Making Digital Twins using the Deep Learning Kit (DLK)

Linyong (Leo) Pang[1], Suhas Pillai[2], Thang Nguyen[2], Mike Meyer[2], Ajay Baranwal[2], Henry Yu[1], Mariusz Niewczas[1], Ryan Pearman[1], Abhishek Shendre[1], Aki Fujimura[1]
[1] D2S, Inc. (US)
[2] Center for Deep Learning in Electronics Manufacturing (US)
[1]4040 Moorpark Avenue, Suite 250, San Jose, CA 95117 USA

ABSTRACT

Deep learning (DL) is one of the fastest-growing fields in artificial intelligence (AI). While still in its early stages of adoption, DL has already shown it has the potential to make significant changes to the lithography and photomask industries through the automation or optimization of equipment and processes. The key element required for application of DL techniques to any process is a large volume of data to adequately train the DL neural networks. The accuracy of the classification or prediction of any DL system is dependent on the depth and breadth of the training data to which it is exposed. For semiconductor manufacturing, finding adequate data – especially for corner cases – can be difficult and/or expensive.

In this paper, we will present two digital twins that are themselves built from DL as a part of a DL Starter Kit. We will demonstrate the creation of DL-based digital twins for a mask scanning electron microscope (SEM) and for curvilinear inverse lithography technology (ILT).

Keywords: Digital Twin, Deep Learning, DL, Artificial Intelligence, AI, simulation, photomask, GPU, CD SEM, Curvilinear ILT, Curvilinear Masks

1. INTRODUCTION

Deep learning is a field of AI that is creating a great deal of excitement in many industries. While it is clear that DL has the potential to be a transformative technology for all phases of the semiconductor manufacturing process, in these early days, many mask shops and wafer fabs are unsure of how to start to apply DL to their processes and challenges. Because DL for the semiconductor industry has some important differences from DL for the everyday world (such as image recognition or transformation), some DL practitioners in the industry may be having difficulty getting sufficient performance out of their DL systems due to lack of sufficient data.

Simulations that generate data for DL applications utilize digital replicas of the systems they are simulating, as well as replicas of other systems that interact with the systems they are simulating. These replicas are called digital twins. The digital representations provide both the elements of the systems and the dynamics of how the systems operate throughout their life cycles [1]. This paper introduces two digital twins that were created using DL. The first is a digital twin of mask SEM that quickly and inexpensively generates images that look like SEM pictures. The second is a digital twin of curvilinear ILT that quickly and inexpensively generates curvilinear mask test data. While a DL-based mimicking of ILT is not good enough for wafer performance, it is very useful for testing the many mask operations such as mask data preparation (MDP), mask writing, mask inspection, mask metrology, mask defect categorization, and evaluation and test of the many processing steps in mask making. Both digital twins are useful for generating training data to use for other DL projects that would work with curvilinear shapes or SEM images as inputs. Combined together, the digital twins can generate images that mimic SEM images of curvilinear mask data from CAD drawings without having manufactured the mask.

Both of these digital twins are a part of a DL Starter Kit which is an AI platform that enables semiconductor manufacturing companies and mask shops to get a head start on all that is needed to apply DL in our industry. The DL Starter Kit includes a GPU-accelerated computation platform, accurate physical models for mask and lithography, a fully distributed DL framework, pre-trained neural networks trained for common mask and wafer problems, and a number of digital twins, including the two described in this paper. In this paper we briefly review the background and general application of DL and digital twins in the context of semiconductor manufacturing. We will then describe the two digital twins.

2. DEEP LEARNING AND DIGITALTWINS FOR SEMICONDUCTOR MANUFACTUING

2.1 What is Deep Learning

DL is a subset of machine learning that uses a many-layered neural network model to recognize, predict or transform things [2]. DL does not “reason”; rather, it does an extremely sophisticated and extensive version of pattern matching. DL is similar to mask and wafer models for mask process correction (MPC) or optical proximity correction (OPC). In these fields, models are made increasingly accurate by a refinement of the model form over time; then the model parameters are calibrated, given a particular model form for a particular process; finally, an application of the particular model parameters in that model form for a particular design at runtime performs the desired function. Similarly, for DL, a neural network structure (model form) appropriate for a particular problem is chosen, then “training” (calibration) chooses parameters, then “inferencing” (runtime) applies DL for a particular problem.

2.2 DL neural network training requires a tremendous amount of data

A tremendous amount of data is required to train a DL model. The accuracy of the classification or prediction of any DL system is dependent on the depth and breadth of the training data to which it is exposed. For example, in the case of using a convolutional neural network (CNN), a popular and powerful deep learning network, to recognize dogs and cats, millions of images are first labeled by a human, using a DL technique called supervised learning. These images are fed into the CNN, where the training of the network occurs by many iterations of controlled trial and error resulting in iterative optimization of the network such that the network’s conclusions about whether a picture is of a cat or a dog matches the provided labels.

Once the CNN is trained, it can be used to recognize new dog or cat images. This is the inference phase. When a new image is fed into the trained CNN, it is doing a sophisticated pattern matching, whereby it extracts different levels of patterns from low-level to high-level, such as edges, texture, color, mouth, head, legs, etc. Once it passes through this sophisticated CNN, it matches what it learned in training, and outputs its conclusion as to whether the picture contains a dog or a cat. With enough training data of sufficient variety provided to the CNN, the likelihood of inferencing a correct answer on a new, previously unseen picture is high enough to be better than humans, particularly when faced with a large number of such tests. Machines are able to repeat the same process over and over again on vast amounts of data tirelessly, maintaining the same error rate on the first hour as on the 10,000th hour.

The strength of DL is the computer’s ability to tirelessly optimize the network by being trained on vast amounts of data. It is the programmer of the DL system’s responsibility to choose the right network for the task, and to choose the right set of data with which to train the network.

In typical DL applications such as image recognition, voice recognition, automatic translation, anomaly detection or automatic categorization in consumer-oriented applications, the distribution of data types in the training set should mimic the distribution pattern of real life. A DL programmer trying to improve her DL network performance so that a dog is correctly recognized 98% of the time instead of 96% doesn’t need to be concerned with an extremely unusual-looking dog that only is 0.01% of the dog population. This is why typical teaching in DL is to match the distribution of the training set to that of real-life distribution of data. Following this teaching will fail in the application of DL for our community. The key difference is the importance of getting the abnormal cases right in our industry. For our industry, that extremely unusual-looking dog – that extremely rare and difficult-to-repeat mask error – is exactly the case we need DL to recognize. But yet, unusual cases are obviously unusual. The mask makers are extremely good at making masks, so unusual situations do not naturally occur. Forcibly generating unusual situations through the manufacturing process is expensive and time consuming. And it is impractical to generate new unusual cases on demand every time a particular weakness is found in the DL training process, requiring a new set of data to feed the DL network in order to improve it.

This is the same reason that the autonomous driving world has an extremely large investment in simulation software to generate anomalous situations. Real driving is amazingly safe and standard. Having millions of miles of driving data on video is insufficient to train a DL network on anomalous situations. Simulation software is used, therefore, to generate abnormal situations and modify existing videos of real driving, for example by inserting a ball being thrown on front of the car, or a child running after that ball. Just as the autonomous driving community quickly discovered that data-augmentation through simulation is essential for them, applications of DL for the semiconductor manufacturing industry require data-augmentation through simulation. DL only does pattern matching, so the system won't know what to do if it hasn't been trained on similar data.

2.4 Digital twins are critical for DL training data generation

Digital twins can be very useful for every industry, because they can be used for development, testing, training, and prediction. They play a central role in producing DL training data, providing the virtual system behaviors, and input and output data required for accurate simulations.

Often, a company working on a DL project is needing to generate input data that comes from its customer to the company. Asking a customer to generate data can be difficult, and existing customer data are usually proprietary and cannot be released. Replicating everything that the customer has to generate the data is also expensive, time-consuming, and the company does not have the right expertise to operate the tools in any case. Having a digital twin of the customer's process that generates the input data to the company enables a fast and inexpensive way to generate that data.

A simulation-based digital twin is probably not the best answer in such situations because a complete simulation suite required to make a digital twin of the customer's process is probably prohibitive. DL provides an answer with a DL-based digital twin. In contrast to a simulation-based digital twin, which simulates every step and component of a process, a DL-based digital twin simply mimics the external behavior of the customer's process.

In general, DL-based digital twins are a promising use of DL for any step in the manufacturing process. When the digital twin is being used to generate training data for another DL process, this is using one DL process to generate data for another DL process. The possibilities for recursive application of DL are infinite.

2.5 Every process/piece of equipment in mask and wafer ecosystem needs three digital twins

In the mask shop and wafer fab as an example, the tape-out design data is first processed by optical proximity correction (OPC) or inverse lithography technology (ILT) to generate the mask design. Then this mask design will go through mask data preparation (MDP), mask writing by mask writer, mask inspection by inspection tools, mask CD metrology by CD SEM, mask review by Aerial Image Measurement System (AIMS^{TM}), and mask repair by repair tools. In the wafer fab, the wafer print will go through wafer CD metrology by wafer CD SEM, wafer inspection by wafer inspection tool, and wafer defect review by wafer review SEM.

The output from the digital twin of one process or piece of equipment can be used as the input of another process or piece of equipment, or its digital twin. For example, output of mask writer digital twin can be used as input data for a mask inspection digital twin, a mask CD SEM digital twin, a mask review digital twin, and/or a mask repair digital twin. On the other hand, the mask writer digital twin also needs a mask CD SEM and mask inspection digital twin to verify that its output can pass mask CD metrology and inspection. In another words, every piece of equipment or process needs three digital twins: its own digital twin, digital twin of its upstream process/equipment for input, and digital twin of its downstream process/equipment for testing and verification.

2.5 DL Starter Kit

In addition to a GPU-accelerated computation platform, the DL Starter Kit has four additional components: GPU-accelerated curvilinear mask and wafer simulations plus curvilinear geometry libraries, GPU-accelerated rigorous simulator, a fully distributed DL framework, and DL neural networks for pre-trained mask and wafer problems (Figure 1).

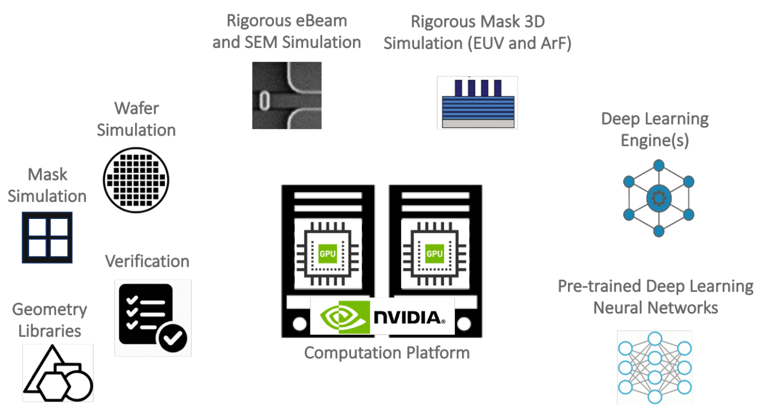


Figure 1: Components of the DL Starter Kit

3. MASK CD SEM DIGITAL TWINS CREATED WITH DL STARTER KIT

There are two types of digital twins that are/will be needed for almost any analysis in mask shops and wafer fabs: a SEM digital twin and a curvilinear ILT digital twin. The next two sections present the processes used to create these essential digital twins using the DL Starter Kit.

3.1 SEM Digital Twin is Required for Almost all Data Analysis in Mask Shops and Wafer Fabs

SEM is a must-have for both mask CD metrology and wafer CD metrology, because it is the only metrology method that can give the nanometer accuracy required by mask shops and wafer fabs. Almost all mask and wafer analysis is based on SEM images. However, getting mask and wafer SEM images from mask shops and wafer fabs is difficult. It usually takes tremendous work to set up the recipe, and it takes a long time to scan. Then if you are doing research and development, you also need to compete with SEM work required by production. When you develop a DL project that is related to mask, your input is a mask SEM image. A mask SEM digital twin can produce the volume of mask SEM images required for DL training.

3.2 Neural Style Transfer DL Technique Provided Initial Approach for SEM image Digital Twin

Our first approach was to use neural style transfer – to use DL techniques to compose images in the style of another image [3]. Neural style transfer is an optimization technique used to take two images a *content* image and a *style reference* image, and blend them together to create an *input* image that is transformed to look like the content image, but "painted" in the style of the style image (Figure 2).

Figure 2: An example of neural style transfer: transfer a dog picture into a new artwork with van Gogh style

The principle of neural style transfer is to define two distance functions, one that describes how different the content of two images are, Dcontent, and one that describes the difference between the two images in terms of their style, Dstyle. Then, given three images, a desired style image, a desired content image, and the input image (initialized with the content image), we try to transform the input image to minimize the content distance with the content image and its style distance with the style image.

When applying style transfer to a SEM digital twin, the idea is that the SEM image signature, these random noises and signal response to edges can be the style, the simulated mask patterns can be the content, then style transfer can translate the simulated mask pattern into a mask SEM image. Figure 3 shows the result of this. The SEM image generated from the neural transfer has some signatures of SEM, but not close.

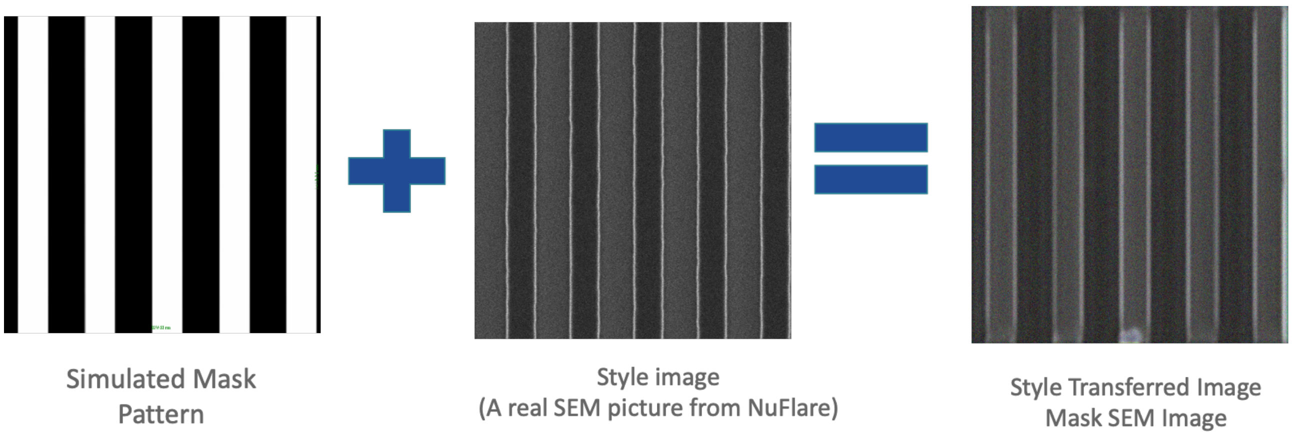


Figure 3: Transfer a simulated mask pattern into a mask pattern SEM image using neural style transfer

3.3 The Final SEM Image Digital Twin Used Generative Adversarial Networks (GAN): Pix2Pix

Generative adversarial networks (GANs) have had a huge success since they were introduced in 2014[4]. In GAN, two neural networks (the generative network and the discriminative network) contest with each other in a game. Given a training set, this technique learns to generate new data with the same statistics as the training set. The generative network generates candidates while the discriminative network evaluates them. The contest operates in terms of data distributions. Typically, the generative network learns to map from a latent space to a data distribution of interest, while the discriminative network distinguishes candidates produced by the generator from the true data distribution. The generative network's training objective is to increase the error rate of the discriminative network (i.e., "fool" the discriminator network by producing novel candidates that the discriminator thinks are part of the true data distribution) [5].

Pix2Pix is a conditional adversarial network used as a general-purpose solution to image-to-image translation problems [6]. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. Pix2Pix is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks (Figure 4).

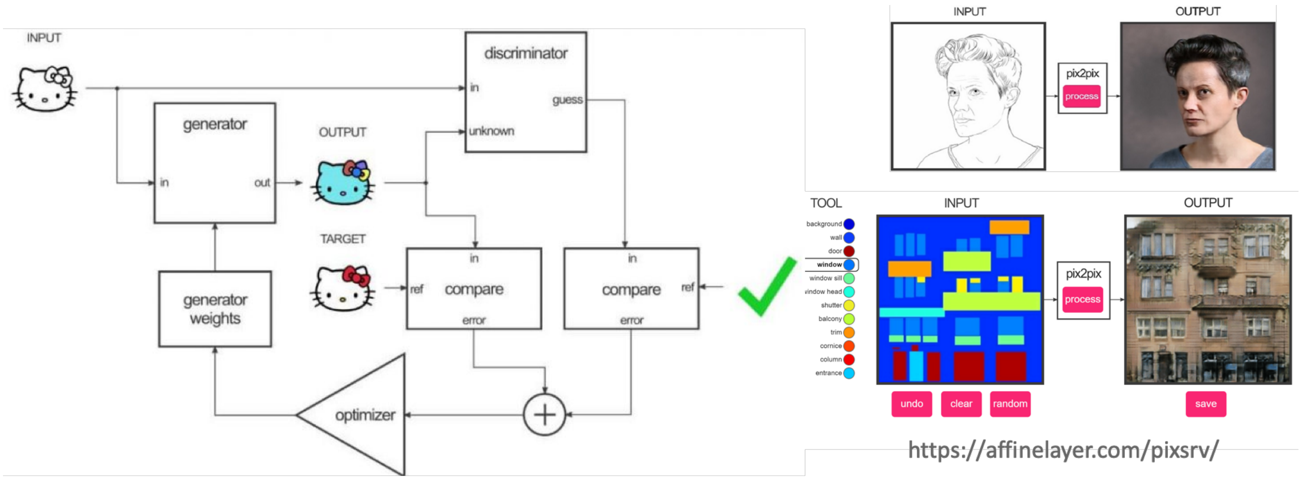


Figure 4: Diagram of GAN, and examples of Pix2Pix using GAN

Pix2Pix is applied to solve the mask CD SEM digital twin problem. In this implementation, the generator network generates a mask SEM image from a simulated mask pattern. The simulated mask pattern is also paired with real SEM image captured on the real mask in the discriminator network, with the instruction that this is the real data. Then the generated mask SEM image is fed into the discriminator network until the discriminator cannot tell it is fake. Then both the generator and discriminator networks are trained with large volumes of such data (Figure 5).

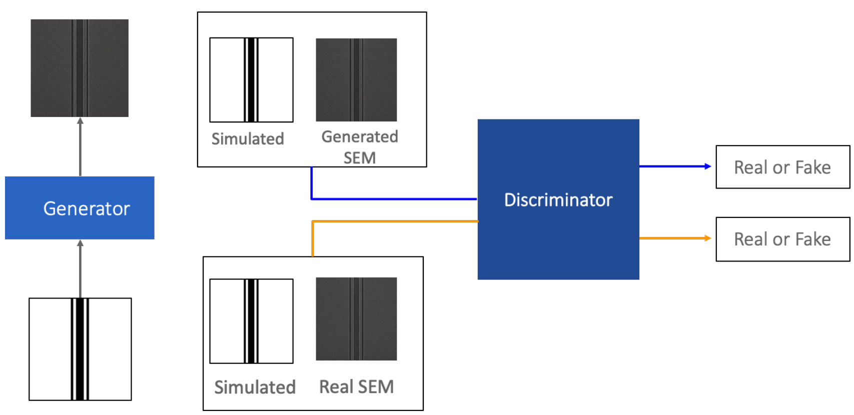
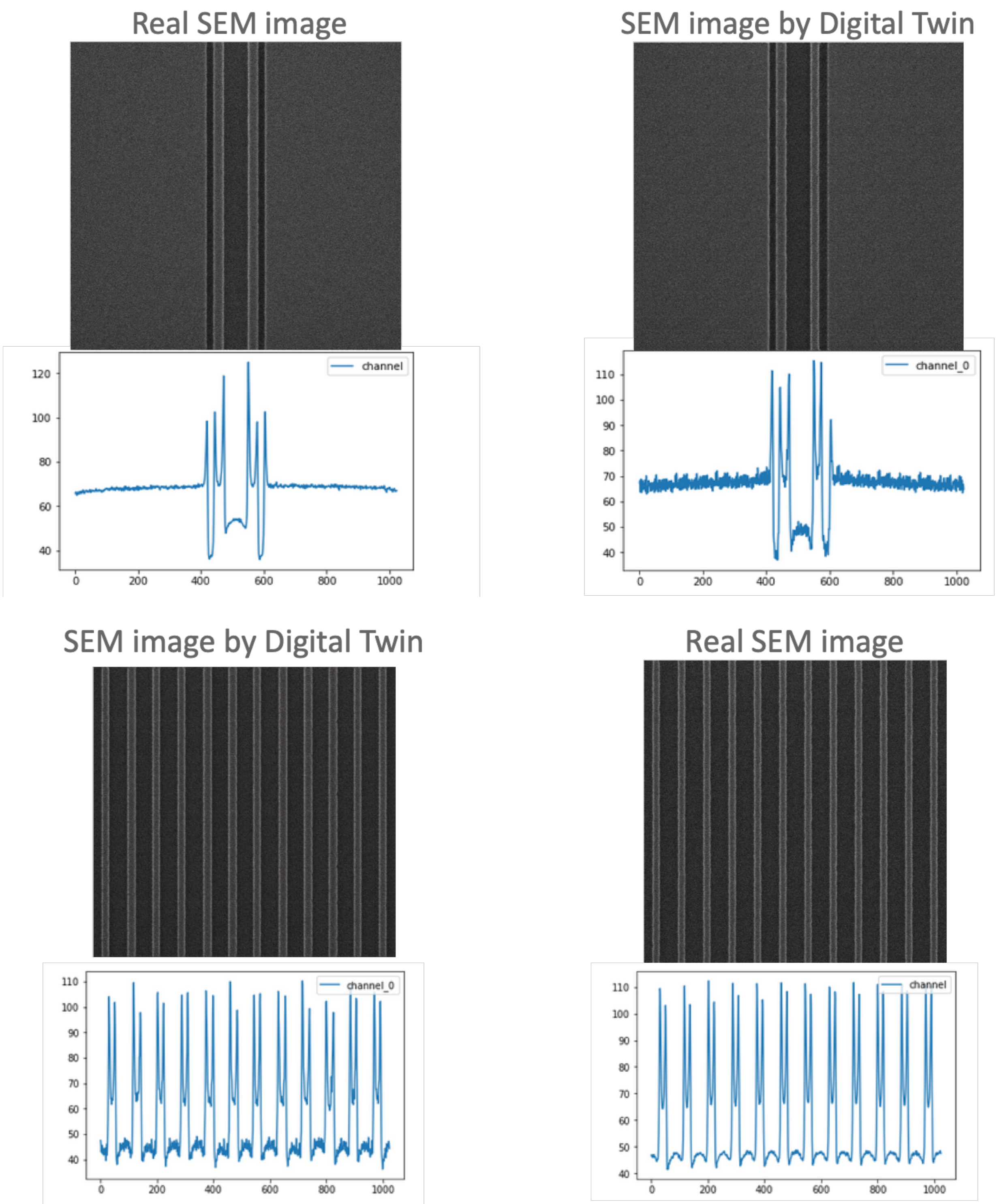


Figure 5: Diagram of training mask SEM digital twin using GAN and Pix2Pix

Figure 6 shows three examples of mask SEM images generated by the SEM digital twin and the real SEM image. The image intensity on a horizontal cutline at the same location are shown as well. Note that not only do the images look very similar, but also the signal response on edges are similar as well.



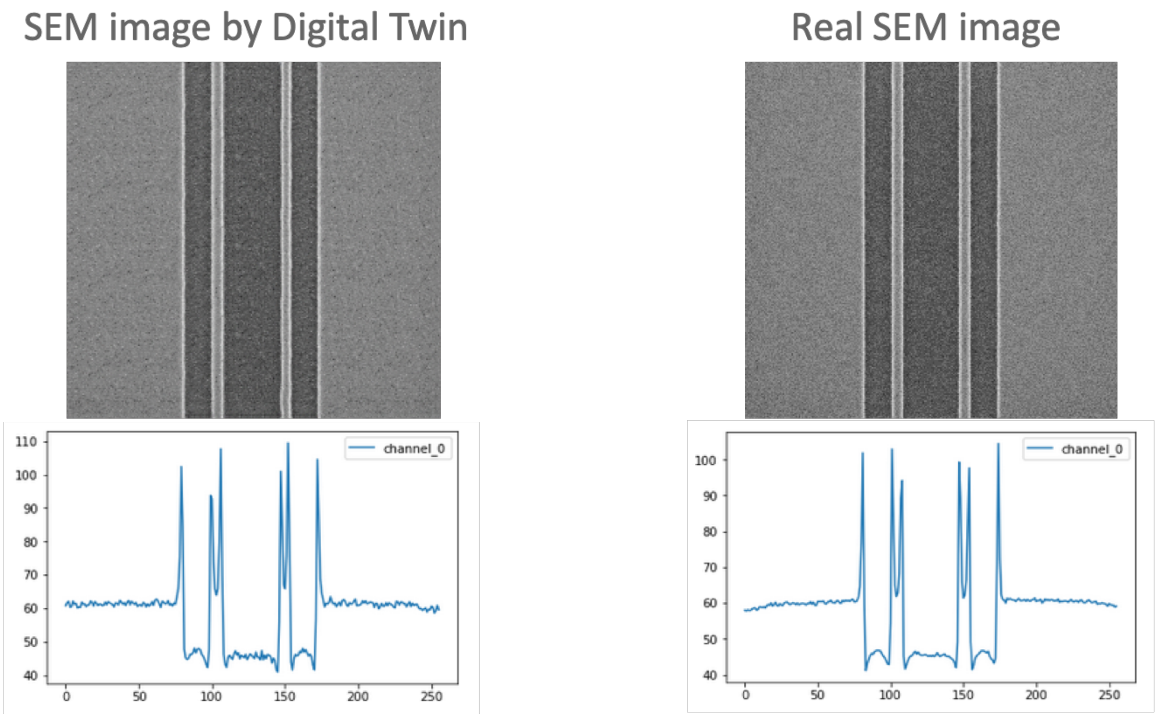


Figure 6: Examples of SEM images generated by mask the SEM digital twin and the real SEM images

Three more examples are shown in Figure 7. It displays simulated mask pattern as input to the mask SEM digital twin, SEM image generated from the mask SEM digital twin, and the corresponding real SEM image. One can see that the SEM images generated by SEM digital twin are so close to the real SEM image that human eyes can barely tell the difference.

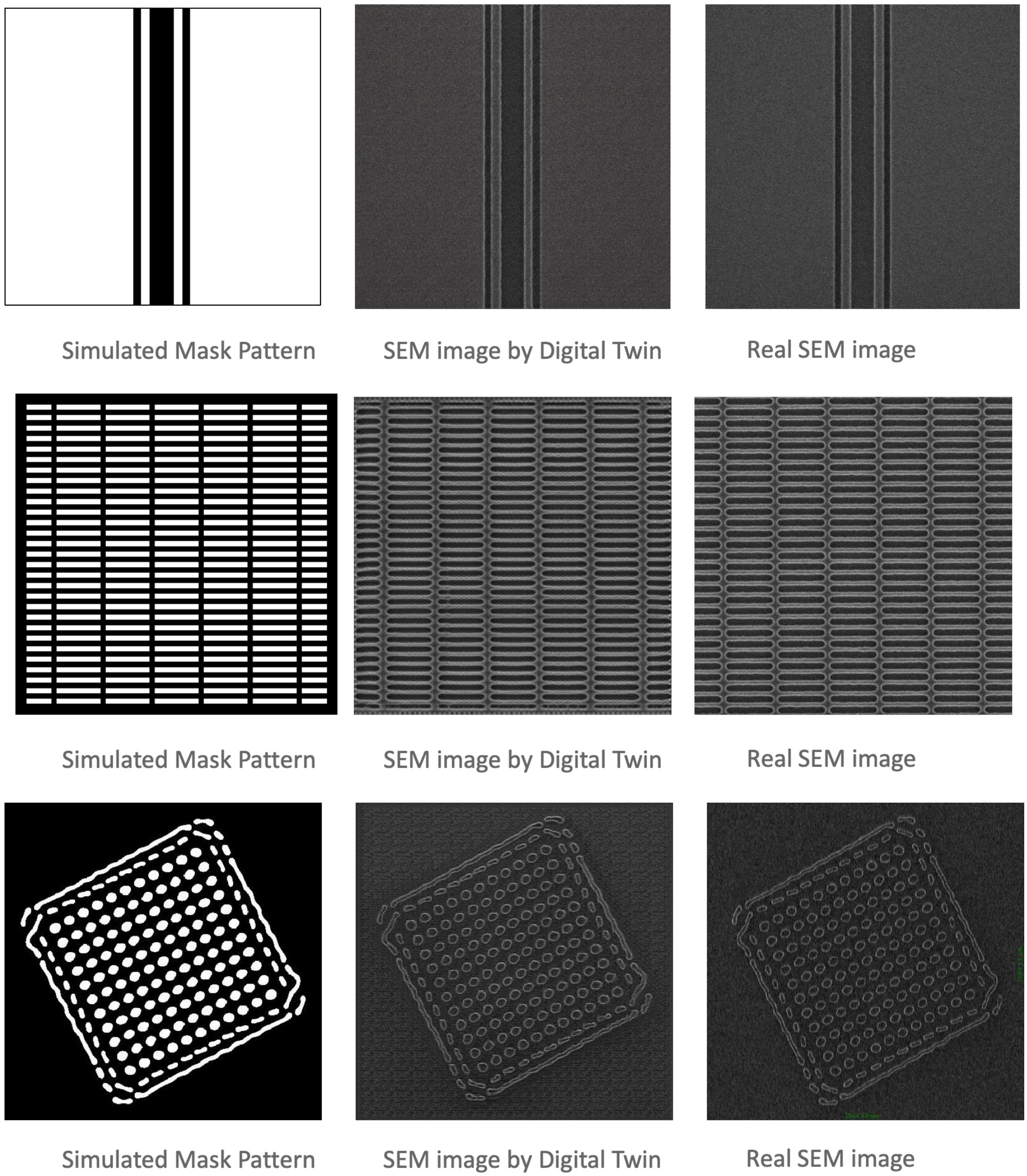


Figure 7: Examples of input and output from the mask SEM digital twin and the corresponding real SEM images

4. CURVILINEAR ILT DIGITAL TWINS CREATED WITH DL STARTER KIT

4.1 DL-Generated Curvilinear ILT Digital Twin: Fast, Cost Effective Way to Create Curvilinear ILT Test Data

The mask industry is facing the transition from Manhattan patterns to curvilinear patterns. This was first enabled by the introduction of the multi-beam mask writer, which can write curvilinear patterns without a write time penalty [7, 8]. Then recently, studies have shown both that curvilinear design can reduce mask layers and process and pack designs in a smaller area [9], and that curvilinear masks can reduce variability on mask and wafer [10]. In addition, the development of full-chip, curvilinear ILT in a day has eliminated the last roadblock to apply full-chip ILT for all critical layers [11].

Figure 8 shows a part of a CAD design and the corresponding curvilinear ILT mask. For the whole mask and wafer ecosystem to be fully ready to handle curvilinear data for high-volume production, each step in the mask and wafer process must be optimized and verified with curvilinear mask designs. As suspected trouble spots in the mask manufacturing flow are uncovered, additional data that stimulate the sensitive aspects of the flow become desirable. Being able to generate realistic curvilinear ILT output shapes quickly and inexpensively from any real or artificial CAD design data allows the process to be resilient to curvilinear input faster to improve time to high volume manufacturing. A curvilinear ILT digital twin is a cost-effective way to generate the curvilinear ILT data needed for this training and verification.

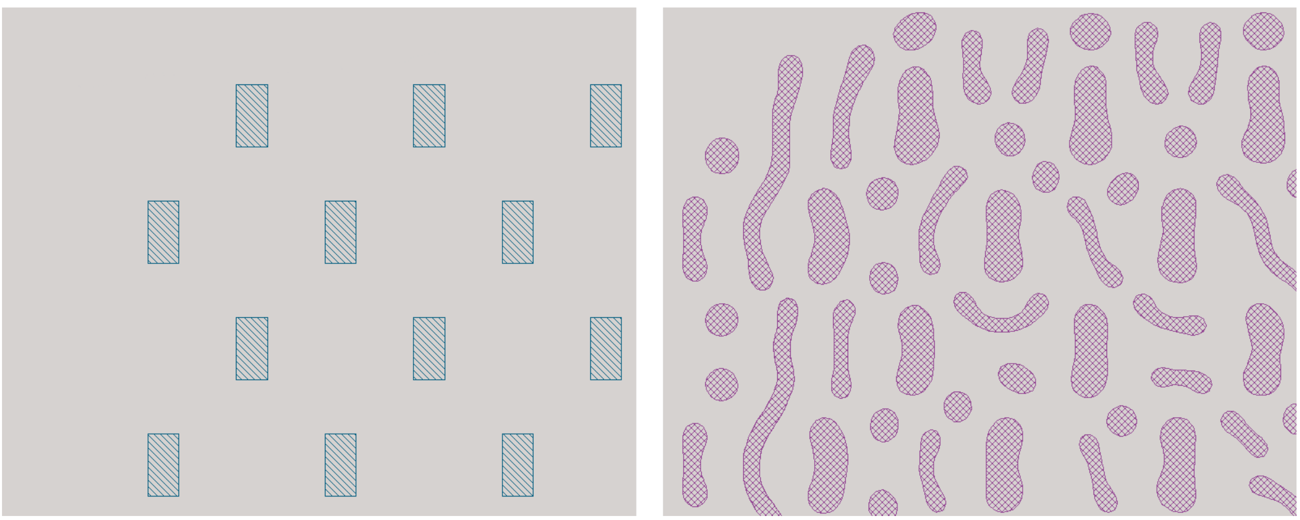


Figure 8: Example of CAD design (left) and a corresponding curvilinear ILT mask (right)

For those that already have or are building DL applications in the mask shop, the ILT digital twin can generate training data to teach these DL networks how to handle curvilinear data. As the DL network performance is tuned by its programmer, the new data that the programmer needs to train for the network's weaknesses can be quickly generated using the ILT digital twin without asking the customer for some specific data, or running the real ILT to generate the data.

4.2 Curvilinear ILT and its Digital Twins

Figure 9 shows three examples of curvilinear ILT and results generated by its digital twins. It should be noted that results of this ILT digital twin cannot be used for wafer print. As DL is a statistical method, its results are not accurate enough to meet edge-placement error (EPE) and process-window requirements on the wafer. But its mask pattern shapes are very close to the curvilinear ILT result, making it useful as a fast, inexpensive generator of curvilinear mask test cases. For example, it can be used to test curvilinear MPC for quality of results or for runtime implications. It may be useful to evaluate various curvilinear data formats to evaluate data volume, encoding and decoding time, or its lossiness. It can be used to evaluate the accuracy in writing various realistic ILT output shapes for multi-beam mask writers, or how various mask processes perform for curvilinear shapes. It can be used by mask inspection tools to test their inspection algorithms for curvilinear masks. It can be used to evaluate repair of curvilinear shapes. And it can be used to evaluate metrology or AIMS.

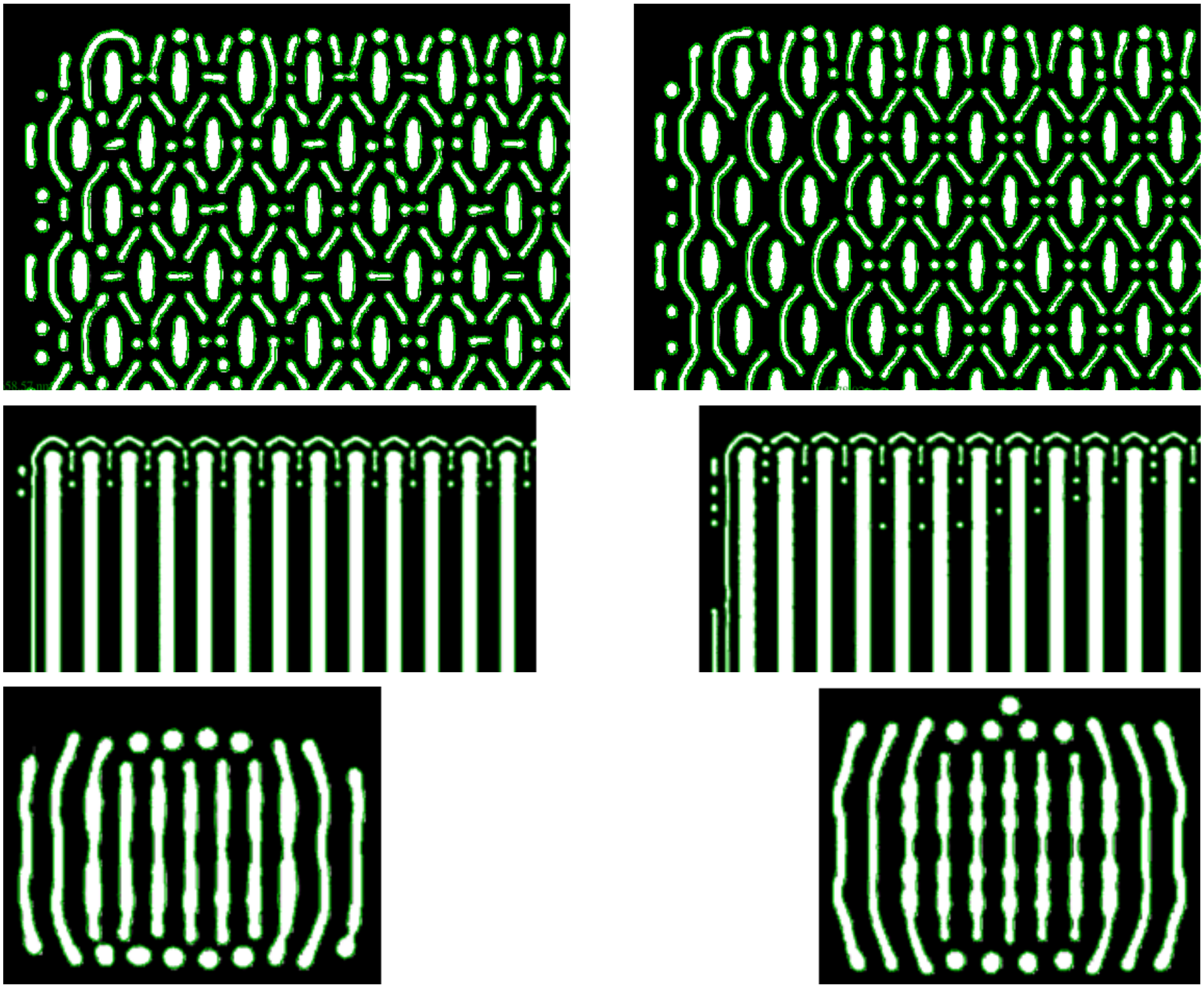


Figure 9: Examples of curvilinear ILT digital twin (left) and the real curvilinear mask patterns generated from curvilinear ILT (right)

Unlike its real twin, the curvilinear ILT digital twins operate on a partition-by-partition basis. It is a workstation-class solution that generates shapes in the range of 100um x 100um areas, not for full-chip. As DL inferencing is extremely fast, the results are computed more than an order of magnitude faster. Just like its real twin, the digital twin is given as an input the target wafer shapes (the CAD design) and outputs the computed mask image required to generate the wafer shape.

We chose U-Net [12], a popular image segmentation network architecture for the curvilinear ILT digital twin. The ILT digital twin didn't need the complexity of a conditional GAN because it is more deterministic than the SEM image digital twin, which must contend with the SEM image noise. U-Net was chosen because it is a fast architecture with a large receptive field, but still has good resolution. A large receptive field is needed so the model can account for distant shapes that influence the curvilinear shapes.

U-Net was originally used to segment cells in light-based microscopic images. The ground truth for segmenting cells was manually labelled, while the ground truth (the curvilinear ILT mask) for training the digital twin is generated by the real twin. The wafer model that was used by the real twin to generate the training data for the digital twin is the model implied in the digital twin. For generating different types of shapes, say for asymmetric light sources, new training data needs to be generated using the different light source in the real curvilinear twin and the same DL network would need to be retrained.

When the digital twin is run, it rasterizes the CAD design and feeds this to the trained DL model, which generates a rasterized curvilinear ILT mask. Because the DL model is a statistical method, it may generate small mask rule check (MRC) violations, which must be corrected before writing the contoured ILT mask. Correcting the MRC violations is important because the intended uses of the digital twin require a MRC-clean mask.

Digital twins can also be used in a chain: for a DL project that requires curvilinear ILT mask SEM images to do training, one can use the curvilinear ILT digital twin to generate curvilinear ILT mask patterns, then use the mask SEM digital twin to generate the curvilinear mask SEM image.

5. SUMMARY AND CONCLUSIONS

The accuracy of the classification or prediction of any DL system is dependent on the depth and breadth of the training data to which it is exposed. Therefore, a very large volume of data covering the broadest-possible set of conditions is required for good results. Any experimental data in mask and wafer manufacturing requires spending hundreds of thousands, if not millions of dollars, to tape out designs, write the masks, process them, print the wafers with scanners, process them, then measure millions of locations on a SEM over days or weeks, sometimes even months. Simulators allow these processes to happen in a digital world, training a DL solution to reach its full potential while saving time and money. Digital twins are digital replicas of real processes or equipment that are used in simulations to create DL data or to train or verify equipment.

We introduced a DL Starter Kit, which includes a GPU-accelerated computation platform, accurate physical models for mask and lithography, a fully distributed DL framework, and pre-trained neural networks trained for common mask and wafer problems. Using this DL Starter Kit, semiconductor manufacturing companies and mask shops can build their deep neural network models much faster with pre-trained neural network models, build digital twins that can be used to generate data or add their own data, and re-train the neural network model in that environment to learn a desired behavior.

We have demonstrated a SEM digital twin and a curvilinear ILT digital twin, both of which are DL-based digital twins. The results from these digital twins are available as a part of the DL Starter Kit and are good enough to be used by the mask and wafer ecosystem for their DL projects and to train and verify their equipment. Digital twins can also be used together; eventually one could build a virtual mask shop or wafer fab from a string of digital twins. Every process or piece of equipment will need three digital twins: the digital twin of its own system for training, diagnosing, and prediction; the digital twin of its upstream process or tool for generating input for its own tool; and digital twin of its downstream process or tool for verifying the output or results of its own tool.

REFERENCES

[1] Wikipedia, Digital Twin, https://en.wikipedia.org/wiki/Digital_twin

[2] Wikipedia, Introduction to Deep Learning, https://www.wikipedia.org/wiki/Deep_learning

[3] Gatys, Leon A., Alexander S. Ecker, Matthias Bethge, and C. V. Sep. "A neural algorithm of artistic style. arXiv 2015." *arXiv preprint arXiv:1508.06576*.

[4] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

[5] Wikipedia, Generative adversarial network, https://en.wikipedia.org/wiki/Generative_adversarial_network

[6] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125-1134. 2017.

[7] Klein, Christof, and Elmar Platzgummer. "MBMW-101: World's 1st high-throughput multi-beam mask writer." In *Photomask Technology 2016*, vol. 9985, p. 998505. International Society for Optics and Photonics, 2016.

[8] Matsumoto, Hiroshi, Hideo Inoue, Hiroshi Yamashita, Hirofumi Morita, Satoru Hirose, Munehiro Ogasawara, Hirokazu Yamada, and Kiyoshi Hattori. "Multi-beam mask writer MBM-1000 and its application field." In *Photomask Japan 2016: XXIII Symposium on Photomask and Next-Generation Lithography Mask Technology*, vol. 9984, p. 998405. International Society for Optics and Photonics, 2016.

[9] Dounde, Amit, Syed Muhammad Yasser Sherazi, Ryoung-Han R. Kim, Darko Trivkovic, Werner Gillijns, and Youssef Drissi. "A study of curvilinear routing in IN5 standard cells: challenges and opportunities." In *Photomask Technology 2019*, vol. 11148, p. 111481C. International Society for Optics and Photonics, 2019.

[10] Pearman, Ryan, Jeff Ungar, Nagesh Shirali, Abishek Shendre, Mariusz Niewczas, Leo Pang, and Aki Fujimura. "How curvilinear mask patterning will enhance the EUV process window: a study using rigorous wafer+ mask dual simulation." In *Photomask Japan 2019: XXVI Symposium on Photomask and Next-Generation Lithography Mask Technology*, vol. 11178, p. 1117809. International Society for Optics and Photonics, 2019.

[11] L. Pang, et al, " Study of Mask and Wafer Co-design That Utilizes a New Extreme SIMD Approach to Computing in Memory Manufacturing--Full-Chip Curvilinear ILT in a Day." In *Photomask Technology 2019*, vol. 11148, (to be published) International Society for Optics and Photonics, 2019.

[12] Ronneberger, Olaf, Philip Fischer, Philip, Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation." In: *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI),* pp. 234–241. 2015.