Comparative Analysis of Grating Reconstruction: Deep Learning versus Levenberg-Marquardt Methods

L. Fu, X.Wang, K. Frenner, and S. Reichelt
Institute of Applied Optics, University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany

ABSTRACT

Model-based optical scatterometry is a widely utilized non-destructive measuring technique in semiconductor manufacturing for retrieving features on wafers. It offers an attractive solution for quality control and process monitoring. However, the increasing complexity of 3D nanoscale device structures presents significant challenges for optical scatterometry. To address these challenges, it is crucial to integrate different methods and create a hybrid metrology approach that could encompass measurements, modeling, and data analysis techniques. To tackle this objective, we explore in this study two alternative approaches for parameter reconstruction, distinct from the conventional library search method. The first approach utilizes a neural network based on a Resnet architecture, while the second approach employs the Levenberg-Marquardt algorithm, a nonlinear least square fitting technique. By performing a comparative analysis of the two methods, we propose a strategy to combine them for accurate and efficient parameter reconstructions.

Keywords: Deep learning, parameter reconstruction, Fourier scatterometry, Levenberg-Marquardt algorithm, hybrid metrology

1. INTRODUCTION

Optical scatterometry is a powerful technique used in the semiconductor industry for measuring the geometry of patterned features on wafers. It provides critical information about the dimensions and profiles of these features, which is essential to ensure the function of the semiconductor devices. One of the advantages of scatterometry is its non-destructive nature, making it attractive for quality control and process monitoring. Another advantage is its high accuracy. It is capable of achieving resolution at a nanometer scale. Recently, there has been a shift with optical scatterometry towards machine learning techniques to analyze large and complex datasets and to extract intricate patterns[1,2]. Nevertheless, it was pointed out that machine-learning based solutions can only provide a complementary solution where modeling is challenging. Thus, the future of scatterometry lies in the combination of different testing tools[3]. To confront this challenge, this report investigates two parameter reconstruction procedures: a neural network and an iterative curve fitting algorithm, specifically the Levenberg-Marquardt algorithm. Our goal is to evaluate and compare the performance of the two methods and propose a strategy to combine them. It is well accepted that Fourier scatterometry surpasses conventional optical ellipsometry in terms of efficiency and sensitivity[4]. We thus focus on parameter reconstruction within the framework of Fourier scatterometry.

Geometrical parameters of a Si grating are shown in Figure 1(a). The grating is covered by a thin SiO_2 layer with a thickness of 3 nm. It has a periodicity of 500 nm for simulation experiments Exp. I - Exp. III, and 100 nm for Exp. IV. The refractive indices are taken from Aspnes and Studna 1983 for Si, and from Lemarchand 2013 for SiO_2[5]. The sidewall angle (SWA) of the grating is defined as the projected length of the slanted sidewall onto the bottom. R_T and R_B characterize the rounding at the top and bottom of the grating line, respectively, and are defined as radius in nanometer. The critical dimension (CD) is the line width at half of the grating height H. The parameters and their variation ranges for different simulation tests are listed in Table I. The range for each parameter is discretized into 15 points. However, only 10 thousand parameter sets were chosen at random for image calculations.

Figure 1(b) illustrates the optical setup for Fourier scatterometry as the basis of our analysis. The grating is illuminated by a polarized laser beam focused by a microscope objective with NA = 0.7. The polarization is either along the *x*- or the *y*-direction in terms of a global coordinate system. When light is passing through the microscope, its polarization undergoes a change which can be described by a rotation matrix[6]. Scattered light from the grating surface is collected by the same objective and the image in the Fourier plane (or pupil plane) in terms of the NA coordinates (NA_x in the *x*-direction and

Modeling Aspects in Optical Metrology IX, edited by Bernd Bodermann, Karsten Frenner,
Proc. of SPIE Vol. 12619, 1261907 · © 2023 SPIE · 0277-786X
doi: 10.1117/12.2688328

NA_y in the *y*-direction) is recorded by a camera. The modeling is carried out by using an in-house-developed software package MicroSim based on rigorous coupled wave analysis (RCWA) method. The pupil images are sampled with 21×21 pixels in a range of [-0.7, 0.7] and 100 Fourier modes are considered. The Mueller matrix is then calculated via the Jones matrix at each incident and azimuth angle, which results in 16 images for each element of the Mueller matrix.

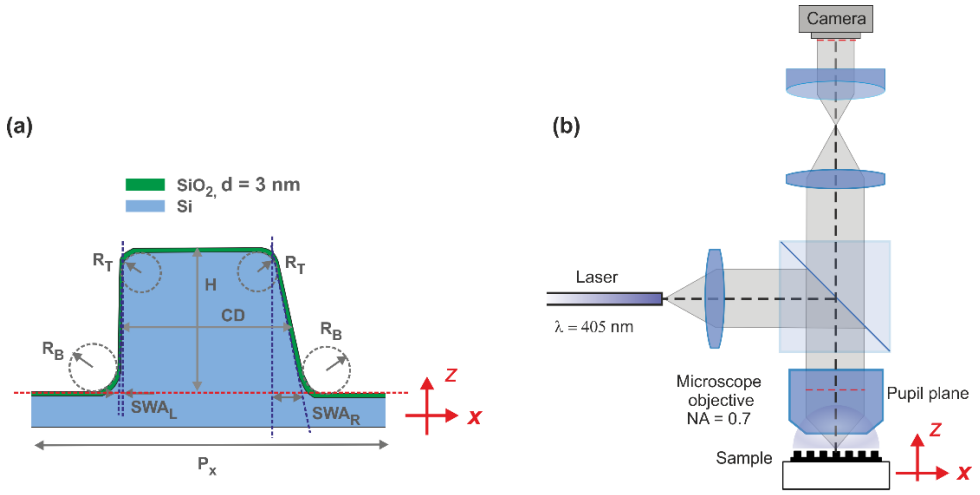


Figure 1. (a) Grating geometry with defined structural parameters studied for this report. (b) Set up of a Fourier scatterometer for generating pupil images and images of Mueller matrices per postprocessing. A laser at a wavelength of 405 nm is collimated and reflected by a beam splitter. It is then focused onto the sample surface via an objective lens with NA of 0.7. The scattered light from the sample is collected by the same objective lens forming a pupil image at the Fourier plane, and is recorded by the camera.

For the machine learning approach, we choose a deep neural network with an architecture of Resnet 18, which has been extensively employed for computer vision and optical metrology tasks[2]. As illustrated in Figure 2, the network consists of two main parts. The first part is a copy of the 18-layer Resnet structure. It is used to capture important features from input images. Within it, the initial convolutional block consists of a convolutional layer, a batch-normalization (BN) layer and a ReLU activation layer. Large kernels and striding are employed to scan and downsample the input images. The subsequent max-pooling layer further downsamples the images. To extract features at different scales, four distinct residual blocks are subsequently utilized. Each comprises two convolutional blocks with a dropout structure. The extracted information is then aggregated through an adaptive average pooling layer. The second part of the network is used to translate the extracted feature information into normalized structure parameters through two fully connected (FC) blocks activated by ReLU function and one FC block with linear activation.

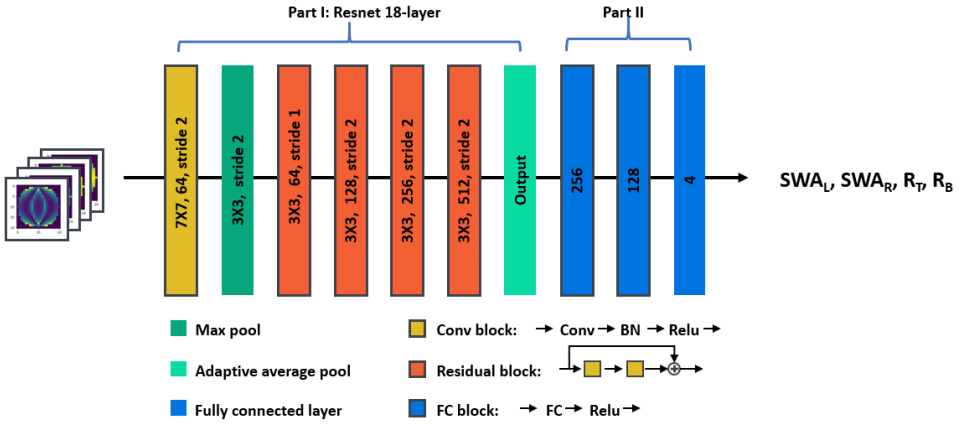


Figure 2.A Resnet architecture with 18 layers.

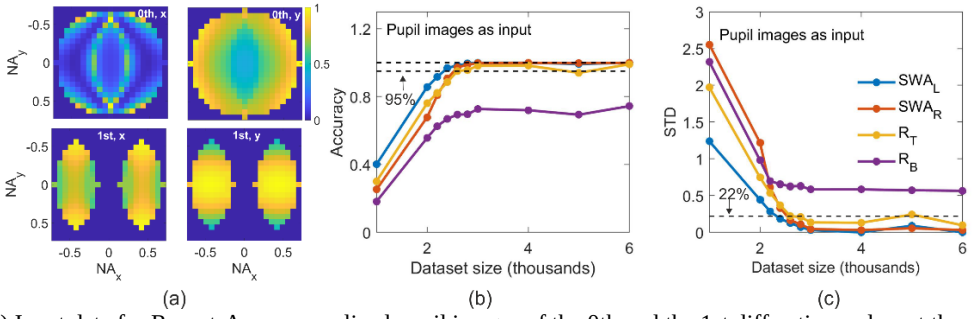
For this Resnet approach, we focus on reconstruction of 4 structural parameters as listed for Exp. I in Table I, and prepare two kinds of images as input datasets. The first kind is pupil images calculated from illumination polarized along the *x*- and *y*-directions, respectively. Two images for each polarization can be obtained, namely the 0th and the 1st order (for P_x = 500 nm at λ = 405 nm). In total, four images for each parameter set are used as input. This network is designated as Resnet-A. The second kind is images of Mueller matrix at the pupil plane, which are calculated from the first-kind datasets[4]. This network is designated as Resnet-B. For this Resnet, only the 0th order images are used, the reason of which will be explained below. In total, 20 thousand datasets from gratings with randomly chosen parameter sets (as labels) within their respective range are generated for both *x*- and *y*-polarizations. During the training of the neural network, all the labels and images are normalized to a value of one.

Table 1. Parameters used for different experiments performed in this report.

	SWA_L	SWA_R	R_T (nm)	R_B (nm)	P_x (nm)	H (nm)	CD (nm)
Exp. I	[0, 14] nm [82°, 90°]	[0, 14] nm [82°, 90°]	[0, 14]	[0, 14]	500	100	250
Exp. II	[0, 14] nm [82°, 90°]	[0, 14] nm [82°, 90°]	[0, 14]	[0, 14]	500	100	250
Exp. III	[0, 14] nm [82°, 90°]		[0, 14]		500	[91, 105]	[246.5, 253]
Exp. IV	[0, 14] nm [82°, 90°]	[0, 14] nm [82°, 90°]	5		[96, 102.5]	100	[46.5, 53]

2. PERFORMANCE OF THE TRAINED NEURAL NETWORK

As an example, Figure 3(a) shows four calculated pupil images from one parameter set, two from the 0th order and the 1st order, respectively, at the *x*-polarization and two from those orders at the *y*-polarization. Such pupil images as input data were used to train the Resnet (Resnet-A).



(a) (b) (c)

Figure 3. (a) Input data for Resnet-A are normalized pupil images of the 0th and the 1st diffraction orders at the *x*- and *y*-polarizations, respectively. (b) Accuracy of the reconstructed parameters from the trained Resnet-A versus the size of dataset. (c) Standard deviation of the Resnet-A versus the size of the dataset.

Figure 3(b) shows the accuracy of the trained Resnet varying with the size of the dataset, and Figure 3(c) shows the standard deviation (STD). Except for the curve for R_B, which is the bottom rounding parameter, all the other three parameters achieve an accuracy of 95% with only 2600 dataset (In total, there are 2600×2 datasets when considering both *x*- and *y*-polarizations). STDs are all smaller than 22% for parameters of SWA_L, SWA_R, and R_T. The problem with R_B will be explained in the following. On the whole, high accuracy and low STD value are demonstrated, although the performance for R_T is slightly poorer compared to SWA_L and SWA_R. It reveals that Resnet-A is more sensitive to the sidewall angles than to R_T and R_B. When only the 0th order images are used as input data, the same performance is obtained, the results of

which are not shown here. This indicates that the 1st order does not provide more structural information to the neural network. For this reason, only the 0th order images are considered for the following experiments from Exp. II to Exp. IV.

In the following, images of Mueller matrix are used as input data to train the Resnet (Resnet-B). The images shown in Figure 4(a) are normalized to m_{11}. Due to symmetry, only 9 element images are employed, including the four diagonal elements and the elements above the diagonal. Without considering the curve for R_B, the network achieves the accuracy of 95% at a dataset size of 2000, which is less than 2600 as shown in Figure 3(a). With this size of dataset, the network also has a STD below 22%. Furthermore, almost identical performance for the three parameters of SWA_R, SWA_L and R_T is demonstrated. Obviously, features on the images of Mueller matrix are more sensitive to grating parameters.

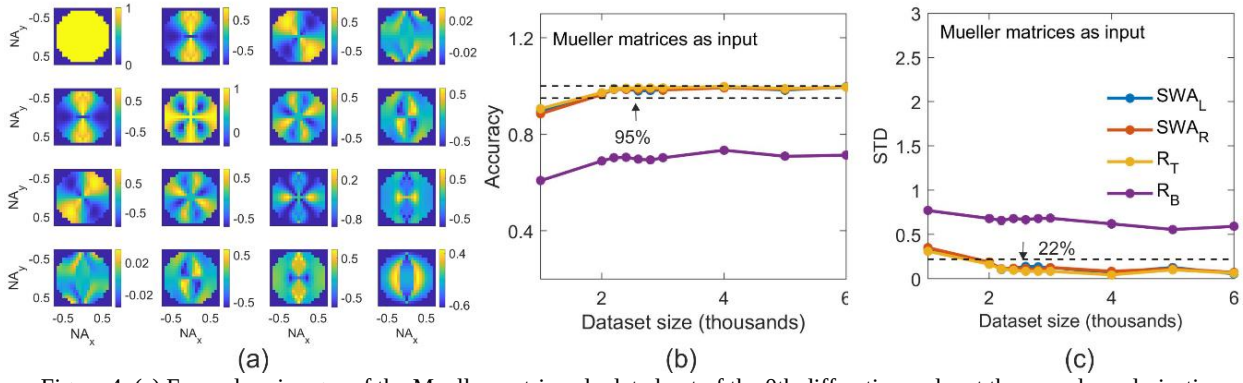


Figure 4. (a) Exemplary images of the Mueller matrix calculated out of the 0th diffraction order at the *x*- and *y*-polarizations. (b) Accuracy of the trained Resnet-B using the images in (a) as input versus the size of dataset. (c) Standard deviation as a function of the size of the dataset for Resnet-B.

To explore why the two Resnets show a low accuracy and high STD for parameter R_B, we plot the true parameters for 20 tests in Figure 5(a), and the relative errors ($\Delta P/P$) between the true and predicted parameters in Figure 5(b) for Resnet-A and in (c) for Resnet-B. The relative error of R_B with both Resnets are extraordinarily high from the tests numbered 6, 10, 12 and 17, with which the true value of R_B is set at 2 nm, as can be seen from Figure 5(a). When R_B is larger than 4 nm, R_B can be properly reconstructed and the accuracy and STD are comparable with other parameters. This explains the overall poor performance of the two Resnets in estimating R_B, and indicates that current network is not sensitive to bottom rounding when R_B is too small. To illustrate this performance clearly, we plot the relative error of R_B (from Figure 5(c)) as a function of the true R_B in Figure 5(d), and a curve with exponential decay (dashed line) can well fit their relation.

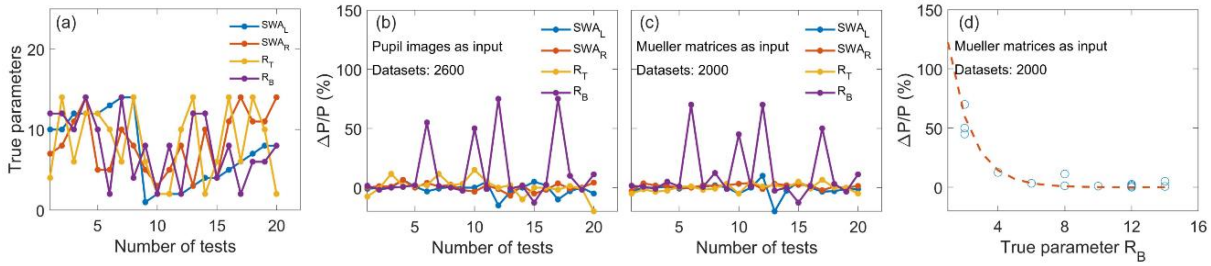


Figure 5. (a) Variation of the true parameters for different tests. (b) The deviation of the predicted parameters with respect to the true parameters (relative error) using pupil images at the *x*- and *y*-polarizations. (c) The same plot for the case using images of Mueller matrix as input to train the neural network. (d) Relative error of predicted R_B as a function of its true values. A curve with exponential decay (dashed line) can roughly describe the relation.

The advantage of the neural network for reconstruction is obvious. When a library search with a resolution of 1 nm is expected, datasets of 50625×2 have to be prepared. Using Resnet-A, only 5% of the datasets are needed and using Resnet-B, only 4% of the datasets are needed. Assuming that the forward simulation for each parameter set on a personal computer is about 20 min., depending on the pupil sampling points, the number of Fourier modes with RCWA calculation and on the slices to approximate the grating structure. Time is tremendously saved to prepare the datasets. This also makes the preparation of datasets in a different parameter space easier. Furthermore, the reconstruction time using the trained neural

network is much faster than a library search (for reconstruction of 986 parameter sets, Resnet used about 4 seconds while library search used about 541 seconds, for instance). However, for current studies noises are not considered yet, and the trained Resnets should be tested for measured images, both of which will be done in the following research.

3. PARAMETER RECONSTRUCTION USING LEVENBERG-MARQUARDT ALGORITHM

The Levenberg-Marquardt (LM) method is a standard technique for solving nonlinear least squares problems. It was developed to fit a parameterized function to a set of measured data points by minimizing the sum of the squares of the errors between the measured data points and the function. The LM algorithm is well known for its accuracy and efficiency in parameter minimization. It achieves this by combining and balancing two optimization methods: the gradient descent method (GDM) and the Gauss-Newton method (GNM). The GDM helps in navigating the parameter space by following the steepest descent direction, while the GNM approximates the Hessian matrix to find a more accurate and efficient update direction[7].

Assuming we have a function $F(\vec{x};\vec{p})$ of an independent variable $\vec{x}$ with m measurement points and a vector of $\vec{p}$ with n parameters ($n = 4$ in our case), and we try to fit it to a set of data points to $F(\vec{x}_i;\vec{p}_i)$. For fitting the pupil image, we transform the image, which has 21×21 points, into an array of y with $m = 317$ effective points falling into the circle of NA = 0.7. For the studies using LM algorithm (Exp. II-Exp. IV), only images of the 0th diffraction order at x-polarization are used. For a fast test, only 30 Fourier modes were taken in calling the RCWA calculation. As in the previous studies, no noise was considered for current study and an initial parameter set of $\vec{p} = [1, 1, 1, 1]$ was used. In Figure 6(a) the predicted y_{fit} and the true y_{true} are plotted together. Good agreement over the points is obtained. Figure 6(b) shows one fitting example to illustrate the retrieval process of the parameters. Although the process shown in Figure 6(b) needs only 5 iterations, the number of iterations for other tests can be larger, which might be explained by sensitive studies as described in Ref.8. Nevertheless, with 5 iterations, the forward simulation has been called 18 times.

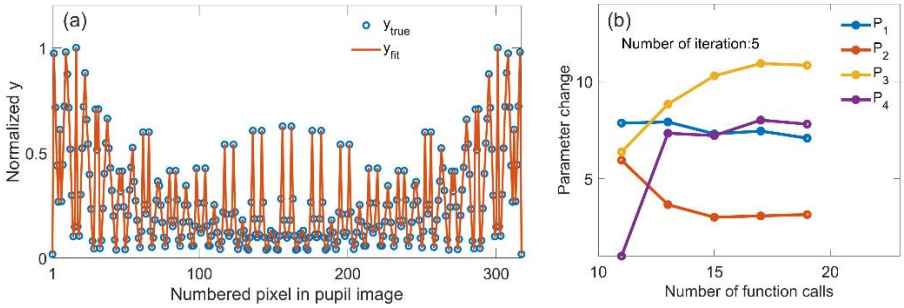


Figure 6. (a) The input pupil image with 21×21 pixels is transformed into an array of y_{true} with 317 points (the points falling into the pupil with NA = 0.7). The y_{fit} is obtained through the LM fitting algorithm. (b) The change of the predicted parameters with iterations for the 19th test shown in Figure 7(b) highlighted by a green arrow.

To explore the statistic performance of this algorithm, 20 parameters sets are randomly chosen as shown in Figure 7(a). Their relative error ($\Delta P/P$) by reconstruction is plotted in Figure 7(b). Contrary to the curves shown in Figure 5(b) and Figure 5(c), parameter R_B has been well reconstructed, also in the case when R_B is smaller than 2 nm. The difference might be attributed to the different regularization procedures of the two methods. With the LM method, curve fitting is applied while with the neural network, the reconstruction is connected to feature extraction.

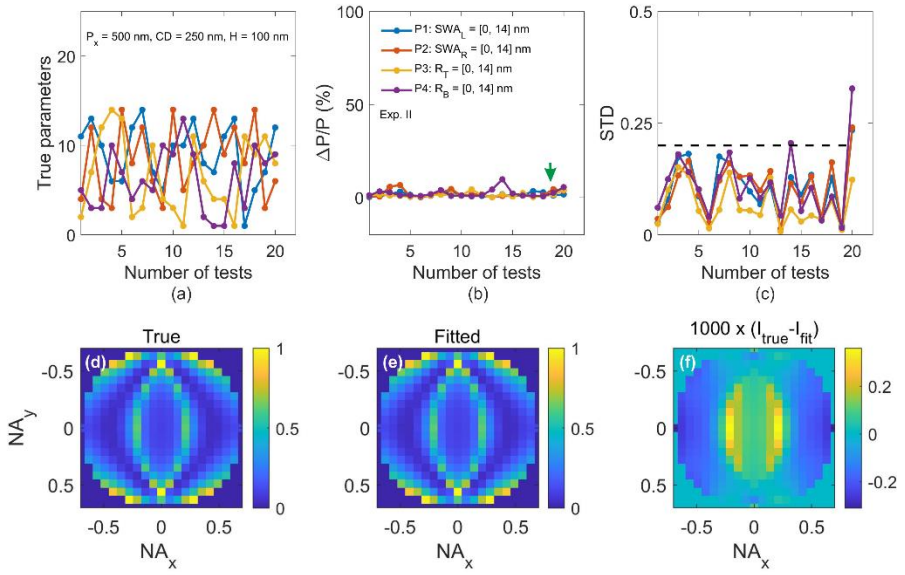


Figure 7. (a) The randomly chosen true parameters in 20 tests with a fixed P_x, H and CD. (b) Relative error of the predicted parameters with respect to the true parameters. (c) Standard deviation of the fitted parameters. (d-f) The 0th order pupil image at the *x*-polarization calculated using the true parameters, the fitted parameters, and the difference of the two images which is multiplied by 1000.

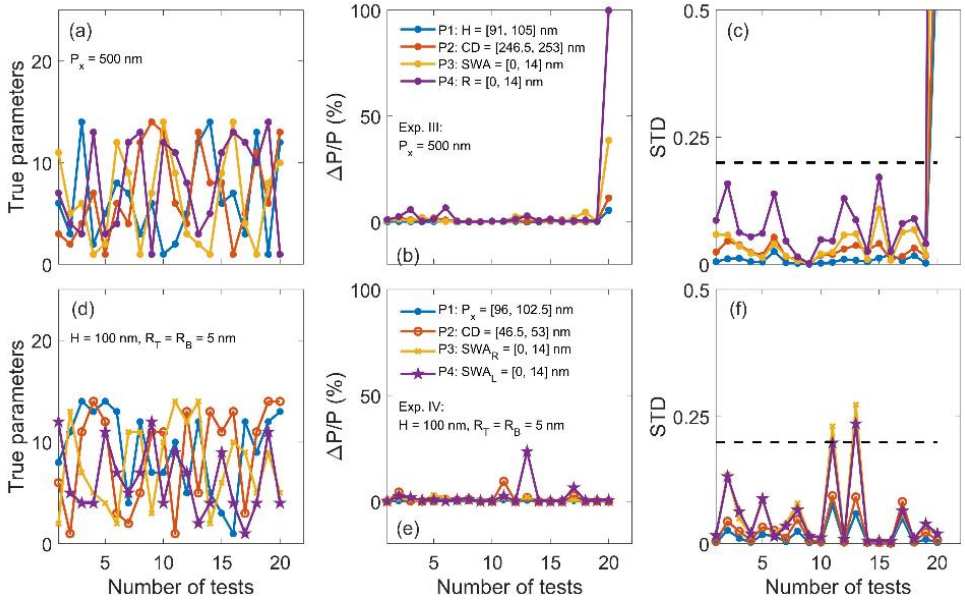


Figure 8. Further reconstruction experiments (Exp. III and Exp. IV) using the LM algorithm. (a-c) are results from a parameter set of [H, CD, SWA, R] with a fixed P_x. (d-f) are results from a parameter set of [P_x, CD, SWA_R, SWA_L] with fixed H=100 nm and R_T= R_B=5 nm. P_x is changed to a scale of 100 nm.

Further reconstruction experiments are performed by using the LM algorithm as shown in Figure 8. Figure 8(a-c) show the results for Exp. III from the grating with P_x = 500 nm. Other structural parameters such as H, CD, SWA (with identical left and right sidewall angles), R (identical top and bottom radius) are varied. Except for the last test, high accuracies are shown through the tests. The reason for the extraordinary behavior of the last point has to be studied in a more detail. Figure 8(d-f) are results for Exp. IV from structures with a periodicity around 100 nm. The grating height is still fixed at 100 nm. The top and bottom roundings have an identical radius fixed at 5 nm. P_x, CD, SWA_R and SWA_L as a parameter

set are retrieved. High accuracy through the tests is also achieved. Comparing the results shown in Figures 8(a-c), we find that the LM algorithm works also well for subwavelength P_x with comparable accuracy for larger P_x.

We have noticed further that using LM algorithm to fit images of Mueller matrix does not improve the performance compared with fitting simply the pupil image at *x*-polarization. Indeed, this observation further highlights the inherent differences between the two methods. The underlying principles of the Levenberg-Marquardt (LM) method and neural network approach underscore their different capabilities in parameter reconstruction. The advantage of the LM algorithm for reconstruction is straightforward: no datasets need to be prepared in advance. Thus, the parameter space to be reconstructed can be changed flexibly.

4. CONCLUSION AND OUTLOOK

Our research demonstrates that a trained neural network built on the Resnet 18 architecture is well-suited for replacing the conventional library search method in parameter reconstruction for Fourier scatterometry. When pupil images are used as input data, not only does it require much fewer datasets (5%) than the conventional approach, but it also offers faster processing, making it more applicable for in-line monitoring applications. Additionally, when training the network using images of the Mueller matrix in the pupil plane, even fewer datasets (4%) are needed. This improvement can be attributed to the high sensitivity of the Mueller matrices to grating parameters.

We also explored the traditional Levenberg-Marquardt (LM) algorithm for parameter reconstruction. Although its accuracy may not be as high as that of the trained neural network (depending on the stop criterion), it achieves convergence within several iterations. Moreover, the LM method offers flexibility as it does not require prepared datasets. However, the LM algorithm is connected to a forward calculation tool through function calls. With the current implementation, the function was called 18 times for 5 iterations. Even if each forward calculation takes only 1 minute, the overall procedure becomes too time-consuming for in-line applications. To accelerate this procedure, we suggest replacing the forward RCWA simulation with a physics-driven network[9]. This substitution would significantly improve the speed of the forward calculation. In anyway, by integrating different techniques into a hybrid metrology, we can expect a more robust and powerful approach for accurate and fast parameter reconstructions.

ACKNOWLEDGMENTS

The authors thank German Research Foundation DFG for supporting the project of FR 3383/2-1.

REFERENCES

[1] Pan, B., “Optical metrology embraces deep learning: keeping an open mind,” Light Sci Appl 11(1), 139 (2022).

[2] Zuo, C., Qian, J., Feng, S., Yin, W., Li, Y., Fan, P., Han, J., Qian, K. and Chen, Q., “Deep learning in optical metrology: a review,” Light Sci Appl 11(1), 39 (2022).

[3] IEEE IRDS., “International roadmap for devices and systems: Metrology,” IEEE Advancing Technology for Humanity (2022 Edition).

[4] Gödecke, M. L., Frenner, K. and Osten, W., “Model-based characterisation of complex periodic nanostructures by white-light Mueller-matrix Fourier scatterometry,” Light: Advanced Manufacturing 2(2), 1 (2021).

[5] “RefractiveIndex.Info.”, https://refractiveindex.info/ (access on July 2023).

[6] Totzeck, M., “Numerical simulation of high-NA quantitative polarization microscopy and corresponding near-fields,” Optik 112(9), 399–406 (2001).

[7] Gavin, H. P., “The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems,” (2022). https://people.duke.edu/~hpgavin/ExperimentalSystems/lm.pdf

[8] Likhachev, D. V., “Parametric sensitivity analysis as an essential ingredient of spectroscopic ellipsometry data modeling: An application of the Morris screening method,” J Appl Phys 126(18), 184901 (2019).

[9] Lim, J. and Psaltis, D., “MaxwellNet: Physics-driven deep neural network training based on Maxwell’s equations,” APL Photonics 7(1), 011301 (2022).