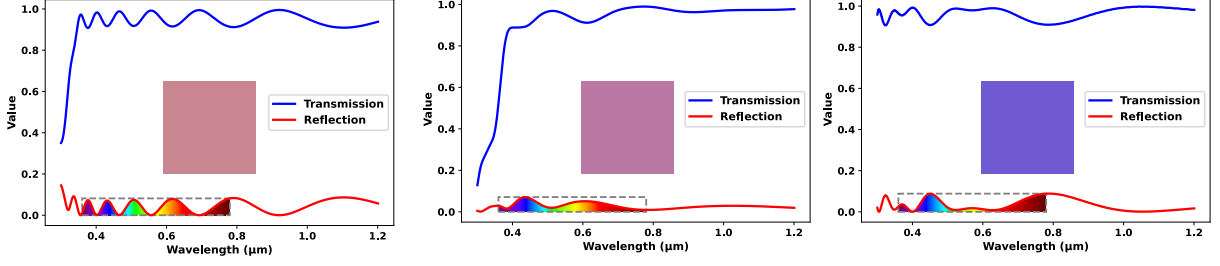


Figure S2: Spectra of three Task 1 designs found by DG-DRL in the 90–100% transmittance interval. Blue and red curves show transmission and reflection, respectively; the insets show the corresponding colors.

S3.2 Task 2: Metagrating near-field beam shaping

Geometry and field comparison

The mirror-symmetric metagrating is described by 13 width indices. Index values 0–4 correspond to strip widths of 0, 200, 400, 600, and 800 nm; Table S7 lists the catalog using unit IDs 1–5. The silicon-strip height is 800 nm. Lengths are specified in nanometers in the MaxwellFDFD implementation, with $\Delta x = \Delta y = 40$ nm. The geometry is extruded over a 10 nm span in the invariant direction for the two-dimensional calculation.

Table S7: Discrete unit-width catalog used in Task 2.

Unit ID	Width (nm)	Description
1	0	empty slot
2	200	narrow strip
3	400	medium strip
4	600	wide strip
5	800	widest strip

Including 400 nm perfectly matched layer padding, the computational window is $x \in [-13.4, 13.4] \mu\text{m}$ and $y \in [-26.4, 3.0] \mu\text{m}$. No absorbing layer is used along the thin invariant direction. The plane-wave source is located in the upper air region at $y = 1.5 \mu\text{m}$, with x polarization and normal incidence along y . Silicon and SiO_2 optical constants use the Palik/Si and Palik/SiO2 entries. Every method uses the same settings and wavelength of 1550 nm.

For all candidate structures, the field map is converted to grayscale and cropped using the same fixed pixel window. Target images use the matching crop. SSIM is calculated for the collimation, dual-beam collimation, and focusing targets. Final performance is the best SSIM found in 5,000 optimization-stage FDFD simulations.

Hybrid initialization and supervised training

The Hybrid SL-RL implementation follows the SL-initialization and RL-optimization arrangement of Ref. (Yeung et al., 2024). Its dataset comprises 5,000 uniformly sampled discrete designs and the corresponding near-field E_x distributions calculated by MaxwellFDFD. Field distributions serve as model inputs and geometry indices as labels. The split is 3,500 training, 750 validation, and 750 test samples. Input images are single-channel grayscale, resized to 128×128 , and normalized to $[-1, 1]$.

The supervised network contains four convolutional blocks with channel widths 64, 128, 256, and 512. Each block consists of a 3×3 convolution, batch normalization, LeakyReLU, and 2×2 average pooling. The features are flattened and passed to a fully connected layer with 100 hidden units and

dropout probability 0.2. The output consists of 13 categorical predictions with five classes each, with logits of shape [Batch, 13, 5].

The network uses cross-entropy loss and the Adam optimizer, with learning rate 10^{-3} and weight decay 10^{-4} . Early stopping is based on validation loss, with a patience of eight epochs. Figure S3 shows the training and validation curves. Hybrid SL-RL and T-DRL use the same Dueling Double DQN network and matched training parameters in the RL stage. The initial design is provided by the supervised network for Hybrid SL-RL and sampled randomly for T-DRL.

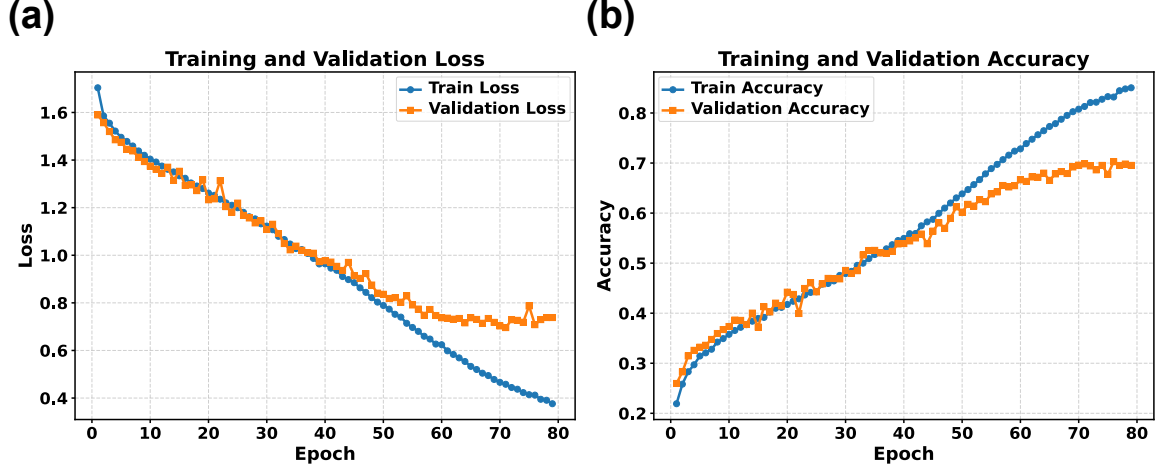


Figure S3: Supervised-model training for Task 2. (a) Training and validation loss. (b) Training and validation accuracy averaged over the 13 parameter indices.

Complete comparison

Table S8 reports the mean and standard deviation of the best SSIM over ten seeds for all five methods and three target patterns.

Table S8: Final design quality in Task 2 after 5,000 optimization-stage FDFD simulations. Mean SSIM and standard deviation are reported over 10 random seeds.

Target pattern	Method	Mean SSIM	Std. dev.
Collimated	T-DRL	0.7346	0.0107
	CEM	0.7196	0.0214
	PSO	0.7233	0.0478
	Hybrid SL-RL (Yeung et al., 2024)	0.8021	0.0362
	DG-DRL	0.9248	0.0090
Dual-collimated	T-DRL	0.7437	0.0506
	CEM	0.8030	0.0351
	PSO	0.7674	0.0487
	Hybrid SL-RL (Yeung et al., 2024)	0.8488	0.0504
	DG-DRL	0.9251	0.0093
Focal	T-DRL	0.6716	0.0283
	CEM	0.7408	0.0192
	PSO	0.7766	0.0635
	Hybrid SL-RL (Yeung et al., 2024)	0.8387	0.0322
	DG-DRL	0.9692	0.0222

S3.3 Task 3: Freeform-grating beam steering

RCWA model and optimizer settings

Reticolo simulates one grating period, which is periodic along x and invariant along the strip direction. The patterned layer is represented by a piecewise-constant permittivity profile over 64 cells, each with width $P/64$. The layer thickness is 325 nm, and the period follows Eq. (6) of the main text. The surrounding media are air and glass, with refractive indices 1 and 1.45, respectively. For the three methods in this study, the wavelength-dependent silicon index is linearly interpolated from the local material table `p_Si.mat`.

The normally incident wave has transverse magnetic polarization. Reticolo returns the efficiencies of the diffraction orders, and the transmitted +1 order supplies η for the reward 100η . The Fourier truncation order is $n_F = 40$, including harmonics from -40 to $+40$. The three methods in this study use the same material data, layer structure, polarization, and order-extraction rules. The nine operating conditions and the limit of 40,000 simulations per seed and condition are defined in the main text.

The DG-DRL environment uses a custom OpenAI Gym interface and calls Reticolo through `matlab.engine`. Its 65-dimensional state contains the 64-cell design and a step indicator. Actions are encoded as 0 for keeping the current cell and 1 for flipping it. CEM and PSO act directly on complete binary candidate structures.

CEM samples independent Bernoulli variables for the 64 cells. Each generation contains 256 candidates, the highest-performing 20% form the elite set, and the distribution is updated with smoothing coefficient 0.7. Each bit probability is clipped to $[0.01, 0.99]$. Binary PSO uses 256 particles, inertia weight $w = 0.729$, cognitive and social coefficients $c_1 = c_2 = 1.49445$, and velocities clipped to $[-4, 4]$. A sigmoid maps the clipped velocities to bit probabilities. Both methods start from random binary designs and stop at the common simulation-cost limit.

Complete controlled results

Figure S4 shows the reported best-so-far curves for the nine conditions. Table S9 lists the mean, standard deviation, and maximum over ten seeds. Tables S10–S12 list the values for individual seeds.

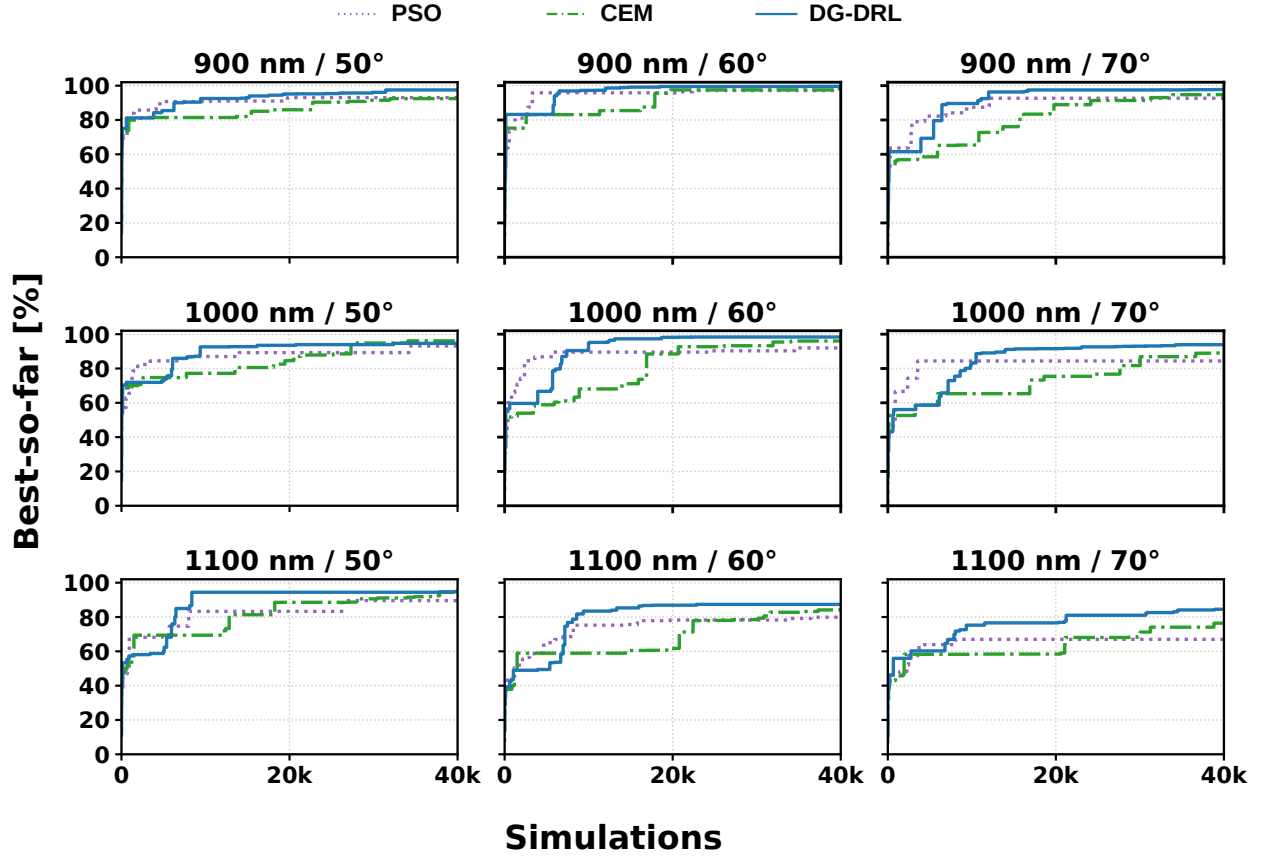


Figure S4: Best-so-far diffraction-efficiency curves for PSO, CEM, and DG-DRL across the nine Task 3 wavelength–angle conditions. The horizontal axis counts completed RCWA simulations.

Table S9: Task 3 controlled comparison under 40,000 RCWA simulations per seed: mean $\pm$ standard deviation and maximum diffraction efficiency (%) over 10 seeds. Values are reported as mean $\pm$ SD (max).

Wavelength (nm)	Angle (deg)	PSO	CEM	DG-DRL
900	50	89.1 $\pm$ 2.3 (93.0)	90.9 $\pm$ 0.8 (92.4)	95.7 $\pm$ 1.6 (97.5)
900	60	94.7 $\pm$ 2.0 (97.2)	96.7 $\pm$ 1.1 (98.2)	99.0 $\pm$ 0.6 (99.6)
900	70	89.6 $\pm$ 3.0 (92.7)	92.7 $\pm$ 1.2 (94.7)	96.3 $\pm$ 1.3 (97.8)
1000	50	88.1 $\pm$ 2.4 (93.1)	91.8 $\pm$ 2.7 (96.1)	93.5 $\pm$ 0.7 (94.7)
1000	60	88.5 $\pm$ 2.0 (92.0)	94.2 $\pm$ 1.1 (96.0)	97.1 $\pm$ 1.2 (98.3)
1000	70	73.6 $\pm$ 6.4 (84.4)	85.8 $\pm$ 1.6 (89.0)	91.2 $\pm$ 2.2 (93.9)
1100	50	81.5 $\pm$ 4.6 (89.6)	91.9 $\pm$ 1.6 (94.8)	92.6 $\pm$ 1.8 (95.2)
1100	60	76.2 $\pm$ 2.6 (79.9)	82.8 $\pm$ 1.1 (84.1)	86.3 $\pm$ 1.1 (87.4)
1100	70	63.2 $\pm$ 3.3 (67.0)	72.9 $\pm$ 1.7 (76.4)	78.2 $\pm$ 3.4 (84.5)

Table S10: Task 3 seed-by-seed best efficiencies (%) for $\lambda = 900$ nm after 40,000 RCWA simulations.

Angle (deg)	Seed	PSO	CEM	DG-DRL
50	1	88.9	90.3	97.4
50	2	86.8	91.4	94.8
50	3	87.6	89.9	96.8
50	4	87.1	90.4	92.3
50	5	91.6	90.5	93.9
50	6	87.3	92.4	97.5
50	7	86.8	90.1	95.8
50	8	91.3	91.6	96.3
50	9	93.0	90.9	95.2
50	10	90.7	91.5	96.9
60	1	94.8	97.8	99.3
60	2	95.7	97.3	98.8
60	3	96.4	95.5	99.6
60	4	92.6	98.2	97.8
60	5	95.5	96.8	99.0
60	6	90.5	96.4	99.1
60	7	95.8	95.5	99.3
60	8	95.3	95.1	99.1
60	9	93.4	97.0	98.2
60	10	97.2	97.5	99.6
70	1	91.9	93.1	97.5
70	2	92.7	94.7	97.5
70	3	82.2	90.3	96.5
70	4	91.7	91.6	94.2
70	5	91.2	93.1	96.9
70	6	90.1	92.4	96.0
70	7	90.2	93.3	97.8
70	8	88.0	93.3	96.2
70	9	88.4	92.5	95.9
70	10	89.9	92.2	94.2

Table S11: Task 3 seed-by-seed best efficiencies (%) for $\lambda = 1000$ nm after 40,000 RCWA simulations.

Angle (deg)	Seed	PSO	CEM	DG-DRL
50	1	85.7	94.9	93.2
50	2	87.0	96.1	93.5
50	3	90.7	91.1	93.9
50	4	88.9	89.9	92.7
50	5	87.0	93.8	93.8
50	6	89.2	93.3	92.5
50	7	86.3	88.2	92.4
50	8	86.1	88.5	93.9
50	9	93.1	92.2	94.7
50	10	87.0	90.6	93.9
60	1	88.8	94.4	98.2
60	2	87.7	94.0	94.2
60	3	88.5	93.6	97.1
60	4	87.7	94.7	97.5
60	5	89.3	93.7	97.9
60	6	84.4	92.9	97.7
60	7	89.5	96.0	98.3
60	8	92.0	92.6	96.7
60	9	90.3	93.9	97.4
60	10	87.1	95.8	96.2
70	1	64.9	84.8	92.2
70	2	66.7	86.7	92.8
70	3	73.9	84.1	91.4
70	4	81.1	85.7	87.9
70	5	77.9	86.3	91.3
70	6	69.7	86.7	93.2
70	7	75.2	83.9	87.8
70	8	74.5	84.0	89.2
70	9	84.4	86.9	93.9
70	10	67.7	89.0	92.2

Table S12: Task 3 seed-by-seed best efficiencies (%) for $\lambda = 1100$ nm after 40,000 RCWA simulations.

Angle (deg)	Seed	PSO	CEM	DG-DRL
50	1	83.4	91.9	90.3
50	2	83.3	94.8	92.2
50	3	80.1	91.9	93.2
50	4	87.7	90.5	91.8
50	5	81.2	92.4	94.4
50	6	77.5	90.8	95.2
50	7	89.6	91.0	93.3
50	8	74.9	91.5	90.1
50	9	78.5	89.8	91.1
50	10	79.1	94.4	94.0
60	1	75.7	83.1	85.8
60	2	78.2	84.1	87.1
60	3	74.9	82.5	86.8
60	4	78.7	82.9	87.4
60	5	73.8	83.7	86.6
60	6	79.9	82.8	86.6
60	7	72.3	80.2	86.9
60	8	75.8	83.0	83.9
60	9	79.2	82.0	84.9
60	10	73.9	83.2	86.9
70	1	67.0	74.0	84.5
70	2	62.5	73.1	76.6
70	3	62.4	73.1	77.2
70	4	66.1	70.7	81.4
70	5	59.0	72.8	78.5
70	6	64.6	74.1	78.1
70	7	66.3	72.3	78.6
70	8	63.5	76.4	79.1
70	9	56.7	71.3	71.5
70	10	63.9	71.3	76.8

Published reference values

Tables S13–S15 collect the published values used for reference in the main text (Jiang and Fan, 2019; Seo et al., 2022; Park et al., 2024). Each table retains the reported number of simulations and efficiency statistics.

Table S13: Task 3 literature reference values for $\lambda = 900$ nm. Efficiencies are reported in percent.

Condition	Year	Reference	Method	Simulations (k)	Max.	Mean $\pm$ SD
900 nm, 50°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	93	64 $\pm$ 16
900 nm, 50°	2019	(Jiang and Fan, 2019)	GLOnet	200	98	90 $\pm$ 10
900 nm, 50°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	81	81 $\pm$ 0
900 nm, 50°	2022	(Seo et al., 2022)	RL	2000	98.7	97.6 $\pm$ 1.1
900 nm, 50°	2024	(Park et al., 2024)	PIRL	200	97.4	96 $\pm$ 1.3
900 nm, 50°	2026	This work	DG-DRL	40	97.5	95.7 $\pm$ 1.6
900 nm, 60°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	93	59 $\pm$ 18
900 nm, 60°	2019	(Jiang and Fan, 2019)	GLOnet	200	97	73 $\pm$ 18
900 nm, 60°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	99.2	99.2 $\pm$ 0
900 nm, 60°	2022	(Seo et al., 2022)	RL	2000	99.7	99.5 $\pm$ 0.2
900 nm, 60°	2024	(Park et al., 2024)	PIRL	200	99.7	99.5 $\pm$ 0.2
900 nm, 60°	2026	This work	DG-DRL	40	99.6	99.0 $\pm$ 0.6
900 nm, 70°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	92	59 $\pm$ 13
900 nm, 70°	2019	(Jiang and Fan, 2019)	GLOnet	200	98	83 $\pm$ 14
900 nm, 70°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	79.7	79.7 $\pm$ 0
900 nm, 70°	2022	(Seo et al., 2022)	RL	2000	98.3	98 $\pm$ 0.3
900 nm, 70°	2024	(Park et al., 2024)	PIRL	200	99.1	98.6 $\pm$ 0.4
900 nm, 70°	2026	This work	DG-DRL	40	97.8	96.3 $\pm$ 1.3

Table S14: Task 3 literature reference values for $\lambda = 1000$ nm. Efficiencies are reported in percent.

Condition	Year	Reference	Method	Simulations (k)	Max.	Mean $\pm$ SD
1000 nm, 50°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	95	55 $\pm$ 16
1000 nm, 50°	2019	(Jiang and Fan, 2019)	GLOnet	200	96	85 $\pm$ 12
1000 nm, 50°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	96.5	96.5 $\pm$ 0
1000 nm, 50°	2022	(Seo et al., 2022)	RL	2000	96.9	96.3 $\pm$ 0.5
1000 nm, 50°	2024	(Park et al., 2024)	PIRL	200	97.9	91.1 $\pm$ 6.2
1000 nm, 50°	2026	This work	DG-DRL	40	94.7	93.5 $\pm$ 0.7
1000 nm, 60°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	92	56 $\pm$ 14
1000 nm, 60°	2019	(Jiang and Fan, 2019)	GLOnet	200	98	85 $\pm$ 17
1000 nm, 60°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	99	99 $\pm$ 0
1000 nm, 60°	2022	(Seo et al., 2022)	RL	2000	98.8	98.7 $\pm$ 0.1
1000 nm, 60°	2024	(Park et al., 2024)	PIRL	200	98.7	97.4 $\pm$ 1.3
1000 nm, 60°	2026	This work	DG-DRL	40	98.3	97.1 $\pm$ 1.2
1000 nm, 70°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	84	62 $\pm$ 12
1000 nm, 70°	2019	(Jiang and Fan, 2019)	GLOnet	200	93	76 $\pm$ 18
1000 nm, 70°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	58.1	58.1 $\pm$ 0
1000 nm, 70°	2022	(Seo et al., 2022)	RL	2000	94.7	94.4 $\pm$ 0.4
1000 nm, 70°	2024	(Park et al., 2024)	PIRL	200	92.6	87.8 $\pm$ 4.3
1000 nm, 70°	2026	This work	DG-DRL	40	93.9	91.2 $\pm$ 2.2

Table S15: Task 3 literature reference values for $\lambda = 1100$ nm. Efficiencies are reported in percent.

Condition	Year	Reference	Method	Simulations (k)	Max.	Mean $\pm$ SD
1100 nm, 50°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	91	49 $\pm$ 10
1100 nm, 50°	2019	(Jiang and Fan, 2019)	GLOnet	200	91	77 $\pm$ 11
1100 nm, 50°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	71.2	71.2 $\pm$ 0
1100 nm, 50°	2022	(Seo et al., 2022)	RL	2000	98.4	98.3 $\pm$ 0.05
1100 nm, 50°	2024	(Park et al., 2024)	PIRL	200	95.4	93.3 $\pm$ 3.5
1100 nm, 50°	2026	This work	DG-DRL	40	95.2	92.6 $\pm$ 1.8
1100 nm, 60°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	79	52 $\pm$ 15
1100 nm, 60°	2019	(Jiang and Fan, 2019)	GLOnet	200	80	59 $\pm$ 17
1100 nm, 60°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	79.8	79.8 $\pm$ 0
1100 nm, 60°	2022	(Seo et al., 2022)	RL	2000	87.1	86.9 $\pm$ 0.2
1100 nm, 60°	2024	(Park et al., 2024)	PIRL	200	94.9	89.6 $\pm$ 4.7
1100 nm, 60°	2026	This work	DG-DRL	40	87.4	86.3 $\pm$ 1.1
1100 nm, 70°	2019	(Jiang and Fan, 2019)	Adjoint-based	200	84	59 $\pm$ 14
1100 nm, 70°	2019	(Jiang and Fan, 2019)	GLOnet	200	84	65 $\pm$ 14
1100 nm, 70°	2022	(Seo et al., 2022)	Greedy (depth=2)	200	58.7	58.7 $\pm$ 0
1100 nm, 70°	2022	(Seo et al., 2022)	RL	2000	78.5	78.3 $\pm$ 0.1
1100 nm, 70°	2024	(Park et al., 2024)	PIRL	200	84.1	77.8 $\pm$ 3.6
1100 nm, 70°	2026	This work	DG-DRL	40	84.5	78.2 $\pm$ 3.4

References

- Jiang, J., Fan, J.A.: Global optimization of dielectric metasurfaces using a physics-driven neural network. *Nano Letters* **19**(8), 5366–5372 (2019) <https://doi.org/10.1021/acs.nanolett.9b01857>
- Park, C., Kim, S., Jung, A.W., Park, J., Seo, D., Kim, Y., Park, C., Park, C.Y., Jang, M.S.: Sample-efficient inverse design of freeform nanophotonic devices with physics-informed reinforcement learning. *Nanophotonics* **13**(8), 1483–1492 (2024) <https://doi.org/10.1515/nanoph-2023-0852>
- Seo, D., Nam, D.W., Park, J., Park, C.Y., Jang, M.S.: Structural optimization of a one-dimensional freeform metagrating deflector via deep reinforcement learning. *ACS Photonics* **9**(2), 452–458 (2022) <https://doi.org/10.1021/acsp Photonics.1c00839>
- Yeung, C., Pham, B., Zhang, Z., Fountaine, K.T., Raman, A.P.: Hybrid supervised and reinforcement learning for the design and optimization of nanophotonic structures. *Optics Express* **32**(6), 9920–9930 (2024) <https://doi.org/10.1364/OE.512159>