Re-examining VLSI Manufacturing and Yield through the Lens of Deep Learning

(Invited Talk)

Mohamed Baker Alawieh, Wei Ye, and David Z. Pan

ECE Department, University of Texas at Austin

***Abstract*—The continued scaling of integrated circuit technologies, along with the increased design complexity, has exacerbated the challenges associated with manufacturability and yield. In today's semiconductor manufacturing, lithography plays a fundamental role in printing design patterns on silicon. However, the growing complexity and variation of the manufacturing process have tremendously increased the lithography modeling and simulation cost. Besides, both the role and cost of resolution enhancement techniques (RETs) – now indispensable in the design process – have increased. Parallel to these developments are the recent advancements in Machine Learning (ML) which have provided a far-reaching data-driven perspective for problem solving. In this work, we shed light on the recent Deep Learning (DL) based approaches that have provided a new lens to examine traditional manufacturability and yield challenges. We present lithography modeling and simulation techniques, leveraging advanced learning paradigms, which have demonstrated unprecedented efficiency. Moreover, we demonstrate the role DL can play in advancing RETs by presenting its successful application in assist feature generation. Also critical to yield is the post fabrication wafer map defect analysis step which our work tackles using a novel confidence-aware deep learning scheme. This paper further discusses the future prospects of DL-based approaches in the scope of circuits manufacturability and yield.**

I. Introduction

As the integrated circuits (IC) technologies continues to scale deep into the sub-micron region, the gap between design expectation and manufacturing capability continues to widen [1]. Hence, the challenges associated with retaining the robustness of state-of-the-art designs continue to exacerbate.

Of particular significance in this regard is the role lithography plays in printing design patterns on silicon [1]. However, the growing complexity and variation of the manufacturing process have tremendously increased the lithography modeling and simulation cost. This has also results in the increase of both the role and cost of resolution enhancement techniques (RETs) which have become indispensable in the design process. These challenges in the manufacturing process have imposed a heavy burden on the post fabrication defect analysis which is a critical step for yield improvement.

In practice, the drive for extreme scaling and advanced chip complexities has been driven in part by the immense computational demand in today's applications, and machine learning (ML) frameworks are definitely topping the list of such applications. Recent advances in machine learning in general, and deep learning in particular, have dramatically altered the perception of computing through providing a far-reaching data-driven perspective for problem solving. Thus, experts in all fields of study have been re-examining, through the new lens of machine learning, different problems that traditional computing paradigms were ill-equipped to handle. In fact, with new successes and adoption in many domains, deep learning has been rapidly infiltrating into diverse fields [2]–[5].

Despite the fact that DL has become among the applications imposing the newest challenges on IC manufacturing process, recent research has shown that this process has as well taken its fair share from the prosperous DL revolution. ML has recently attracted a lot of attention in the IC manufacturing community, a fact that is clearly reflected by the numerous applications of DL recently proposed to address challenges in the field. These applications span different tasks and leverage every branch of ML towards advancing the field [2], [6], [7].

In this paper, we present our recent applications of DL for VLSI manufacturability and yield. At the lithography level, and to bypass the cost-intensive and time-consuming experimental verification, the semiconductor industry has relied on lithography simulation for process development and performance verification [8], [9]. However, the steady decrease of the feature sizes along with the growing complexity and variation of the manufacturing process have tremendously increased the lithography modeling complexity and prolonged the already-slow simulation procedure. Considering the fact that machine learning based approaches have demonstrated superior efficacy in a particular stage during lithography modeling [10], [11], we have recently proposed LithoGAN [12] as a novel end-to-end lithography modeling framework based on conditional generative adversarial network (CGAN) that has demonstrated tremendous success in computer vision over the past few years [13]–[17].

LithoGAN has demonstrated impressive efficiency for lithography modeling considering a thin mask model for topography effects. This scope has been further extended with our TEMPO framework [18] that was proposed as a novel thick mask effect modeling framework using a single, one-fits-all model capable of predicting aerial image intensity at different resist heights.

Moreover, the stringent lithography criteria have made the RET techniques indispensable, yet more challenging and computationally expensive. Here also, DL can be used to advance RETs by presenting its successful application in sub-resolution assist feature (SRAF) generation which is a key RET adopted to improve the target pattern quality and lithographic process window. In our recent work [19], a novel formulation for SRAF generation was proposed to cast the problem as a domain transfer task. Then, GAN-SRAF

ICCAD '20, November 2–5, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8026-3/20/11…$15.00
https://doi.org/10.1145/3400302.3415779

was proposed for efficient SRAF insertion leveraging recent advancements in generative adversarial learning.

At the post fabrication stage, wafer defect detection is an instrumental step in yield analysis where missed defects can significantly affect the production yield. Once again, machine learning has been proposed to address this challenge. In our work, we tackle this challenge with a new focus on trust-awareness predictions by proposing a deep selective learning framework for wafer defect detection [20]. The key idea is to equip the model with an integrated reject option which can be leveraged to reduce the misclassification risk for the model, new defect detection, data change detection, and resource allocation.

The details of the aforementioned deep learning frameworks for VLSI manufacturability and yield will be covered in this paper along with future directions. We first review the necessary background about machine learning models and applications in Section II. Next, we present our contributions at the lithography level, LithoGAN and TEMPO, in Section III and then present our SRAF generation scheme, GAN-SRAF, in Section IV. Section V covers the detail of our trust-aware model for wafer map defect detection. In Section VI, we reflect on the impact of these applications, future works, and conclusions.

II. Background

In this section, we review the background of the state-of-the-art learning models that have been recently adopted to address challenges in manufacturability and yield. We first review the current status of machine learning applications in electronic design automation (EDA), and then present the necessary background for some machine learning models used in our work.

A. Machine Learning in EDA

In addition to its applications in VLSI manufacturability and yield, which is the main scope of this work, ML has been infiltrating into almost all stages of the VLSI design cycle. At the circuit synthesis level, ML has been used for circuit modeling and optimization especially for analog and mixed signal designs [21]–[23].

It was also adopted in physical design for better design quality and faster convergence [24]. For placement, ML has been used for datapath placement [25], and routability optimization [26]–[28]. Besides, the ML-inspired placement framework, DREAMPlace [29], has demonstrated unprecedented acceleration for the placement process. Moreover, ML has been used for routing guidance [30] and routing violation detection [28].

This wide range of applications of ML in EDA demonstrates the key role ML is playing in the field through providing a new perspective to re-examine traditional problems in VLSI design. In this paper, we focus on applications of DL in VLSI manufacturability and yield which have demonstrated impressive results in terms of efficiency and performance.

B. Generative Adversarial Networks

Generative adversarial networks have been adopted in a wide range of applications in VLSI design. In this section, we review the necessary background for these models that we will build upon in the following sections of this paper.

Generative adversarial networks (GANs) were proposed as generative models that learn a mapping from a random noise vector z to an output y, $G : z \rightarrow y$ [13]. The architecture of a GAN is composed of two main components: the *generator* and the *discriminator*. The generator G is trained to produce samples based on an input noise vector z that cannot be distinguished from "real" images by an adversarially trained discriminator, D, which is trained to do as well as possible at detecting the generator "fakes" [13].

The conventional generator in a GAN is basically an encoder-decoder scheme similar to that in an AE where the input is passed through a series of layers that progressively downsamples it (i.e, encoding), until a bottleneck layer, at which point the process is reversed (i.e, decoding) [13], [31], [32]. On the other hand, the discriminator is a convolutional neural network whose objective is to classify 'fake' and 'real' images. Hence, its structure differs from that of the generator and resembles a typical two-class classification network [13], [31], [32]. This adversarial scheme is represented in the objective function given as:

$$\min_G \max_D \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))], \quad (1)$$

where $D(\bullet)$ represents the probability of a sample being real; i.e., not generated by G.

After training, the generator part of the GAN is used to generate new samples using random noise vectors while the discriminator is discarded as it is only needed for the training process [13], [31], [32].

In literature, different versions of GANs, tailored towards specific domain and applications, were proposed especially for image related tasks. Among these are the CGANs which, in contrast with original GANs, learn a mapping from an observed image x and random noise vector z, to y, $G : \{x, z\} \rightarrow y$. Technically, CGANs have changed the objective from a pure generative one to a domain-transfer task capable of establishing a mapping between images in different domains. Its applications span different domains ranging form image coloring to aerial to map, edge to photo translations, and medical applications among others [12], [30], [33]–[36]

As an example, the architecture of the CGAN used in the LithoGAN framework is shown in fig. 1 where G translates an image from the layout domain to the resist shape domain, and D examines image pairs to detect fake ones (further details about this application are presented in III-A). Mathematically, one form of a loss function used for training the CGAN can be given as [15], [31]:

$$\begin{aligned}\mathcal{L}_{CGAN}(G, D) = &\mathbb{E}_{x,y}[\log D(x, y)] \\ &+ \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \\ &+ \lambda \mathbb{E}_{x,z,y} l(y, G(x, z)),\end{aligned} \quad (2)$$

where x is a sample in the input domain and y is its corresponding sample in the output domain. Comparing equations (1) and (2), one can notice the addition of the loss term which penalizes the difference between the generated sample $G(x, z)$ and its corresponding golden reference y. Different loss functions are adopted in different CGAN models including ℓ_1-norm and ℓ_2-norm.

III. Lithography Modeling and Simulation

Lithography is one of the key stages in VLSI manufacturing. In fact, it was among the first venues were machine learning has been introduced in the VLSI flow with machine learning based hotspot detection techniques [37]–[41]. However, with recent advancement in DL research, new

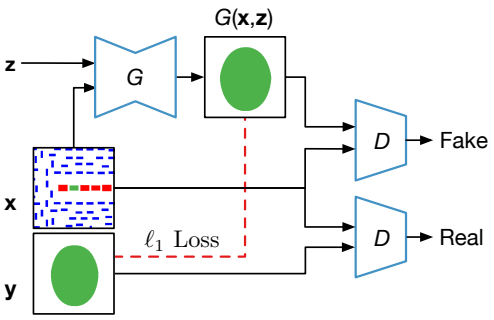


Fig. 1 CGAN for lithography modeling [12]

applications for DL in lithography have been recently studied. In particular, lithography modeling and simulation are of particular importance due to their exorbitant computational cost. In this section, we will introduce how deep learning can help accelerate lithography steps and facilitate design closure.

A. LithoGAN

During the lithography process, a designed mask pattern is transferred into a resist pattern on the top surface of a semiconductor wafer [1], [42]. The semiconductor industry has relied on lithography simulation for process development and performance verification. Rigorous lithography simulation precisely simulates the physical effects of materials but is computationally expensive. Therefore, compact models stand as a speedup alternative to rigorous computation with a small sacrifice in accuracy, which enables its wide application for lithography verification.

Fig. 5 shows a typical flow of lithography simulation. First, an aerial image is calculated from a mask pattern using a compact optical model. Then a resist model is used to determine the locally varying slicing thresholds [43]. The thresholds are processed through extrapolation together with the corresponding aerial image to evaluate the critical dimension (CD) of the printed patterns.

Machine learning-based techniques have been proposed as a substitute for compact models for better simulation quality [2], [10], [11], [44]. [44] proposed an artificial neural network (ANN) for resist height prediction. [10] proposed a convolutional neural network (CNN) model that predicts the slicing thresholds in aerial images accurately. Recently, [11] proposed a transfer learning model to cope with the deficiency in the manufacturing data at advanced technology nodes.

These machine learning-based resist modeling techniques still suffer from an exorbitant computational cost while providing partial modeling schemes that rely heavily on pre- and post-processing procedures. For this purpose, we propose an end-to-end lithography modeling framework, LithoGAN, to directly map mask patterns to resist patterns by utilizing CGAN [12]. The input domain is the mask designs converted to red-green-blue (RGB) images, where the target contact of interest is encoded into the green channel, neighboring

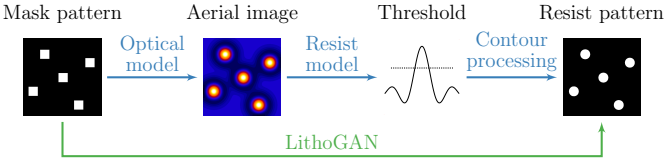


Fig. 2 Conventional lithography simulation flow consisting of multiple stages and the proposed LithoGAN flow [12]

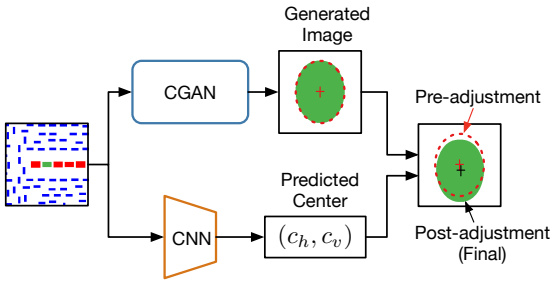


Fig. 3 LithoGAN framework [12].

contacts are encoded into the red channel, and SRAFs are encoded into the blue channel as shown in fig. 3. The output of CGAN is the zoomed-in resist patterns corresponding to the center contact.

For traditional computer vision tasks, the locations of the objects in the generated image are not a major concern. For example, when trained on car images, the output of the GAN is judged upon based on the quality of an image as seen by a human while neglecting the exact location of the car in the image. However, for the lithography modeling task, the location of the generated resist pattern is as important as the shape of the pattern.

Therefore, as illustrated in fig. 3, there are two data paths in LithoGAN, where the shape and the location of the resist pattern are predicted separately. In the first path, a trained CGAN model is utilized to predict the shape of the resist pattern. During training, the golden pattern is re-centered at the center of the image, and the coordinates of the original center are saved for CNN training. In other words, the model is trained to predict resist patterns that are always centered at the center of the images. On the other hand, the second path is composed of a convolutional neural network (CNN) trained to predict the center of the resist pattern based on the mask image. Here the center refers to the center of the bounding box enclosing the resist pattern. They are combined in the last step before output: the image generated by CGAN is adjusted by recentering the resist shape based on the coordinates predicted from the CNN.

Our experimental results reported in [12] demonstrates that LithoGAN can achieve $\sim$1800$\times$ runtime reduction when compared to rigorous simulation, while obtaining resist pattern results that fall within the accepted lithography range. Sample results are shown in fig. 4 where the resist patterns are accurately predicted.

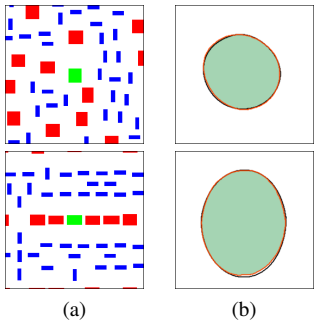


Fig. 4 (a) Mask pattern input and (c) LithoGAN output. Each row represents one clip example. The golden contour is outlined in black. The prediction pattern is filled with green and outlined in red.

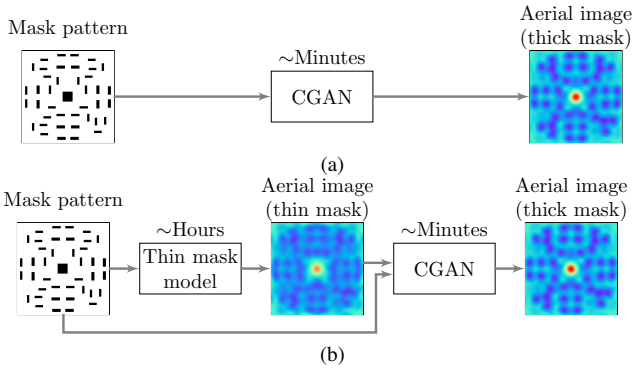


Fig. 5 (a) Proposed Scheme 1 for high efficiency and (b) Scheme 2 for high accuracy in TEMPO.

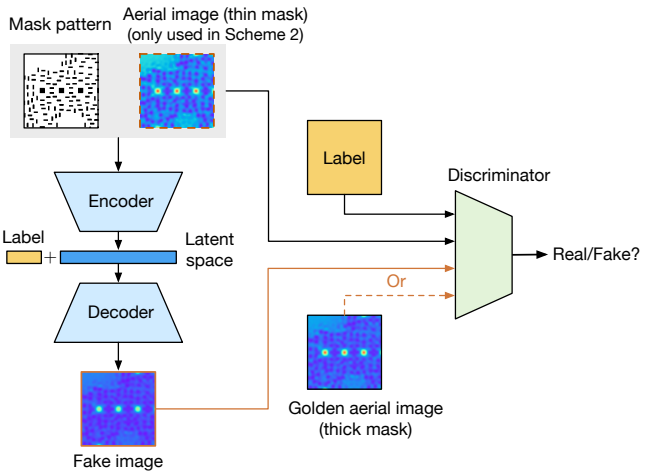


Fig. 6 Overview of the TEMPO model.

B. TEMPO

The continuous device scaling has posed the mask topography effects among the major challenges in lithography modeling. When the feature sizes start to be comparable to the wavelength, the widely-used thin mask approximation is no longer adequate with the increasingly pronounced impacts of thick mask effects on the lithography imaging [45]–[47]. The failure to consider mask topography effects in lithography modeling could lead to critical dimension (CD) error and focus shift, resulting in the shrinkage of process window and the decrease of the image quality and the process robustness.

In a rigorous thick mask simulation flow, the simulator takes as input a mask pattern and generates the corresponding aerial image. While such an approach is the common practice today, its inordinate runtime hinders its application in the early stages of the process development and mask optimizations. Recently, advances in machine learning have been leveraged to devise new mask modeling techniques [48], [49]. These machine learning approaches rely on conventional modeling techniques that require intensive feature engineering and depend heavily on post-processing methods which affect the model accuracy.

In [18], we propose TEMPO as a fast modeling framework that can significantly speed up the thick mask modeling. In practice, TEMPO provides in one of its schemes a CGAN model capable of mimicking the rigorous simulation process with orders of magnitude speedup as shown in Figure 5a. For applications with high accuracy requirements, TEMPO provides an alternative framework, namely Scheme 2 shown in Figure 5b. As a first step, TEMPO in Scheme 2 runs a fast thin mask model to generate aerial images assuming no mask topography effect, and the output aerial image is used along with the mask pattern as the input to the CGAN model. In this way, the aerial image given by the thin mask model provides the CGAN model with additional information not present in the mask pattern image, and hence improves its accuracy.

In lithography simulation, an accurate 3D view of aerial images at different resist heights is crucial to evaluate cross-section views of the resist pattern. Image translation using CGAN was proposed as a means for domain transfer between two distinct domains. However, aerial image generation requires domain transfer from the single mask pattern domain to multiple resist height domains. For aerial image generation and other similar tasks, the most straightforward option is to train multiple domain-to-domain models. So, for m target domains, m such models are needed. Clearly, the approach of building individual models has multiple drawbacks. Most evident is the size of the model that scales with the number of target domains. Besides, when assuming that different target domains are independent, an opportunity for information sharing between those slightly different tasks for different resist heights is missed. In fact, there usually exist global features that can be learned from images of all domains, especially for the case of 3D aerial image generation, where differences in target domains are only slight shifts in image intensity. Without information sharing, each generator cannot fully utilize the entire training data and can only learn from mask and one target domain out of m available.

We propose within TEMPO a new one-fits-all model where a one-hot encoding vector of length m carrying the target domain information is appended to the latent space representation in the bottleneck layer, as shown in Figure 6. This way, the information is appended at a critical location in the generator where it can guide the output image generation while having a compact representation. For the design of discriminator, compared to that used in StarGAN [50] where each one extra input channel is needed for each domain for multi-domain translation, the encoding scheme in TEMPO requires only a single channel for all the domains. This can significantly improve the scalability of TEMPO when faced with a significant increase in the number of target domains.

Compared to other models with the same objective, our proposed model is the most compact given the novel target domain encoding used in both the generator and discriminator networks. Compared to having individual bi-domain models, our model is better positioned to learn global features effectively through across-domain information sharing. Besides, success to fully utilize training data in TEMPO enhances the quality of generated images and introduces a lower risk of overfitting. Hence, with the information-sharing scheme in TEMPO, a significant accuracy improvement is achieved.

Our results presented in [18] shows that TEMPO gives smaller root mean square error (RMSE) and CD errors when compared with the baseline with multiple individual GAN models, which further demonstrates the advantages of our one-fits-all approach. Besides, the two schemes in TEMPO obtain $1170\times$ and $27\times$ speedup when compared with rigorous simulation while achieving satisfactory performance in aerial image quality and critical dimension fidelity.

IV. GAN-SRAF

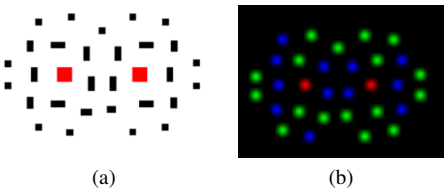
Sub-resolution assist feature generation is a key RET to improve the target pattern quality and lithographic process window. These assist features are not actually printed; instead, the SRAF patterns would deliver light to the positions of target patterns at proper phase which can improve the robustness of target printing to lithographic variations [51].

In literature, different SRAF generation approaches have been proposed and adopted. On one hand, there are rule-based approaches that can achieve acceptable accuracy within short execution time for simple designs and regular target patterns; yet fall short of handling complex shapes [52], [53]. On the other hand, model-based SRAF generation methods have been proposed relying on either simulated aerial images to seed the SRAF generation [54], [55], or inverse lithography technology (ILT) to compute the image contour and guide the SRAF generation [56]. Despite better lithographic performance compared to the rule-based approach, the model-based SRAF generation is very time-consuming [51].

Xu et al [51] introduced machine learning to tackle the problem of SRAF insertion more efficiently [2], [51]. The proposed method relies on SRAF features extraction with local sampling scheme to obtain the optimal SRAF map. This approach has achieved 10x speedup compared to model-based approaches with comparable quality [51].

Although this approach has demonstrated significant speedup compared to model-based approaches while achieving comparable results in terms of process variation band, there is still significant room for improvement by leveraging recent advancement in the field of image processing in computer vision [15], [31]. Recently, we proposed two GAN schemes to address the SRAF generation task [19]. In the first, we propose to use CGAN for SRAF generation by casting the problem into an image translation task where the two images domains are: (i) original layout and (ii) layout with SRAFs. Hence, generating an SRAF scheme for a particular layout can be seen as translating the layout image from the first domain (i.e., original layout) to the second domain (i.e., layout with SRAFs). Towards this goal, a set of paired images (i.e. original layout with no SRAF paired with the corresponding layout after SRAF insertion) is provided for the network to learn the desired translation.

While this approach is adequate for cases where paired data is available, we also propose an alternative GAN scheme that handles the case where data is available but is not necessarily paired. In practice, the availability of adequate training datasets is one of the major challenges facing machine learning models. Therefore, and knowing that paired data may not be available especially at the early stages in IC technology nodes, we propose an SRAF insertion scheme featuring an unpaired image-to-image translation. Our proposed model, inspired by the Cycle Generative Adversarial architecture (CyGAN) [17] learns simultaneously a two-way image translation using unpaired data. Unlike the CGAN scheme where paired data is used to learn a one-way translation, CyGAN - as the name implies - uses a cycle scheme to learn two-way translation using unpaired data. In practice, two parallel translation models are trained where the objective is to reconstruct an image after undergoing two translations: (i) from native domain to the other domain, then (ii) back to the native domain. Such reconstruction will be accurate for both domains when both translation tasks are accurate.



(a) (b)

Fig. 7 Multi-channel heatmaps encoding process where (a) shows an original layout representation and (b) shows the encoded representation [35].

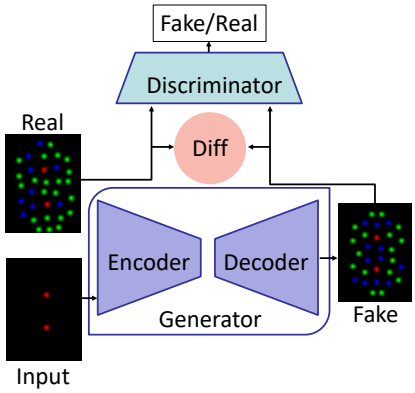


Fig. 8 CGAN-based GAN-SRAF flow [19].

Hence, this learning scheme uses a cycle translation to learn the mapping using unpaired images.

Both of the proposed GAN schemes require casting the layout information into image format. However, direct image representation of layout is not suitable for the SRAF generation using GANs due to two major limitations. First, GANs exhibit inherent limitation in detecting sharp edges and are not guaranteed to generate 'clean' rectangular shapes for the SRAFs [57]. In addition, extracting the SRAF information from the image to be mapped back to the layout file can be prohibitively expensive. Hence, a special encoding scheme, typically used in keypoint estimation [58], [59], is proposed in [35] to overcome the aforementioned limitations. The proposed scheme is based on multi-channel heatmaps which associates each object type with one channel in the image [59], [60]. An example of such encoding is shown in fig. 7 where an original layout is shown in fig. 7a and the multi-channel heatmap representation is shown in fig. 7b. In this example, the number of channels is set to 3 to visualize the encoded representation through an RGB image: (i) target patterns (in red), (ii) horizontal SRAFs (in green) and (ii) vertical SRAFs (in blue).

This encoding has two main advantages: (i) no sharp edges in the image representation, and (ii) images generated by the GAN models can be easily mapped back to layout files using a fast custom CUDA accelerator for the decoding scheme [35]. The CGAN model and CyGAN model used in GAN-SRAF are shown in fig. 8 and fig. 9. In fig. 9, red and green arrows represent the two arcs in the cycle translation with each arc representing a translation from one domain to another.

Our results presented in [19] show that GAN-SRAF can achieve 14× reduction in runtime compared to the work in [51] (LS_SVM) and 144× when compared to model based (MB) approaches while achieving comparable results. A summary of the results is shown in table I where process variation (PV) band and edge placement error (EPE) results

	No SRAF	MB	LS_SVM	CGAN	CyGAN
PV band (*0.001 um^2)	3.354	2.845	3.009	2.916	**2.773**
PV band ratio	1	0.848	0.897	0.869	**0.827**
EPE (nm)	3.9287	0.5270	**0.5067**	0.5410	0.5721
EPE ratio	1	0.134	**0.129**	0.138	0.146
Runtime (sec)	-	6910	700	**48**	**45**

TABLE I The comparison of evaluation metrics and run time across different SRAF generation schemes is shown.

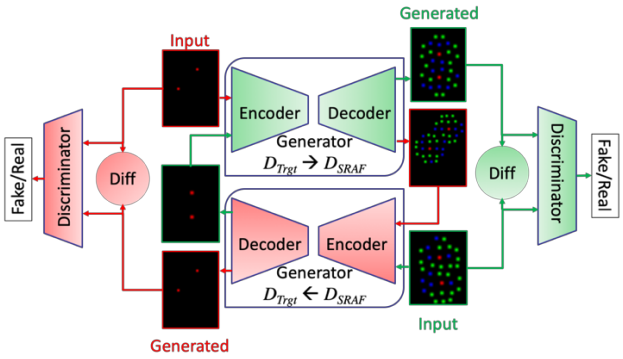


Fig. 9 CyGAN-based GAN-SRAF flow [19].

are reported in addition to the runtime.

V. Wafer Map Defect Classification

A critical first step towards improving yield during the IC design cycle is to identify the underlying factors that contribute most to yield loss, and for that, wafer map analysis is a key. Traditionally, wafer inspection was performed by experienced engineers who can identify the failure cause based on the wafer defect pattern. However, such process is tedious and an automated alternative is desired [61].

Machine learning techniques have been recently proposed to tackle the job using both unsupervised and supervised learning paradigms [62]–[64]. With unsupervised learning, clusters of wafer maps are constructed, and experienced engineers then label each of them with its defect pattern [62], [63]. On the other hand, supervised learning techniques rely on features extracted from the wafer maps to build a classification model that is capable of classifying new wafer maps based on their defect type [61], [65]. The aforementioned approaches rely on a set of features to capture the spatial properties of wafers. However, these wafers can be instinctively perceived as images with defect patterns being spatial features of these images. Hence, the native spatial characteristics of the defect patterns can be best preserved by using the natural representation of wafers as images.

In [20], we proposed a novel framework for wafer map defect pattern classification using deep selective learning. Beside achieving superior accuracy compared to conventional approaches by leveraging the intrinsic image representation of a wafer, our proposed approach exhibits unique features that are tailored to address two challenges accompanying the task. One major challenge arises from the fact that some wafers may exhibit new defect patterns that were not previously seen by the model during training. In such a case, the model is expected to give a wrong label which can mask a new type of defects. Moreover, some wafer maps may exhibit more than one defect pattern which can overwhelm the classification model. To handle these cases, we propose using a convolutional neural network with an integrated *reject option* [66],

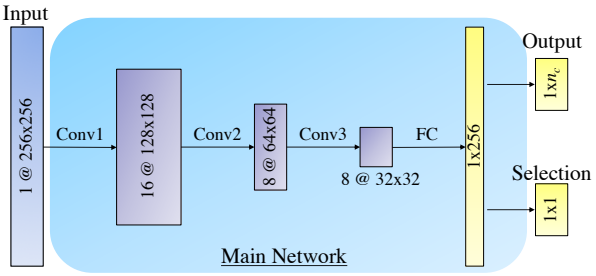


Fig. 10 The CNN network architecture showing both the prediction and selection heads.

[67]. In other words, given a user set compromise between risk and coverage during training, the model is trained to optimize for classification and rejection simultaneously. With this option, the model can choose to discard predictions with high risk of misclassification; i.e, the model abstains from prediction for some samples to maintain a low risk level. Clearly, the reject option can further improve the accuracy of the model on the selected samples.

Compared to a traditional image classification network, the proposed network architecture includes two output heads. The first is the prediction head implementing the main classification function f, while the other is a selection head consisting of a single neuron with a sigmoid activation to implement the selection function g. These two heads depart at the end of the network architecture after the main blocks including convolutional fully connected layers as shown in Fig. 10.

The new objective of the training process is to minimize the selective loss while meeting the coverage constraints. This can be expressed mathematically as new loss function [66]:

$$\mathcal{L}_{(f,g)} = r(f,g) + \lambda\Psi(c_0 - c(g)) \quad \text{where} \quad \Psi(z) = \max(0,z)^2. \tag{3}$$

Here, $r(f,g)$ is the classification loss computed on selected samples only, c(g) is the average coverage, c_0 is the user set target coverage, λ is hyper-parameter reflecting the importance of the coverage constraint, and Ψ is a quadratic penalty function.

The second challenge we address in this work is class imbalance. Defect classes have different frequencies of occurrence which typically result in an imbalanced training process where some minority classes are dominated by other majority ones. In this work, we propose using data augmentation to generate synthetic samples from the under-represented classes. In particular, we train a convolutional auto-encoder to generate samples from the distribution of the target class and use synthetic samples alongside the original ones in the training process [68].

Our experimental results, using the the WM-811k industrial wafer dataset [69] with 9 classes, shown in 11, have demonstrated that our approach can achieve superior accuracy when compared to conventional approach with 99% accuracy under selective learning framework and 94% under full coverage setting. Besides, The proposed selective learning scheme for wafer map defect detection has many advantages on the application side. We list here three of these applications: (i) detection of new defect class(es), (ii) resource allocation for human in the loop setup, and (iii) detection of changes in the data distribution.

One major advantage of the selective learning scheme is

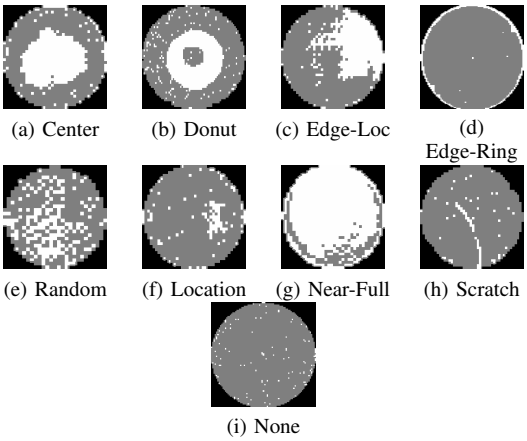


Fig. 11 Sample wafer examples for different pattern types: (a) Center, (b) Donut, (c) Edge-Location, (d) Edge-Ring, (e) Random, (f) Near-Full and (g) Scratch.

that it allows detecting a new defect class when it shows up. Intuitively, if a new defect occurs, the model should abstain from labeling the new defect samples because they are associated with high risk. To validate this utility, we set an experiment where one class was excluded from our training process and all its samples were used during testing. This is done to test whether selective learning will label the samples from the unseen class. Results presented in [20] show that, with selective learning, the model abstained from predicting a label for all samples belonging to the new class.

Another application is for resource allocation. Such a model is developed to reduce the cost associated with having experienced engineers manually label the wafer. Selective learning provides a perfect allocation of resources as the model is predicting with high confidence and the high risk samples, which are typically the most interesting for the engineers, are automatically detected and passed for examination. Finally, the proposed scheme helps in detecting concept shifts or major changes in the distribution of the data.

VI. Discussion & Conclusion

As shown in the aforementioned applications, the adoption of recent advancement of deep learning in VLSI design flow has revolutionized the problem solving approaches replacing many conventional schemes with data-driven solutions. The aforementioned applications of deep learning in the field of manufacturability and yield represent a part of a new paradigm in the field based on machine learning with the driving motivation being enhancing performance and speeding up design closure.

In practice, machine leaning models are now generating solutions that can parallel those of conventional tools, and are doing that much faster. With LithoGAN and TEMPO's unprecedented speedup of more than 1000×, a major disruptive approach is being introduced to the lithography simulation field. Similar improvements have been also seen with remarkable SRAF insertion speedup.

On the other side, trust in machine learning models has always been a major concern facing the wide scale adoption of such models in industry. The fact that this issue is now attracting research attention is a good indication about the progress being made to adopt such solutions. With our selective learning scheme, we have proposed a trust-aware model that can address the trust concern. In fact, a hybrid system of machine and human intelligence, such as the one used in our selective learning scheme, can be the best option to move forward. With machine intelligence doing the bulk of the time-consuming work human efforts can be tailored towards better usage while still supervising the machine's performance and interfering when needed.

In this paper, recent deep learning applications in the field of manufacturability and yield are presented. While conventional approaches to address these tasks are data intensive and computationally expensive, machine learning is emerging as an alternative framework that can substitute them while improving performance and/or efficiency, eventually contributing to fast design closure and good manufacturability.

With the proven success in many applications, wider adoption of these models still faces some challenges which relates to data scarcity, confidence and trust, and high resolution/accuracy requirements in many applications. With more success at this front in the near future, wider adoption of machine learning in design for manufacturability in particular is expected.

VII. Acknowledgement

This work is supported in part by NSF under Award No. 1718570 and Kioxia Corporation.

References

[1] C. Mack, *Fundamental Principles of Optical Lithography: The Science of Microfabrication*. John Wiley & Sons, 2008.

[2] Y. Lin, M. B. Alawieh, W. Ye, and D. Pan, "Machine learning for yield learning and optimization," in *Proc. ITC*, 2018.

[3] W. Shi, M. B. Alawieh, X. Li, and H. Yu, "Algorithm and hardware implementation for visual perception system in autonomous vehicle: A survey," *Integration*, vol. 59, pp. 148–156, 2017.

[4] H. Yu, W. Shi, M. B. Alawieh, C. Yan, X. Zeng, X. Li, and H. Yu, "Efficient statistical validation of autonomous driving systems," in *Safe, Autonomous and Intelligent Vehicles*. Springer, 2019, pp. 5–32.

[5] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[6] I. A. M. Elfadel, D. S. Boning, and X. Li, "Machine learning in VLSI computer-aided design," 2018.

[7] M. B. Alawieh, Y. Lin, W. Ye, and D. Z. Pan, "Generative learning in VLSI Design for Manufacturability: Current status and future directions," *Journal of Microelectronic Manufacturing*, vol. 2, no. 4, 2019.

[8] C. A. Mack, *Field Guide to Optical Lithography*. SPIE Press Bellingham, 2006, vol. 6.

[9] X. Ma and G. R. Arce, *Computational lithography*. John Wiley & Sons, 2011, vol. 77.

[10] T. M. S. N. Yuki Watanabe, Taiki Kimura, "Accurate lithography simulation model based on convolutional neural networks," in *Proc. SPIE*, vol. 10147, 2017.

[11] Y. Lin, M. Li, Y. Watanabe, T. Kimura, T. Matsunawa, S. Nojima, and D. Z. Pan, "Data efficient lithography modeling with transfer learning and active data selection," *IEEE TCAD*, 2018.

[12] W. Ye, M. B. Alawieh, Y. Lin, and D. Z. Pan, "LithoGAN: End-to-end lithography modeling with generative adversarial networks," in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 107.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[14] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, vol. 2, no. 3, 2017, p. 4.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[18] W. Ye, M. B. Alawieh, Y. Watanabe, S. Nojima, Y. Lin, and D. Z. Pan, "TEMPO: Fast mask topography effect modeling with deep learning." in *Proc. ISPD*, 2020, pp. 127–134.

[19] M. B. Alawieh, Y. Lin, Z. Zhang, M. Li, Q. Huang, and D. Z. Pan, "Gan-sraf: Sub-resolution assist feature generation using generative adversarial networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2020.

[20] M. B. Alawieh, D. Boning, and D. Z. Pan, "Wafer map defect patterns classification using deep selective learning," in *Proc. DAC*, 2020.

[21] M. B. Alawieh, F. Wang, and X. Li, "Efficient hierarchical performance modeling for analog and mixed-signal circuits via bayesian co-learning," *IEEE TCAD*, pp. 1–13, 2018.

[22] F. Wang, P. Cachecho, W. Zhang, S. Sun, X. Li, R. Kanj, and C. Gu, "Bayesian model fusion: large-scale performance modeling of analog and mixed-signal circuits by reusing early-stage data," *IEEE TCAD*, vol. 35, no. 8, pp. 1255–1268, 2015.

[23] B. Xu, K. Zhu, M. Liu, Y. Lin, S. Li, X. Tang, N. Sun, and D. Z. Pan, "Magical: Toward fully automated analog ic layout leveraging human and machine intelligence," in *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2019, pp. 1–8.

[24] A. B. Kahng, "Machine learning applications in physical design: Recent results and directions," in *Proceedings of the 2018 International Symposium on Physical Design*, 2018, pp. 68–73.

[25] S. Ward, D. Ding, and D. Z. Pan, "PADE: a high-performance placer with automatic datapath extraction and evaluation through high dimensional data learning," in *Proc. DAC*, 2012, pp. 756–761.

[26] W.-T. J. Chan, P.-H. Ho, A. B. Kahng, and P. Saxena, "Routability optimization for industrial designs at sub-14nm process nodes using machine learning," in *Proceedings of the 2017 ACM on International Symposium on Physical Design*, 2017, pp. 15–21.

[27] M. B. Alawieh, W. Li, Y. Lin, L. Singhal, M. Iyer, and D. Z. Pan, "High-definition routing congestion prediction for large-scale FPGAs," in *ASPDAC*, 2020.

[28] Z. Xie, Y.-H. Huang, G.-Q. Fang, H. Ren, S.-Y. Fang, Y. Chen, and J. Hu, "Routenet: Routability prediction for mixed-size designs using convolutional neural network," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.

[29] Y. Lin, Z. Jiang, J. Gu, W. Li, S. Dhar, H. Ren, B. Khailany, and D. Z. Pan, "Dreamplace: Deep learning toolkit-enabled gpu acceleration for modern vlsi placement," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2020.

[30] K. Zhu, M. Liu, Y. Lin, B. Xu, S. Li, X. Tang, N. Sun, and D. Z. Pan, "GeniusRoute: A new analog routing paradigm using generative neural network guidance," in *Proc. ICCAD*, 2019.

[31] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[32] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*.

[33] C. Yu and Z. Zhang, "Painting on placement: Forecasting routing congestion using conditional generative adversarial nets," in *DAC*, 2019.

[34] B. Xu, Y. Lin, X. Tang, S. Li, L. Shen, N. Sun, and D. Z. Pan, "Wellgan: Generative-adversarial-network-guided well generation for analog/mixed-signal circuit layout," in *Proceedings of the 56th Annual Design Automation Conference 2019*. ACM, 2019, p. 66.

[35] M. B. Alawieh, Y. Lin, Z. Zhang, M. Li, Q. Huang, and D. Z. Pan, "GAN-SRAF: Sub-resolution assist feature generation using conditional generativeadversarial networks," in *Proc. DAC*, 2019.

[36] H. Yang, S. Li, Y. Ma, B. Yu, and E. F. Young, "Gan-opc: mask optimization with lithography-guided generative adversarial nets," in *Proceedings of the 55th Annual Design Automation Conference*. ACM, 2018, p. 131.

[37] D. Ding, X. Wu, J. Ghosh, and D. Z. Pan, "Machine learning based lithographic hotspot detection with critical-feature extraction and classification," in *Proc. ICICDT*, 2009, pp. 219–222.

[38] W. Ye, Y. Lin, M. Li, Q. Liu, and D. Z. Pan, "LithoROC: Lithography hotspot detection with explicit roc optimization," in *Proceedings of the 24th Asia and South Pacific Design Automation Conference*. ACM, 2019, pp. 292–298.

[39] W. Ye, M. B. Alawieh, M. Li, Y. Lin, and D. Z. Pan, "Litho-GPA: Gaussian process assurance for lithography hotspot detection," in *Proc. DATE*, 2019.

[40] H. Yang, L. Luo, J. Su, C. Lin, and B. Yu, "Imbalance aware lithography hotspot detection: a deep learning approach," *JM3*, vol. 16, no. 3, p. 033504, 2017.

[41] J. Chen, Y. Lin, Y. Guo, M. Zhang, M. B. Alawieh, and D. Z. Pan, "Lithography hotspot detection using a double inception module architecture," *Journal of Micro/Nanolithography, MEMS, and MOEMS*, vol. 18, no. 1, p. 013507, 2019.

[42] H. Levinson, "Principles of Lithography," in *Proc. SPIE*, pp. 261–263.

[43] T. M. A.-M. G. M. E. John Randall, Kurt G. Ronse, "Variable-threshold resist models for lithography simulation," in *Proc. SPIE*, vol. 3679, 1999.

[44] Y. S. Seongbo Shim, Suhyeong Choi, "Machine learning-based 3d resist model," in *Proc. SPIE*, vol. 10147, 2017.

[45] A. K. Wong and A. R. Neureuther, "Mask topography effects in projection printing of phase-shifting masks," *IEEE TED*, vol. 41, no. 6, pp. 895–902, 1994.

[46] R. L. Gordon and C. A. Mack, "Mask topography simulation for EUV lithography," in *Proc. SPIE*, vol. 3676, 1999, pp. 283–297.

[47] J. Ruoff, "Impact of mask topography and multilayer stack on high NA imaging of EUV masks," in *Photomask Technology 2010*, vol. 7823, 2010, p. 78231N.

[48] V. Agudelo, T. Fühner, A. Erdmann, and P. Evanschitzky, "Application of artificial neural networks to compact mask models in optical lithography simulation," *JM3*, vol. 13, no. 1, p. 011002, 2013.

[49] X. Ma, X. Zhao, Z. Wang, Y. Li, S. Zhao, and L. Zhang, "Fast lithography aerial image calculation method based on machine learning," *Applied Optics*, vol. 56, no. 23, pp. 6485–6495, 2017.

[50] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. ICCV*, 2018, pp. 8789–8797.

[51] X. Xu, Y. Lin, M. Li, T. Matsunawa, S. Nojima, C. Kodama, T. Kotani, and D. Z. Pan, "Sub-Resolution Assist Feature Generation with Supervised Data Learning," *IEEE TCAD*, vol. PP, no. 99, 2017.

[52] J.-H. Jun, M. Park, C. Park, H. Yang, D. Yim, M. Do, D. Lee, T. Kim, J. Choi, G. Luk-Pat *et al.*, "Layout optimization with assist features placement by model based rule tables for 2x node random contact," in *Proc. SPIE*, 2015, pp. 94 270D–94 270D.

[53] C. Kodama, T. Kotani, S. Nojima, and S. Mimotogi, "Sub-resolution assist feature arranging method and computer program product and manufacturing method of semiconductor device," Aug. 19 2014, US Patent 8,809,072.

[54] K. Sakajiri, A. Tritchkov, and Y. Granik, "Model-based sraf insertion through pixel-based mask optimization at 32nm and beyond," in *Proc. SPIE*, 2008, pp. 702 811–702 811.

[55] R. Viswanathan, J. T. Azpiroz, and P. Selvam, "Process optimization through model based sraf printing prediction," in *Proc. SPIE*, 2012, pp. 83 261A–83 261A.

[56] B.-S. Kim, Y.-H. Kim, S.-H. Lee, S.-I. Kim, S.-R. Ha, J. Kim, and A. Tritchkov, "Pixel-based sraf implementation for 32nm lithography process," in *Proc. SPIE*, 2008, pp. 71 220T–71 220T.

[57] B. Wu, H. Duan, Z. Liu, and G. Sun, "SRPGAN: perceptual generative adversarial network for single image super resolution," *CoRR*, vol. abs/1712.05927, 2017.

[58] X. Zhou, A. Karpur, C. Gan, L. Luo, and Q. Huang, "Unsupervised domain adaptation for 3d keypoint prediction from a single depth scan," *CoRR*.

[59] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[60] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NIPS*, 2014.

[61] M.-J. Wu, J.-S. R. Jang, and J.-L. Chen, "Wafer map failure pattern recognition and similarity ranking for large-scale data sets," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 1, pp. 1–12, 2014.

[62] M. B. Alawieh, F. Wang, and X. Li, "Identifying wafer-level systematic failure patterns via unsupervised learning," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 37, no. 4, pp. 832–844, 2017.

[63] C.-F. Chien, S.-C. Hsu, and Y.-J. Chen, "A system for online detection and classification of wafer bin map defect patterns for manufacturing intelligence," *International Journal of Production Research*, vol. 51, no. 8, pp. 2324–2338, 2013.

[64] J. Y. Hwang and W. Kuo, "Model-based clustering for integrated circuit yield enhancement," *European Journal of Operational Research*, vol. 178, no. 1, pp. 143–153, 2007.

[65] K. W. Tobin Jr, S. S. Gleason, T. P. Karnowski, S. L. Cohen, and F. Lakhani, "Automatic classification of spatial signatures on semiconductor wafer maps," in *Metrology, Inspection, and Process Control for Microlithography XI*, vol. 3050. International Society for Optics and Photonics, 1997, pp. 434–444.

[66] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *Proc. ICML*, 2019.

[67] C.-K. Chow, "An optimum character recognition system using decision functions," *IRE Transactions on Electronic Computers*, no. 4, pp. 247–254, 1957.

[68] D. Holden, J. Saito, T. Komura, and T. Joyce, "Learning motion manifolds with convolutional autoencoders," in *SIGGRAPH Asia 2015 Technical Briefs*. ACM, 2015, p. 18.

[69] "WM-811K wafer map," https://www.kaggle.com/qingyi/wm811k-wafer-map, accessed: 2019-07-30.