RESEARCH ARTICLE | JANUARY 02 2026

Vision Transformer network for optical overlay metrology on semiconductor wafers

L. de Wolf ; M. Lipp ; M. Cochez ; A. den Boef ; L. V. Amitonova

APL Mach. Learn. 4, 016101 (2026)
https://doi.org/10.1063/5.0301749

Vision Transformer network for optical overlay metrology on semiconductor wafers

Cite as: APL Mach. Learn. 4, 016101 (2026); doi: 10.1063/5.0301749
Submitted: 10 September 2025 • Accepted: 10 December 2025 •
Published Online: 2 January 2026

L. de Wolf,[1,2] M. Lipp,[1,3,a)] M. Cochez,[2,4] A. den Boef,[1,3,5] and L. V. Amitonova[1,3]

AFFILIATIONS

[1] Advanced Research Center for Nanolithography (ARCNL), Amsterdam, The Netherlands
[2] Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[3] Department of Physics and Astronomy, and LaserLaB, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[4] ELLIS Institute Finland & Abo Akademi University, Turku, Finland
[5] ASML Netherlands B.V., Veldhoven, The Netherlands

[a)] **Author to whom correspondence should be addressed:** m.lipp@arcnl.nl

ABSTRACT

Fast and high-precision wafer metrology is critical for the semiconductor industry. In this work, we explore the use of simple and cost-effective optical sensors in combination with data-driven algorithms. We propose and compare three data-driven approaches with varying complexity that can directly infer sub-nanometer metrology parameters from low-numerical-aperture optical coherent microscope images with the focus on precision, noise robustness, and data efficiency. In particular, we apply Vision Transformers (ViTs), Convolutional Neural Networks, and Multilayer Perceptrons to simulated datasets with varying aberrations. We report sub-nanometer measurement accuracy and precision for all models in the presence of strong optical aberrations and noise also. Furthermore, we find that ViTs consistently achieve low errors and excel under limited data regimes compared to other models.

I. INTRODUCTION

Metrology is essential for semiconductor manufacturing processes. The industry increasingly relies on a larger number of smaller metrology marks while simultaneously demanding higher precision and speed. Hence, increasing the information yield from individual measurements and being robust to non-ideal measurement conditions is key to conform to industry demands. Data-driven machine learning is a promising solution to this problem as it allows utilizing complex relation in the data and is potentially correct for experimental noise and aberrations.

Diffraction-based overlay (DBO) metrology is a very promising technique for wafer metrology. μDBO metrology target consists of gratings patterned on successive process layers, as presented in Fig. 1. One mark consists of four grating pairs: two pairs for horizontal and two pairs for vertical overlay measurements. In the following, we will focus on only one direction as the procedure for the orthogonal direction follows analogously. We consider two sizes of μDBO targets: C10 ($10 \times 10\ \mu m^2$) and C16 ($16 \times 16\ \mu m^2$). Under coherent illumination, the intensity difference between +1st and −1st diffraction orders, coming from these two biased grating pairs, is highly sensitive to their relative lateral displacement. This shift-dependent modulation enables accurate extraction of overlay error, making DBO a powerful approach for high-precision, non-destructive in-line metrology.[1–7] More details on this specific configuration are provided in Appendix C.

However, currently optical overlay (OV) metrology techniques face significant challenges.[9–11] To effectively manage a broad range of materials used in modern chip manufacturing, an OV metrology tool should cover a very wide wavelength range from visible to near-infrared wavelengths. Furthermore, high-numerical-aperture (high-NA) imaging is essential to resolve small metrology marks embedded within complex device layouts. However, high-NA optics over a broadband wavelength range inherently suffer from increased sensitivity to aberrations, which, therefore, degrade measurement accuracy.

Various methods have been proposed to mitigate aberrations before determining the OV and mainly rely on optical setup optimizations.[12–18] For instance, spherical aberrations have been compensated using a tube lens in telecentric configuration.[12] Despite the effectiveness of hardware optimization, these methods introduce additional complexity to the optical setup, require precise

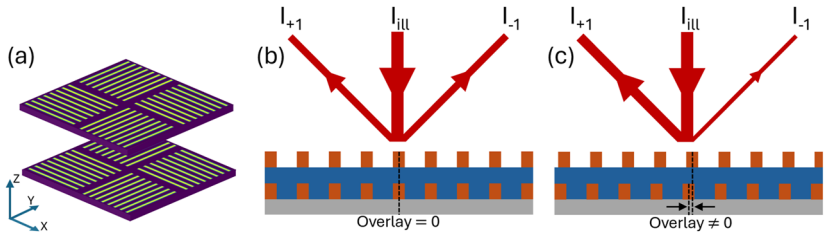

FIG. 1. Illustrations[8] in panel (a) show the 3D view of the μDBO metrology marks printed on consecutive layers of the sample and those in panels (b) and (c) show the cross section of the same marks in a 2D side view. The symmetric (b) and asymmetric (c) scattering is shown for the grating with and without overlay bias, respectively.

calibration, and can limit the flexibility or speed of the imaging system.

Computational imaging is a novel imaging paradigm based on joint optimization of hardware and software. Numerical methods leveraging the full electromagnetic field information to computationally correct a wider range of aberrations have been proposed.[19–21] However, this still requires more complex and stable measurement systems. Another major challenge in DBO metrology is the inherently low diffraction efficiency of the marks. As overlay precision approaches the sub-nanometer scale, the required relative intensity measurement precision reaches the level of 0.01%.[22] Achieving this level of sensitivity demands bright, coherent light sources to ensure a sufficient photon budget. However, such sources introduce coherent noise, which degrades image quality. Several methods have been proposed to suppress coherent noise artifacts. Averaging multiple holograms were obtained through various optical methods (also known as hologram multiplexing),[23–25] mean/median filtering,[26] bandpass Fourier filtering,[27] and window functions.[22,28] However, combining approaches and correcting for stronger, more complex aberrations is difficult. Machine learning could offer a precise and efficient solution without changes to the sensor hardware.

Machine learning tools, specifically Convolutional Neural Network (CNN), have been increasingly popular in optical microscopy.[29–33] Most methods rely on a phase and amplitude measurement to estimate the phase front aberrations, which are computationally removed to retrieve the cleaned image. For example, Ref. 32 used U-Net CNN architecture to predict background regions in distorted phase maps, from which the Zernike coefficients are derived and used to compensate the wavefront aberrations in the frequency domain. Similarly, Ref. 29 also uses a CNN to predict Zernike coefficients, but performs the coefficient extraction and aberration compensation before phase unwrapping. A different method is presented in Ref. 33, where an end-to-end deep learning framework called HRNet is used to reconstruct amplitude, phase, and two-sectional objects. Aberration correction using a ResNet50 module to first identify the different types of aberrations and their coefficients, before using a U-Net module to reconstruct the undistorted image, has been demonstrated in Ref. 30. In Ref. 34, a U-Net architecture with residual connections is trained to reduce speckle noise, using noisy and noise-free DHM phase image pairs. The work of Ref. 35 also uses a CNN to suppress speckle noise from noisy and noise-free image pairs, but applies their model on the wrapped phase and evaluates on both simulations and experimental data. Self-supervised learning has also been applied to coherent noise reduction in digital holography and has the advantage that no noise-free image has to be provided during training. The work of Ref. 36 outlines an algorithm to train a denoising model by maximizing the maximum likelihood estimate of pairs of images with random noise distributions. All these studies aimed to reconstruct the corrected image from raw data, which is not actually needed in the metrology task. Focusing directly on metrology parameters, such as OV, allows for more flexibility, robustness, and speed.

In this work, we focus on an angle-resolved diffraction-based scatterometer for overlay metrology. We aim to determine the OV with the highest precision and speed at the lowest (hardware) cost. Therefore, instead of high-NA scatterometry, we directly image the diffracted light with a low-NA single lens onto a camera. We consider a simple optical microscope that captures only the amplitude image without built-in optical aberration corrections and deploy front-to-end machine learning models to estimate the overlay value directly from the raw data. We make use of the fact that we know the design of the metrology mark and do not need to work with an arbitrary sample. Finally, we go beyond the well-established CNN architecture and develop a new model using the more recent Vision Transformer (ViT). We compare this model to other architectures from the literature on six different simulated datasets, ranging from ideal to strongly distorted images. For each of the models, we perform quantitative hyperparameter tuning and extensive performance comparisons in terms of computational cost, speed, and reconstruction accuracy and precision. We demonstrate overlay accuracy in the order of 10^{-2} nm even for distorted images and identify the Vision Transformer as the most efficient and robust architecture, which still achieves top overlay performance.

II. DATA

To offer a comprehensive analysis of machine learning tools for optical overlay metrology, we consider multiple datasets. It allows for capturing different application scenarios. To enable fast and flexible access to labeled measurement data, we utilize an advanced DBO metrology simulator implemented in MATLAB that uses the beam propagation method.[37]

First, the simulator generates the Gaussian illumination with an 800 nm wavelength and initializes the sample as a stack of two gratings with a pitch of 600 nm. The transmission and reflection coefficients are calculated from the diffraction efficiency of the

grating and the refractive index of the substrate material. The transmitted field is propagated to the bottom grating, taking into account parabolic phase profiles because of defocus. Then, the diffraction from the bottom μDBO target is calculated. Finally, the field is transmitted back through the top grating and superimposed with the diffracted light from the top grating.

The read and shot noise is drawn from $\mathcal{N}(0,1)$ and $\mathcal{N}(0,I)$, respectively, where I is the intensity of the field. In addition, all data are randomly shifted by ±5 px in horizontal and vertical directions to emulate non-perfect sensor alignment.

To be able to use realistic optical aberrations of the metrology sensor in our simulations, aberrations of a custom-designed lens for the OV sensor have been experimentally measured using digital holography.[38] These wavefront aberrations are parameterized using the Zernike polynomials,[39] which allows for the introduction of specific effects such as defocus or μDBO target tilt by choosing distinct values for certain coefficients or the application of arbitrary distortions with random coefficients.

We investigate two commonly used μDBO metrology targets, C10 ($10 \times 10\ \mu m^2$) and C16 ($16 \times 16\ \mu m^2$), and the following three cases for each of the two μDBO targets (six datasets total).

- Wavefront 0 (W0): no wavefront aberration.
- Wavefront 1 (W1): stochastic wavefront aberration.
- Wavefront 2 (W2): constant experimentally measured wavefront aberration.

In Fig. 2, an example from each dataset is shown. W0 corresponds to the nearly ideal measurement case, where high-quality optics without aberrations is used or computational correction is applied prior. In this regime, the images are affected only by relatively low spatial resolution due to the diffraction limit. The W2 datasets contain constant experimentally measured aberrations, corresponding to the realistic case, for example, when the measurements are taken from the same sensor in relatively stable measurement conditions. In contrast, the W1 dataset has random aberrations for each measurement, which corresponds, for example, to highly unstable measurement conditions or when the same model is used across different sensors. Manufacturing errors such as grating pitch variations or line roughness are not considered during simulation but could play a role during the experimental application, specifically for the smaller μDBO target sizes. However, the sensitivity to these parameters is an order of magnitude smaller since the grating size and pitch is of the order of μm while overlay is in nm.

Each of these datasets contains a total of 2048 simulated measurements, where each datapoint is a two-channel 256×256 px dimensional tensor, representing ±1st diffraction orders, and the label is a programmed OV in nanometers drawn from a uniformly distributed $\mathcal{U}(-10, 10)$.

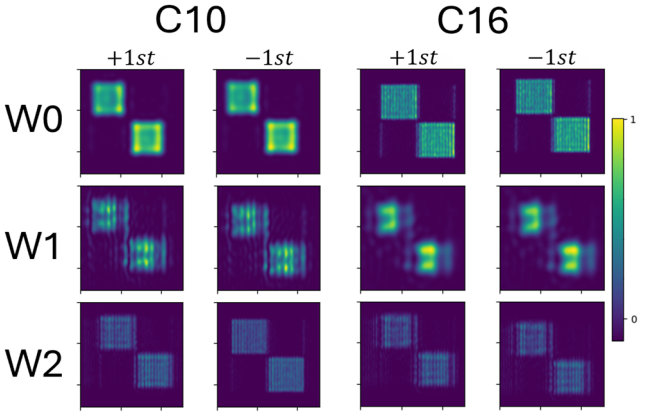

FIG. 2. Examples of image pairs from each dataset. C10 corresponds to the smaller $10 \times 10\ \mu m^2$ μDBO target and C16 to the bigger $16 \times 16\ \mu m^2$ μDBO target. W0, W1, and W2 represent different types of aberrations, as introduced in Sec. with W0 having the least amount of distortions, W1 having different aberration for each data point, and W2 having fixed aberrations across all images in the dataset. The color scale represents the normalized intensity.

III. MODELS

We investigated three established deep learning architectures and designed corresponding models based on the requirements of optical OV metrology. For each class of models, we scale the number of parameters to create differently sized models and consider a total of nine models ranging from 100k to 30M parameters. We scale up all parts of the model evenly to ensure the best operability of every model.

A. Multilayer Perceptron (MLP)

The Perceptron belongs to the oldest machine learning architectures and is universally used in many different applications.[20,40,41] The Multilayer Perceptron is characterized by stacking multiple

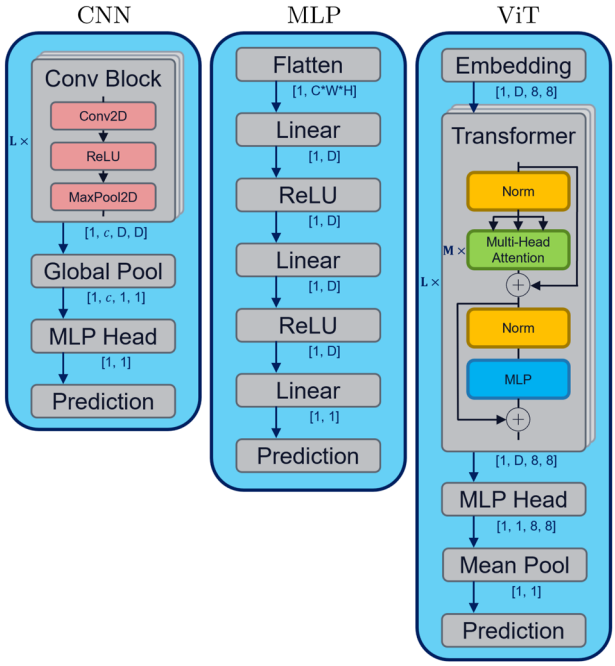

FIG. 3. Illustration of our CNN (left), MLP (middle), and ViT (right) models; also see Tables I–III for parameter choices. The `MLP Head` consists of two dense layers, which are separated by `ReLU` activation for the CNN and `GeLU` activation for the ViT, matching the activation function of their respective architecture.

TABLE I. MLP variations with *D* being the latent dimension.

Name	*D*
mlp_15M	140
mlp_30M	240

dense layers that fully connect all input values to all output values, followed by a non-linear activation function. While its general design allows for applications in many scenarios, its lack of structure also hinders the construction of very deep and large networks.

We construct a simple MLP (as shown in Fig. 3, middle) that consists of three layers with ReLU activation functions. Before the first layer, we convert our input image into a column vector by flattening it. This column vector is then projected to the specified latent dimension *D* by the first layer, continued by a projection that maintains the dimensionality, and then, followed by a final projection to a scalar value. Table I shows our parameter choices.

B. Convolutional Neural Network (CNN)

Convolutional Neural Networks were originally designed for computer vision and are used to extract image features by applying convolutions over (multi-dimensional) data using learnable kernels. Each kernel is thereby associated with an image feature and performs local linear transformations over sliding windows to create a map of this feature, often called a feature map. Since the kernel size is usually much smaller than the image size, CNNs are more parameter-efficient than MLPs.

Our CNN stacks 3–5 `Conv Blocks`, which halve the spatial resolutions of the input and doubles the amount of feature maps after each block. A `Conv Block` applies a two-dimensional convolution, a ReLU activation function, and a two-dimensional max-pooling operation. Convolutions use a kernel size of 3 and add 1 padding pixel near the borders of the inputs, which ensures spatial dimensionalities are retained after this operation and the max pooling operation halves the spatial resolutions at the end of the block. After the `Conv Blocks`, we apply average pooling over the remaining spatial dimensions and use an `MLP head` for prediction, which contains two dense layers with `ReLU` activation. The CNN architecture is shown in the left panel of Fig. 3, and our parameter choices are shown in Table II.

C. Vision Transformer (ViT)

Vision Transformer is an adaptation of the Transformer[42] model for the application in Computer Vision. Analogously to the original Transformer, which was developed for textual data and handles sentences as sequences of tokens, the Vision Transformer treats images as a sequence of non-overlapping patches. These patches are embedded into *D*-dimensional vectors (tokens) but first a linear transform and then `learned position embeddings`,[43] to serve as input to a conventional Transformer. The Transformer remains unchanged and consists of multiple layers, which contain a multi-head self-attention block and an MLP block. To stabilize training, layer normalization is applied before each component and residual connections are added after each component.

Our implementation of the ViT uses non-overlapping patches that are 32 × 32 px in size and 4–10 Transformers with `GeLU` activations. We then project the output to a single feature map by using an `MLP Head`, which consists of two dense layers with a `GeLU` activation function. Finally, we apply average pooling over the remaining feature map to obtain a single value. An overview of the ViT architecture is shown in Fig. 3 (right), and our parameter choices are shown in Table III.

TABLE II. CNN variations with *L* being the number of `Conv Blocks`, *c* being the number of feature maps, and *D* being the dimension of the feature maps.

name	*L*	*c*	*D*	MLP head dim.
conv_100k	3	128	32	256
conv_500k	4	256	16	512
conv_2M	5	512	8	1024

TABLE III. ViT variations, with *D* indicating the latent dimension, *M* being the number of Heads, and *L* being the number of Transformer layers.

Name	*D*	*M*	*L*	MLP head dim.
vit_500k	96	2	4	192
vit_1M	120	4	6	240
vit_5M	240	6	8	720
vit_15M	360	8	10	1440

D. Training

We train all our models from scratch for 250 epochs, minimizing the mean squared error (MSE) between predicted and ground truth overlay values. We split each dataset in 1024 data points for training and for evaluation. We update our weights using the standard configuration for the AdamW optimizer, which we combine with a cosine annealing learning rate scheduler. For best comparability, we perform hyper parameter scanning for each model individually and compare performance with their respective optical parameters (see Appendix A). In addition, we clip gradients for weights with a magnitude higher than one and use mixed-precision training to prevent exploding gradients and reduce memory overhead, respectively. All models were implemented in Python using the PyTorch (Lightning) library and trained on NVIDIA RTX 2070 Super and/or NVIDIA A100 GPUs. For all models, the training time per epoch is roughly ~1 s and the entire training is complete in less than 5 min (see Appendix D).

IV. RESULTS

A. Full dataset

In the first set of experiments, we train each model on each of the datasets separately and then evaluate its performance. Figure 4 presents the average MAE across all models on all test datasets, with the error bars showing the variance over five repetitions. The MAE directly gives the average difference between the predicted and true overlay values in nanometers and stays significant sub-nanometer

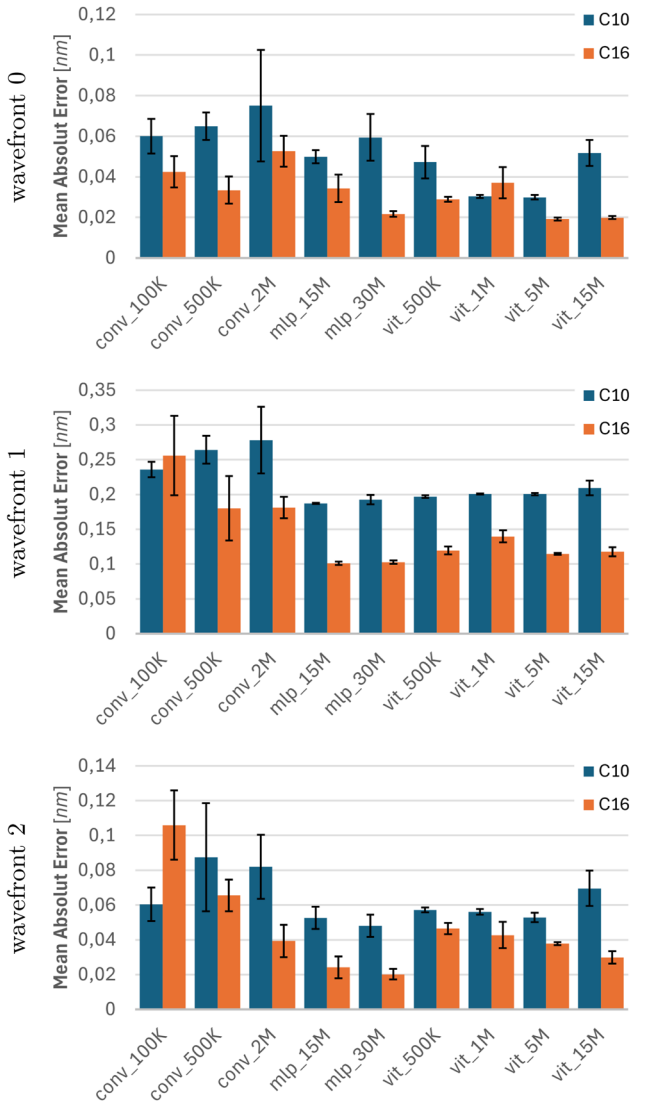

FIG. 4. OV reconstruction error for all model and dataset combinations in nanometer. The error bar shows the standard deviation over five repetitions. The name of the model is written below each bar, with Vision Transformer as vit, Multilayer Perceptron as mlp, and Convolutional Neural Network as conv. The specifications for each model can then be found in Table III for all Vision Transformers, in Table II for all Convolutional Neural Networks, and in Table I for all Multilayer Perceptrons.

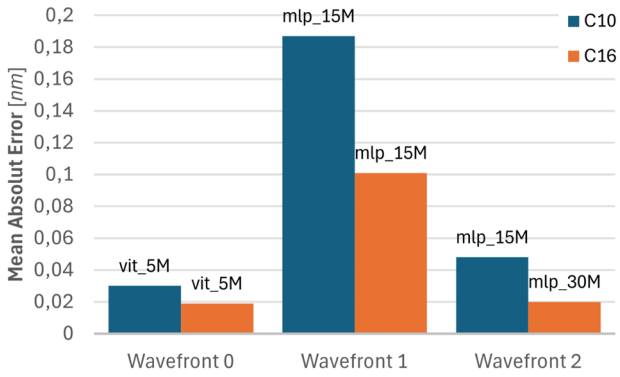

FIG. 5. Mean absolute error in nanometer of the best performing model for each datasets. Wavefront 0, 1, and 2 represent different types of aberrations, and C10 and C16 correspond to the $10 \times 10\ \mu m^2$ and $16 \times 16\ \mu m^2$ μDBO targets, respectively. The name of the best performing model is written above each bar, with Vision Transformer as vit, Multilayer Perceptron as mlp, and Convolutional Neural Network as conv. The specifications for each model can then be found in Table III for all Vision Transformers, in Table II for all Convolutional Neural Networks, and in Table I for all Multilayer Perceptrons.

in all cases. This shows the great potential of AI powered overlay reconstruction in general.

First, comparing the effect of wavefront 0, 1, and 2 scenarios on the performance of all models reveals a significantly higher reconstruction error for the random W1 case, while W0 and W2 stay mostly on the same level, although the W2 dataset is exposed to strong aberrations in contrast to the W0 dataset. This shows that the models can completely compensate for the effect of the wavefront aberrations if they remain constant. The performance for the W0 case, especially for the C16 μDBO target, fluctuates randomly because of the extreme uniformity of the dataset.

Next, comparing the overall performance between the larger, less-distorted C16 μDBO target and the smaller, more-volatile C10 μDBO target shows an increased error for the C10 μDBO target. These results are summarized in Fig. 5, which shows the MAE of the best models for each of the μDBO targets per wavefront aberration.

Finally, we compare the different models and find that convolutional models perform worse than ViT and MLP models in general. It has the largest MAE and fluctuates the most among all models. ViT and MLP models perform similarly, while the ViT models contain fewer parameters and scale better for larger input sizes. In addition, the MLP models mainly excel in easy scenarios, namely, the C16 μDBO target in the W0 and W1 cases. In the more challenging C10 and W2 cases, the performance of the ViT and MLP is nearly identical. Concerning the number of parameters per model, the C16 μDBO target generally benefits from larger models, while the C10 μDBO target prefers smaller models.

B. Limited data budget

In the second set of experiments, we train the models on a randomly selected subset of the training dataset and then evaluate their performance. Since, in practice, experimentally measured labeled data are usually sparse, the performance of most ML applications is limited by the size of the training dataset. However, some model architectures are more data-efficient than others, which makes them more suitable for certain applications. Here, we investigate how the performance and OV accuracy of the models scale with training dataset size.

We limit the amount of training data logarithmically between 512 and 32 and compare the models on the wavefront 1 and 2 datasets with the C16 and C10 μDBO targets, as we already witnessed similar performance between the wavefront 0 and wavefront 2 datasets in Sec. IV A. Figure 6 shows the MAE of each model

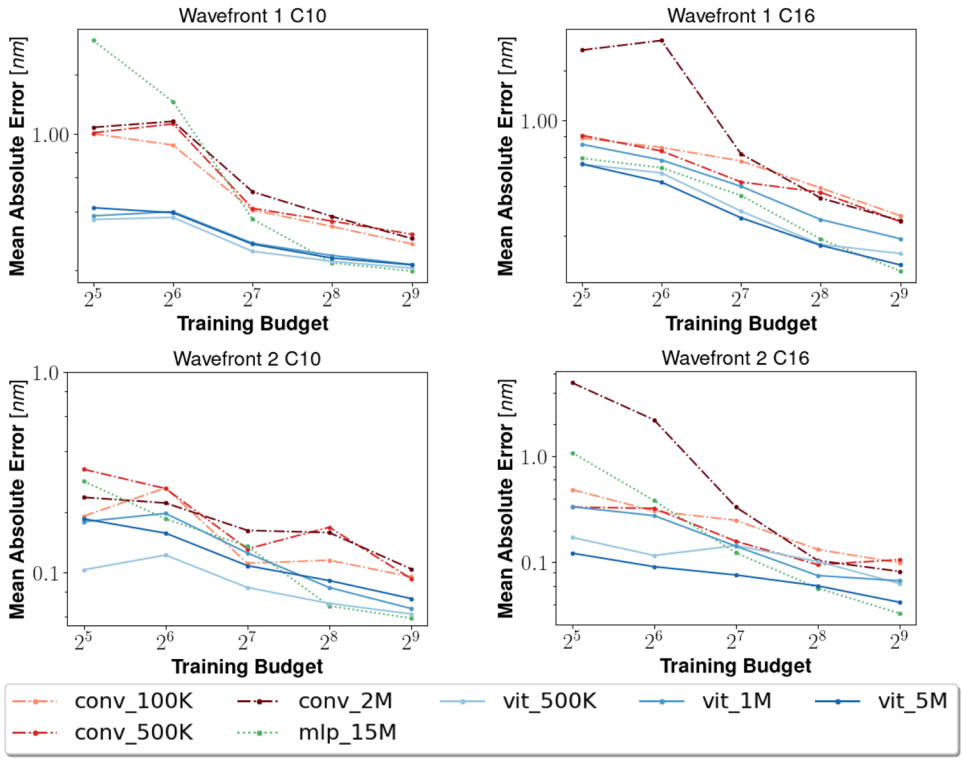

FIG. 6. OV reconstruction error for different training budgets in double log scale. Each model is visualized using a different color, with Vision Transformer as vit (blue), Multilayer Perceptron as mlp (green), and Convolutional Neural Network as conv (red). The specifications for each model can then be found in Table III for all Vision Transformers, in Table II for all Convolutional Neural Networks, and in Table I for all Multilayer Perceptrons. Some ViT and MLP models are removed from this plot because their training would not converge for these small budgets. The color of the plot represents the architecture and the darker the color, the more parameters the model has.

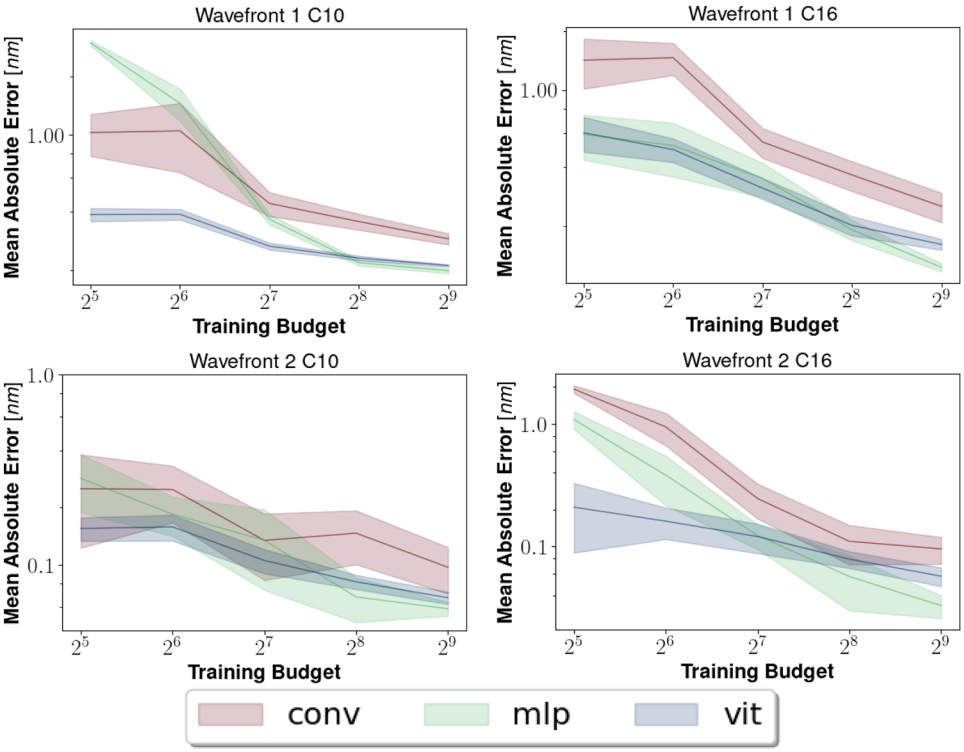

FIG. 7. Plotting the averaged OV error and averaged standard deviation in a double log plot. The largest ViT and MLP models are removed from this plot because their training would not converge for these small budgets.

except for the largest ViT and MLP models, which did not converge reliably during training with this reduced number of training points. The corresponding trend that smaller models perform better on a smaller training budget than their larger counterparts can be observed for all models, except for the ViT where the performance of all models is comparable. We also observe in general that models benefit from more training data, which not only reduces the average error achieved but also decreases the variance of the results between models that share the same architecture and model size (see Fig. 7).

Overall, we are able to sustain sub-nanometer precision on average across datasets, while the ViTs, especially the smaller-sized models, outperform the CNNs and MLPs in terms of reconstruction accuracy and consistency for a limited data budget. Smaller-sized models perform better compared to their larger counterparts, which, combined with their reduced computational cost, makes them the more efficient choice for real-life implementations.

V. CONCLUSION

To summarize, we have conducted a quantitative comparison of several deep learning architectures for μDBO metrology using different sets of simulated optical metrology data. All models show sub-nanometer precision even in the presence of experimental noise and strong wavefront aberrations, while the best models achieve precisions of even single percents of a nanometer. We identified Vision Transformers as the leading architecture because of their good performance, scalability, and resilience to aberration. The ViT achieves leading OV accuracy in almost all test cases. It achieves consistent sub-nanometer precision over a wide range of training budgets and model sizes, allowing for maximal flexibility. Finally, the ViT offers the best computing cost efficiency and scalability for larger metrology input data and a higher number of extracted parameters. The reconstruction of key metrology parameters directly from strongly distorted images could reduce the demand on the quality of the optical sensor and allow for faster and less costly measurement systems while increasing the wafer metrology accuracy.

ACKNOWLEDGMENTS

This work was conducted at the Advanced Research Center for Nanolithography, a public-private partnership between the University of Amsterdam, Vrije Universiteit Amsterdam, University of Groningen, the Netherlands Organization for Scientific Research (NWO), and the semiconductor equipment manufacturer ASML and was partly financed by a contribution from the National Growth Fund program NXTGEN HIGHTECH through the "(Nano) Metrology Systems" project.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

L. de Wolf and M. Lipp contributed equally to this work.

L. de Wolf: Data curation (equal); Formal analysis (lead); Methodology (lead); Software (lead); Validation (equal); Writing – original draft (supporting); Writing – review & editing (equal). **M. Lipp**: Conceptualization (lead); Data curation (equal); Methodology (supporting); Software (supporting); Supervision (equal); Validation (equal); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **M. Cochez**: Project administration (equal); Supervision (equal); Writing – review & editing (equal). **A. den Boef**: Methodology (equal); Supervision (equal); Writing – original draft (equal). **L. V. Amitonova**: Funding acquisition (lead); Project administration (equal); Resources (lead); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The Python code for the neural network models implemented in this paper is available on GitHub: https://github.com/MLippARCNL/Vision-Transformer-for-Optical-Wafer-Metrology.

APPENDIX A: HYPER-PARAMETER TUNING

To ensure equal evaluation conditions, we conduct extensive hyper-parameter search for all model variants, optimizing both the learning rate and weight decay. We refrain from optimizing for all datasets and instead lay our focus on the W0 since optimizing on the other wavefront aberrations may result in over-specific hyper-parameter configurations. In addition, by only focusing on the aberration-free case, we allow ourselves to assess whether the hyper parameters generalize well to different types of distortions. However, we do differentiate between target sizes and search for hyper parameters for both the C10 and C16 targets separately since early testing indicated significant differences between target sizes. Optimization was performed using Bayesian optimization; more specifically, we used tree-structured Parzen estimaton (TPE) algorithm, as implemented in Optuna. For each trial, we train each model for 75 epochs on random subsets (75%) of the training data and use the rest for validation. Final parameter configurations were determined after 25 trials.

APPENDIX B: MODEL VARIANCE WITH LIMITED TRAINING BUDGET

Figure 7 shows the trend of the averaged spread of the MAE when changing the training budget for all three architectures, both targets and the W1 and W2 datasets. As expected, the spread and MAE generally increases for smaller budgets and for the W1 compared to the W2 dataset. The different scales of the y axis and the logarithmic scaling makes it challenging to precisely compare the graphs, but the ViT model generally exhibits the lowest error and smallest spread. This matches the analysis in Sec. IV B.

APPENDIX C: CONVENTIONAL OVERLAY APPROACH

In order to explain the signal formation in DBO (e.g., μDBO), we use a simple plane-wave propagation model. A schematic drawing of the signal formation is shown in Figs. 1 and 8. The metrology target consists of two stacked gratings, whose ±1st diffraction orders are captured by a camera. Figure 1 shows that the intensities of both orders are equal, if both gratings are aligned, e.g., zero overlay. However, in case of a small shift between the gratings, an asymmetry in

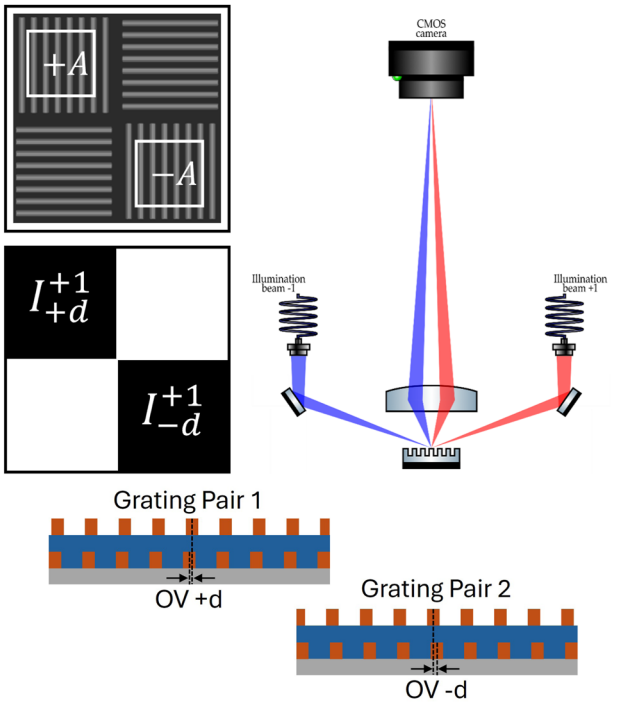

FIG. 8. On the left, the metrology mark with two regions of interest and a schematic of the retrieved +1st diffraction order are shown; on the right, the DBO experimental setup is sketched, and below, one can see the side view of two pads with programmed overlay.[38]

the intensities ΔI arises, which scales linearly with the overlay value OV,

$$\Delta I = I^{+1} - I^{-1} = K \times OV, \tag{C1}$$

where I^{+1}, I^{-1} are the intensities of the ±1st diffraction orders and K is the unknown sample constant.

To remove this unknown constant from the equation, we consider two pairs of grating with programmed overlay $\pm d$, such that

$$\Delta I_{-d} = I^{+1}_{-d} - I^{-1}_{-d} = K \times (OV - d), \tag{C2}$$

$$\Delta I_{+d} = I^{+1}_{+d} - I^{-1}_{+d} = K \times (OV + d). \tag{C3}$$

Combining both equations and solving for OV yields

$$OV = d\left(\frac{\Delta I_{+d} + \Delta I_{-d}}{\Delta I_{+d} - \Delta I_{-d}}\right). \tag{C4}$$

Figure 8 shows top and side views of the resulting DBO metrology target and a schematic of the expected +1st diffraction order with indicated regions of interest as well as the measurement setup.

APPENDIX D: RUN TIME EVALUATION

Figure 9 shows a comparison between the run times of our models on our computing infrastructure (NVIDIA RTX 2070 with 8 GB VRAM, 128 GB RAM). For the measurement, we disabled logging, excluded the data loading time, and averaged over 100 epochs with 16 batches of 64 images. The time is remeasured in ms and plotted with the logarithmical scale to highlight the differences for shorter run times.

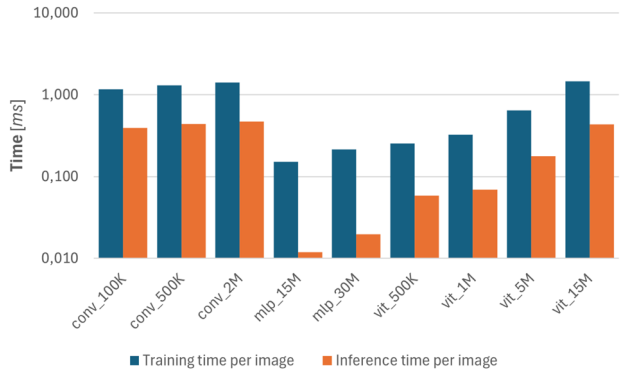

FIG. 9. Time for training and evaluation per image in batch mode with the y-axis plotted in ms in the logarithmical scale.

It is directly visible that inference is faster than training for all models, which is expected since during inference the backpropagation is not calculated. Similarly, larger models have longer run times than their smaller versions for the same architecture, which is also expected.

The CNN architecture has the slowest times compared to the other models, specifically when taking into account their small number of parameters. However, they scale best for larger image sizes since the kernel size, and therefore, the number of parameters of a convolutional layer, does not depend on it.

In contrast, the MLP network has the fastest times, which is mainly caused by the small pixel size we considered during this study. Because a dense layer connects all inputs and outputs, the number of parameters scales exponentially with the input size. Hence, in terms of run time and computational cost, MLPs are only applicable for relatively small sizes.

In between both previous architectures, the ViT combines fast execution, good performance, and good scalability in terms of image size as well as number of training images. As such, it is the most versatile contender with the option to adjust its size to the specific requirement.

REFERENCES

[1] A. J. den Boef, "Optical wafer metrology sensors for process-robust cd and overlay control in semiconductor device manufacturing," Surf. Topogr.: Metrol. Prop. **4**(2), 023001 (2016).

[2] P. Leray, D. Laidler, S. Cheng, M. Coogans, A. Fuchs, M. Ponomarenko, M. van der Schaar, and P. Vanoppen, "Achieving optimum diffraction based overlay performance," Proc. SPIE **7638**, 76382B (2010).

[3] M. Adel, D. Kandel, V. Levinski, J. Seligson, and A. Kuniavsky, "Diffraction order control in overlay metrology: A review of the roadmap options," Proc. SPIE **6922**, 692202 (2008).

[4] I. Abdulhalim, M. Adel, M. Friedmann, and M. Faeyrman, "Periodic patterns and techniques to control misalignment between two layers," US Patents No. 7,656,528 (February 2, 2010).

[5]Y. Xu and I. Abdulhalim, "Periodic patterns and techniques to control misalignment between two layers," U.S. Patents No. 6,483,580 (November 19, 2002).
[6]J. Bischoff, R. Brunner, J. J. Bauer, and U. Haak, "Light-diffraction-based overlay measurement," Proc. SPIE **4344**, 222–233 (2001).
[7]K. Bhattacharyya, C.-M. Ke, G.-T. Huang, K.-H. Chen, H.-J. H. Smilde, A. Fuchs, M. Jak, M. van Schijndel, M. Bozkurt, M. van der Schaar, S. Meyer, M. Un, S. Morgan, J. Wu, V. Tsai, F. Liang, A. den Boef, P. ten Berge, M. Kubis, C. Wang, C. Fouquet, L. G. Terng, D. Hwang, K. Cheng, T. S. Gau, and Y. C. Ku, "On-product overlay enhancement using advanced litho-cluster control based on integrated metrology, ultra-small DBO targets and novel corrections," Proc. SPIE **8681**, 868104 (2013).
[8]C. Messinis, "Dark-field digital holographic microscopy for advanced semiconductor metrology," Ph.D. thesis (Research and Graduation Internal, Vrije Universiteit, Amsterdam, 2022).
[9]L. Miccio, D. Alfieri, S. Grilli, P. Ferraro, A. Finizio, L. De Petrocellis, and S. D. Nicola, "Direct full compensation of the aberrations in quantitative phase microscopy of thin objects by a single digital hologram," Appl. Phys. Lett. **90**(4), 041104 (2007).
[10]J. Min, B. Yao, P. Gao, B. Ma, S. Yan, F. Peng, J. Zheng, T. Ye, and R. Rupp, "Wave-front curvature compensation of polarization phase-shifting digital holography," Optik **123**(17), 1525–1529 (2012).
[11]G. E. Moore, "Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, April 19, 1965, pp.114 ff," IEEE Solid-State Circuits Soc. Newsl. **11**(3), 33–35 (2006).
[12]E. Sánchez-Ortiga, P. Ferraro, M. Martínez-Corral, G. Saavedra, and A. Doblas, "Digital holographic microscopy with pure-optical spherical phase compensation," J. Opt. Soc. Am. A **28**(7), 1410–1417 (2011).
[13]A. Stadelmaier and J. H. Massig, "Compensation of lens aberrations in digital holography," Opt. Lett. **25**(22), 1630–1632 (2000).
[14]C. J. Mann, L. Yu, C.-M. Lo, and M. K. Kim, "High-resolution quantitative phase-contrast microscopy by digital holography," Opt. Express **13**(22), 8693–8698 (2005).
[15]W. Qu, C. O. Choo, V. R. Singh, Yu. Yingjie, and A. Asundi, "Quasi-physical phase compensation in digital holographic microscopy," J. Opt. Soc. Am. A **26**(9), 2005–2011 (2009).
[16]J. Garcia-Sucerquia, "Noise reduction in digital lensless holographic microscopy by engineering the light from a light-emitting diode," Appl. Opt. **52**(1), A232–A239 (2013).
[17]C. Messinis, T. T. M. van Schaijk, N. Pandey, V. T. Tenner, S. Witte, J. F. de Boer, and A. den Boef, "Diffraction-based overlay metrology using angular-multiplexed acquisition of dark-field digital holograms," Opt. Express **28**(25), 37419–37435 (2020).
[18]T. T. M. van Schaijk, C. Messinis, N. Pandey, A. Koolen, S. Witte, J. F. de Boer, and A. Den Boef, "Diffraction-based overlay metrology from visible to infrared wavelengths using a single sensor," J. Micro/Nanopatterning, Mater., Metrol. **21**(01), 014001 (2022).
[19]T. Colomb, J. Kühn, F. Charrière, C. Depeursinge, P. Marquet, and N. Aspert, "Total aberrations compensation in digital holographic microscopy with a reference conjugated hologram," Opt. Express **14**(10), 4300–4306 (2006).
[20]P. Ferraro, S. De Nicola, A. Finizio, G. Coppola, S. Grilli, C. Magro, and G. Pierattini, "Compensation of the inherent wave front curvature in digital holographic coherent microscopy for quantitative phase-contrast imaging," Appl. Opt. **42**(11), 1938–1946 (2003).
[21]T. Colomb, E. Cuche, F. Charrière, J. Kühn, N. Aspert, F. Montfort, P. Marquet, and C. Depeursinge, "Automatic procedure for aberration compensation in digital holographic microscopy and applications to specimen shape compensation," Appl. Opt. **45**(5), 851–863 (2006).
[22]C. Messinis, M. Adhikary, T. Cromwijk, T. T. M. van Schaijk, S. Witte, J. F. de Boer, and A. den Boef, "Pupil apodization in digital holographic microscopy for reduction of coherent imaging effects," Opt. Continuum **1**(5), 1202–1217 (2022).
[23]F. Pan, W. Xiao, S. Liu, F. J. Wang, L. Rong, and R. Li, "Coherent noise reduction in digital holographic phase contrast microscopy by slightly shifting object," Opt. Express **19**(5), 3862–3869 (2011).
[24]L. Rong, W. Xiao, F. Pan, S. Liu, and R. Li, "Speckle noise reduction in digital holography by use of multiple polarization holograms," Chin. Opt. Lett. **8**, 653–655 (2010).
[25]W. Xiao, J. Zhang, L. Rong, F. Pan, S. Liu, F. Wang, and A. He, "Improvement of speckle noise suppression in digital holography by rotating linear polarization state," Chin. Opt. Lett. **9**, 060901 (2011).
[26]J. Garcia-Sucerquia, J. A. H. Ramírez, and D. V. Prieto, "Reduction of speckle noise in digital holography by using digital image processing," Optik **116**(1), 44–48 (2005).
[27]J. Maycock, B. M. Hennelly, J. B. McDonald, Y. Frauel, A. Castro, B. Javidi, and T. J. Naughton, "Reduction of speckle in digital holography by discrete Fourier filtering," J. Opt. Soc. Am. A **24**(6), 1617–1622 (2007).
[28]Y. Morimoto, M. Toru, M. Fujigaki, and N. Kawagishi, "Subnanometer displacement measurement by averaging of phase difference in windowed digital holographic interferometry," Opt. Eng. **46**(2), 025603 (2007).
[29]L. Huang, J. Tang, L. Yan, J. Chen, and B. Chen, "Wrapped phase aberration compensation using deep learning in digital holographic microscopy," Appl. Phys. Lett. **123**(14), 141109 (2023).
[30]Z. Lin, S. Jia, Y. C. Xu, B. Wen, H. Zhang, L. Wang, and M. Han, "Fast phase distortion identification and automatic distortion compensated reconstruction for digital holographic microscopy using deep learning," Opt. Lasers Eng. **185**, 108718 (2025).
[31]S. Ma, R. Fang, Y. Luo, Q. Liu, S. Wang, and X. Zhou, "Phase-aberration compensation via deep learning in digital holographic microscopy," Meas. Sci. Technol. **32**(10), 105203 (2021).
[32]T. Nguyen, V. Bui, V. Lam, C. B. Raub, L.-C. Chang, and G. Nehmetallah, "Automatic phase aberration compensation for digital holographic microscopy based on deep learning background detection," Opt. Express **25**(13), 15043–15057 (2017).
[33]Z. Ren, Z. Xu, and E. Y. Lam, "End-to-end deep learning framework for digital holographic reconstruction," Adv. Photonics **1**(1), 016004 (2019).
[34]W. Jeon, W. Jeong, K. Son, and H. Yang, "Speckle noise reduction for digital holographic images using multi-scale convolutional neural networks," Opt. Lett. **43**(17), 4240–4243 (2018).
[35]K. Yan, L. Chang, M. Andrianakis, V. Tornari, and Y. Yu, "Deep learning-based wrapped phase denoising method for application in digital holographic speckle pattern interferometry," Appl. Sci. **10**(11), 4044 (2020).
[36]D. Yin, Z. Gu, Y. Zhang, F. Gu, S. Nie, S. Feng, J. Ma, and C. Yuan, "Speckle noise reduction in coherent imaging based on deep learning without clean data," Opt. Lasers Eng. **133**, 106151 (2020).
[37]J. Van Roey, J. Van der Donk, and P. E. Lagasse, "Beam-propagation method: Analysis and assessment," J. Opt. Soc. Am. **71**(7), 803–810 (1981).
[38]T. van Gardingen-Cromwijk, S. G. J. Mathijssen, M. Noordam, S. Witte, J. F. de Boer, and A. den Boef, "Enhancing diffraction-based overlay metrology capabilities in digital holographic microscopy using model-based signal separation," J. Micro/Nanopatterning, Mater., Metrol. **23**(4), 044006 (2024).
[39]S. van Haver, W. M. J. Coene, K. D'havé, N. Geypen, P. van Adrichem, L. de Winter, A. J. E. M. Janssen, and S. Cheng, "Wafer-based aberration metrology for lithographic systems using overlay measurements on targets imaged from phase-shift gratings," Appl. Opt. **53**(12), 2562–2582 (2014).
[40]I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. P. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Curran Associates, Inc., 2021).
[41]Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learn. Res. **3**, 1137–1155 (2003).
[42]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need (2023).
[43]A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale (2021).