Moore and More (2025) 2:15
https://doi.org/10.1007/s44275-025-00032-5

ORIGINAL ARTICLE

Recent advances in computational lithography technology

Qian Jin[1] · Qingsong Peng[1] · Yumeng Liu[1] · Xiaotian Qiu[1] · Qi Sun[1]

Received: 20 December 2024 / Revised: 18 February 2025 / Accepted: 3 March 2025
© The Author(s) 2025

Abstract
As semiconductor manufacturing technology advances, lithography faces increasingly severe challenges from resolution limitations and process variability. Computational lithography has become a critical technology that ensures manufacturability and yield in advanced semiconductor nodes. This paper comprehensively reviews recent advancements in computational lithography, focusing on three core areas: lithography modeling, hotspot detection, and mask optimization. Particular attention is given to innovative applications of machine learning in these domains, along with an analysis of the technical challenges in full-chip optimization. By summarizing the latest technologies and applications, this paper aims to offer valuable insights into the trends and future directions of computational lithography in advanced semiconductor manufacturing.

Keywords Optical lithography · Computational lithography · Machine learning · Artificial intelligence · Semiconductor manufacturing · Hotspot detection · Mask optimization

1 Introduction

Lithography is a core process in integrated circuit (IC) manufacturing that transfers complex circuit patterns from masks onto silicon wafers. It is central to defining the resolution, accuracy, and functionality of microstructures in chips. By leveraging focused beams of light, lithography enables the production of transistors, interconnects, and other fundamental components at nanometer scales, shaping the performance and manufacturability of modern semiconductors [1]. However, lithography faces significant challenges as the IC technology node shrinks to meet technological demands. Diffraction, optical aberrations, and process variability have increasingly hindered its ability to deliver precise and reliable patterning.

To address these limitations, computational lithography has become a pivotal enabler of advanced semiconductor manufacturing. It integrates advanced computational techniques with photolithographic processes to model, optimize, and correct imaging inaccuracies, thereby overcoming the physical barriers of traditional lithography [2]. Computational lithography ensures manufacturability and high yield at advanced nodes by improving resolution, compensating for process-induced distortions, and expanding the process window. Furthermore, it reduces development time and cost, making it an essential tool for accelerating innovation in semiconductor manufacturing.

However, the increasing complexity of IC layouts, the demand for greater precision, and the need to manage massive computational workloads create bottlenecks that traditional methods struggle to overcome [3]. Three key topics—lithography modeling, hotspot detection, and mask optimization (MO)—are particularly critical to addressing these challenges. Lithography modeling provides accurate simulations of optical and resist interactions, enabling precise predictions of how design layouts will translate into wafer patterns. Hotspot detection identifies potential design regions prone to manufacturing defects, facilitating preemptive design adjustments. MO uses insights from modeling and hotspot detection to refine mask patterns, improving printability and robustness. These three highly related topics, collectively form the backbone of computational lithography by addressing the most pressing pain points in the lithographic process.

In recent years, machine learning (ML) has been widely used in the field of IC manufacturing [4–7], and its rapid development has further promoted the advancement of computational lithography technology. These methods excel in handling complex datasets, uncovering hidden patterns, and achieving high computational efficiency compared

✉ Qi Sun
qisunchn@zju.edu.cn

[1] Zhejiang University, Hangzhou 310027, China

Published online: 13 November 2025

to traditional algorithms [8]. For instance, deep learning models can learn from vast datasets to identify error-prone regions with high accuracy in hotspot detection. Similarly, in MO, learning-based approaches can bypass iterative, time-consuming simulations, accelerating the design process while maintaining high fidelity. These innovations highlight the transformative potential of ML in the lithography workflow, which this survey explores in detail through specific applications and case studies.

While previous surveys on ML in electronic design automation (EDA) have focused on applications in the design phase, such as optimizing circuit performance, predicting routing congestion, and enhancing timing accuracy [9–11], the integration of ML in computational lithography—the critical bridge between design and manufacturing—has not been comprehensively reviewed. These previous works primarily address ML's role in the initial design stages, leaving a gap in understanding its application in the final manufacturing phase, particularly in lithography modeling, hotspot detection, and MO. This paper fills that gap by providing a dedicated exploration of ML applications within computational lithography, with a particular focus on overcoming the unique challenges of advanced semiconductor nodes. Unlike the broader EDA surveys, we delve into how ML techniques are transforming the lithography process, specifically improving resolution, compensating for process-induced distortions, and refining mask patterns. Through detailed case studies and the latest advancements, we demonstrate how ML accelerates full-chip lithography optimization, offers innovative solutions to traditional bottlenecks, and enables more efficient manufacturing at increasingly smaller technology nodes.

The remainder of this paper is organized as follows: Section 2 reviews advancements in lithography modeling, Section 3 discusses hotspot detection, and Section 4 explores MO techniques. Finally, Section 5 concludes with a summary and discussion of future directions for research and innovation in this field.

2 Lithography modeling

Lithography comprises two fundamental steps: exposure and development. To accurately model these processes, lithography employs the optical model and the resist model, as depicted in Fig. 1. Optical modeling maps a mask image to light intensity (aerial image) that is projected on a silicon wafer. Resist modeling deals with the interaction between light intensity and resist materials and forms the final shape on the silicon wafer.

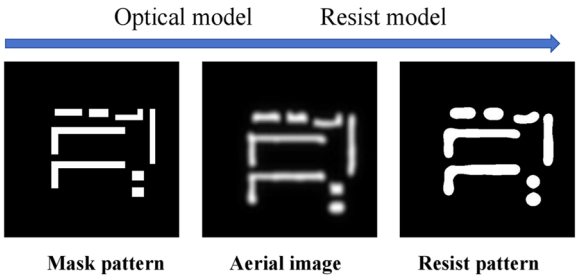


Fig. 1 Lithography modeling flowchart. Reproduced with permission [3] Copyright 2021, Association for Computing Machinery

2.1 Optical modeling

The foundation of the optical model is grounded in scalar diffraction theory within Fourier optics [12], which facilitates the calculation of the light intensity distribution $I(x, y)$ on the resist surface, commonly referred to as the aerial image:

$$I(x,y) = \iint\iint\iint J(\xi,\eta)\mathcal{F}(M)(\xi',\eta')\mathcal{F}(M)^*(\xi'',\eta'') \\ H(\xi+\xi',\eta+\eta')H^*(\xi+\xi'',\eta+\eta'') \\ e^{-i2\pi((\xi'-\xi'')x+(\eta'-\eta'')y)}\,d\xi\,d\eta\,d\xi'\,d\eta'\,d\xi''\,d\eta'', \tag{1}$$

where (x, y) denotes the spatial coordinates of the image plane, (ξ,η), (ξ',η'), and (ξ'',η'') refer to the frequency coordinates of the source, mask spectrum, and its conjugate, respectively. J is the lithographic illumination source, $\mathcal{F}$ denotes the Fourier inverse transform, H is the optical transfer function (OTF) of the system, and * is the Hermitian transpose operator.

This equation applies to the ideal case of a fully coherent light source and a simplified optical system model. However, in practical high-resolution lithography systems, complex effects such as diffraction, coherence, and aberrations should be considered, necessitating the introduction of more sophisticated models. There are two possible numerical methods for calculating light intensity distribution, namely the Abbe imaging formula and the Hopkins imaging formula.

The Abbe imaging formula considers spatial images as the incoherent sum of imaging results for each source point. The contribution from each source point toward the final aerial image is calculated independently and then summed. For convenience, the illumination cross-coefficients (ICC) is introduced as:

$$\mathrm{ICC}(x,y;\xi,\eta) = \left|\iint H(\xi+\xi',\eta+\eta')\mathcal{F}(M)(\xi',\eta')e^{-i2\pi(\xi'x+\eta'y)}\,d\xi'\,d\eta'\right|^2, \tag{2}$$

the Abbe imaging formula rewrites the Eq. 1 as:

$$I_{\text{Abbe}}(x,y) = \iint J(\xi,\eta)\text{ICC}(x,y;\xi,\eta)d\xi\, d\eta. \tag{3}$$

Since Abbe imaging theory applies to low-resolution, large-scale systems, its calculation speed is relatively slow, which limits its use in high-resolution applications. In contrast, the Hopkins imaging formula pre-calculates the contributions from the source and the mask, representing them as a transmission cross coefficient (TCC) matrix, thereby avoiding the need to calculate each diffraction order individually [13]. The light intensity distribution can be expressed as:

$$I_{\text{Hopkins}}(x,y) = \iint\iint \mathcal{F}(M)(\xi',\eta')\mathcal{F}(M)^*(\xi'',\eta'')\text{TCC} \\ e^{-i2\pi((\xi'-\xi'')x+(\eta'-\eta'')y)}\, d\xi'\, d\eta'\, d\xi''\, d\eta'', \tag{4}$$

where TCC is the cross transfer coefficient matrix, which is calculated as follows:

$$\text{TCC} = \iint J(\xi,\eta)H(\xi+\xi',\eta+\eta')H^*(\xi+\xi'',\eta+\eta'')\, d\xi\, d\eta. \tag{5}$$

By separating the mask from the optical system, the Hopkins imaging formula can effectively accelerate the computation process and is suitable for high-resolution lithography systems.

2.2 Resist modeling

Exposure is a critical step that converts the optical imaging result (aerial image) into the actual pattern on the wafer, and the resist model simulates this process. Based on the computational precision and complexity, resist models are typically classified into strict models and compact models [14]. The strict model is based on detailed physical and chemical mechanisms to simulate the exposure, development, and etching processes of the resist. It can accurately describe the dissolution process within the resist but demands a significant computational load and has a slower simulation speed. A common form of the strict model is:

$$D(x,y,z) = I(x,y) \otimes G, \tag{6}$$

where $D(x, y, z)$ is the dissolution rate of the resist, $I(x, y)$ is the intensity distribution of the exposure image, and G is a Gaussian kernel to simulate chemical reactions.

Compact models simplify the resist processes to improve computational efficiency and simplify the simulation process. Common compact models include the constant threshold resist (CTR) model and the Variable Threshold Resist (VTR) model, and their differences are illustrated in Fig. 2. The CTR model assumes that the exposure reaction of the resist is based on a fixed exposure threshold. The resist will be developed away when the exposure dose exceeds this threshold. The final resist shape R is defined as:

$$R(x,y) = \begin{cases} 1, & I(x,y) > I_{\text{threshold}} \\ 0, & I(x,y) \le I_{\text{threshold}} \end{cases}. \tag{7}$$

Compared to the CTR model, the VTR model accounts for variations in exposure threshold with light dosage and process parameters [15], providing a more accurate simulation of resist behavior:

$$R(x,y) = \frac{1}{1+e^{-\alpha(I(x,y)-I_{\text{threshold}}(x,y))}}, \tag{8}$$

where α controls the sensitivity of the threshold variation.

3 Lithography hotspot detection

As the process nodes continue to shrink, graphic fidelity and imaging accuracy become key challenges. During lithography, a phenomenon known as the Optical Proximity Effect (OPE) occurs, causing deviations in the mask pattern's imaging on the wafer. This discrepancy is especially

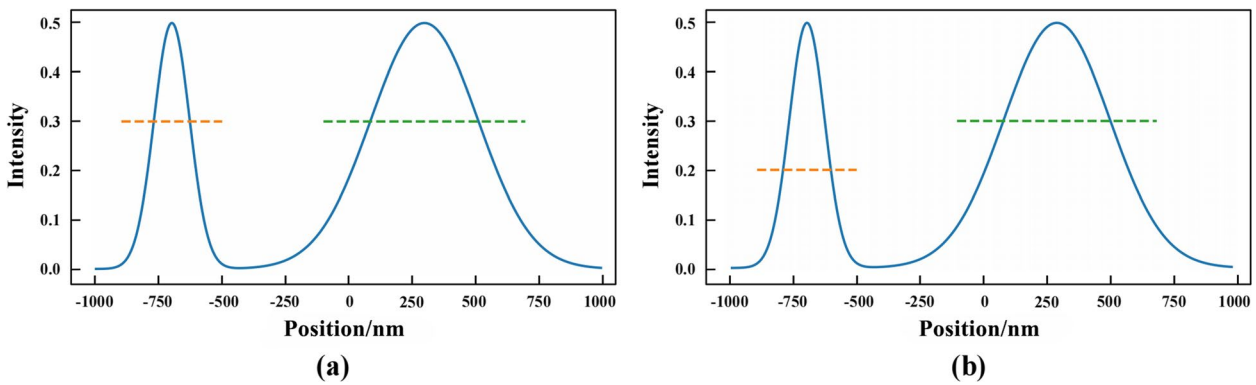


Fig. 2 Two types of compact resist model: **a** CTR model and **b** VTR model. Reproduced with permission [15]. Copyright 2024, IEEE

problematic for small-scale structures, as it can impact device performance and reliability. Certain layout patterns exhibit variations that may lead to critical defects, commonly known as photolithographic hotspots. Efficient and accurate detection of these hotspot patterns is essential to prevent defects from arising before printing. This is crucial for completing layouts, finalizing designs, and improving yield during the physical verification. The conventional methodology entails the execution of a comprehensive lithography simulation of the entire chip, utilizing the lithography modeling delineated in Section 2. This approach enables the precise identification of lithography hotspots. However, it is inherently time-consuming and resource-intensive. In contrast, methods based on pattern matching and ML have significant advantages in lithographic hotspot detection.

3.1 Pattern-matching-based hotspot detection

The pattern-matching-based lithography hotspot detection method employs known hotspot feature templates to identify potential hotspots in a design through rapid matching and plate comparison. The fundamental principle of this method is the construction of a comprehensive hotspot library, which enables the identification of patterns in the design that are susceptible to lithographic defects. Pattern-matching methods are typically classified into two principal categories: rule-based matching and template-based matching.

3.1.1 Rule matching

As hotspot patterns are frequently attributable to particular geometric characteristics, such as geometric widths and spacing that are insufficiently generous, a naive approach would be to rely on a set of predefined design rules to expeditiously identify patterns that deviate from the design rules. Some of these design rules are illustrated in the Fig. 3. However, the evolution of technology nodes has resulted in photolithography effects involving more neighboring feature relationships. Consequently, it is becoming increasingly challenging to represent the relationships between a large group of non-adjacent objects with a small set of rules. This presents a significant obstacle to customizing design rules.

Yao et al. [16] introduced a method for defining design rules based on range patterns of lithographic hotspots and developed an efficient algorithm to identify layout areas that match these patterns. Range patterns consist of two-dimensional geometric configurations with variable shapes, offering more flexibility than traditional design rule checking (DRC). Xu et al. [17] later introduced via range patterns and incompletely specified patterns to represent new classes of process hotspots. Yu et al. [18] proposed a DRC-based hotspot detection method by extracting only the key topological features of hotspot patterns and converting them into design rules. Park et al. [19] developed a simplified edge-driven DRC rule to create rectangles for hotspot graphical descriptions and utilized a mathematical vector concept to minimize the iterations needed for matching patterns in various orientations, resulting in significant speed improvements. Wang et al. [20] focused on the partial pattern-matching problem, extracting only three types of spacing and edge length rules to quickly identify patterns that satisfy the specified rules. Qiu et al. [21] introduced an innovative approach to the representation of hotspot pattern features, combining edge-pair representation with space-filling curves and hash technology. This approach enables the rapid identification and location of hotspots in a layout by matching the edge-pair geometric features.

3.1.2 Template matching

In contrast to rule-based matching, template matching employs existing hotspot patterns as templates for matching patterns within the layout. This is based on the assumption that a pattern that resembles a hotspot pattern is also likely to be a hotspot. The most critical element of template matching involves the extraction of relevant hotspots from existing data points, which may subsequently be utilized as templates. It is essential to assess the similarity between different patterns for effective matching. Ma et al. [22] posit that the DRC-based hotspot detection method is unwieldy to maintain as technology nodes advance, and put forth a novel hotspot detection system. The system first identifies the locations of hotspots through process simulation, then extracts small layout "clips" centered on the hotspots and calculates the similarity between these clips by calculating their overlapping regions. Subsequently, the hotspots are clustered using an incremental clustering algorithm, which

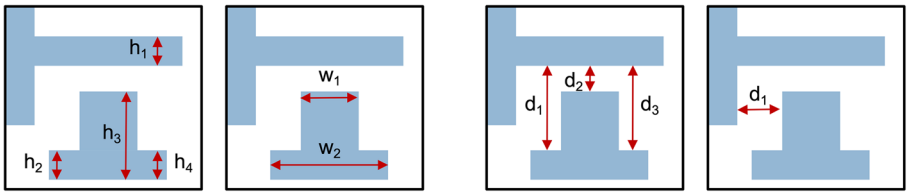


Fig. 3 Design rules based on geometric width and spacing

Rule 1: the width and height of a block tile

Rule 2: the spacing between two adjacent block tiles

identifies representative hotspots within each cluster and feeds them into a pattern-matching tool for the detection of similar hotspots in the newly designed layout. Therefore, the effectiveness of template matching greatly depends on the performance of the layout pattern classification algorithm.

To minimize the time needed for pattern matching, it is essential to use as few templates as possible to effectively represent all hotspots. Chen et al. [23] used clip shifts to provide more clustering opportunities for layout patterns and promote cluster size minimization. Chang and Jiang [24] proposed a fast layout pattern classification algorithm called iClaire that uses Markov clustering to obtain natural groupings of layout patterns and uses clip shifting and center-of-mass reconstruction to further reduce the number of clusters efficiently. The geometric similarity of all the above methods is calculated using area matching and edge displacement of a pair of hotspot patterns, which are sensitive to small changes or shifts in shape and may lead to errors in hotspot detection. Guo et al. [25] suggested a metric based on improved tangent space (ITS), which is computationally efficient and can accommodate noise, while also demonstrating enhanced accuracy.

Although pattern matching has been shown to be extremely accurate and fast in hotspot detection tasks, the method is only capable of identifying hotspot types that have already been seen.

3.2 Machine learning-based hotspot detection

Lithography hotspot detection can be formulated as a binary classification problem, where the goal is to determine if a given layout pattern is a lithography hotspot. After training on a labeled dataset, the ML model can efficiently predict hotspots in new layout patterns. The detailed workflow of the methodology is presented in Fig. 4. A key advantage of this approach is its ability to detect unknown hotspots by learning the relationship between layout patterns and their printability. Nagase et al. [26] were the first to explore using artificial neural networks (ANNs) for hotspot detection by converting photolithography simulation contour images into bitmaps. These bitmaps were fed into a four-layer neural network, which successfully converged and was then used to predict unseen layouts. Ding et al. [27] used key features instead of pixel-based layout as input vectors to effectively reduce the learning/detection noise during training, resulting in an effective increase in detection accuracy. Wuu et al. [28] proposed a density-encoded feature extraction method and employed support vector machine (SVM) techniques to detect context-sensitive hotspots in the standard cell library. They further introduced a two-stage hotspot pattern classification method [29], which examines both the central pattern and the surrounding pattern to minimize false alarms.

False alarms refer to the incorrect prediction of non-hotspot patterns as hotspots, and it is a major challenge in applying ML techniques to hotspot detection, as false non-hotspot patterns require expensive post-processing hotspot removal. However, while improving the accuracy of hotspot detection, the false alarm rate inevitably increases, which guides researchers to develop more advanced algorithms to enhance the performance of hotspot detection. Ding et al. [31] proposed a high-performance lithography hotspot detection method using multi-level ML models (ANN and SVM) to gradually refine the detection results. Different levels of hotspot detection models sequentially filter and refine the detection results, reducing the false alarm rate and improving the overall detection accuracy. Building on their previous work, Ding et al. [32] further proposed EPIC, an efficient and effective predictor for IC manufacturing hotspots, leveraging quadratic programming to configure and optimize multiple classifiers, thereby integrating them into a unified meta-classifier. Yu et al. [33] first clustered the patterns based on topology, subsequently extracting key features from each cluster to train an SVM kernel tailored to each. This multi-kernel SVM detection technique significantly reduced the false alarm rate.

Training data is a critical component of ML techniques, as data quality directly impacts model prediction accuracy. In practical lithography processes, the number of patterns with hotspots is significantly smaller than non-hotspot patterns, creating a serious data imbalance issue for ML-based hotspot detection. Yang et al. [34] applied hotspot upsampling and random mirror flipping techniques to balance the training dataset prior to network training. They subsequently developed a bias learning technique for unbalanced datasets, using the discrete cosine transform (DCT) to generate

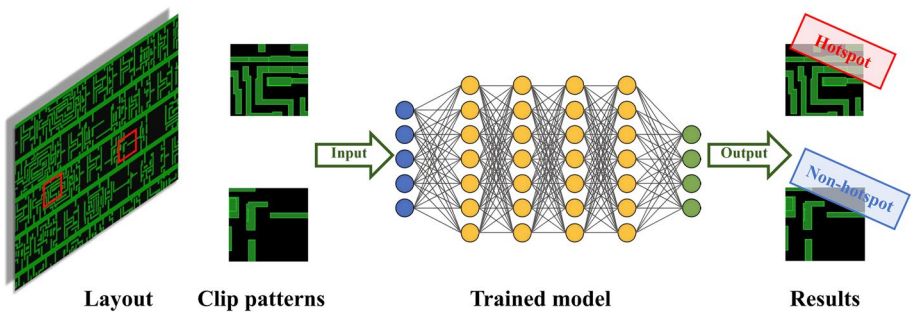


Fig. 4 Hotspot detection via machine learning. Reproduced with permission [30]. Copyright 2022, Multidisciplinary Digital Publishing Institute

feature tensors and enhance the accuracy of CNNs at lower depths [35]. Ye et al. [36] utilized a loss function for direct area under curve (AUC) maximization during training, systematically enhancing the model's performance in handling class imbalance issues in lithography hotspot detection. Wang et al. [37] proposed a two-stage classifier for hotspot detection. In the first stage, an inverse unbalanced dataset is constructed to train an iterative neural network that thoroughly learns hotspot features while progressively discarding non-hotspot samples to reduce the false alarm rate. In the second stage, a balanced dataset is created through oversampling, enabling a neural network to accurately predict the optimal class of samples.

Moreover, the acquisition of labeled data during chip design and fabrication incurs significant costs, necessitating a meticulous selection of layout patterns to ensure a sufficiently comprehensive training dataset. Chen et al. [38] proposed a semi-supervised hotspot detection framework that uses self-paced multitask learning. It assigns pseudo-labels to unlabeled samples with a classification model, checks the confidence of these labels using weights, and includes unlabeled data in training to reduce the need for labeled data. Yang et al. [39] proposed an interactive optimization technique that combines diversity-aware bulk sampling with learning models to efficiently select patterns exhibiting high uncertainty and inter-sample variance for labeling. This approach requires only a limited number of labeled patterns to achieve high detection accuracy, significantly reducing the computational overhead of photolithography simulation. Xiao et al. [40] enhanced uncertainty and diversity metrics by applying an entropy-weighting method and introduced a novel batch-mode selection strategy, further reducing lithography simulation overhead and runtime.

Shallow learning models rely heavily on manual feature extraction, making it challenging to capture complex pattern information and often leading to information loss. In contrast, deep learning, as an end-to-end approach, enables automatic extraction of image features through convolutional neural networks with stronger pattern representation capabilities, making it increasingly popular in lithographic hotspot detection [34, 35, 41]. Shin and Lee [42] were the first to introduce CNNs into lithographic hotspot detection, using shift and inversion variations to enhance the dataset and applying batch normalization to support CNN training. This approach achieved an improved balance between recall and precision. Chen et al. [43] enhanced the CNN by adding two inception modules, enabling the extraction of pattern features at multiple scales and improving hotspot detection accuracy. Geng et al. [44] proposed a novel end-to-end hotspot detection process that combines layout feature embedding with hotspot detection. They introduced an attention module within the hotspot detector, allowing the backbone network to adaptively focus on the most relevant hotspot information.

Due to the large number of network parameters, deep neural network (DNN) models are memory-intensive and require extended inference times. Jiang et al. [45] proposed a hotspot detector based on binary neural networks (BNNs), leveraging the fact that layouts can be treated as binary images. This approach achieves an 8× speedup in computational efficiency compared to floating-point neural networks. Additionally, they used an ensemble learning method to combine the BNN-based model with two of its shallower variants, further enhancing hotspot detection performance. Pei et al. [46] introduced a bit-level quantization algorithm for hotspot detection, reducing network runtime by half with an int8 network while Maintaining inference accuracy within a 1% margin.

Early lithographic hotspot detection methods focused on small layout clips, predicting hotspot probability only at the center of each clip. This approach becomes highly time-consuming when applied to large, full-chip layouts. To enable full-chip hotspot detection, Chen et al. [47] proposed a region-based hotspot detector that uses a feature extractor built on an encoder–decoder structure and an inception–based module for efficient feature extraction across the entire layout. Additionally, they implemented a novel classification and regression strategy to reduce the detection area, allowing for effective multi-hotspot detection in large-scale layouts. Gai et al. [48] introduced an end-to-end hotspot detector based on a fully convolutional network, replacing the CNN's fully connected layer with convolutional layers that use kernels covering the entire input region. This design allows the model to accept inputs of any size. Additionally, they proposed a transfer learning scheme for hotspot detectors across different technology nodes, reducing the data needed to develop detectors for advanced nodes. Zhu et al. [49] proposed an end-to-end single-stage hotspot detection framework that utilizes a ResNet-50 and feature pyramid backbone to generate multilevel feature maps. Through multitask learning, the model captures diverse topological features and includes center and corner heads to identify the centroids and corners of hotspot regions. Additionally, a transformer encoder-based module is employed to capture global feature dependencies, resulting in higher accuracy and faster inference for large-area hotspot detection.

In the latest research, efforts have been made to apply large language models (LLMs) to hotspot detection tasks, harnessing their advanced capabilities in pattern recognition and semantic processing. Chen et al. [50] introduced LLM-HD, a hotspot detection model utilizing an LLM capable of directly processing GDS (graphic database system, a standard file format for IC layout data) binaries without conversion to pixel-based images. This innovative approach retains crucial spatial and structural information often lost

in traditional image conversion. Through GDS-based hierarchical semantic encoding, LLM-HD leverages a pre-trained language model architecture, effectively adapting natural language processing (NLP) techniques to hotspot detection in semiconductor layouts. The model achieves notable gains in both accuracy and efficiency.

In summary, ML has emerged as a powerful tool for hotspot detection, offering significant advantages over traditional pattern-matching methods. While conventional approaches rely on predefined rules and struggle to recognize novel or complex patterns, ML-based methods adaptively learn from diverse and extensive datasets, enabling data-driven detection with greater accuracy. ML models streamline the feature extraction and classification process, while deep learning techniques, especially convolutional neural networks, automatically perform complex feature extraction, capturing intricate spatial details often missed by rule-based systems. Future models are expected to support large-scale, real-time analysis of semiconductor layouts and seamlessly accommodate multiple formats, including the standard GDS format and the more compact OASIS (open artwork system interchange standard) format.

4 Mask optimization

MO aims to refine mask structure to ensure accurate pattern transfer onto wafers by compensating for optical diffraction and process variability. Among the key techniques, optical proximity correction (OPC) was developed to address imaging errors caused by OPE by adjusting the local geometric structures of mask patterns, as illustrated in Fig. 5.

Traditional OPC techniques primarily include rule-based OPC (RBOPC) and model-based OPC (MBOPC), which have played important roles across different periods and technology nodes. However, as node sizes shrink and structural complexity increases, these methods struggle to meet the high-precision demands of modern lithography processes. Consequently, inverse lithography technology (ILT) and ML-based intelligent OPC methods have been introduced to achieve higher correction accuracy and efficiency. Additionally, source mask optimization (SMO), which simultaneously optimizes both the light source and mask structure, further enhances imaging quality. In this section, we focus on OPC, ILT, and SMO as they represent the core of MO and are critical to addressing the challenges of modern lithography while leveraging the potential of emerging computational techniques.

4.1 Foundational OPC techniques

RBOPC and MBOPC form the foundation of traditional OPC methods. RBOPC is suited for larger process nodes, quickly correcting mask patterns through a rule library to meet the requirements of early lithography processes. In contrast, MBOPC achieves fine correction by segmenting and optimizing the details of mask edges, offering greater precision and flexibility, The comparison of these two methods is shown in Fig. 6.

RBOPC was one of the earliest OPC methods applied in IC Manufacturing, primarily used in Lithography processes for 180 nm and larger nodes [54]. The core concept of RBOPC involved establishing a rule library based on geometric parameters and using a lookup method to correct features such as line widths, edges, and corners on the mask. During the design phase, the system matched mask pattern characteristics to correction parameters in the rule library, enabling quick local adjustments to the mask layout [55]. The RBOPC rule library typically included four primary rule types: line width, line end, corner, and via corrections. These rules were generated based on optical simulation and experimental data to accommodate mask correction requirements

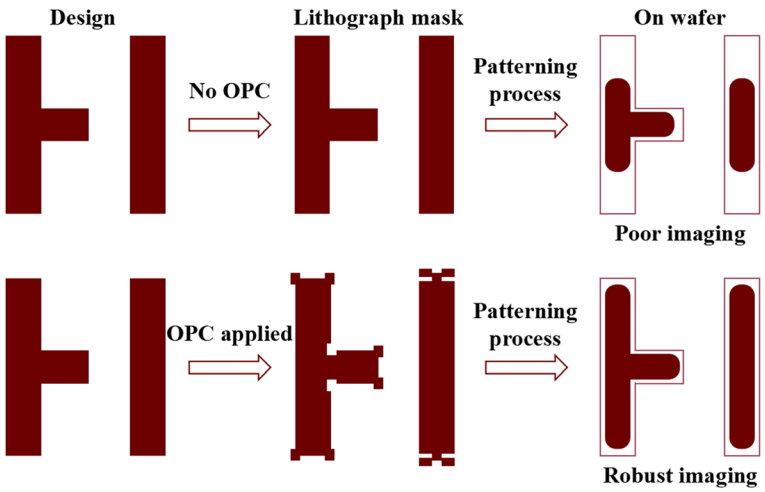


Fig. 5 Reducing OPE with OPC. Reproduced with permission [51]. Copyright 2023, SignOff Semiconductors

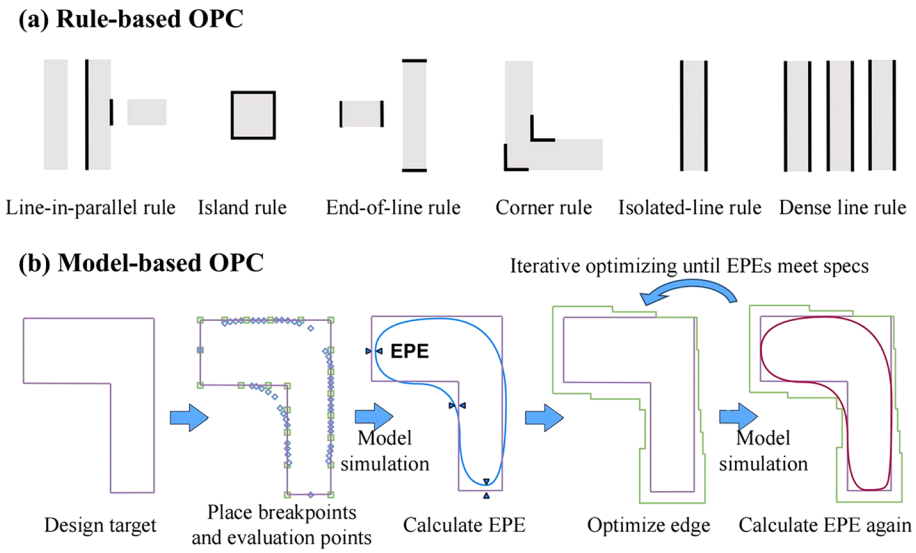


Fig. 6 **a** Rule-based OPC. Reproduced with permission [52]. Copyright 2017, IEEE and **b** model-based OPC. Reproduced with permission [53]. Copyright 2022, Chinese Optical Society

under different process conditions [56]. Due to this pre-established rule and lookup approach, RBOPC was characterized by fast computational speed and system simplicity, making it suitable for the rapid processing requirements in large-scale IC manufacturing [57].

However, the limited content and approximate nature of the rule library made it challenging to compensate for complex proximity effects or achieve global optimization. As nodes shrink below 90 nm, lithography processes demand significantly higher precision and resolution. RBOPC could not meet high-precision requirements for small dimensions and complex patterns, prompting MBOPC to be widely used in the industry. Unlike RBOPC, MBOPC utilizes mathematical modeling and numerical optimization algorithms to simulate and correct mask patterns accurately, which can achieve global optimization and provide enhanced resolution and fidelity [58]. Its core principle involved segmenting the mask edge pattern into smaller fragments and adjusting each fragment's position to compensate for optical and process deviations during lithographic imaging. The optimization was achieved by minimizing edge placement error (EPE). Although MBOPC had higher computational complexity, its optimization flexibility was significantly enhanced, enabling much higher precision at deep submicron nodes [59].

To further improve the efficiency and accuracy of MBOPC, researchers proposed various enhancements. Cobb et al. [60] introduced the mask error enhancement factor (MEEF) matrix to account for the interactive effects among multiple edge segments, enabling collaborative adjustment of the EPE at target locations. Chen et al. [61] proposed an optimization algorithm for the intensity distribution function based on the Jacobian matrix, using sparse matrix control techniques that focused only on regions with a significant impact on control point intensity, thereby avoiding global calculations. Lin et al. [62] advanced this by introducing the "sensitive region" concept, restricting calculations to specific areas that influenced control point intensity and thus reducing redundant computations. Furthermore, Lei et al. [63] optimized the MEEF matrix by developing a sparse matrix-based feedback control mechanism, which improved EPE correction in complex layouts through cross-segment optimization.

As the requirements for lithography processes continued to increase, research on the process adaptability of MBOPC gradually attracted attention. Kuang et al. [64] proposed a process-variation-aware MO method that enhanced the robustness of mask design through phased iterative optimization. This approach included edge fragmentation, assist feature insertion, EPE optimization, and process window optimization. Awad et al. [65] introduced a method based on intensity difference maps that used adaptive assist feature insertion to optimize EPE control. By incorporating fragment shifting and "hammer" structure design, this method reduced the computational load of lithographic simulation.

The intricate polygonal patterns in deep submicron nodes, particularly complex structures with small feature sizes, such as corner rounding and edge "ripple" effects, posed challenges for traditional MBOPC under high-precision requirements. To address these issues, researchers proposed adaptive segmentation and localized optimization methods. Yang et al. [66] introduced a dynamic adaptive segmentation algorithm that adjusted segmentation methods and sampling point locations during the correction loop, reducing pattern distortion caused by fixed segmentation approaches. Huang et al. [67] suggested the addition of specially designed shapes at corners to mitigate rounding effects and enhance the accuracy of small pixel patterns. Additionally, Weisbuch et al. [68] incorporated SEM images enriched with 2D and 3D contour information, improving model calibration accuracy and enhancing adaptability to complex photoresist structures.

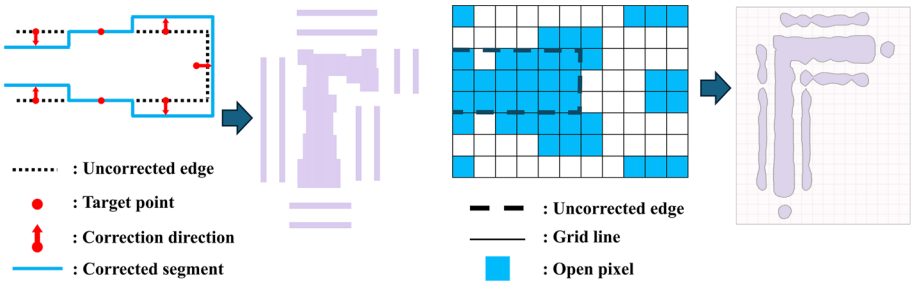


Fig. 7 Comparison of OPC (left) and ILT (right). Reproduced with permission [73] Copyright 2022, SemiWiki

4.2 Inverse lithography technology

Traditional OPC methods, with their inherently limited solution space, struggle to meet the precision demands of modern lithography at advanced technology nodes, presenting significant challenges for lithographic processes. In contrast, ILT generates high-precision masks by inversely solving optical problems [69], Making it a powerful candidate for OPC in 32 nm and smaller nodes [70].

ILT is a technique that utilizes mathematical optimization methods to design photomask patterns, aiming to achieve highly accurate and desired patterns on the wafer. Its core definition focuses on solving an inverse problem. Traditional lithography methods typically rely on a "forward" process, where patterns are sequentially imaged from the mask onto the wafer. However, as chip dimensions continue to shrink and circuit complexity increases, traditional approaches struggle to meet the demands for finer patterns and higher fidelity. Against this backdrop, ILT targets the creation of more complex and precise patterns on the wafer through inverse problem-solving.

Mathematically, ILT is typically constructed as a framework for solving an inverse problem, where the optimization of the mask design aims to meet specific imaging requirements. The inputs to this process are the target wafer pattern (the desired pattern on the wafer, commonly denoted as Φ) and the known lithographic forward operation (the transformation process from the mask pattern to the wafer pattern, typically denoted as f). The objective of ILT is to find an optimal mask pattern ψ, such that the resulting wafer pattern ω produced under the lithography process closely matches the design target Φ, thereby satisfying the approximate relationship:

$$\omega = f(\psi), \tag{9}$$

$$\psi = f^{-1}(\Phi). \tag{10}$$

Since the forward operation f in the lithography process involves complex physical phenomena, such as light diffraction, photoresist response, and focal adjustment, the problem lacks a straightforward analytical solution. To address this challenge, ILT reformulates the inverse problem as an optimization task: by defining a target or cost function (often referred to as the Hamiltonian), it measures the difference between the mask pattern and the resulting wafer pattern. This cost function serves to quantify the "error" between the wafer pattern produced by the mask and the target pattern. By optimizing this function, ILT can incrementally approach the optimal solution.

Specifically, within the ILT framework, the energy or cost function is typically defined based on the pixel-level differences between the wafer pattern ω and the target pattern Φ [71, 72]. For instance, this function can be constructed by calculating the sum of squared errors between the wafer and target patterns, ensuring that the error is minimized. To accelerate the approach to the optimal solution ψ, ILT often employs gradient-based optimization methods. These gradient optimization techniques calculate the gradient of the cost function concerning the current mask design, providing a direction to adjust the mask pattern such that the cost function moves toward minimization. Common gradient optimization techniques include gradient descent, conjugate gradient methods, and others. Through these iterative methods, the mask design is gradually updated, with each iteration reducing the error in the cost function, ultimately reaching an optimal mask design.

ILT and OPC differ significantly in their underlying principles. ILT is a global optimization-based design method that defines a cost function and seeks an optimal solution, enabling the mask pattern to closely replicate the target design with high fidelity. Unlike OPC, which relies on predefined rules or physical models to make local adjustments to mask edges for OREs, ILT's optimization is unrestricted by geometric shapes, allowing for complex, non-Manhattan structures that can better counteract various optical distortions. In contrast, OPC is limited to generating Manhattan geometries composed of rectangles and straight lines and lacks a global optimization perspective, as it typically focuses on correcting specific hotspot areas only, as shown in Fig. 7.

ILT demonstrates significant advantages in flexibility and structural complexity. With its global optimization capability and freedom in shape design, ILT effectively compensates for optical effects and provides higher fidelity in pattern imaging along with a broader process window [74]. This makes ILT particularly suitable for scenarios with extreme nodes and highly complex pattern requirements, meeting the

precision demands of advanced lithography. In comparison, OPC, which relies on preset rules, offers limited flexibility and struggles to address irregular or complex optical effects, thus restricting its precision at smaller design nodes. ILT's high flexibility allows it to generate intricate nonlinear curves and curved structures, which align better with the optical needs of advanced nodes, making it exceptionally well suited for complex process requirements.

The global optimization and precise, pixel-level adjustments involved in ILT lead to higher computational complexity, typically necessitating advanced computing resources such as graphics processing unit (GPU) acceleration and parallel processing to ensure timely optimization [75, 76]. Although ILT's computational demands are substantial, advances in technology continue to improve its efficiency, enabling expanded applications at advanced nodes. While ILT incurs higher computational costs than OPC, it significantly extends the process window and enhances imaging fidelity, rendering it highly valuable for advanced lithography processes. By contrast, OPC only requires local edge adjustments, making it computationally lighter and faster, suitable for rapid, large-scale production at conventional nodes. However, in high-precision and complex scenarios, OPC's capabilities are limited compared to ILT.

4.3 Machine learning-based OPC

In the pursuit of higher precision and lower defect rates in lithography, traditional OPC methods face bottlenecks in efficiency and computational cost. Conventional iterative optimization models are not only time-consuming but also struggle to meet the precise mask generation demands required by complex designs. With the development of ML, particularly deep learning, both the speed and quality of the OPC process have improved significantly. ML provides flexible and efficient tools to optimize the mask generation process, supporting precise lithographic patterning. The following sections discuss the application of ML in both local and full-chip MO.

4.3.1 Deep learning enhanced OPC

Compared to traditional ML, deep learning offers significantly more powerful capabilities in data processing, feature extraction, and nonlinear prediction [77], enabling numerous new possibilities for OPC. By replacing traditional iterative processes with end-to-end deep learning models, computational time can be greatly reduced, achieving notable improvements in print quality and mask manufacturability. This section integrates various types of deep learning models to explore their applications in local MO.

In the early OPC processes, researchers adopted ML-based printability prediction and edge displacement prediction models to improve mask generation efficiency. Matsunaga et al. [78] proposed a hierarchical Bayesian model (HBM) that utilizes Bayesian inference and Markov Chain Monte Carlo (MCMC) methods to construct a multi-level prediction framework, effectively reducing the iteration demands and computational costs of traditional models. For printability prediction, Jiang et al. [79] introduced a ML framework based on matrix concentric circle sampling (MCCS) and a second-order circular subset selection algorithm, which accelerates the OPC process through efficient feature extraction, enabling rapid prediction without reliance on specific neural network architectures.

With the rapid development of deep learning, the application of generative models in MO has gradually gained traction. These methods reformulate the MO problem as an image generation task. These approaches leverage generative models, such as generative adversarial networks (GANs), to capture the mapping relationship between target patterns and masks, enabling the efficient generation of optimized mask patterns. Yang et al. [80] first proposed a GAN-OPC, where the generator is designed as an autoencoder structure to generate initial mask patterns from target patterns, while the discriminator evaluates the quality of the generated masks. To address the challenge of GAN model convergence, they employ an ILT-guided pretraining strategy during training to effectively initialize the generator. This approach can generate more precise mask patterns while simplifying the complex iterative processes of traditional OPC. Furthermore, Alawieh et al. [81] formulated the SRAF generation problem as an image-to-image translation task, utilizing conditional GANs (CGANs) to generate SRAF layouts. To overcome the limitations of GANs in generating sharp-edged shapes, they adopt a multi-channel heatmap encoding scheme to capture the global distribution characteristics of SRAFs, producing solutions that more closely align with the target SRAF layouts. Zheng et al. [82] introduced EMOGen, a co-evolution framework combining variational autoencoder-based pattern generation and DNN-based ILT. By generating topology matrices and geometry vectors under design rule constraints, EMOGen produces challenging patterns to expose ILT weaknesses, enabling iterative improvement of both models and enhancing mask optimization efficiency.

Recently, CNN has been studied to help select the best decomposition candidates and determine the appropriate OPC engine. Zhong et al. [83] proposed the layout decomposition and mask optimization (LDMO) framework, which utilized CNN-based clustering and sampling methods to filter layouts in the early stages of optimization, focusing resources on the most promising candidate layouts, thus enhancing MO efficiency. In the same year, Yang et al. [84] introduced a heterogeneous OPC framework that used a classification model to automatically select the most suitable

OPC engine for different patterns, achieving adaptive optimization for complex layouts.

To push further advancements in ILT, researchers have developed several novel end-to-end optimization methods. Chen et al. [85] proposed labelfree neural network-based inverse lithography technology (LF-NNILT), which leveraged an unsupervised autoencoder network and lithography system layer to enable MO without labeled data. Yang and Ran [86] further developed the implicit learning of inverse lithography technologies (ILILT) framework, using an implicit learning layer to directly generate optimized masks and a weight-bound sequence model to simulate the optimization trajectory, thus reducing the iterative demands of traditional ILT. Chen et al. [87] introduced the DiffOPC method, which combined edge- and pixel-level optimization within ILT. This method employed a differentiable design to achieve continuous edge segment optimization, with acceleration using NVIDIA's compute unified device architecture (CUDA) and mask rule constraints ensuring high manufacturability of the generated masks. Recently, Xu et al. [88] proposed the ERFNet-ILT framework—an approach for inverse lithography technology that leverages ERFNet, an efficient residual decomposition convolutional network. This framework uses the ERFNet architecture and a feature fusion module to generate mask labels without pre-labeled data. Only a set of layout patterns and compact lithography simulation models are required.

4.3.2 Full-chip optimization

As IC design scale and complexity increase, localized MO no longer meets the requirements, making efficient and accurate full-chip optimization a key demand for new technology nodes. Most existing studies focus on the optimization of small mask clips. However, applying these results at the full-chip scale requires seamlessly integrating the fragment optimizations into a complete layout. This poses significant technical challenges in balancing large-scale processing, boundary continuity, pattern accuracy, and computational efficiency. These challenges have made full-chip MO a focal point of research in recent years.

Inspired by GAN-OPC, Chen et al. [89] proposed the deep agile mask optimization (DAMO) framework, which employs a coarse-to-fine window segmentation algorithm. This approach uses DBSCAN (density-based spatial clustering of applications with noise) clustering to identify high via density regions and further refines these regions with KMeans++ to generate fixed-size windows, ensuring that deep learning algorithms can efficiently handle full-chip layouts. The core innovation of DAMO lies in the shared DCGAN model architecture between the lithography simulator and the mask generator, enhanced with UNet++ and residual blocks. The lithography simulator backpropagates lithographic error gradients to guide the mask generator in optimizing mask designs, effectively addressing the computational bottlenecks and optimization challenges of traditional methods at the full-chip scale. Following this, Ciou et al. [90] introduced a pix2pix conditional GAN for mask generation. By employing high-precision image stitching techniques, this approach achieves seamless detail generation and merging in full-chip masks, meeting high accuracy requirements.

Researchers have proposed pattern reuse and adaptive optimization strategies to further reduce computational redundancy associated with repetitive patterns in full-chip optimization. AdaOPC, a method introduced by Zhao et al. [91], introduced an adaptive solver selection mechanism, combined with a pattern library reuse technique, to identify and reuse similar pattern optimizations, significantly enhancing full-chip processing efficiency. The improved AdaOPC 2.0 [92] incorporates online updates of hierarchical libraries and supports variation with displacement, enabling adaptive solver selection in regions of varying complexity, further reducing computational redundancy and effectively reusing large-scale repetitive patterns such as via layers, increasing optimization efficiency.

Learning to optimize (L2O) methods and open-source platforms also provide flexible environments for algorithm development and testing in full-chip MO. In L2O-ILT, a method introduced by Zhu et al. [93] proposed L2O-ILT, which unfolded the iterative optimization process of ILT into a neural network structure, with each layer representing an iteration of the ILT algorithm. This alternative optimization strategy accelerates convergence, reducing the resource demands for full-chip optimization. Meanwhile, Zheng et al. [94] developed OpenILT as an open-source platform based on the PyTorch framework, which integrates GPU acceleration and Fast Fourier Transform (FFT) operations, supporting parallel processing of full-chip data. This allows users to quickly develop and validate custom ILT algorithms, greatly enhancing the flexibility of full-chip MO. Yang et al. [95] proposed a method that combines convolutional neural networks with Fourier neural operators (FNO) and uses litho-guided self training technology to achieve seamless large-scale MO. This method captures long-range dependencies in the design, ensuring global information is processed effectively in large-scale MO.

Recently, reinforcement learning methods have achieved remarkable success in MO, particularly in automated strategy selection and multi-boundary collaborative adjustment. Liang et al. [96] proposed RL-OPC, a method leveraging deep Q-learning to directly optimize key performance metrics, including EPE and process variation band (PVB). By training local edge movement strategies, this method processes full-chip mask design piece by piece, achieving efficient global optimization. Building on

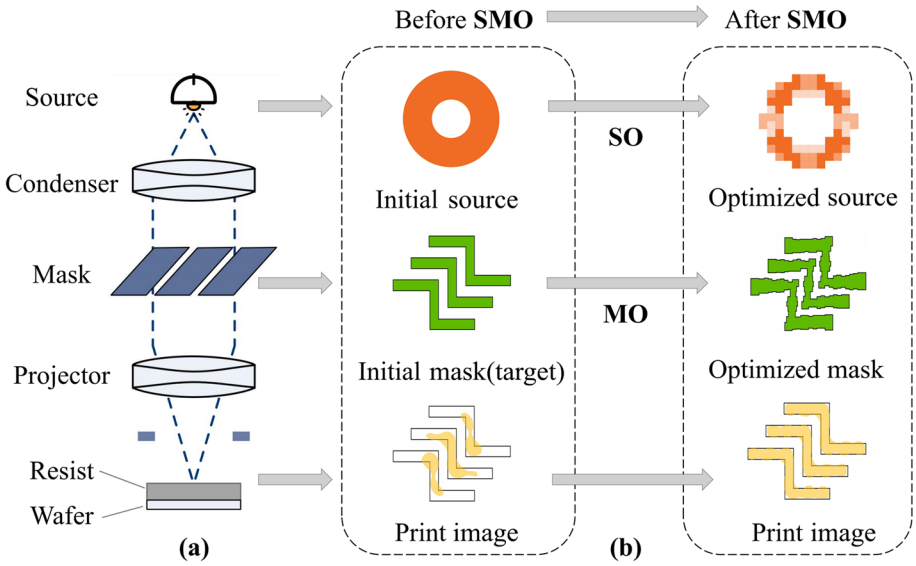


Fig. 8 SMO technology schematic. Reproduced with permission [100]. Copyright 2024, IEEE

this work, they introduced CAMO [97], an extension of RL-OPC, which integrates graph neural networks (GNN) and recurrent neural networks (RNN) into a reinforcement learning framework. This extension captures spatial correlations in mask boundaries, significantly enhancing spatial consistency in multi-segment boundary optimization. To further minimize human intervention, Chen et al. [98] proposed the intelligent OPC system, which combines reinforcement learning with LLMs. Using RL for global exploration and LLM to generate the final OPC recipe, this system fully automates mask design, significantly reducing computational costs and labor demands.

4.4 Source mask optimization

While OPC and ILT technologies have made significant strides in MO, their focus on the mask alone overlooks the critical influence of source design on lithographic performance. Lithographic imaging is a synergistic process involving the source, mask, and optical system, with the source's shape and properties playing a decisive role in determining optical transmission efficiency and resolution. As process nodes shrink further, MO becomes inadequate for achieving higher pattern resolution. This limitation is particularly pronounced in extreme ultraviolet (EUV) lithography and advanced multi-exposure techniques, where the coordinated optimization of the mask and source has become a key area of research [99].

4.4.1 Principles of SMO

SMO technology offers greater optimization flexibility than traditional lithography optimization techniques, as it can simultaneously optimize both the illumination source and mask patterns, as shown in Fig. 8. This technology fully leverages the characteristics of off-axis illumination, effectively enhancing lithographic resolution.

However, the high optimization flexibility comes with challenges such as increased computational load and high time costs. This is particularly evident in full-chip SMO applications, where the layout contains thousands of patterns. In such cases, directly implementing SMO could lead to unacceptable time costs, so full-chip SMO is typically divided into several steps [101, 102]: first, key patterns are selected from the layout, and SMO is performed on these patterns to obtain an optimized illumination source suitable for the entire chip. Next, this optimized illumination source performs MO for the full-chip patterns. During this process, the effectiveness of the key pattern selection technique directly determines the illumination source obtained through SMO, which in turn affects the final optimized process window. Additionally, the characterization methods for the source and mask also significantly influence the optimization flexibility and efficiency of SMO.

Based on optimization principles, SMO methods can be classified into gradient-based algorithms [103–105] and heuristic algorithms [106, 107]. Gradient-based algorithms leverage the gradient of the cost function, primarily optimizing by minimizing the objective function (such as imaging quality or process window) during the lithographic imaging process. These algorithms demonstrate good convergence under specific conditions and can converge effectively, but they may become trapped in local solutions. In contrast, heuristic algorithms, such as particle swarm optimization (PSO) and genetic algorithms (GA), have a strong ability to search for the global optima without relying on gradient

information, however, they tend to be less efficient and incur relatively high computational costs.

4.4.2 Machine learning-based SMO

With the rapid development of ML technologies, researchers have proposed a series of innovative methods to enhance optimization efficiency and computational capabilities. Among these, Chen et al. [108] developed SoulNet, which utilizes generative neural networks for ultra-fast source optimization (SO), demonstrating advantages in handling high-dimensional data. Wang et al. [109] proposed a method for fast pixelated photolithographic source and mask joint optimization, leveraging compressed sensing for efficient information extraction, significantly improving the processing capabilities for large-scale designs. This method ensures photolithographic precision while reducing the computational resources required. Sun et al. [110] introduced a sampling-based imaging model that combines heuristic algorithms to enhance the efficiency of source and MO in immersion lithography. Despite still relying on partial partitioning techniques, its flexible processing capabilities and rapid generation of feasible solutions allow it to perform excellently in complex tasks. Diff-SMO [100] adopts a conditional discrete diffusion model to fully accelerate the entire SMO process through GPU utilization based on Abbe imaging theory. This method not only enhances the overall optimization efficiency but also introduces prior knowledge to guide the optimization process, enabling the generation of high-quality source patterns and further improving the efficiency of the SMO process. The introduction of the BiSMO framework [111] represents a breakthrough in SMO. This framework enhances lithographic precision and computational speed by simultaneously calculating the gradients for SO and MO. By addressing the constraints of lower-level SO on upper-level MO, it provides a global perspective for the optimization process, helping to surpass local minima and improving the exploration of the solution space, thus replacing traditional SMO methods. Furthermore, in recent years, traditional SMO methods have focused on evaluating imaging quality at the focal plane, while ignoring the process window. To address this, researchers have proposed a multi-objective optimization SMO method based on the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [112]. This method combines a rapid focus-varying aerial image model (NSGA-SMO) to directly optimize PW performance while maintaining imaging quality, thereby enhancing the robustness of SMO results.

In summary, the application of ML in SMO is continually deepening, encompassing a variety of methods ranging from compressive sensing and sampling models to generative networks. These studies significantly enhance the efficiency and accuracy of SMO and provide new insights for the future development of lithography technology.

5 Conclusion and future challenges

In conclusion, computational lithography has become the cornerstone of modern semiconductor manufacturing, addressing the resolution and pattern fidelity challenges faced by traditional photolithographic techniques. The key components of this field—optical modeling, hotspot detection, and MO—work synergistically to achieve the overarching goal of transferring high-precision patterns onto silicon wafers. Together, these technologies ensure that photolithography can continue to scale to increasingly smaller technology nodes, which are critical for the ongoing development of semiconductor devices.

The integration of ML into computational lithography has been transformative, automating prediction and correction processes, accelerating design cycles, and improving lithographic accuracy and reliability. However, significant challenges remain, particularly in managing computational complexity and acquiring high-quality training datasets. Future advancements must address these limitations through more efficient data utilization, improved model interpretability, and enhanced computational frameworks.

ML models heavily depend on large-scale, high-quality datasets, but data collection in production environments is prohibitively expensive and time-consuming. Integrating computational lithography principles with advanced data processing techniques, such as compressed sensing and transfer learning, could alleviate this dependency, reducing data acquisition costs while maintaining predictive accuracy.

While deep learning offers powerful predictive capabilities, its black-box nature limits physical interpretability, making it difficult to diagnose errors or refine models when deviations occur. Future research should focus on physics-informed ML, where domain-specific constraints and theoretical formulations enhance model transparency. By integrating prior knowledge from optical and resist models, ML-driven lithographic solutions can transition from approximate predictions to more precise, explainable, and manufacturable outcomes.

As technology nodes continue to shrink, lithographic models must evolve accordingly. EUV lithography, though critical for advanced semiconductor manufacturing, presents new challenges such as photon shot noise, stochastic defects, and increased process variability. Future research should refine multiexposure and SMO strategies while exploring new modeling techniques that accurately capture complex three-dimensional device structures like FinFETs and Gate-All-Around (GAA) transistors.

The increasing computational demands of lithographic simulations require breakthroughs in high-performance computing. Parallelizing ILT using GPU acceleration and specialized hardware architectures could enable full-chip ILT implementation, significantly improving computational efficiency. Further advancements in differentiable inverse design algorithms may also streamline optimization workflows, reducing turnaround times while enhancing lithographic precision.

The convergence of IC design and manufacturing necessitates deeper integration of design-technology co-optimization (DTCO). Embedding lithography-friendly constraints into EDA workflows would enable automated detection and avoidance of high-risk design patterns at earlier stages, improving manufacturability. Additionally, incorporating real-time feedback mechanisms, such as defect distribution data from the manufacturing process, into iterative design-rule adjustments could further enhance yield. The emergence of ML-driven DTCO frameworks, including LLMs, offers new possibilities for automating lithography-aware design evaluation, reducing costly design iterations, and improving process predictability.

Authors' contributions Qian Jin was responsible for the overall content planning, research, and drafting of the manuscript. Qingsong Peng surveyed the related works in Section 3. Yumeng Liu conducted the research for lithography modeling presented in Section 2. Xiaotian Qiu focused on the research related to ILT technologies within Section 4 on mask optimization. Qi Sun, the corresponding author, reviewed and approved the manuscript. All authors have read and approved the final manuscript.

Funding This work was funded by the Zhejiang University Education Foundation Qizhen Scholar Foundation.

Data availability The data supporting this study's findings are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate Ethical approval was not required as this study did not involve human or animal subjects.

Consent for publication Consent for publication is granted as the manuscript does not contain any personally identifiable information.

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Thompson LF (1983) An introduction to lithography. ACS Publications. https://doi.org/10.1021/bk-1983-0219.ch001
2. Lam EY, Wong AK (2009) Computation lithography: virtual reality and virtual virtuality. Opt Express 17(15):12259–12268. https://doi.org/10.1364/oe.17.012259
3. Lin Y (2021) Deep learning for mask synthesis and verification: a survey. In: Proceedings of the 26th Asia and South Pacific Design Automation Conference, ACM, Virtual, pp 825–832. https://doi.org/10.1145/3394885.3431624
4. Jiang Y, Lu X, Jin Q, Sun Q, Wu H, Zhuo C (2024) FabGPT: an efficient large multimodal model for complex wafer defect knowledge queries. arXiv preprint arXiv:2407.10810
5. Jin Q, Qiao Y, Chen Y, Zhuo C, Sun Q (2024) A selectivenet-based method for defect classification in semiconductor manufacturing. In: 2024 Conference of Science and Technology for Integrated Circuits (CSTIC), IEEE, Shanghai, pp 1–3. https://doi.org/10.1109/CSTIC61820.2024.10531871
6. Jin Q, Jiang Y, Lu X, Liu Y, Chen Y, Gao D et al (2024) SEM-CLIP precise few-shot learning for nanoscale defect detection in scanning electron microscope image. In: 2024 IEEE/ACM International Conference On Computer Aided Design (ICCAD), IEEE, New Jersey, pp 1–8. https://doi.org/10.1145/3676536.3676752
7. Qiao Y, Xie W, Lou S, Jin Q, Zeng L, Chen Y et al (2024) Minimizing labeling, maximizing performance: a novel approach to nanoscale scanning electron microscope (SEM) defect segmentation. In: Proceedings of the 61st ACM/IEEE Design Automation Conference, ACM, San Francisco, pp 1–6. https://doi.org/10.1145/3649329.3655939
8. Rapp M, Amrouch H, Lin Y, Yu B, Pan DZ, Wolf M et al (2021) MLCAD: a survey of research in machine learning for CAD keynote paper. IEEE Trans Comput-Aided Des Integr Circ Syst 41(10):3162–3181. https://doi.org/10.1109/TCAD.2021.3124762
9. Huang G, Hu J, He Y, Liu J, Ma M, Shen Z et al (2021) Machine learning for electronic design automation: a survey. ACM Trans Des Autom Electron Syst (TODAES) 26(5):1–46. https://doi.org/10.1145/3451179
10. Yu B (2023) Machine learning in EDA: when and how. In: 2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD), IEEE, Snowbird, pp 1–6
11. Chen T, Zhang GL, Yu B, Li B, Schlichtmann U (2022) Machine learning in advanced IC design: a methodological survey. IEEE Des Test 40(1):17–33. https://doi.org/10.1109/MDAT.2022.3216799
12. Goodman JW (2005) Introduction to Fourier Optics, 3rd edn. Roberts and Company Publishers, New York
13. Hopkins, HH (1953) On the diffraction theory of optical images. In: Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 217(1130), 408-432. https://doi.org/10.1098/rspa.1953.0071
14. Yang H, Lin Y, Yu B (2022) Machine learning for mask synthesis and verification. Machine Learning Applications in Electronic Design Automation. Springer International Publishing, Cham, pp 425–470. https://doi.org/10.1007/978-3-031-13074-8_15
15. Lin Z, Ren K, Gao D, Wu Y, Xu S, Lu M (2024) End-to-end lithography modeling based on process parameters and deep learning. In: 2024 2nd International Symposium of Electronics

Design Automation (ISEDA), IEEE, Xi'an, pp 524–529. https://doi.org/10.1109/ISEDA62518.2024.10617568
16. Yao H, Sinha S, Chiang C, Hong X, Cai Y (2006) Efficient process-hotspot detection using range pattern matching. In: Proceedings of the 2006 IEEE/ACM international conference on Computer aided design, ACM, San Jose, pp 625–632. https://doi.org/10.1145/1233501.1233630
17. Xu J, Sinha S, Chiang CC (2007) Accurate detection for process-hotspots with vias and incomplete specification. In: 2007 IEEE/ACM International Conference on Computer-Aided Design, IEEE, San Jose, pp 839–846. https://doi.org/10.1109/ICCAD.2007.4397369
18. Yu YT, Chan YC, Sinha S, Jiang IHR, Chiang C (2012) Accurate process-hotspot detection using critical design rule extraction. In: Proceedings of the 49th Annual Design Automation Conference, ACM, San Francisco, pp 1167–1172. https://doi.org/10.1145/2228360.2228565
19. Park JW, Todd R, Song X (2016) Geometric pattern match using edge driven dissected rectangles and vector space. IEEE Trans Comput Aided Des Integr Circ Syst 35(12):2046–2055. https://doi.org/10.1109/TCAD.2016.2535908
20. Wang SH, Chen YJ, Wang TC, Chen O (2020) An algorithm for rule-based layout pattern matching. In: Proceedings of the 39th International Conference on Computer-Aided Design, ACM, Virtual, pp 1–8. https://doi.org/10.1145/3400302.3415606
21. Qiu Q, Zhang Y, Ge W, Wang C (2024) Edge pair-based layout pattern matching using space-filling curve. In: 2024 2nd International Symposium of Electronics Design Automation (ISEDA), IEEE, Xi'an, pp 404–409. https://doi.org/10.1109/ISEDA62518.2024.10617982
22. Ma N, Ghan J, Mishra S, Spanos C, Poolla K, Rodriguez N et al (2008) Automatic hotspot classification using pattern-based clustering. In: Design for Manufacturability through Design Process Integration II, vol 6925. SPIE, pp 35–44. https://doi.org/10.1117/12.772867
23. Chen KJ, Chuang YK, Yu BY, Fang SY (2017) Minimizing cluster number with clip shifting in hotspot pattern classification. In: Proceedings of the 54th Annual Design Automation Conference 2017, ACM, Austin, pp 1–6. https://doi.org/10.1145/3061639.3062283
24. Chang WC, Jiang IHR (2019) iClaire: a fast and general layout pattern classification algorithm with clip shifting and centroid recreation. IEEE Trans Comput Aided Des Integr Circ Syst 39(8):1662–1673. https://doi.org/10.1109/TCAD.2019.2917849
25. Guo J, Yang F, Sinha S, Chiang C, Zeng X (2012) Improved tangent space based distance metric for accurate lithographic hotspot classification. In: Proceedings of the 49th Annual Design Automation Conference, ACM, San Francisco, pp 1173–1178. https://doi.org/10.1145/2228360.2228577
26. Nagase N, Suzuki K, Takahashi K, Minemura M, Yamauchi S, Okada T (2007) Study of hot spot detection using neural networks judgment. In: Photomask and Next-Generation Lithography Mask Technology XIV, vol 6607. SPIE, San Jose, pp 393–398. https://doi.org/10.1117/12.728959
27. Ding D, Wu X, Ghosh J, Pan DZ (2009) Machine learning based lithographic hotspot detection with critical-feature extraction and classification. In: 2009 IEEE international conference on IC design and technology, IEEE, Austin, pp 219–222. https://doi.org/10.1109/ICICDT.2009.5166300
28. Wuu JY, Pikus FG, Torres A, Marek-Sadowska M (2009) Detecting context sensitive hot spots in standard cell libraries. In: Design for Manufacturability through Design-Process Integration III, vol 7275. SPIE, San Jose, pp 312–320. https://doi.org/10.1117/12.814316
29. Wuu JY, Pikus FG, Torres A, Marek-Sadowska M (2011) Rapid layout pattern classification. In: 16th Asia and South Pacific Design Automation Conference (ASP-DAC 2011), IEEE, Yokohama, pp 781–786
30. Liao L, Li S, Che Y, Shi W, Wang X (2022) Lithography hotspot detection method based on transfer learning using pre-trained deep convolutional neural network. Appl Sci 12(4):2192. https://doi.org/10.3390/app12042192
31. Ding D, Torres JA, Pan DZ (2011) High performance lithography hotspot detection with successively refined pattern identifications and machine learning. IEEE Trans Comput Aided Des Integr Circ Syst 30(11):1621–1634. https://doi.org/10.1109/TCAD.2011.2164537
32. Ding D, Yu B, Ghosh J, Pan DZ (2012) EPIC: efficient prediction of IC manufacturing hotspots with a unified meta-classification formulation. In: 17th Asia and South Pacific Design Automation Conference, IEEE, Sydney, pp 263–270. https://doi.org/10.1109/ASPDAC.2012.6164956
33. Yu YT, Lin GH, Jiang IHR, Chiang C (2013) Machine-learning-based hotspot detection using topological classification and critical feature extraction. In: Proceedings of the 50th annual design automation conference, ACM, Austin, pp 1–6. https://doi.org/10.1145/2463209.2488816
34. Yang H, Luo L, Su J, Lin C, Yu B (2017) Imbalance aware lithography hotspot detection: a deep learning approach. J Micro/Nanolithography MEMS MOEMS 16(3):033504–033504. https://doi.org/10.1117/1.JMM.16.3.033504
35. Yang H, Su J, Zou Y, Yu B, Young EF (2017) Layout hotspot detection with feature tensor generation and deep biased learning. In: Proceedings of the 54th Annual Design Automation Conference 2017, ACM, Austin, pp 1–6. https://doi.org/10.1145/3061639.3062270
36. Ye W, Lin Y, Li M, Liu Q, Pan DZ (2019) LithoROC: lithography hotspot detection with explicit ROC optimization. In: Proceedings of the 24th Asia and South Pacific Design Automation Conference, ACM, Tokyo, pp 292–298. https://doi.org/10.1145/3287624.3288746
37. Wang B, Jiang L, Zhu W, Guo L, Chen J, Chang YW (2021) Two-stage neural network classifier for the data imbalance problem with application to hotspot detection. In: 2021 58th ACM/IEEE Design Automation Conference (DAC), IEEE, pp 175–180. https://doi.org/10.1109/dac18074.2021.9586237
38. Chen Y, Lin Y, Gai T, Su Y, Wei Y, Pan DZ (2019) Semi-supervised hotspot detection with self-paced multi-task learning. In: Proceedings of the 24th Asia and South Pacific Design Automation Conference, ACM, Tokyo, pp 420–425. https://doi.org/10.1145/3287624.3287685
39. Yang H, Li S, Tabery C, Lin B, Yu B (2020) Bridging the gap between layout pattern sampling and hotspot detection via batch active learning. IEEE Trans Comput Aided Des Integr Circ Syst 40(7):1464–1475. https://doi.org/10.1109/TCAD.2020.3015903
40. Xiao Y, Su M, Yang H, Chen J, Yu J, Yu B (2021) Low-cost lithography hotspot detection with active entropy sampling and model calibration. In: 2021 58th ACM/IEEE Design Automation Conference (DAC), IEEE, San Francisco, pp 907–912. https://doi.org/10.1109/DAC18074.2021.9586273
41. Yang H, Lin Y, Yu B, Young EF (2017) Lithography hotspot detection: from shallow to deep learning. In: 2017 30th IEEE International System-on-Chip Conference (SOCC), IEEE, Munich, pp 233–238. https://doi.org/10.1109/SOCC.2017.8226047
42. Shin M, Lee JH (2016) Accurate lithography hotspot detection using deep convolutional neural networks. J Micro/Nanolithography MEMS MOEMS 15(4):043507–043507. https://doi.org/10.1117/1.JMM.15.4.043507

43. Chen J, Lin Y, Guo Y, Zhang M, Alawieh MB, Pan DZ (2019) Lithography hotspot detection using a double inception module architecture. J Micro/Nanolithography MEMS MOEMS 18(1):013507–013507. https://doi.org/10.1117/1.JMM.18.1.013507
44. Geng H, Yang H, Zhang L, Miao J, Yang F, Zeng X et al (2020) Hotspot detection via attention-based deep layout metric learning. In: Proceedings of the 39th International Conference on Computer-Aided Design, ACM, Virtual, pp 1–8. https://doi.org/10.1145/3400302.3415661
45. Jiang Y, Yang F, Yu B, Zhou D, Zeng X (2020) Efficient layout hotspot detection via binarized residual neural network ensemble. IEEE Trans Comput Aided Des Integr Circ Syst 40(7):1476–1488. https://doi.org/10.1109/TCAD.2020.3015918
46. Pei Z, Zhao W, He Z, Yu B (2023) Bit-level quantization for efficient layout hotspot detection. In: 2023 International Symposium of Electronics Design Automation (ISEDA), IEEE, Nanjing, pp 465–470. https://doi.org/10.1109/ISEDA59274.2023.10218502
47. Chen R, Zhong W, Yang H, Geng H, Zeng X, Yu B (2019) Faster region-based hotspot detection. In: Proceedings of the 56th Annual Design Automation Conference 2019, ACM, Las Vegas, pp 1–6. https://doi.org/10.1145/3316781.3317824
48. Gai T, Qu T, Wang S, Su X, Xu R, Wang Y et al (2021) Flexible hotspot detection based on fully convolutional network with transfer learning. IEEE Trans Comput Aided Des Integr Circ Syst 41(11):4626–4638. https://doi.org/10.1109/TCAD.2021.3135786
49. Zhu B, Chen R, Zhang X, Yang F, Zeng X, Yu B et al (2021) Hotspot detection via multi-task learning and transformer encoder. In: 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), IEEE, pp 1–8. https://doi.org/10.1109/ICCAD51958.2021.9643590
50. Chen Y, Wu Y, Wang J, Wu T, He X, Yu J et al (2024) LLM-HD: layout language model for hotspot detection with GDS semantic encoding. In: Proceedings of the 61st ACM/IEEE Design Automation Conference, ACM, San Francisco, pp 1–6. https://doi.org/10.1145/3649329.3658479
51. SignOff (2023) Optical-proximity-correction (OPC). [EB/OL]. https://signoffsemiconductors.com. Accessed 27 Oct 2023
52. Dong X, Zhang L (2017) Analog layout retargeting with process-variation-aware rule-based OPC. In: 2017 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, Baltimore, pp 1–4. https://doi.org/10.1109/ISCAS.2017.8051014
53. Shi W, Yu Z, Jiang J, Che Y, Li S (2022) Computational lithography technology in the context of chip manufacturing. Laser Optoelectron Prog 59(9):0922001. https://doi.org/10.3788/LOP202259.0922001
54. Park JS, Park CH, Rhie SU, Kim YH, Yoo MH, Kong JT et al (2000) An efficient rule-based OPC approach using a DRC tool for 0.18/spl mu/m ASIC. In: Proceedings IEEE 2000 First International Symposium on Quality Electronic Design (Cat. No. PR00525), IEEE, San Jose, pp 81–85. https://doi.org/10.1109/ISQED.2000.838858
55. Shi R, Cai Y, Hong X, Wu W, Yang C (2001) The selection and creation of the rules in rules-based optical proximity correction. In: ASICON 2001. 2001 4th International Conference on ASIC Proceedings (Cat. No. 01TH8549), IEEE, Shanghai, pp 50–53. https://doi.org/10.1109/ICASIC.2001.982496
56. Cai Y, Zhou Q, Hong X, Shi R, Wang Y (2007) Optical proximity correction (OPC) technology and its application. Sci China Ser E 37(12):1607–1619
57. Otto OW, Garofalo JG, Low K, Yuan CM, Henderson RC, Pierrat C et al (1994) Automated optical proximity correction: a rules-based approach. In: Optical/Laser Microlithography VII, vol 2197. SPIE, San Jose, pp 278–293. https://doi.org/10.1117/12.175422
58. Ma X, Zhang SE, Pan YH (2022) Computational lithography research and progress. Laser Optoelectron Prog 59(9):0922008
59. Spence C (2005) Full-chip lithography simulation and design analysis: how OPC is changing IC design. In: Emerging Lithographic Technologies IX, vol 5751. SPIE, San Jose, pp 1–14. https://doi.org/10.1117/12.608020
60. Cobb NB, Granik Y (2002) Model-based OPC using the MEEF matrix. In: 22nd Annual BACUS Symposium on Photomask Technology, vol 4889. SPIE, San Jose, pp 1281–1292. https://doi.org/10.1117/12.467435
61. Chen Y, Wu K, Shi Z, Yan X (2007) A feasible model-based OPC algorithm using Jacobian matrix of intensity distribution functions. In: Optical Microlithography XX, vol 6520. SPIE, San Jose, pp 1506–1515. https://doi.org/10.1117/12.711763
62. Lin B, Yan XL, Shi Z, Yang YW (2011) A sparse matrix model-based optical proximity correction algorithm with model-based mapping between segments and control sites. J Zhejiang Univ Sci C 12:436–442. https://doi.org/10.1631/jzus.C1000219
63. Lei J, Hong L, Lippincott G, Word J (2014) Model-based OPC using the MEEF matrix II. In: Optical Microlithography XXVII, vol 9052. SPIE, San Jose, pp 170–178. https://doi.org/10.1117/12.2046635
64. Kuang J, Chow WK, Young EF (2015) A robust approach for process variation aware mask optimization. In: 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE), IEEE, Grenoble, pp 1591–1594
65. Awad A, Takahashi A, Tanaka S, Kodama C (2014) A fast process variation and pattern fidelity aware mask optimization algorithm. In: 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), ACM, San Jose, pp 238–245
66. Yang YW, Shi Z, Yan XL (2008) Model-based dynamic dissection in OPC. J Semicond 29(7):1422–1427
67. Huang Z, Ni L, Fujii K, Huang X (2019) A model based OPC method to add serifs for corner rounding design of CMOS image sensor. In: Photomask Technology 2019, vol 11148. SPIE, San Jose, pp 219–225. https://doi.org/10.1117/12.2536495
68. Weisbuch F, Naranaya AS (2014) Assessing SEM contour based OPC models quality using rigorous simulation. In: Advances in Patterning Materials and Processes XXXI, vol 9051. SPIE, San Jose, pp 34–49. https://doi.org/10.1117/12.2047826
69. He Z, Zuo Y, Jiang J, Zheng H, Ma Y, Yu B (2023) OpenDRC: An efficient open-source design rule checking engine with hierarchical GPU acceleration. In: 2023 60th ACM/IEEE Design Automation Conference (DAC), IEEE, San Francisco, pp 1–6. https://doi.org/10.1109/DAC56929.2023.10247734
70. Lai GG, Fussell DS, Wong D (1996) Hinted quad trees for VLSI geometry DRC based on efficient searching for neighbors. IEEE Trans Comput Aided Des Integr Circ Syst 15(3):317–324. https://doi.org/10.1109/43.489102
71. Gao JR, Xu X, Yu B, Pan DZ (2014) MOSAIC: mask optimizing solution with process window aware inverse correction. In: Proceedings of the 51st Annual Design Automation Conference, ACM, San Francisco, pp 1–6. https://doi.org/10.1145/2593069.2593163
72. Cecil T, Braam K, Omran A, Poonawala A, Shu J, Vandam C (2020) Establishing fast, practical, full-chip ILT flows using machine learning. In: Optical Microlithography XXXIII, vol 11327. SPIE, San Jose, pp 12–29. https://doi.org/10.1117/12.2551425
73. SemiWiki (2022) Inverse lithography technology – a status update from TSMC. [EB/OL]. [2022-02-06] (2022). https://semiwiki.com. Accessed 06 Feb 2022
74. Pang L (2021) Inverse lithography technology: 30 years from concept to practical, full-chip reality. J Micro/Nanopatterning Mater Metrol 20(3):030901–030901. https://doi.org/10.1117/1.JMM.20.3.030901

75. Yu Z, Chen G, Ma Y, Yu B (2022) A GPU-enabled level-set method for mask optimization. IEEE Trans Comput Aided Des Integr Circ Syst 42(2):594–605
76. Chen G, Yu Z, Liu H, Ma Y, Yu B (2023) DevelSet: deep neural level set for instant mask optimization. IEEE Trans Comput Aided Des Integr Circ Syst 42(12):5020–5033. https://doi.org/10.1109/TCAD.2023.3286262
77. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
78. Matsunawa T, Yu B, Pan DZ (2015) Optical proximity correction with hierarchical Bayes model. In: Optical Microlithography XXVIII, vol 9426. SPIE, San Jose, pp 238–247. https://doi.org/10.1117/12.2085787
79. Jiang B, Zhang H, Yang J, Young EF (2019) A fast machine learning-based mask printability predictor for OPC acceleration. In: Proceedings of the 24th Asia and South Pacific Design Automation Conference, Tokyo, pp 412–419. https://doi.org/10.1145/3287624.3287682
80. Yang H, Li S, Ma Y, Yu B, Young EF (2018) GAN-OPC: mask optimization with lithography-guided generative adversarial nets. In: Proceedings of the 55th Annual Design Automation Conference, ACM, San Francisco, pp 1–6. https://doi.org/10.1145/3195970.3196056
81. Alawieh MB, Lin Y, Zhang Z, Li M, Huang Q, Pan DZ (2019) GAN-SRAF: sub-resolution assist feature generation using conditional generative adversarial networks. In: Proceedings of the 56th Annual Design Automation Conference 2019, ACM, Las Vegas, pp 1–6. https://doi.org/10.1145/3316781.3317832
82. Zheng S, Ma Y, Yu B, Wong M (2024) EMOGen: enhancing mask optimization via pattern generation. In Proceedings of the 61st ACM/IEEE Design Automation Conference 2024, ACM, San Francisco, pp. 1-6. https://doi.org/10.1145/3649329.3655680
83. Zhong W, Hu S, Ma Y, Yang H, Ma X, Yu B (2021) Deep learning-driven simultaneous layout decomposition and mask optimization. IEEE Trans Comput Aided Des Integr Circ Syst 41(3):709–722. https://doi.org/10.1109/TCAD.2021.3061494
84. Yang H, Zhong W, Ma Y, Geng H, Chen R, Chen W et al (2021) VLSI mask optimization: from shallow to deep learning. Integration 77:96–103. https://doi.org/10.1016/j.vlsi.2020.11.001
85. Chen JT, Zhao YY, Zhang Y, Zhu JX, Duan XM (2022) Label-free neural networks-based inverse lithography technology. Opt Express 30(25):45312–45326. https://doi.org/10.1364/OE.472495
86. Yang H, Ren H (2024) ILILT: implicit learning of inverse lithography technologies. arXiv preprint arXiv:2405.03574
87. Chen G, Yang H, Ren H, Yu B, Pan DZ (2024) Differentiable edge-based OPC. arXiv preprint arXiv:2408.08969
88. Xu H, Tang F, Qi P, Yuan Y, Liang H, Huang Z (2024) A self-training end-to-end mask optimization framework based on semantic segmentation network. Integration 96:102150. https://doi.org/10.1016/j.vlsi.2024.102150
89. Chen G, Chen W, Ma Y, Yang H, Yu B (2020) DAMO: deep agile mask optimization for full chip scale. In: Proceedings of the 39th International Conference on Computer-Aided Design, pp 1–9. https://doi.org/10.1145/3400302.3415705
90. Ciou W, Hu T, Tsai YY, Hsuan CT, Yang E, Yang TH, et al (2022) Machine learning optical proximity correction with generative adversarial networks. J Micro/Nanopatterning Mater Metrol 21(4):041606–041606. https://doi.org/10.1117/1.JMM.21.4.041606
91. Zhao W, Yao X, Yu Z, Chen G, Ma Y, Yu B et al (2022) AdaOPC: a self-adaptive mask optimization framework for real design patterns. In: Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design, ACM, San Diego, pp 1–9. https://doi.org/10.1145/3508352.3549468
92. Zhao W, Yao X, Yin S, Bai Y, Yu Z, Ma Y et al (2024) AdaOPC 2.0: enhanced adaptive mask optimization framework for via layers. IEEE Trans Comput Aided Des Integr Circ Syst. 43(9):2674–2686. https://doi.org/10.1109/TCAD.2024.3378600
93. Zhu B, Zheng S, Yu Z, Chen G, Ma Y, Yang F et al (2023) L2O-ILT: learning to optimize inverse lithography techniques. IEEE Trans Comput Aided Des Integr Circ Syst. 43(3):944–955. https://doi.org/10.1109/TCAD.2023.3323164
94. Zheng S, Yu B, Wong M (2023) OpenILT: an open source inverse lithography technique framework. In: 2023 IEEE 15th International Conference on ASIC (ASICON), IEEE, Nanjing, pp 1–4. https://doi.org/10.1109/ASICON58565.2023.10396314
95. Yang H, Li Z, Sastry K, Mukhopadhyay S, Anandkumar A, Khailany B, Singh V, Ren H (2022) Large scale mask optimization via convolutional Fourier neural operator and litho-guided self training. arXiv preprint arXiv:2207.04056
96. Liang X, Ouyang Y, Yang H, Yu B, Ma Y (2023) RL-OPC: mask optimization with deep reinforcement learning. IEEE Trans Comput Aided Des Integr Circ Syst. 43(1):340–351
97. Liang X, Yang H, Liu K, Yu B, Ma Y (2024) CAMO: correlation-aware mask optimization with modulated reinforcement learning. arXiv preprint arXiv:2404.00980
98. Chen G, Yang H, Yu B, Ren H (2024) Intelligent OPC engineer assistant for semiconductor manufacturing. arXiv preprint arXiv:2408.12775
99. Wang XZ, Dai FZ, Li SK (2020) Integrated circuit and lithographic tool. Science Press, Beijing
100. Chen G, Wang Z, Yu B, Pan DZ, Wong MD (2024) Ultra-fast source mask optimization via conditional discrete diffusion. IEEE Trans Comput Aided Des Integr Circ Syst. 43(7):2140–2150. https://doi.org/10.1109/TCAD.2024.3361400
101. Zhang DQ, Chua G, Foong Y (2012) Source mask optimization methodology (SMO) and application to real full chip optical proximity correction. In Optical Microlithography XXV, SPIE, San Jose, 8326: 637-647. https://doi.org/10.1117/12.916614
102. Pei J, Shao F, ElSewefy O (2013) Compatibility of optimized source over design changes in the foundry environment. In Optical Microlithography XXVI, SPIE, San Diego 8683:532–539
103. Shen Y, Peng F, Huang X et al (2019) Adaptive gradient-based source and mask co-optimization with process awareness. Chin Opt Lett 17(12):121102. https://doi.org/10.3788/COL201917.121102
104. Li S, Wang X, Bu Y (2013) Robust pixel-based source and mask optimization for inverse lithography. Opt Laser Technol 45:285–293. https://doi.org/10.1016/j.optlastec.2012.06.033
105. Shen Y, Peng F, Zhang Z (2019) Semi-implicit level set formulation for lithographic source and mask optimization. Opt Express 27(21):29659–29668. https://doi.org/10.1364/OE.27.029659
106. Wang L, Li S, Wang X et al (2017) Source mask projector optimization method of lithography tools based on particle swarm optimization algorithm. Acta Opt Sin 37(10):1022001. https://doi.org/10.3788/AOS201737.1022001
107. Fühner T, Erdmann A (2005) Improved mask and source representations for automatic optimization of lithographic process conditions using a genetic algorithm. Proc. SPIE 5754:41–426. https://doi.org/10.1117/12.599410
108. Chen Y, Lin Y, Dong L, Gai T, Chen R, Su Y et al (2019) SoulNet: ultrafast optical source optimization utilizing generative neural networks for advanced lithography. J Micro/Nanolithography MEMS MOEMS 18(4):043506–043506. https://doi.org/10.1117/1.JMM.18.4.043506
109. Wang Z, Ma X, Chen R, Zhang S, Arce GR (2020) Fast pixelated lithographic source and mask joint optimization based on

compressive sensing. IEEE Trans Comput Imag 6(1):981–992. https://doi.org/10.1109/TCI.2020.3000010
110. Sun Y, Li Y, Liao G, Yuan M, Wei P, Li Y et al (2022) Sampling-based imaging model for fast source and mask optimization in immersion lithography. Appl Opt 61(2):523–531. https://doi.org/10.1364/AO.437655
111. Chen G, He H, Xu P, Geng H, Yu B (2024) Efficient bilevel source mask optimization. In: Proceedings of the 61st ACM/IEEE Design Automation Conference, ACM, San Francisco, pp 1–6. https://doi.org/10.1145/3649329.3656252
112. Zhang Q, Liu J, Sun H, Zhou J, Jin C, Wang J et al (2024) Multi-objective and multi-solution source mask optimization using NSGA-II for more direct process window enhancement. Opt Express 32(4):5301–5322. https://doi.org/10.1364/OE.515546

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Qian Jin received her B.S. degree from Wuhan University of Technology, China, in 2022. She is currently pursuing her Ph.D. degree at the College of Integrated Circuits, Zhejiang University. Her research focuses primarily on the automatic detection and analysis of defects, large language model applications in integrated circuit manufacturing, and computational lithography.



Qingsong Peng received his master's degree in Electronic Science and Technology from Hefei University of Technology, China, in 2024. He is currently pursing a Ph.D. degree at the College of Integrated Circuits, Zhejiang University. His research interests include hardware security and layout pattern classification.



Yumeng Liu received her B.S. degree from Shandong University, China, in 2024. Currently, she is pursuing a master's degree at the College of Integrated Circuits, Zhejiang University. Her research primarily focuses on the generation and detection of wafer defect images.



Xiaotian Qiu received his B.Eng. degree in Information Engineering from South China University of Technology, China, and his M.Eng. degree in Electronic Engineering from the National University of Singapore, Singapore, in 2023. He is currently a Research Assistant at the College of Integrated Circuits, Zhejiang University, China. His research interests include the integration of artificial intelligence in semiconductor manufacturing.



Qi Sun (Member, IEEE) is now a ZJU100 Young Professor at Zhejiang University. He received his doctorate degree from the Dept. of CSE at CUHK in 2022. He then engaged in postdoctoral research at Cornell University. His research areas include ML for EDA, Deep Learning, DNN acceleration, RISC-V SoC, etc. He has been awarded two ICCAD Best Paper Awards, a DATE Best Paper Nomination Award, etc.