LithoSim: A Large, Holistic Lithography Simulation Benchmark for AI-Driven Semiconductor Manufacturing

Hongquan He[1] Zhen Wang[1] Jingya Wang[1] Tao Wu[1]
Xuming He[1] Bei Yu[2] Jingyi Yu[1†] Hao Geng[1†]
[1]ShanghaiTech University [2]The Chinese University of Hong Kong

Abstract

Lithography orchestrates a symphony of light, mask and photochemicals to transfer the integrated circuit patterns onto the wafer. Lithography simulation serves as the critical nexus between circuit design and manufacturing, where its speed and accuracy fundamentally govern the optimization quality of downstream resolution enhancement techniques (RETs). While machine learning promises to circumvent computational limitations of lithography process through data-driven or physics-informed approximations of computational lithography, existing simulators suffer from inadequate lithographic awareness due to insufficient training data capturing essential process variations and mask correction rules. We present LithoSim, the most comprehensive lithography simulation benchmark to date, featuring over 4 million high-resolution input-output pairs with rigorous physical correspondence. The dataset systematically incorporates alterable optical source distributions, metal and via mask topologies with optical proximity correction (OPC) variants, and process windows reflecting fab-realistic variations. By integrating domain-specific metrics spanning AI performance and lithographic fidelity, LithoSim establishes a unified evaluation framework for data-driven and physics-informed computational lithography. The data (https://huggingface.co/datasets/grandiflorum/LithoSim), code (https://dw-hongquan.github.io/LithoSim), and pre-trained models (https://huggingface.co/grandiflorum/LithoSim) are released openly to support the development of hybrid ML-based and high-fidelity lithography simulation for the benefit of semiconductor manufacturing.

1 Introduction

Simulation stands as a cornerstone of modern artificial intelligence (AI), enabling data-driven emulation of complex physical system, from protein folding dynamics [1] to climate modeling [2, 3, 4]. These AI-powered simulation not only accelerate computational cost [5, 6] but also unlock closed-loop optimization paradigms by bridging synthetic data simulation with differentiable physical models [7, 8]. A critical application of this paradigm is in semiconductor manufacturing, specifically lithography simulation [9, 10]. Lithography is a optical and chemical system of transferring intricate circuit patterns $\mathbf{M}$ onto silicon wafers using light $\mathbf{J}$ with a fixed projector $\mathbf{H}$ depicted in Figure 1(b). However, at nanometer scales, fundamental physics of optical diffraction and unavoidable manufacturing variations (*a.k.a.* process variations) distort the intended patterns. Techniques called resolution enhancement technologies (RETs) [11, 12, 13], such as optical proximity correction (OPC) and source mask optimization (SMO) shown in Figure 2, are used to pre-distort the design masks to

†Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

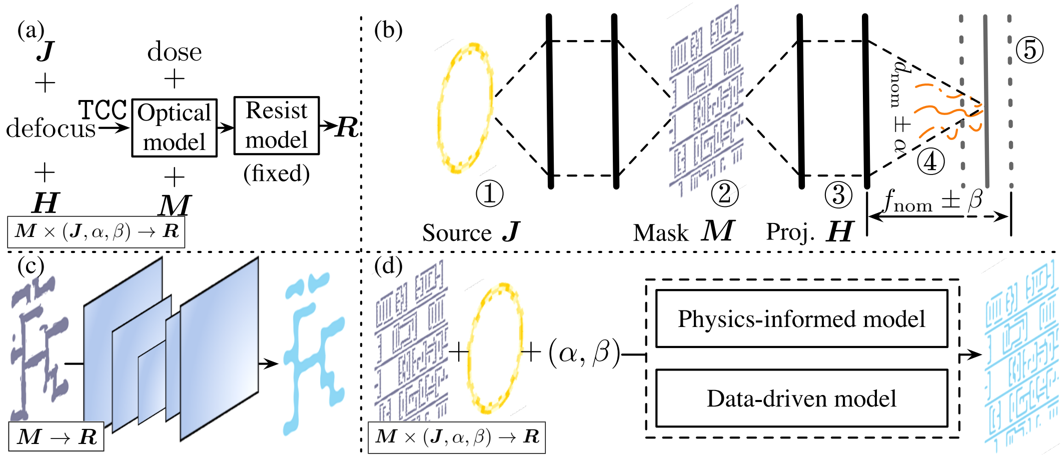

Figure 1: (a). Lithography simulation tools by combining source and defocus on a fixed projector to create an optical model (TCC in Appendix A.1), then using mask and dose inputs to generate resist. (b). Physical lithography setup comprises 5 primary elements: ① An adjustable illumination system $\mathbf{J}$. ② The mask setup $\mathbf{M}$ with a basic binary design made of see-through and non-see-through sections. ③ An aligned and fixed projection module $\mathbf{H}$. ④ An exposure mechanism with 2 critical process variation (α, β). ⑤ A resist station to yield the end producted resist $\mathbf{R}$. (c). Previous benchmark [14] at 45nm node considering source and process variations as constants, using DNNs for limited surrogate models $\mathbf{M} \rightarrow \mathbf{R}$. (d). Our benchmark at sub-28nm node considers simulation across larger mask ranges with source and process variations, using data-driven or physics-informed generative models for holistic simulation $\mathbf{M} \times (\mathbf{J}, \alpha, \beta) \rightarrow \mathbf{R}$ with all the elements ① to ⑤.

Table 1: Comparison of Lithography Simulation Benchmarks.

Items		CAD13 [15]	ISPD19 [16]	N14 [17]	LithoBench [14]	**LithoSim**
# of Var.	source	×	×	×	×	620
	dose	3	×	×	×	13
	defocus	2	×	×	×	5
Mask Config.	Type	M	V	V	M/V	M/OPC-M/V/OPC-V
	Num.	5k	21k	1.6k	16k/115k	1210 for each
	Size	4	4	4	4/1	16 for all
Tech. Node	Metal	32	–	–	32	$14 \sim 28$
	Via	–	40	14	45	$14 \sim 40$
# of Output	resist	30k	21k	1.6k	131k	$>$ **4M**

Mask size and Tech. node measurements in μm^2 and nm, respectively. $\text{k} = 1,000$, $\text{M} = 1,000,000$.
In mask type, M: `Metal`, V: `Via`, OPC: mask optimization for compensating optical diffraction.

compensate for these effects, aiming to print the desired pattern accurately. RET workflows heavily rely on simulating the lithography process.

Traditional lithography simulators use complex physical models like Figure 1(a) that are computationally extremely expensive with $> 10^3$ CPU-hours per square millimeter. This bottleneck makes simulation and RET optimization slow and impractical. Machine Learning (ML) offers a promising path by learning differentiable image-to-image translation between the input mask design $\mathbf{M}$ and the final printed resist pattern $\mathbf{R}$ in Figure 1(c), bypassing the costly physics solvers to create fast surrogate models [18].

However, current public datasets for ML-based lithography such as CAD13 [15], ISPD19 [16], N14 [17], and LithoBench [14] listed in Table 1 are inadequate for developing models that meet RET requirements, suffering from 3 key limitations. First, datasets are outdated scaling, primarily covering older $32 \sim 45\text{nm}$ technology nodes, not the cutting-edge sub-28nm nodes used in advanced lithography. Second, mask scales, typically $\leq 4\mu\text{m}^2$, are too small to capture crucial optical proximity effects. Third, They lack 4 essential variations that RET must handle: different types of mask $\mathbf{M}$ with

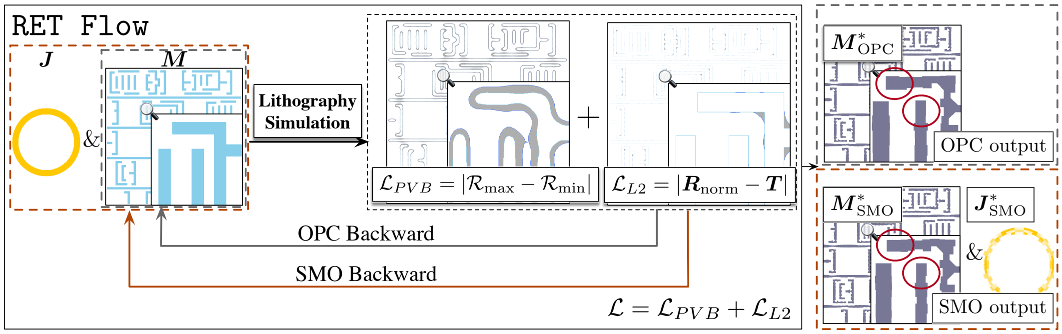

Figure 2: Overview of RETs (OPC and SMO).

or without OPC [19], variations in the light source shape and intensity $\mathbf{J}$, fluctuations in exposure dose α. deviations in defocus β of fixed projector $\mathbf{H}$ shown in Figure 1(b). This lack of realism severely limits the practical lithography of models trained on current datasets for differentiable optimization applications [11, 12, 13, 19, 20, 21, 22, 23, 24, 25, 26, 27] in Figure 2, where accurately simulating the interaction of all these physical variables is essential.

For real semiconductor manufacturing, Lithography simulation forms the computational backbone of modern RETs. As illustrated in Figure 2, RET relies on iterative, physics-aware feedback from the simulator to optimize mask $\mathbf{M}$ and illumination source $\mathbf{J}$ listed in Appendix A.2. The goal is to minimize a composite loss function that includes the contour fidelity under nominal conditions and robustness across process variations. This necessitates a simulator that is not only fast but also accurately models the interaction of all physical variables, a capability absent in existing benchmarks.

To address those critical gaps, we introduce LithoSim illustrated in Figure 1(d), a comprehensive benchmark designed to enable the development and evaluation of ML models for practical lithography simulation and differentiable RET optimization. LithoSim provides:

- Masks at sub-28nm nodes, both with and without OPC, covering larger scales with $16\mu\text{m}^2$ to capture proximity effects.
- Extensive parametric combinations listed in Table 1 covering over 600 distinct source configurations with `annular`, `quadrupole`, and `dipole` illuminations, 13 dose variation levels spanning from -12% to $+12\%$ of the nominal value, and 5 defocus offsets over a range of $\pm 80\text{nm}$, mirroring the key variables in RET flow.
- The first unified benchmark and evaluation framework to assess modern deep learning architectures (CNNs [28], Vision Transformers [29], physics-informed models like FNO-based flows [21, 17], and SOCS [30, 10]) on their ability to simulate lithography patterns with metrics both in ML and lithography domain.
- A unique out-of-distribution (OOD) evaluation for model generalization under varying mask conditions, a core requirement for RET. Specifically, it allows testing models trained on OPC'ed masks on completely non-OPC'ed masks, as well as the reverse. This directly mimics the flow where mask $\mathbf{M}$ is iteratively modified during differentiable optimization like OPC and SMO, while a robust simulator must remain accurate even as the input mask changes significantly between iterations.

2 Related Work

The advancement of ML-based lithography simulators is hampered by the limited scope of existing metal [15] and via [16, 17] datasets. Current metal layer data, such as CAD13 [15], relies predominantly on only 10 base patterns at the 32nm node, artificially augmented via rotation and reflection to generate $4,875$ synthetic variants under identical design rules [11]. Similarly, via layer datasets comprise fragmented sub-regions of full-chip layouts [16, 17], simulated under idealized and fixed process conditions. While these resources have facilitated initial research into data-driven architectures [28, 31, 32] and physics-informed models [17, 30], they fundamentally lack critical manufacturing parameters, especially notably, realistic source illuminator profiles and process varia-

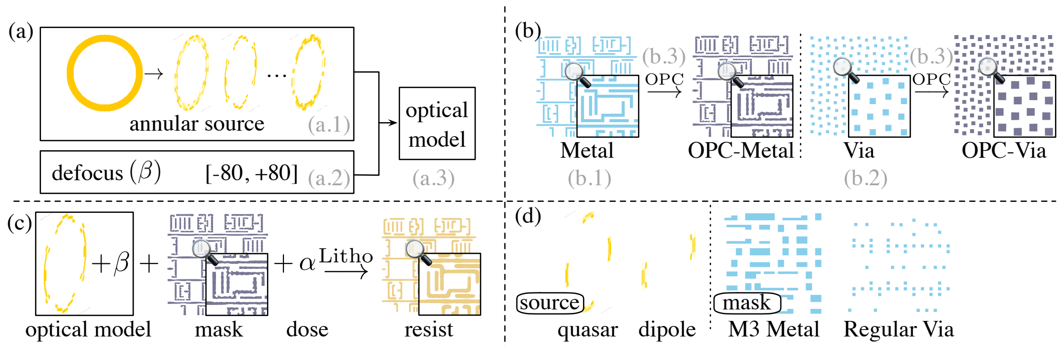

Figure 3: LithoSim Dataset Collection Pipeline. (a). Source & Optical Model Generation: (a.1) generates sources $\mathbf{J}$ with normalized intensity distributions, (a.2) applies defocus levels per source, (a.3) output >1,800 optical models via simulation tool. (b). Mask Preparation: (b.1) extracts layout clips from full-chip M1 design. (b.2) synthesizes via layouts with foundry design rules. (b.3) processes original designs with OPC to generate OPC'ed masks. (c). Resist Synthesis: simulates resists $\mathbf{R}$ via simulation tool. (d).Out-of-Distribution (OOD) Benchmark: generates for testing model generalization to unseen conditions.

tions used in rigorous simulation [9, 10, 33]. This omission ultimately renders the resulting models unsuitable for RET-oriented optimization.

Recent advances in fabrication-aware neural lithography, such as bilevel optimization [26, 13] and end-to-end differentiable neural pipeline [24, 34], highlight the critical role of manufacturing-digital twins in closing design-to-fabrication gaps. While these approaches [26, 24] excel at modeling post-lithography 3D topography, LithoSim addresses a complementary challenge: it focus on predicting resist contours under optical and process variations directly supports emerging differentiable ILT like [20, 23, 22]. By providing standardized evaluation of PV-band generalization critical for mask optimization in [25, 35, 27], LithoSim bridges the gap between high-fidelity physical emulation and optimizers requiring differentiable surrogates.

3 LithoSim Dataset Construction

3.1 Experiment Outline

The experimental framework formulates lithography simulation as a high-dimensional regression problem with four complementary input modalities: source, mask, dose, and defocus. Mask inputs $\mathbf{M} \in \{0,1\}^{W\times H}$ represent binary patterns at sub-28nm resolutions, with dimensions scaling as $16\mu\text{m}^2$ ($W = H = 4096$) to capture proximity effects. Source configurations $\mathbf{J} \in \mathbb{R}^{N\times 3}$ describe particular light source distribution through N discrete source points j_i ($i \in [0, N)$), each defined by normalized intensity $v_i \in [0, 1]$ and Cartesian coordinates $(x_i, y_i) \in [-1, 1]^2$. Dose and defocus parameters (α, β) introduce controlled process variations, where dose modulates exposure energy $\alpha \in [-0.12, 0.12]$ while defocus emulates lens aberrations $\beta \in [-80, 80]$, conforming commonly used variations in real simulations.

All the experiments is trained and tested with 4 H100 Graphics cards with Intel Core Xeon Platinum 8462Y+ processors with Adam optimizer and a 10^{-4} learning rate of 10^{-5} weight decay. Either a linear combination of BCE and Dice loss, or only MSE is used as loss fuction.

3.2 Dataset Collection

The dataset is generated through a scalable lithography simulation pipeline executed on 100 parallelized CPUs, following the simulation flow illustrated in Figure 1 (a). To ensure diversity and physical fidelity, the source and mask integrates manufacturing-specific design rules and rigorous computational lithography principles. Following advanced lithography, we set NA is 1.35 and wavelength is 193nm, incorporating Zernike lens aberrations up to 37 terms. Optical model needs to be built before the simulation and each requires approximately 40 minutes to complete the process.

Table 2: Details of LithoSim Benchmark.

Dataset	Train	Val	Test	Total
OPC-Metal	$693,330$	$99,000$	$198,000$	$990,330$
Metal	$903,672$	$129,096$	$258,423$	$1,291,191$
OPC-Via	$655,842$	$93,654$	$187,341$	$936,837$
Via	$607,365$	$86,757$	$173,514$	$867,636$
OOD	–	–	$1,580$	$1,580$

Subsequent resist simulations consume 15 seconds per pattern, generating final 4096×4096 images with $1\text{nm}/\text{pixel}$ resolution.

Source and Optical Model Generation. A total of more than 600 annular illumination sources with $0 \sim 1$ normalized intensity distributions are first synthesized based on the central symmetry of off-axis illumination as well as the classical values of the inner and outer radium. For each source, three defocus values (-40nm, 0nm, $+40\text{nm}$) are applied to simulate process variations, yielding more than $1,800$ unique optical models using the rigorous lithography simulator following Figure 3 (a).

Mask Preparation. As illustrated in Figure 3 (b), two types of mask are constructed: (1) **Metal Layer:** $1,200$ layout clips (16nm^2 each) are extracted from a full-chip M1 layer design. These clips are processed through optical proximity correction (OPC) using the optical models, generating paired `Metal` (original) and `OPC-Metal` (corrected) mask sets; (2) **Via Layer:** $1,200$ via layouts adhering to foundry design rules are synthesized and similarly corrected via OPC, producing `Via` and `OPC-Via` mask sets.

Resist Synthesis. 4 types of mask datasets (`Metal`, `OPC-Metal`, `Via`, `OPC-Via`) are combined with $\pm 10\%$ normalized dose variations and calculated through corresponding optical model to simulate resist profiles. This cross-condition sampling strategy produces a comprehensive in-distribution dataset capturing multi-physics interactions across source distributions, mask types, and process variations shown in Figure 3 (c).

Out-of Distribution (OOD) Dataset. To evaluate model generalization, 20 additional illumination sources (10 dipole, 10 quasar) are designed. These sources with ± 80 defocus and $\pm 12\%$ dose are paired with 20 layout clips from M3 and via layers of a distinct CPU design illustrated in Figure 3. The OOD dataset is generated using identical simulation pipelines but exhibits structural and process condition disparities compared to the primary dataset.

Following the above data collection guidelines, LithoSim benchmark in Table 2 combines high-throughput computational lithography with AI-oriented data diversity, producing widely distributed multi-parameters (source, mask, dose, defocus) to resists mappings. Figure 4 visualizes all the datasets with different lithography conditions in LithoSim. OPC'ed masks (OPC-Metal and OPC-Via) yield resists with smaller edge placement error compared with non-OPC'ed masks. Compared to nominal dose, a positive deviation expands resist area while a negative deviation induces the undercut of resist. Defocus perturbations introduce subtler but critical effects, inducing $< 1\text{nm}$ resist contour shifts that ML-based models must capture to enable robust RET. A slight bias also occupies in the correction of the same mask by different light sources. The systematic variation of optical models, mask corrections, and essential process parameters establishes a robust foundation for data-driven and physics-informed lithography modeling.

3.3 Dataset Split

LithoSim benchmark in Table 2 is partitioned to evaluate simulation performance across in-distribution and out-of-distribution (OOD) scenarios. For each mask category (`Metal`, `OPC-Metal`, `Via`, `OPC-Via`), the corresponding data samples are stratified into training (70%), validation (10%), and testing (20%) subsets. The validation set serves for hyperparameter tuning and early stopping, while the test set quantifies in-distribution predictive accuracy. Crucially, splits are performed independently per mask type to prevent cross-contamination between original (`Metal`, `Via`) and OPC-corrected (`OPC-Metal`, `OPC-Via`) layouts, mitigating biases in learning mask-correction synergies.

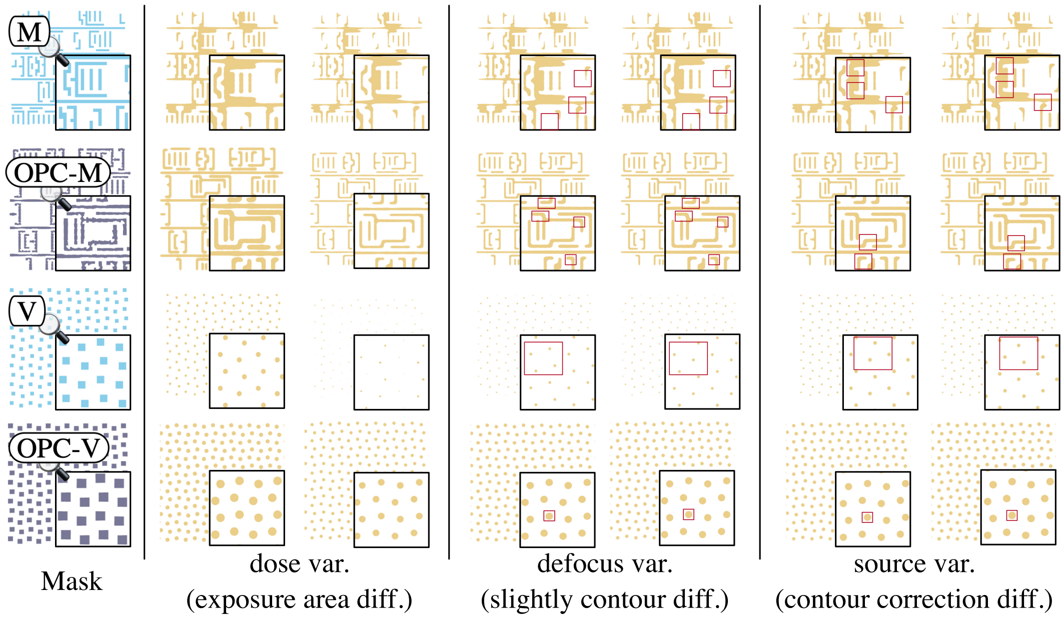

Figure 4: The splits of LithoSim benchmark and comparison of different lithography conditions.

To assess generalization beyond training distributions, the OOD dataset, comprising M3/via layer clips and unconventional illumination (dipole, quasar), is reserved exclusively for testing. This separation ensures that OOD evaluation reflects real-world scenarios where models encounter unseen design rules, optical conditions, or process variation drifts.

4 Experiments

4.1 Baseline Architectures

We establish six baseline architectures for lithography simulation, comprising 2 data-driven models (ED-CNN, ED-Trans) and 4 physics-informed variants (RFNO, CFNO, MFNO, SOCS). We also add an **electromagnetic (EM) approximation method** as an upper bound. ML models should asymptotically approach this white-box results. Crucially, a viable ML model must demonstrate robust performance not only on in-distribution data but also on out-of-distribution cases. This requirement stems from the industry's stringent criterion that edge placement error (EPE) should remain below $1\mathrm{nm}$ in lithography simulation regardless of mask variations to qualify for small-scale industrial testing. Implementation details appear in Appendix A.4.

All baselines share unified conditional encoding schemes: 2D continuous positional encoding and chunk-based compression with dynamic query generation as well as hierarchical attention (intra-chunk local attention followed by cross-chunk global aggregation) for source coordinates ($[B, N, 3] \rightarrow [B, N, D] \rightarrow [B, K, D], K \ll N$) while 1D encoding for dose and defocus variations ($[B, 1] \rightarrow [B, D]$). All condition is embedded into backbones using chunked litho-aware attention, which enables memory-efficient cross-attention between masks and lithography parameters (source, dose, defocus) through compressed conditions and chunk-wise computation. Implementation details appear in Appendix A.3.

Encoder-Decoder CNN (ED-CNN). It implements hierarchical encoder-decoder processing following CNN-based [28, 31, 36] flow. The encoder employs cascaded ResNet blocks with channel-wise multipliers and non-local attention at specified resolutions. Chunked cross-attention fuses physical parameters (source, dose, and defocus) at the bottleneck. The decoder uses transposed convolutions and residual attention blocks for detail-preserving upsampling.

Encoder-Decoder Transformer (ED-Trans). It introduces spatial-domain transformers [37] through patch embedding and sequence processing. Input masks are projected and reshaped to enable standard

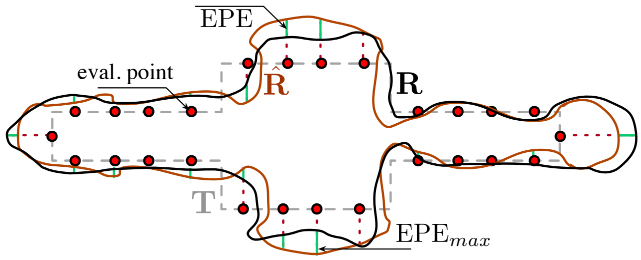

Figure 5: Illustration of EPE calculation.

transformer operations in sequence. Then, sequence-based masks are processed by four transformer layers with learnable positional encoding. Following cross-attention fusion of source, dose, and defocus, depth-wise transformers enable hierarchical abstraction before spatial reconstruction via inverse projection.

Reduced FNO (RFNO). It employs a reduced Fourier neural operator [17, 38] in Figure 6 that operates masks in the frequency domain through parameterized low-rank kernel convolutions in Fourier space, utilizing truncated mode interactions to capture global low frequency. The architecture explicitly bridges spectral (RFNO) and spatial (CNN) representations, then fuses physical parameters to reconstruct final resists.

Convolutional FNO (CFNO). It introduces a convolutional Fourier neural operator [21] in Figure 7 that synergizes spectral transformations with spatial convolutions for efficient operator learning. CFNO first decomposes high-dimensional masks into localized patches, projects them into the Fourier domain via FFT, and applies a parameterized linear transformation to capture global interactions. These spectral features are then mapped back to the spatial domain through inverse FFT and restructured into geometric masks. The architecture also cat spectral CFNO and spatial CNN, then fuses physical parameters and generate resists via a decoder.

Mixed FNO (MFNO). It leverages both global frequency-domain correlations [39] through CFNO and spatial locality via RFNO [17]. MFNO in Figure 8 processes masks through spectral decomposition in localized chunks, employing separate parameterized weights for low-frequency and high-frequency components to enable multiscale frequency modulation. Masks are partitioned into spatial chunks where Fourier transforms extract frequency features, with truncated modes reducing computational complexity while preserving dominant spectral patterns. A CNN enhances local feature interactions after inverse Fourier reconstruction.

Sum of Coherent Sources (SOCS). It introduces a physics-inspired framework, drawing parallels to the optical lithography process [30, 10] (see Appendix A.1). SOCS first transforms masks into frequency domain through FFT, followed by a dedicated complex-valued encoder comprising cascaded complex-ResNet blocks and complex-attention mechanisms to preserve phase-aware representations. Lithography parameters are adaptively integrated through chunked complex litho-aware attention, enabling parameter-conditioned feature fusion. Spatial details are subsequently recovered via a complex decoder architecture that synergistically combines complex transposed convolution operators with complex-ResNet. The final resist profile prediction is achieved through inverse FFT (IFFT) followed by sum of mask decomposition to reconstruct the spatial-domain results.

4.2 Evaluation Metrics

We provide popular deterministic metrics in the machine learning and lithography simulation on semiconductor manufacture, including MSE, PA, mIOU, EPE$_{max}$ and EPE$_{avg}$. Details of metrics are list in Appendix A.5.

- **Mean Squared Error (MSE)** is sensitive to penalize outliers, which is critical for evaluating generalization ability of lithography simulation (Eq. 19).
- **Pixel Accuracy (PA)** is used to evaluate overall accuracy of resists (Eq. 20).
- **Intersection Over Union (IOU)** is used to evaluate detailed pixel differences of resists (Eq. 21).

Table 3: Comparison of multi-scale ML-based lithography simulation.

Data	Method	MSE	PA	IOU	$\mathrm{EPE}_{\mathrm{max}}$	$\mathrm{EPE}_{\mathrm{avg}}$	TAT
		$\times 10^{-3}(\downarrow)$	%(↑)	%(↑)	nm(↓)	nm(↓)	ms(↓)
OPC-Metal	ED-CNN	11.51 ± 8.39	98.85 ± 0.84	91.06 ± 6.30	1.75 ± 0.49	1.47 ± 0.28	8.94 ± 0.24
	ED-Trans	19.67 ± 9.32	98.03 ± 1.33	85.12 ± 9.15	2.03 ± 0.69	1.81 ± 0.72	11.64 ± 0.25
	RFNO	9.72 ± 6.49	99.03 ± 0.65	92.42 ± 5.00	1.70 ± 0.56	1.12 ± 0.38	9.91 ± 0.31
	CFNO	20.15 ± 9.24	97.98 ± 1.42	84.84 ± 1.42	2.96 ± 0.72	2.36 ± 0.59	10.00 ± 0.42
	MFNO	**6.28** ± 3.84	**99.37** ± 0.38	**95.29** ± 2.44	**1.29** ± 0.28	**1.02** ± 0.27	9.98 ± 0.31
	SOCS	7.94 ± 4.30	99.18 ± 0.51	93.17 ± 5.67	1.55 ± 0.32	1.07 ± 0.58	**8.51** ± 0.20
	EM	5.83	99.51	97.32	1.02	0.74	289.62×10^3
Metal	ED-CNN	**8.06** ± 5.17	**99.09** ± 0.52	**92.32** ± 4.62	**1.64** ± 0.24	**1.30** ± 0.40	9.12 ± 0.27
	ED-Trans	9.89 ± 5.14	99.01 ± 0.51	91.69 ± 5.14	1.71 ± 0.30	1.37 ± 0.25	11.30 ± 0.24
	RFNO	13.59 ± 8.94	98.64 ± 0.89	88.58 ± 7.29	1.84 ± 0.36	1.53 ± 0.39	9.49 ± 0.29
	CFNO	13.08 ± 7.13	98.69 ± 0.71	89.11 ± 5.81	1.71 ± 0.48	1.50 ± 0.48	10.00 ± 0.25
	MFNO	8.39 ± 5.38	99.03 ± 0.54	92.18 ± 3.62	1.65 ± 0.33	1.33 ± 0.29	10.55 ± 0.34
	SOCS	9.35 ± 7.39	99.01 ± 0.70	91.95 ± 6.31	1.82 ± 0.38	1.35 ± 0.38	**8.20** ± 0.22
	EM	7.05	99.33	96.05	1.31	0.99	281.07×10^3
OPC-Via	ED-CNN	5.36 ± 3.96	99.46 ± 0.40	90.68 ± 5.65	1.96 ± 0.37	1.59 ± 0.42	9.03 ± 0.24
	ED-Trans	5.56 ± 3.63	99.44 ± 0.36	89.99 ± 5.72	1.94 ± 0.30	1.67 ± 0.33	11.14 ± 0.22
	RFNO	**3.91** ± 2.28	**99.61** ± 0.23	**92.68** ± 4.32	**1.85** ± 0.27	**1.42** ± 0.28	9.70 ± 0.31
	CFNO	6.87 ± 5.14	99.31 ± 0.51	87.67 ± 8.09	2.10 ± 0.41	1.84 ± 0.35	10.19 ± 0.27
	MFNO	6.04 ± 3.97	99.40 ± 0.39	90.99 ± 4.54	1.99 ± 0.25	1.50 ± 0.23	10.72 ± 0.34
	SOCS	5.28 ± 3.61	99.49 ± 0.52	91.12 ± 7.02	1.98 ± 0.63	1.79 ± 0.58	**8.02** ± 0.27
	EM	3.52	99.71	95.65	1.15	0.92	276.19×10^3
Via	ED-CNN	**4.65** ± 2.95	**99.54** ± 0.30	81.39 ± 8.95	1.07 ± 0.29	0.93 ± 0.34	8.92 ± 0.24
	ED-Trans	5.69 ± 4.20	99.43 ± 0.42	77.93 ± 9.54	1.36 ± 0.53	0.97 ± 0.49	11.30 ± 0.23
	RFNO	4.77 ± 4.02	**99.54** ± 0.40	**83.10** ± 3.30	**1.03** ± 0.11	**0.89** ± 0.10	9.23 ± 0.28
	CFNO	5.94 ± 4.42	99.41 ± 0.44	76.19 ± 9.70	1.37 ± 0.38	1.01 ± 0.40	9.58 ± 0.24
	MFNO	6.39 ± 1.24	99.36 ± 4.84	73.52 ± 4.73	1.47 ± 0.13	1.02 ± 0.12	9.97 ± 0.24
	SOCS	5.09 ± 4.20	99.58 ± 0.33	80.47 ± 4.46	1.24 ± 0.41	0.99 ± 0.40	**7.82** ± 0.22
	EM	3.41	99.79	89.20	0.75	0.64	274.40×10^3

- **Edge Placement Error ($\mathrm{EPE}_{\mathrm{max}}$/$\mathrm{EPE}_{\mathrm{avg}}$)** is a critical indicator for assessing alignment discrepancies in semiconductor manufacturing. As illustrated in Figure 5, it evaluates the reliability of the lithography simulation by calculating the distance between the predicted resist contour and the ground truth after selecting evaluation points on the layout.
- **Turn Around Time (TAT)** is the total amount of time spent by simulation process from coming in the ready state for the first time to its completion.

4.3 Baseline Model Results

Table 3 summarizes the lithography simulation efficiency of all 6 baseline models across 4 mask categories of LithoSim. Each metric in Table 3 is followed by the standard deviation of the corresponding dataset. More experimental settings is list in Appendix A.6.

Overall simulation accuracy and speed. While Transformer-based baseline (ED-Trans) incur the highest turnaround time (TAT) due to hybrid global attention operations, resulting in substantial computational and memory requirements. SOCS achieves minimal latency by strictly adhering to the Hopkins-based frequency-domain encoding-decoding paradigm. Physics-informed models generally exhibit comparable TATs, with data-driven approaches (ED-CNN, ED-Trans) demonstrating overall competitive lithographic awareness when trained on the large scale of LithoSim, a testament to the dataset's capacity to compensate for inductive biases of lithography through sheer data volume.

Data-driven baseline comparison. ED-CNN, leveraging its CNN backbone enhanced with spatial-channel attention mechanisms, marginally outperforms ED-Trans across all metrics, particularly excelling on Metal datasets (*i.e.* OPC-Metal: 91.06% IOU, Metal: 92.32% IOU). This superiority stems from hierarchical capacity of ED-CNN to resolve local mask critical features while modeling long-range optical interactions via attention-based context aggregation. In contrast, the global

Table 4: Generalization ability comparison of baseline models.

Train	Test	Method	MSE($\times 10^{-3}$)	PA(%)	IOU(%)	EPE_{max}(nm)	EPE_{avg}(nm)
OPC-Metal	Metal	ED-CNN	33.99($\uparrow$ 25, 93)	96.60($\downarrow$ 2.49)	74.58($\downarrow$ 17.74)	3.52($\uparrow$ 1.88)	2.79($\uparrow$ 1.49)
		ED-Trans	40.89($\uparrow$ 31.00)	95.91($\downarrow$ 3.1)	69.89($\downarrow$ 21.80)	3.76($\uparrow$ 2.05)	2.95($\uparrow$ 1.58)
		RFNO	30.45($\uparrow$ 16.86)	96.95($\downarrow$ 1.69)	76.98($\downarrow$ 11.60)	3.20($\uparrow$ 1.36)	2.52($\uparrow$ 0.99)
		CFNO	39.30($\uparrow$ 26.22)	96.07($\downarrow$ 2.57)	70.63($\downarrow$ 18.48)	3.72($\uparrow$ 2.01)	2.94($\uparrow$ 1.44)
		MFNO	39.28($\uparrow$ 30.89)	96.07($\downarrow$ 2.96)	70.64($\downarrow$ 21.54)	3.66($\uparrow$ 32.01)	2.89($\uparrow$ 1.56)
		SOCS	**20.31**($\uparrow$ 10.95)	**97.03**($\downarrow$ 1.98)	**85.57**($\downarrow$ 6.38)	**2.82**($\uparrow$ 1.00)	**2.22**($\uparrow$ 0.87)
OPC-Via	Via	ED-CNN	11.62($\uparrow$ 6.97)	98.84($\downarrow$ 0.70)	62.58($\downarrow$ 18.81)	1.46($\uparrow$ 0.39)	1.05($\uparrow$ 0.12)
		ED-Trans	12.04($\uparrow$ 6.35)	98.80($\downarrow$ 0.63)	62.18($\downarrow$ 15.75)	1.47($\uparrow$ 0.11)	1.10($\uparrow$ 0.13)
		RFNO	11.15($\uparrow$ 6.38)	98.88($\downarrow$ 0.66)	63.39($\downarrow$ 19.71)	1.44($\uparrow$ 0.41)	1.07($\uparrow$ 0.18)
		CFNO	14.23($\uparrow$ 0.11)	98.58($\downarrow$ 0.11)	59.45($\downarrow$ 0.11)	1.52($\uparrow$ 0.15)	1.21($\uparrow$ 0.20)
		MFNO	12.75($\uparrow$ 6.36)	98.73($\downarrow$ 0.63)	60.75($\downarrow$ 12.77)	1.50($\uparrow$ 0.03)	1.16($\uparrow$ 0.14)
		SOCS	**7.71**($\downarrow$ 2.62)	**99.43**($\downarrow$ 0.15)	**66.39**($\downarrow$ 14.08)	**1.29**($\uparrow$ 0.05)	**1.02**($\uparrow$ 0.03)
All	OOD	ED-CNN	5.97	99.40	74.13	1.39	0.90
		ED-Trans	6.81	99.32	73.04	1.51	1.04
		RFNO	5.34	99.47	74.71	1.35	0.87
		CFNO	13.89	98.61	63.12	2.03	1.55
		MFNO	6.27	99.37	73.43	1.43	0.94
		SOCS	**4.13**	**99.79**	**80.24**	**0.91**	**0.60**

self-attention of ED-Trans prioritizes mask-wide pattern correlations, achieving suboptimal edge placement error (EPE) compared with ED-CNN in dense layout regions.

Physics-informed baseline comparison. MFNO dominates OPC-Metal simulations (95.29% IOU, 0.69nm EPE) by synergistically capturing global low-frequency optical kernels and local mask topology modulations, a critical requirement for modeling OPC-induced mask feature. The performance of MFNO degrades on OPC-Via and Via layers with 90.99% and 73.52% IOU respectively, where localized low-frequency scattering dominates, favoring RFNO's reduced Fourier domain focus with 92.68% and 83.10% IOU on OPC-Via and Via. The exclusive global spectral processing of CFNO proves least effective for lithography, particularly on OPC-Metal with 20.15×10^{-3} MSE and $> 2\text{nm}$ EPE, as well as only 97.98% PA and 84.84% IOU, as it disregards detailed mask-level edge variations. SOCS delivers stable performance across all mask categories by rigorously encoding Hopkins' partial coherent imaging principles in Appendix A.1, matching top baselines performances.

The baseline results of LithoSim highlight dataset-specific architectural preferences: Metal/OPC-Metal simulations demand concurrent global-local frequency feature learning, while Via layers benefit from localized frequency-space constraints. Also, the parity between data-driven and physics-informed models on LithoSim underscores the dataset's role as an equalizer, providing sufficient physical constraints through data diversity to compensate for missing litho-aware information.

4.4 Baseline Model Generalization Capabilities

The out-of-distribution (OOD) evaluation in Table 4 rigorously assesses baseline models' ability to generalize across various mask distribution. Models trained exclusively on OPC'ed datasets (OPC-Metal and OPC-Via) are tested on uncorrected counterparts (Metal and Via), simulating real-world scenarios where optimized masks must predict uncorrected lithographic outcomes firstly. Additionally, models trained on the full LithoSim dataset are evaluated on the OOD benchmark, which introduces different illumination (dipole/quasar), mask (M3 metal and regular via layer), and process variation distributions.

Data-driven model generalization capabilities. They exhibit significant sensitivity to OPC-induced topology changes on Metal. ED-Trans suffers a 31.00% MSE increase and 17.74% IOU decrease. In contrast, ED-CNN's hybrid architecture, combining convolutional locality with channel-spatial attention, achieves marginally better robustness (MSE $+25.93\%$, IOU -17.74%), outperforming all physics-informed models except RFNO and SOCS. This suggests chunked mask feature extraction, when augmented with attention-based optical context modeling, can partially compensate for missing physics constraints in data-driven approaches.

physics-informed model generalization capabilities. SOCS achieves superior stability with minimal performance degradation: MSE increases by only 10.95×10^{-3} (vs. ED-CNN's 25.93×10^{-3}) and IOU drops by 6.38% when trained on OPC-Metal and tested on Metal. Its physics-grounded Hopkins formulation inherently compensates for mask distribution shifts, maintaining $< 3\text{nm}$ maximum edge placement error (EPE) even under OOD conditions. RFNO achieves the second best performance on both Metal and Via datasets with more low frequency aware in a local range of masks. The full-dataset training paradigm further highlights the OOD supremacy of SOCS, achieving 4.13×10^{-3} MSE, 80.24% IOU, and 0.6nm average EPE on novel M3/via layouts with a 0.44nm margin under ED-Trans.

These results collectively affirm that while data-driven models benefit from LithoSim's diversity, physics-constrained architectures such as RFNO and SOCS remain indispensable for reliable OOD generalization, which is a critical requirement for production-grade RET integration.

5 Limitations

Idealizations. LithoSim leverages rigorous lithography simulator to achieve comprehensive lithographic variation coverage, with two approximations: fixed chemical kinetics assuming ideal resist chemistry modeling during PEB/development illustrated as fixed resist model in Figure 1 (a), and homogeneous resist-substrate optical constants (neglecting wavelength-dependent refractive indices n and interfacial reflectivity k). Despite these simplifications, LithoSim preserves dominant physics governing optical imaging—notably the coupled impacts of source polarization, defocus-dependent aberration, and OPC-induced mask modifications on resist exposure. Future extensions could integrate resist chemistry models while maintaining compatibility with foundational optical-mask-process variability of LithoSim. Currently, we are collaborating with fab partners to conduct LithoSim verification utilizing actual production line data.

Downstream testing. A critical challenge is to integrate learned lithography simulators as modular components into downstream RET flows, such as optical proximity correction (OPC) [11, 27], source mask optimization (SMO) [12, 13], and sub-resolution assist feature (SRAF) insertion [22]. While LithoSim provides foundational losses (*e.g.* L_2, process variation bands in Appendix A.2) and co-optimizable source-mask pairs essential for differentiable optimization, its current formulation remains a standalone tool, lacking the systems-level engineering required for seamless integration into tool chains. Future work could merge its variation-resilient predictions with topography-aware models [26, 23] to co-optimize manufacturability across the lithography stack. Our ultimate goal is to embed LithoSim within these RET flows, wherein it bridges the first critical gap: enabling ML models to supply all differentiable losses outlined in Figure 2 through CUDA-accelerated computations.

6 Conclusions

LithoSim establishes a comprehensive and physically-grounded benchmark for advancing AI-driven lithography simulation in semiconductor manufacturing. By integrating diverse optical sources, mask rules, and realistic process variations, it enables robust training and evaluation of both data-driven and physics-informed models. The benchmark not only bridges critical gaps in existing datasets but also provides a unified framework for assessing model accuracy, generalization, and readiness for downstream resolution enhancement techniques. Through open access to data and code, LithoSim lays a foundational step toward scalable, high-fidelity, and differentiable computational lithography, essential for next-generation design for manufacturing flows.

Acknowledgments and Disclosure of Funding

This work is sponsored by Natural Science Foundation of Shanghai (Project No.25JD1403000) and by MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence.

References

[1] K. Yi, B. Zhou, Y. Shen, P. Liò, and Y. G. Wang, "Graph denoising diffusion for inverse protein folding," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[2] S. Yu, W. M. Hannah, L. Peng, J. Lin, M. A. Bhouri, and et. al., "ClimSim: a large multi-scale dataset for hybrid physics-ML climate emulation," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[3] J. Nathaniel, Y. Qu, T. Nguyen, S. Yu, J. Busecke, A. Grover, and P. Gentine, "ChaosBench: A Multi-Channel, Physics-Based Benchmark for Subseasonal-to-Seasonal Climate Prediction," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[4] G. Morio and C. D. Manning, "An NLP benchmark dataset for assessing corporate climate policy engagement," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[5] H. He, G. Kuang, Q. Sun, and H. Geng, "PaLM: Point Cloud and Large Pre-trained Model Catch Mixed-type Wafer Defect Pattern Recognition," in *IEEE/ACM Proceedings Design, Automation and Test in Eurpoe (DATE)*, 2024.

[6] K. Ma, Z. Wang, H. He, Q. Xu, T. Chen, and H. Geng, "LMM-IR: Large-Scale Netlist-Aware Multimodal Framework for Static IR-Drop Prediction," in *ACM/IEEE Design Automation Conference (DAC)*, 2025.

[7] Y.-L. Qiao, J. Liang, V. Koltun, and M. C. Lin, "Differentiable simulation of soft multi-body systems," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[8] S. M. S. Hassan, A. Feeney, A. Dhruv, J. Kim, Y. Suh, J. Ryu, Y. Won, and A. Chandramowlishwaran, "BubbleML: a multiphase multiphysics dataset and benchmarks for machine learning," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[9] P. Evanschitzky, A. Erdmann, and T. Fuehner, "Extended Abbe approach for fast and accurate lithography imaging simulations," in *European Mask and Lithography Conference (EMLC)*, 2009.

[10] S. Yin, W. Zhao, L. Xie, H. Chen, Y. Ma, T.-Y. Ho, and B. Yu, "FuILT: Full Chip ILT System With Boundary Healing," in *ACM International Symposium on Physical Design (ISPD)*, 2024.

[11] H. Yang, S. Li, Z. Deng, Y. Ma, B. Yu, and E. F. Y. Young, "GAN-OPC: Mask Optimization With Lithography-Guided Generative Adversarial Nets," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2020.

[12] H. Yu, Y. Zhang, B. Jiang, S. Yu, and Z. Mao, "Research of SMO process to improve the imaging capability of lithography system for 28nm node and beyond," in *China Semiconductor Technology International Conference (CSTIC)*, 2017.

[13] G. Chen, H. He, P. Xu, H. Geng, and B. Yu, "Efficient Bilevel Source Mask Optimization," in *ACM/IEEE Design Automation Conference (DAC)*, 2024.

[14] S. Zheng, H. Yang, B. Zhu, B. Yu, and M. D. Wong, "LithoBench: Benchmarking AI Computational Lithography for Semiconductor Manufacturing," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[15] S. Banerjee, Z. Li, and S. R. Nassif, "ICCAD-2013 CAD Contest in Mask Optimization and Benchmark Suite," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2013.

[16] W.-H. Liu, S. Mantik, W.-K. Chow, Y. Ding, A. Farshidi, and G. Posser, "ISPD 2019 Initial Detailed Routing Contest and Benchmark with Advanced Routing Rules," in *ACM International Symposium on Physical Design (ISPD)*, 2019.

[17] H. Yang, Z.-Y. Li, K. Sastry, S. Mukhopadhyay, M. J. Kilgard, A. Anandkumar, B. Khailany, V. Singh, and H. Ren, "Generic lithography Modeling with Dual-band Optics-Inspired Neural Networks," in *ACM/IEEE Design Automation Conference (DAC)*, 2022.

[18] R. Sharma and V. Shankar, "Accelerated training of physics-informed neural networks (PINNs) using meshless discretizations," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[19] L. Pang, "Inverse lithography technology: 30 years from concept to practical, full-chip reality," *Journal of Micro/Nanopatterning, Materials, and Metrology*, 2021.

[20] X. Liang, H. Yang, K. Liu, B. Yu, and Y. Ma, "CAMO: Correlation-Aware Mask Optimization with Modulated Reinforcement Learning," in *ACM/IEEE Design Automation Conference (DAC)*, 2024.

[21] H. Yang and H. Ren, "Enabling Scalable AI Computational Lithography with Physics-Inspired Models," in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)*, 2023.

[22] Z. Yu, P. Liao, Y. Ma, B. Yu, and M. D. F. Wong, "CTM-SRAF: Continuous Transmission Mask-Based Constraint-Aware Subresolution Assist Feature Generation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2023.

[23] S. Zheng, Y. Ma, B. Yu, and M. D. F. Wong, "EMOGen: Enhancing Mask Optimization via Pattern Generation," in *ACM/IEEE Design Automation Conference (DAC)*, 2024.

[24] K. Wei, H. A. Jimenez-Romero, H. Amata, J. Sun, Q. Fu, F. Heide, and W. Heidrich, "Large-Area Fabrication-aware Computational Diffractive Optics," 2025.

[25] S. Sun, F. Yang, B. Yu, L. Shang, D. Zhou, and X. Zeng, "Efficient ILT via Multigrid-Schwartz Method," in *ACM/IEEE Design Automation Conference (DAC)*, 2024.

[26] C. Zheng, G. Zhao, and P. So, "Close the design-to-manufacturing gap in computational optics with a 'real2sim' learned two-photon neural lithography simulator," in *SIGGRAPH Asia*, 2023.

[27] S. Zheng, B. Yu, and M. Wong, "OpenILT: An Open Source Inverse Lithography Technique Framework," in *IEEE International Conference on ASIC (ASICON)*, 2023.

[28] G. Chen, W. Chen, Q. Sun, Y. Ma, H. Yang, and B. Yu, "DAMO: Deep Agile Mask Optimization for Full Chip Scale," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2022.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[30] G. Chen, Z. Pei, H. Yang, Y. Ma, B. Yu, and M. Wong, "Physics-Informed Optical Kernel Regression Using Complex-valued Neural Fields," in *ACM/IEEE Design Automation Conference (DAC)*, 2023.

[31] W. Ye, M. B. Alawieh, Y. Lin, and D. Z. Pan, "LithoGAN: End-to-End Lithography Modeling with Generative Adversarial Networks," in *ACM/IEEE Design Automation Conference (DAC)*, 2019.

[32] M. Liu, H. Yang, B. Khailany, and H. Ren, "An Adversarial Active Sampling-based Data Augmentation Framework for AI-Assisted Lithography Modeling," in *ACM/IEEE Design Automation Conference (DAC)*, 2023.

[33] H. Tanabe, S. Sato, and A. Takahashi, "Fast 3D lithography simulation by convolutional neural network: POC study," in *Photomask Technology*. SPIE, 2020.

[34] J.-R. Gao, X. Xu, B. Yu, and D. Z. Pan, "MOSAIC: Mask optimizing solution with process window aware inverse correction," in *ACM/IEEE Design Automation Conference (DAC)*, 2014.

[35] B. Jiang, L. Liu, Y. Ma, B. Yu, and E. F. Y. Young, "Neural-ILT 2.0: Migrating ILT to Domain-Specific and Multitask-Enabled Neural Network," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2022.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[37] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[38] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier Neural Operator for Parametric Partial Differential Equations," in *International Conference on Learning Representations (ICLR)*, 2021.

[39] Z. Li, H. Zheng, N. Kovachki, D. Jin, H. Chen, B. Liu, K. Azizzadenesheli, and A. Anandkumar, "Physics-Informed Neural Operator for Learning Partial Differential Equations," *ACM / IMS J. Data Sci.*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: LithoSim introduces lithography simulation benchmark with more than 4 million rigorously curated input-output pairs, integrating optical variations, mask corrections, and process variations to establish a unified evaluation flow for ML-based simulation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: LithoSim discusses limitations in Section 5, covering idealized assumptions and downstream integration challenges. For idealizations, LithoSim employs two approximations: fixed chemical kinetics and homogeneous resist-substrate optical constants, though it retains core physics governing optical imaging. Future work could incorporate dynamic resist chemistry models. Regarding downstream testing, while LithoSim provides multi-scale representations and source-mask pairs critical for RET workflows (*e.g.*, OPC/SMO/SRAF), it currently lacks systems-level engineering for seamless integration into EDA toolchains. The authors emphasize the need of LithoSim to bridge this gap through CUDA-accelerated RET operationalization.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: LithoSim describes theoretical formulations for optical lithography in Section A.1 and FNO-based architectures Section A.4. For the SOCS approximation in lithography modeling, assumptions include decomposing source/projector/mask interactions via TCC with SVD truncation (retaining dominant eigenvalues) and approximating the imaging integral via Eq. 5. The FNO framework assumes learnable spectral weights ($\mathcal{W}_\theta$) can approximate lithography kernels by truncating high-frequency modes (e.g., $|k| \leq m$ in RFNO). The validity of these approximations is implicitly supported by alignment with lithography physics (e.g., frequency-domain interactions) and benchmarking results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: LithoSim introduce experiment settings briefly in Appendix A.6. The Hugging Face dataset has been divided into `opc_mtal`, `opc_via`, `metal`, and `via` with a `train_val_test` split. In Github repo, LithoSim also gives a detailed guideline.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: LithoSim provides open access to the dataset, code, and pre-trained models via Hugging Face (https://huggingface.co/datasets/grandiflorum/LithoSim; https://huggingface.co/grandiflorum/LithoSim) and a project website (https://dw-hongquan.github.io/LithoSim), including pre-trained models for reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The data splits are list in Table 2. The hyperparameters and type of optimizer are announced in A.6.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: LithoSim report error bars in Table 3, which is the standard deviation of each metric in the certain dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In A.6, LithoSim introduces computer resources on 4 H100 Graphics cards with Intel Core Xeon Platinum $8462\text{Y}+$ processors.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

Justification: LithoSim follows the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: LithoSim dose not have societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: LithoSim dose not provide any data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The code of LithoSim uses `lightning` framework, which is based on Apache 2.0 license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The data, code, and pre-trained models are released (see in Abstract).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: LithoSim does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: LithoSim does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LithoSim does not involve LLMs as research methods.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Optical Lithography Approach

Typical optical lithography model comprises 3 essential components: source, mask, and projector, as illustrated in Figure 1 (b). The light rays propagate through the projector and the mask and produce diffracted light with layout pattern information. The intensity of optical imaging can be formulated as,

$$\begin{aligned}\mathbf{I}(x,y) = \int\int\int\int\int\int_{-\infty}^{\infty} & \mathbf{J}(f,g)\mathcal{F}(\mathbf{M})\left(f',g'\right)\mathcal{F}(\mathbf{M})^{*}\left(f'',g''\right) \\ & \mathbf{H}\left(f+f',g+g'\right)\mathbf{H}^{*}\left(f+f'',g+g''\right) \\ & \exp\left(-j2\pi((f'-f'')x+(g'-g'')y)\right) \\ & \mathrm{d}f\,\mathrm{d}g\,\mathrm{d}f'\,\mathrm{d}g'\,\mathrm{d}f''\,\mathrm{d}g''\end{aligned}, \tag{1}$$

where $\mathbf{I}$ is the imaging intensity, $\mathbf{J}$ is the source, $\mathbf{H}$ is the optical transfer function (OTF) of projector, and $\mathcal{F}(\mathbf{M})$ is the frequency of the mask $\mathbf{M}$; (f,g), (f',g'), and (f'',g'') represent the normalized frequency-domain coordinates of $\mathbf{H}$, $\mathcal{F}(\mathbf{M})$, and $\mathcal{F}(\mathbf{M})^*$. This formulation does not have an analytical solution but only an approximate solution.

A fast approach method, SOCS approach separating source $\mathbf{J}$ and projector $\mathbf{H}$ from mask $\mathbf{M}$ as,

$$\begin{aligned}\mathbf{I}(x,y) = \int\int\int\int_{-\infty}^{\infty} & \mathcal{T}\left(f',g';f'',g''\right)\mathcal{F}(\mathbf{M})\left(f',g'\right)\mathcal{F}(\mathbf{M})^{*}\left(f'',g''\right) \\ & \exp\left(-j2\pi((f'-f'')x+(g'-g'')y)\right)\mathrm{d}f'\mathrm{d}g'\mathrm{d}f''\mathrm{d}g''\end{aligned}, \tag{2}$$

where $\mathcal{T}$ is the transmission cross-coefficients (TCC) given by,

$$\mathrm{TCC}\left(f',g';f'',g''\right) = \iint_{-\infty}^{\infty} J(f,g)H\left(f+f',g+g'\right)H^{*}\left(f+f'',g+g''\right)\mathrm{d}f\mathrm{d}g. \tag{3}$$

Applying SVD decomposition, Eq. 3 can be approximated by Sum of coherent source (SOCS),

$$\mathrm{TCC}\left(f',g';f'',g''\right) \approx \sum_{q=1}^{\infty}\kappa_q\Phi_q\left(f',g'\right)\Phi_q^{*}\left(f'',g''\right), \tag{4}$$

where, κ_q and Φ_q are q-th eigenvalue and eigenvector of TCC. For fast calculation, we can keep the Q largest eigenvalues and obtain final SOCS approach as,

$$\mathbf{I}(x,y) = \sum_{q=1}^{Q}\kappa_q||\Phi_q(x,y)\otimes M(x,y)||^2, \tag{5}$$

where $\phi_q(x,y)$ and $M(x,y)$ are the spatial distribution of Φ_q and $\mathcal{F}(\mathbf{M})$ respectively.

A.2 Relationship between lithography simulation and RET

As illustrated in Figure 2, lithography simulation forms the computational backbone of modern resolution enhancement techniques (RET) [19], enabling the optimization of sources $\mathbf{J}$ and masks $\mathbf{M}$ through iterative physics-aware feedback.

The simulator maps $(\mathbf{J},\mathbf{M})$ to resists $\mathbf{R}$, which are evaluated via two critical metrics: L2 contour fidelity (geometric deviation from target layout $\mathbf{T}$ under normalized condition) and process variation band (PVB) robustness across dose (α) and focus β conditions as,

$$\begin{aligned}\mathcal{L}_2 &= ||\mathbf{R}_{\text{norm}} - \mathbf{T}||_2^2 \\ \mathcal{L}_{PVB} &= ||\mathbf{R}_{\text{max}} - \mathbf{R}_{\text{min}}||_2^2\end{aligned}, \tag{6}$$

where $\mathbf{R}_{\text{norm}}$ is the resist under $(\alpha,\beta) = (0,0)$, $\mathbf{R}_{\text{max}}$ and $\mathbf{R}_{\text{min}}$ are resists under $(\alpha,\beta) = (-0.1,-40)$ and $(\alpha,\beta) = (0.1,40)$ respectively.

Consequently, the comprehensive RET loss is formulated as,

$$\mathcal{L}_{RET} \equiv \mathcal{L}_{OPC} \equiv \mathcal{L}_{SMO} = \gamma\mathcal{L}_2 + \eta\mathcal{L}_{PVB}, \tag{7}$$

where γ and η are weighting factors for the respective loss components.

In optical proximity correction (OPC) mode, the illumination source $\mathbf{J}$ is fixed, and the simulator guides mask optimization through gradient-based updates:

$$\mathbf{M}^* = \underset{\mathbf{M}}{\text{argmin}} \mathcal{L}_{OPC}(\mathbf{J}, \mathbf{M}). \tag{8}$$

Source mask optimization (SMO) extends this framework by co-optimizing $\mathbf{J}$ and $\mathbf{M}$ in a coupled parameter space as,

$$(\mathbf{J}^*, \mathbf{M}^*) = \underset{(\mathbf{J},\mathbf{M})}{\text{argmin}} \mathcal{L}_{SMO}(\mathbf{J}, \mathbf{M}). \tag{9}$$

LithoSim contains all the input parameters required by Eq. 7, not only the source and mask as the optimization subjects, but also dose and defocus involved in the loss calculation. This makes it possible to train lithography simulation using LithoSim and thereby achieve CUDA-accelerated RET.

A.3 Litho-condition Embedding

Process Variations Embedding:. LithoSim incorporates critical process variations (PV) through a physics-informed positional encoding scheme. For dose and defocus inputs (normalized to [-1, 1]), we employ a continuous positional encoding that transforms scalar PV into a spectral representation through logarithmic frequency bands. For a given PV $v \in [-1, 1]$, the encoding generates $d_{pv}/2$ frequency components with wavelengths logarithmically spaced following,

$$\begin{aligned} PE(v)_{2k} &= sin(v \cdot e^{-k \cdot ln(10^4/d_{pv})}) \\ PE(v)_{2k+1} &= cos(v \cdot e^{-k \cdot ln(10^4/d_{pv})}) \end{aligned}. \tag{10}$$

Source Positional Embedding:. The source spatial characteristics are encoded through a multi-frequency 2D positional encoding that preserves optical reciprocity and illumination coherence properties. For source coordinates $(x, y) \in [-1, 1]^2$ and backbone model dimension d_s,

$$\begin{aligned} PE(x,y)_{4k} &= sin(x \cdot e^{-k \cdot ln(10^4/d_s)}) \\ PE(x,y)_{4k+1} &= cos(x \cdot e^{-k \cdot ln(10^4/d_s)}) \\ PE(x,y)_{4k+2} &= cos(y \cdot e^{-k \cdot ln(10^4/d_s)}) \\ PE(x,y)_{4k+3} &= cos(y \cdot e^{-k \cdot ln(10^4/d_s)}) \end{aligned}. \tag{11}$$

Source Compression:. LithoSim implements a multi-scale attention mechanism that preserves critical optical characteristics while enabling efficient processing of high-dimensional source patterns. The compression occurs through 3 physics-aware stages.

(1). *Coherent Chunk Processing* splits source $\mathbf{J}$ into $C = 64$ chunks matching optical cross-effect size,

$$\mathcal{B}_k = \{\mathbf{J}_i\}_{i=d_sC}^{(d_s+1)C} \in \mathbb{C}^{C \times D}, \tag{12}$$

where positional encoding in Eq. 11 maintains inter-pixel phase relationships critical for diffraction modeling of every chunks.

(2). *Intra-chunk Self-attention* models local interference within coherence area as,

$$\mathcal{C}_{\mathcal{B}_i} = softmax\left(\frac{Q_{local}K_{\mathcal{B}_i}^T}{\sqrt{d}} \otimes M_{valid}\right) \cdot V_{\mathcal{B}_i} \tag{13}$$

where M_{valid} represents the valid position of source (*e.g.* radium $\sigma \in [0.68, 0.83$ for annular source), $K_{\mathcal{B}_i}$ and $V_{\mathcal{B}_i}$ is the key and value of i-th chunked source, Q_{local} is local learnable query.

(3). *Inter-chunk Self-attention* captures global source contribution blending as,

$$\mathcal{C} = \sum_{i=1}^{K} w_i \mathcal{C}_i, \quad w_i \propto e^{\langle Q_{global}, \mathcal{C}_i \rangle}, \tag{14}$$

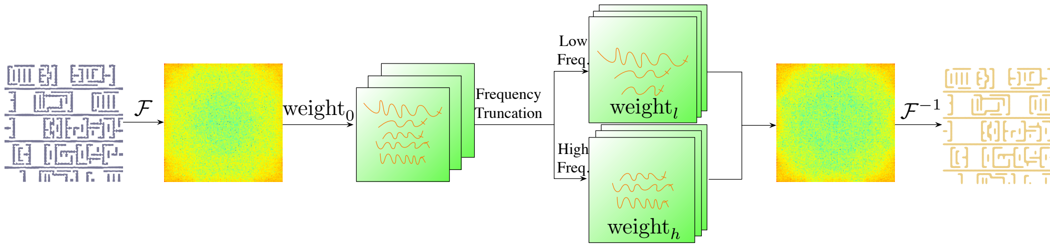

Figure 6: Reduced Fourier Neural Operator (RFNO).

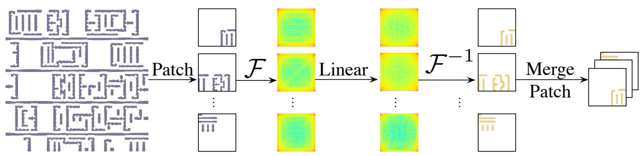

Figure 7: Convolutional Fourier Neural Operator (CFNO).

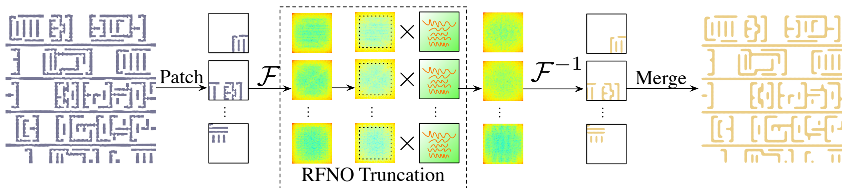

Figure 8: Mixed Fourier Neural Operator (MFNO).

where $\mathcal{C}$ is final compressed source feature, Q_{global} is global query of total chunked source, w_i is the weight of i-th chunked source.

Chunked Litho-aware Cross-attention:. LithoSim implements condition embedding into mask feature $\mathbf{M} \in \{0,1\}^{H\times W}$ for source and process variations (PV) respectively based on chunked cross-attention. $\mathbf{M}$ is partitioned into $N = \frac{H\times W}{d_m^2}$ local chunks of size $d_m \times d_m$.

(1). *PV cross-attention:* For i-th chunked mask region $\mathcal{M}_i \in \{0,1\}^{d_m\times d_m}$ and embedded process variations $\mathcal{V} = PE(v)$ after positional embedding, the PV cross-attention is given by,

$$\mathcal{M}' = \sum_{i=1}^{N} softmax\left(\frac{Q_i K^T}{\sqrt{d}}\right) V, \tag{15}$$

where $Q_i = W_Q \mathcal{M}_i$, $K = W_K \mathcal{V}$, and $V = W_V \mathcal{V}$. W_Q, W_K, and W_V is learnable projection parameters for each chunked mask and process variation. By decomposing the mask into $d_m \times d_m$ optical proximity correction (OPC) regions and computing multi-head attention between chunked mask features (queries) and process-encoded variations (keys/values), it models dose-dependent resist thresholding and defocus-induced blur as spatially varying modulation operators.

(2). *Source cross-attention* For j-th chunked mask region $\mathcal{M}_j \in \{0,1\}^{d_m\times d_m}$ and compressed source $\mathcal{C}$ in Eq. 14, the source cross-attention is given by,

$$\mathcal{M}'' = \sum_{j=1}^{N} softmax\left(\frac{Q_j K^T}{\sqrt{d}} \otimes M_{valid}\right) V, \tag{16}$$

where $Q_i = W_Q \mathcal{M}_j$, $K = W_K \mathcal{C}$, and $V = W_V \mathcal{C}$. W_Q, W_K, and W_V is learnable projection parameters for each chunked mask and compressed source, M_{valid} is the valid region of source.

A.4 Baseline Architecture

The fusion between FNOs, including RFNO, CFNO, and MFNO, and lithography simulation from their shared reliance on spectral representations for efficient physical process approximation. In optical lithography, Eq. 5 can be simplified to formulate as,

$$\mathbf{I} = |\mathcal{F}^{-1}|\mathcal{F}(\mathbf{M}) \otimes \mathcal{F}(\mathbf{K})||^2 \tag{17}$$

where $\mathbf{K}$ is the lithography kernel which is dependent on source and defocus (Figure 1 (a)). FNOs natively operate in the spectral domain through learnable truncated mode interactions $\mathcal{W}_\theta \in \mathbb{C}^{m \times m}$ approximating optical kernel by,

$$\text{FNO}(k) = \mathcal{F}^{-1}\left[\mathcal{W}_\theta(\mathbf{k}) \cdot \mathcal{F}(M)(\mathbf{k})\right], \tag{18}$$

where $\mathcal{W}_\theta$ is local learnable complex-valued spectral weights truncated at modes $|k| \le m$ in **RFNO** (Figure 6), is a patched global complex-valued linear layer in **CFNO** (Figure 7), is patched learnable complex-valued spectral weights in **MFNO** (Figure 8), which aligns with lithography's inherent frequency-space physics in Eq. 17.

FNO-based models typically requires the combination of FNO with an CNN encoder-decoder structure [38, 21, 17] to achieve the purpose of extracting low-frequency mask features at different scales. Unlike FNO, **SOCS** rigorously adheres to the methodology of Eq. 5. The mask is first transformed into the frequency domain via FFT, with all encoding, decoding, and condition interactions executed exclusively in the spectral domain, before directly outputting the resist profile through IFFT.

A.5 Evaluation Metrics Details

AI performance metrics: Given the predicted resist $\hat{\mathbf{R}}$ and the ground truth resist $\mathbf{R}$, the pixel number is N, MSE, PA, IOU is defined respectively as,

$$\text{MSE} = \frac{1}{N}||\mathbf{R} - \hat{\mathbf{R}}||^2 \tag{19}$$

$$\text{PA} = \frac{\mathbf{R} \cap \hat{\mathbf{R}}}{\mathbf{R}} \tag{20}$$

$$\text{IOU} = \frac{\mathbf{R} \cap \hat{\mathbf{R}}}{\mathbf{R} \cup \hat{\mathbf{R}}} \tag{21}$$

Lithographic fidelity metrics: As illustrated in Figure 5, given the predicted resist contour $\mathbf{C}_{\hat{\mathbf{R}}}$, the ground truth resist $\mathbf{C}_{\mathbf{R}}$, and original layout countour $\mathbf{C}_{\mathbf{T}}$. First, sample evaluation points $\mathbf{P}$ at regular intervals (typically 20nm) along $\mathbf{C}_{\mathbf{T}}$. For each point $P_i \in \mathbf{P}$, construct a perpendicular line that intersects both $\mathbf{C}_{\mathbf{R}}$ and $\mathbf{C}_{\hat{\mathbf{R}}}$ at points $P_{i,\mathbf{R}}$ and $P_{i,\hat{\mathbf{R}}}$ respectively. The edge placement error (EPE) at P_i is then defined as the length of the vertical segment $\overline{P_{i,\mathbf{R}}P_{i,\hat{\mathbf{R}}}}$. The maximum EPE across all evaluation points $\mathbf{P}$ is denoted as EPE_{max}, while the average EPE is calculated as EPE_{avg}.

A.6 Experiment Settings

LithoSim is trained and tested with 4 H100 Graphics cards with Intel Core Xeon Platinum 8462Y+ processors. All the baselines is trained with `Adam` optimizer and a 10^{-4} learning rate of 10^{-5} weight decay.

LithoSim uses a linear combination of BCE and Dice loss in Eq. 22 for ED-CNN, ED-Trans, RFNO, CFNO, and MFNO, as well as MSE loss for SOCS.

$$\mathcal{L} = \alpha\mathcal{L}_{dice} + \beta\mathcal{L}_{BCE}, \tag{22}$$

where LithoSim sets $\alpha = \beta = 1$.

In condition embeddings (A.3), LithoSim uniformly sets output dimension of source positional embedding $d_s = 8$, compressed factor $K = 16$, and chunk size 256 for each source. In process variation embedding, the output dimension of value positional embedding is also set as $d_v = 8$. Mask chunk size is set as 64 to capture proximity optical effects.