Machine Vision and Applications (2022) 33:56
https://doi.org/10.1007/s00138-022-01314-w

ORIGINAL PAPER

Depth estimation from a single SEM image using pixel-wise fine-tuning with multimodal data

Tim Houben[1,2] · Thomas Huisman[2] · Maxim Pisarenco[2] · Fons van der Sommen[1] · Peter H. N. de With[1]

Received: 12 January 2022 / Revised: 21 May 2022 / Accepted: 15 June 2022 / Published online: 1 July 2022

Abstract

To support the ongoing size reduction in integrated circuits, the need for accurate depth measurements of on-chip structures becomes increasingly important. Unfortunately, present metrology tools do not offer a practical solution. In the semiconductor industry, critical dimension scanning electron microscopes (CD-SEMs) are predominantly used for 2D imaging at a local scale. The main objective of this work is to investigate whether sufficient 3D information is present in a single SEM image for accurate surface reconstruction of the device topology. In this work, we present a method that is able to produce depth maps from synthetic and experimental SEM images. We demonstrate that the proposed neural network architecture, together with a tailored training procedure, leads to accurate depth predictions. The training procedure includes a weakly supervised domain adaptation step, which is further referred to as pixel-wise fine-tuning. This step employs scatterometry data to address the ground-truth scarcity problem. We have tested this method first on a synthetic contact hole dataset, where a mean relative error smaller than 6.2% is achieved at realistic noise levels. Additionally, it is shown that this method is well suited for other important semiconductor metrics, such as top critical dimension (CD), bottom CD and sidewall angle. To the extent of our knowledge, we are the first to achieve accurate depth estimation results on real experimental data, by combining data from SEM and scatterometry measurements. An experiment on a dense line space dataset yields a mean relative error smaller than 1%.

Keywords SEM images · Scatterometry · Optical critical dimension · Monocular depth estimation · Domain adaptation · Weakly supervised learning

1 Introduction

In the semiconductor industry, critical dimension scanning electron microscopes (CD-SEMs) are used to measure the spatial lateral dimensions of structures on a microchip. These measurements are important for controlling the fabrication process, which enables yield optimization of a produced wafer. Currently, SEM is the fastest way of measurement that provides local geometry information. However, the obtained SEM images are a two-dimensional (2D) representation of the electron interactions with the surface. In practice, detailed metrology that provides the true 3D geometry of this structure is desired for various reasons. It is expected that 3D metrology will become crucial in the semiconductor industry's quest to keep up with the requirements of Moore's Law [1].

Depth estimation from 2D images has been studied thoroughly in the field of computer vision [2] and is nowadays applied to robotics [3], autonomous driving [4], medical imaging [5] and many other scene understanding tasks. Traditionally, these techniques relied on stereo pairs of input images [2], but more recently the subfield of monocular depth estimation has emerged [6]. Here, the depth estimation task is constrained with a single image available per scene during

✉ Tim Houben
t.houben@tue.nl

Thomas Huisman
thomas.huisman@asml.com

Maxim Pisarenco
maxim.pisarenco@asml.com

Fons van der Sommen
fvdsommen@tue.nl

Peter H. N. de With
p.h.n.de.with@tue.nl

[1] Eindhoven University of Technology, Eindhoven, The Netherlands

[2] ASML Netherlands B.V., Veldhoven, The Netherlands

the inference phase. This paper concentrates on performing depth estimation on SEM data to analyze and predict the semiconductor's surface.

Monocular depth estimation is challenging, as it is an ill-posed problem. This challenge results from the fact that multiple 3D scenes can be projected onto the same 2D scene. Currently, many state-of-the-art modeling techniques heavily rely on deep neural networks [7]. These models can perform inference on various types of data, by setting up a high-dimensional non-linear regression or classification problem. Deep neural networks have been applied to many computer vision tasks such as image classification, object detection and semantic segmentation, achieving remarkable results. One reason for these results is the networks' ability to understand a geometric configuration by not only taking local cues into consideration, but also by employing global context such as shape or layout of the scene, which is extremely helpful for solving non-trivial computer vision problems.

Neural networks require large-scale datasets with (manually) annotated ground-truth labels, which can be a difficult operation. In the case of monocular depth estimation from SEM images, the ground-truth data can only be obtained from other sources, such as atomic force microscopy (AFM) [8], transmission electron microscopy (TEM) [9] and scatterometry, also referred to as optical critical dimension (OCD) metrology [10]. The first two sources provide highly accurate and local depth information. However, they commonly provide data in one dimension and are notoriously slow and labour-intensive. Alternatively, OCD metrology is extremely fast, much faster than SEM, but provides measurements averaged over a larger area on the wafer, typically 25 μm^3 or more.

One possibility to circumvent the labeling problem is to generate a synthetic dataset, containing representative geometries, with an electron scattering simulator. Open source implementations based on Monte-Carlo methods are currently available [11] and provide highly accurate simulations of propagating electrons through a material. However, the results of these simulations are not fully accurate. The electron beam and the detector are simplistically approximated, which negatively impacts the image quality. Thereby, physical phenomena, like electron-beam-induced charging and damage, are excluded, while models for the generation of so-called secondary electrons are hard to validate. Therefore, this approach forms only one part of the solution. A second training step is required where the model will be adapted to experimental (real) data.

Machine learning can be a helpful tool for deriving the above models. Domain adaptation is a sub-field of machine learning, where the goal is to maximize prediction performance on a target domain without (complete) labels, with the help of a related and well-labeled source domain, while the prediction task in both domains is identical [12]. In this case, we have the sole availability of coarse-grained labels in the target domain (average depth from OCD), so we can classify this as a weakly supervised domain adaptation problem. More specifically, the goal is to fine-tune a pre-trained network with a limited set of experimental SEM data, paired with OCD metrology measurements. For doing so, an accurately aligned dataset of these modalities is required.

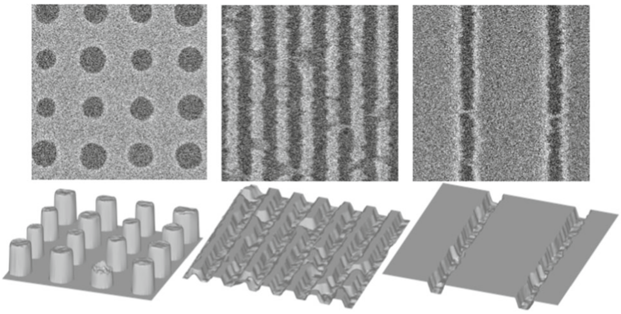

Fig. 1 Qualitative results of the proposed method. Input SEM images are depicted at the top row and corresponding depth maps predictions at the bottom row. From left to right: synthetic contact holes, real experimental dense lines, real experimental isolated trenches. Predictions of the contact holes are inverted in order to improve visualization

The objective of this work is to extract useful 3D information from SEM images, using advanced modeling techniques based on deep neural networks. First, a depth estimation method on synthetic data is explained. Next, this method is extended to work on measured experimental data, without any local ground-truth depth information. Example results of the proposed method are displayed in Fig. 1. This research work presents two contributions. First, we present a method that is capable of predicting a detailed height map and corresponding semiconductor metrics of synthetic SEM images under realistic noise conditions. Second, we demonstrate a weakly supervised domain adaptation technique, in order to incorporate the OCD data into the training procedure. We refer to this technique as pixel-wise fine-tuning.

The paper is organized as follows. After a survey of related work in Sect. 2, Sect. 3 discusses the proposed method in detail. Then, Sect. 4 provides the results and discusses the results in Sect. 5. Finally, the paper concludes in Sect. 6. Additional implementation details are provided in "Appendix."

2 Related work

2.1 Depth Estimation from SEM Images

Several techniques have already been developed to extract depth information from SEM images. A well-known method obtains depth information from observing disparities at descriptive points from a stereo image pair [13,14]. The stereo pair is acquired by tilting the specimen. Unfortunately, this method is not suitable for a SEM, since tilting is not possible due to geometric constraints imposed by the objective lens above the specimen (300-mm wafer). One way to overcome this issue is to tilt the beam (not the specimen) with deflectors [15]. But this tilt angle is limited to less than a degree in typical high-resolution SEMs. Another technique uses a four-channel secondary electron (SE) detector [16]. By combining these four SE intensity maps, it is possible to create a depth profile of the surface. However, this method is not compatible with the magnetic objective lenses that are typically used in a SEM. Moreover, all aforementioned techniques require a different hardware platform, which puts high demands on the system costs.

Also methods based on a single SEM image with conventional hardware have been proposed. In [17] SEM images are compared against a model library with physical models. This method predicts shape approximations interpolated from multiple models in the library. It has only been validated with line space patterns and so far seems to be hard to generalize to various geometries, materials and SEM settings. Alternatively, landing energy is exploited to extract depth information in top-down SEM images [18]. In certain conditions, the SE yield is sensitive to depth, while unresponsive to other shape parameters. The results obtained with synthetic SEM images were verified by experiments on an inverted pyramid shape with unit step depth transitions, but can be extended to more complex structures according to the authors. The main limitation of this approach is the requirement to change landing energies, which is typically undesirable for continuous measurement systems. Another recent work uses a neural network to predict 1D SEM-profile depths from synthetic 1D back-scattered electron (BSE) profiles [19]. A custom-weighted loss function was designed to train the network, which improved the results significantly.

2.2 Monocular depth estimation from natural images

Monocular depth estimation has been an active field of research over the years. Initially, supervised techniques were proposed [6], where ground-truth depth is available during training. Later, self-supervised techniques have become popular as well [20]. Here, depth is inferred by cleverly exploiting information from stereo data [21] or video data [22] during training. This paper will be focused on supervised methods because of the ready availability of ground-truth data for the simulated SEM images and the hardware limitations of stereo imaging.

Starting with [6], supervised depth estimation techniques evolved over the years [23–25], but along with the major improvements on established benchmarks [26,27], the networks became also quite complex [28]. Recent work [29] rephrased the depth estimation problem as an image-to-image translation [30], based on conditional generative adversarial networks (cGANs) [31]. These frameworks add a second network to the training process, which enforces an adversarial loss term, resulting in global consistency of the output. These networks show impressive results, even with a relatively straightforward prediction network [32].

2.3 SEM and deep learning

Deep learning is successfully applied to other tasks in SEM imaging. For instance, deep neural networks are used for line roughness estimation and Poisson denoising [33]. They also seem beneficial for removing artifacts without the need of paired training data [34]. Both works promise great potential for these kind of models in the field of SEM. Similarly, these applications are also established research fields with other use cases, for example, image denoising [35–37] on natural images and contouring [38] on medical images.

3 Methodology

Our approach consists of the following steps. First, a synthetic dataset is generated and pre-processed. Then, a neural network is pre-trained with the generated data. Next, the network weights are adapted using experimental data. After the training process, a diverse test set is used for validation, by comparing key semiconductor performance metrics. Information about the implementation is found in “Appendix B.”

3.1 Synthetic data generation

For the development of the methods in this work, we developed datasets with two types of structures: contact holes (CHs) and line-based spaces (LSs). These datasets are explained in detail in the next sections. The resulting constructed geometries are used as input for a Monte-Carlo particle simulator. For this, we have adopted Nebula [39], which is an open source, GPU-accelerated, solution for simulating the electron-scattering processes in materials. This simulator is currently one of the most accurate solutions available and produces partially realistic SEM images corresponding to the input geometries.

3.1.1 Contact holes dataset

CHs are cylindrical holes inside a layer of material. A hole should span the entire layer along the axial (depth) dimension to ensure contact to the next layer. We have chosen CHs for several reasons. First, the geometry contains non-trivial shape information in two lateral dimensions (circular), in contrast with line-based spaces, where only one lateral

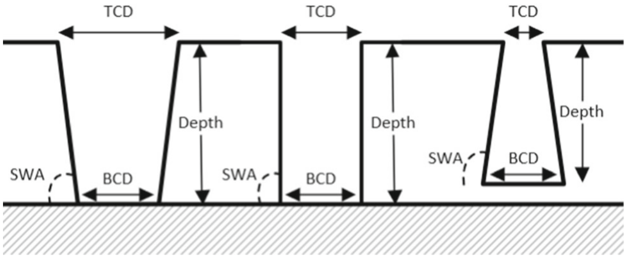

Fig. 2 Side view of the contact hole (CH) geometry, parameterized by: Depth, Top Critical Dimension (TCD), Bottom Critical Dimension (BCD). The Sidewall Angle (SWA) can be inferred from the TCD, BCD and depth of the CH. The left CH has edge-width because SWA < 90°. The middle CH has no edge-width because the wall is perfectly straight. The right CH has overhang because SWA > 90°, and it is not opened because the depth value is insufficient to reach the bottom layer (shaded area)

dimension exhibits significant depth variation. Second, CHs are heavily used in the semiconductor industry, since they enable connecting subsequent layers in a device. Third, from an industry perspective, it is attractive to obtain a proper estimation of the depth value of every CH, in order to determine whether the CH is open or not. Unopened CHs result in failures of the device.

For the creation of randomized CH geometries, the parametric model displayed in Fig. 2 is used. All CHs are generated in a two-dimensional (x, y) grid of unit cells. The total size of the grid is 1024×1024 (nm) and contains 16×16 unit cells, which results in an average pitch of 64 nm. For individual CH generation, we distinguish two type of process deviations. Normal distributions are used to mimic the intra-field (local) process deviations as realistically as possible. Furthermore, inter-field (between image) deviations are applied to the parameters that influence the height prediction the most (depth and sidewall angle). Here, a uniform distribution is used to ensure the network is robust for all possible combinations. More specifically, the center point of a CH within a unit cell deviates from the center with $\Delta_x, \Delta_y \sim \mathcal{N}(0, 1)$ nm. The top critical dimension (TCD) and bottom critical dimension (BCD), both in nm, are defined by:

$$\begin{aligned} \text{TCD} &\sim \max(\mathcal{N}(35, 4), 25), \\ \text{BCD} &\sim \text{TCD} + \Delta_{\text{rand}} + \Delta_{\text{shift}}, \end{aligned} \tag{1}$$

where $\Delta_{\text{rand}} \sim \max(\mathcal{N}(0, 2), -10)$ and $\Delta_{\text{shift}} \sim \mathcal{U}(-5, 2)$. The same Δ_{shift} value is applied to all CHs within the grid. The skew of this distribution was chosen because the patterning process gives rise to a preference of tapered CHs (SWA < 90°). Also line-edge roughness (LER) is applied in the x- and y-direction to perturb the perfectly circle-shaped edge of the CH. More details are available in “Appendix A.” Furthermore, the numerical values are derived from relevant experimental data.

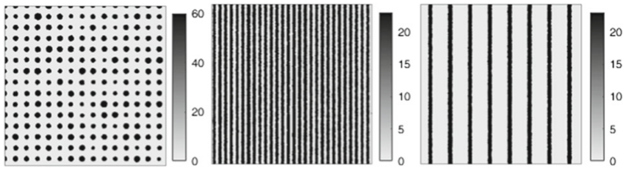

Fig. 3 Top views of the generated geometries. From left to right: contact holes, dense lines and isolated trenches, all with roughness. Pixel color represents the depth value

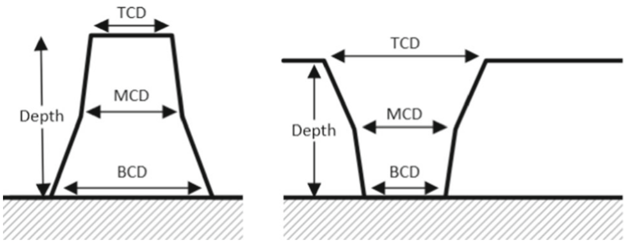

Fig. 4 Left: side view of line-space (LS) geometry. Right: side view of Isolated trench geometry. Both are parameterized by: depth, top critical dimension (TCD), middle critical dimension (MCD) and bottom critical dimension (BCD). The interior of a LS is filled with material. The isolated trench has material everywhere except in the trench. The bottom layer (shaded area) commonly consists of a different material

The depth of the CHs is varied between 20 and 100 nm with steps of 1 nm. One depth value is applied to all CHs in the grid, in order to mimic a lithographic process as close as possible. CHs chosen at random with probability $p = 0.005$ are unopened (filled with extra material), as shown in the rightmost CH in Fig. 2. One example of a resulting geometry is visualized in the leftmost image of Fig. 3. For the simulations, we have used SiO_2 (Silicon dioxide) as top material and Si (Silicon) as bottom material. For the settings of the electron beam, we employed a Gaussian distributed spot-profile, defined by its Full Width Half Maximum (FWHM) of 2.0 nm, a dose of 100 electrons per pixel and a landing energy of 800 eV. These settings are chosen to mimic common CD-SEM operation, except that currently a FWHM around 3.5 nm is more common. In total, we have simulated four geometry realizations per depth value, resulting in 320 images of 1024×1024 pixels, with a pixel size of 1 nm^2.

3.1.2 Line space datasets

LSs are vertical or horizontal strokes of material in a regular fashion, separated by trenches (Fig. 4). Because of the presence of stochastic effects of the fabrication process, the LSs have non-smooth edges. In extreme cases, LSs can have interruptions or get (partly) connected to adjacent structures, often called micro-bridges. LSs are heavily used in devices, since they are the building blocks of transistors, as well as wiring between components.

The geometries should be roughly matched with the experimental data (examples in Fig. 1), which consists of dense lines (16 nm) with 32-nm pitch and isolated trenches (16 nm) with 112-nm pitch. The TCD, MCD and BCD are independently varied from 13 to 20 nm and the depth is varied from 15 to 30 nm and kept equal within one image. The 1D LER is applied to the line contours by an improved variant of the Thorsos method [40]. More details are found in "Appendix A." All parameter ranges were chosen a bit larger than the ranges of the measured data. This makes the simulated data a superset of the actual data, which ensures that all possible cases are covered by the simulated data. Defects such as micro-bridges are not modeled in the synthetic dataset. In total, 550 dense line geometries were constructed together with 1650 isolated trench geometries. Figure 3 shows one example for both.

For the simulator, the same settings were used as for the previous experiment, except for the landing energy (500 eV) and pixel size (0.64 nm^2), to obtain a better match with the experimental data. In total, we have simulated one SEM image per geometry, with a field of view (FOV) of 1024 × 1024 pixels.

3.2 Pre-processing

The fact that some parts of the CD-SEM system are not modeled in the simulator creates a distribution shift between the synthetic and experimental domains. In this Section, we elaborate on the steps taken for decreasing this domain shift. Furthermore, data augmentation techniques are discussed.

3.2.1 Noise

A simplified noise model of a CD-SEM system is displayed in Fig. 5. The first noise contribution is shot noise from the electron gun. The number of primary electrons (PEs) originating from the gun is Poisson distributed. When a PE hits the specimen, it may become a secondary electron (SE), which experiences a stochastic electron cascade (scattering) through the material. This results in a compound Poisson noise distribution. Both effects are accounted for in the simulator. The third noise contribution is from the detector, where dark current is assumed to be dominant. Dark current intrinsically behaves as shot noise (Poisson), but for large numbers, the Poisson distribution will approach a Normal distribution. Therefore, this detector noise is modeled as additive Gaussian $\mathcal{N}(0, \sigma^2)$ with $\sigma \in [0.1, 0.2]$ for normalized image values in the unity interval. A standard deviation of 0.1 appeared to be a realistic result based on experiments in which the beam current was measured with and without the beam blocked. The value of $\sigma = 0.2$ serves as a worst-case upper bound.

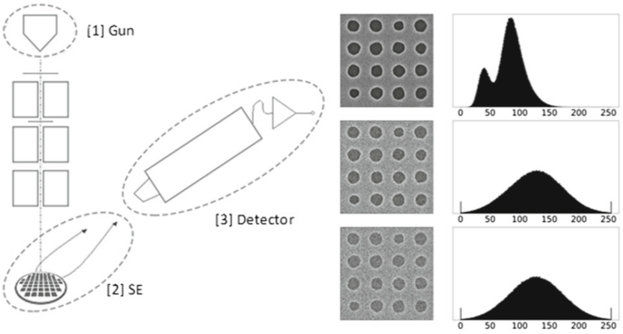

Fig. 5 Left: simplified noise model of a CD-SEM system. Noise from the electron gun (1) and the random walk of secondary electrons (2) are incorporated in Nebula. Detector noise (3) is modeled as additive Gaussian distribution ($\mu = 0, \sigma \in [0.1, 0.2]$). Right: examples of a pre-processed CD-SEM image and their corresponding histograms. From top to bottom: Original image from the simulator, image with added detector noise ($\sigma = 0.1$) and histogram correction, image with added detector noise ($\sigma = 0.2$) and histogram correction

3.2.2 Histogram correction

CD-SEM systems work with a detector current which will be translated into a gray value. This value depends on various CD-SEM aspects, such as the electronics, signal gain, landing energy, etc. Scaling all gray values of an image to use the full dynamic range prevents saturation effects, while changing settings of the CD-SEM. We have implemented this by snapping the lowest 0.2% pixels to the lowest value possible, the highest 0.2% pixels to the highest value possible and scaling everything in between accordingly. Images are stored in 8-bit unsigned integer format. Eight bits typically provide sufficient dynamic range, while maintaining the memory load of millions of images acceptable.

3.2.3 Data augmentation

Additional data augmentation is performed on the fly when training the network. A smaller patch of 256 × 256 pixels is cropped from the generated image at a random location. Detector noise and histogram correction are applied next. Further augmentation may be horizontal flipping, vertical flipping and rotating, with a probability of 0.5 per event. With experimental data, this probability is set to zero, since important aberrations, like charging, are not symmetrical and dependent on the fast-scan direction of the SEM. Examples of pre-processed synthetic images are displayed in Fig. 5.

During inference, an entire image is processed at once by the network, so the only augmentation steps that stay relevant are adding noise and histogram correction. With inference of experimental data, no augmentation step is required.

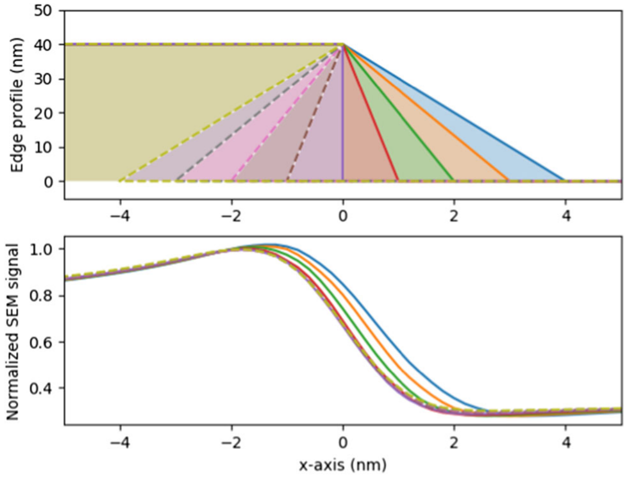

Fig. 6 Top: geometry cross sections of depth changes with steps of different SWAs. Bottom: the corresponding secondary electron-yield signals. Values are averaged over 50 measurements to obtain clean results. Dashed lines correspond to overhanging structures

3.3 Depth estimation from synthetic data

This section involves model selection, network architecture, loss functions and explaining the training process in more detail.

3.3.1 Model selection

There are many ways to represent a 3D structure, e.g., a polygon mesh, a voxel grid or a depth map. To determine what data type is most suitable for this application, an initial experiment was performed for examining the SEM signal, using a simple geometry with a varying SWA, see Fig. 6. We observe no distinctive signal for overhanging structures and conclude that distinguishing them is not possible, with the chosen landing energy only. This implies that only one depth value per pixel location of the SEM image is sufficient to capture all depth information present in the image signal. True 3D data types, like voxel grids, would therefore be redundant. Instead, we have adopted to use depth maps, which directed the research into depth estimation models.

Recent literature on depth estimation uses standardized benchmarks to compare the performance of different approaches [4]. Supervised methods still have the best overall performance. Most supervised methods use a pixel-wise loss function. However, recent work [29] proposes adding an adversarial (non-local) loss term to the depth prediction network. This approach outperforms pixel-wise losses with a relatively simple prediction network and triggers the interest for conducting an extensive loss function evaluation study. This will be elaborated in a separate section.

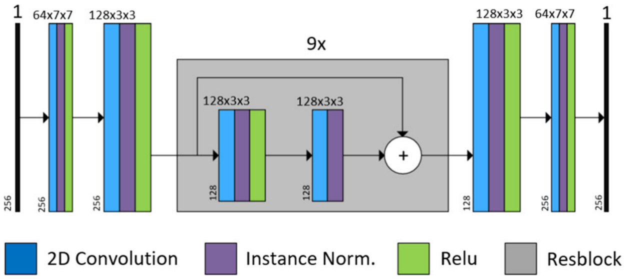

Fig. 7 Architecture of the prediction network, consisting of a convolutional front-end, 9 residual blocks and a transposed convolutional back-end. The number of channels and kernel size are displayed above the convolutional blocks. The width and height of the inputs during training are displayed at the left bottom. The stride of the convolutional layers is unity, except for the layer before (2) and after (1/2) the series of Resblocks. Reflection padding is applied prior to each convolutional block to reduce border artifacts

3.3.2 Network architecture

The network used is based on recent work [41] for image-to-image translation. We denote A_s and A_d as the SEM image and depth map domains, respectively, while a_s and a_d refer to training examples in both domains. The actual prediction network learns a mapping function $G : A_s \rightarrow A_d$ which takes a SEM image as input and outputs a depth map. Furthermore, depending on the loss function, we use a discriminator network with a mapping function D. This network takes a SEM image and a corresponding predicted depth map as input and outputs an error-parameter score that quantifies the quality of the realism.

A detailed overview of the prediction network is found in Fig. 7. It consists of 9 stacked residual blocks [42], together with a convolutional front- and back-end. All residual blocks have two convolutional layers and an identity connection to the next block. This connection is attractive because the convolutional layers only have to learn the difference between the input and the output, which is in many cases less demanding for the network. These skip connections also enable the construction of deeper nets, since they do not suffer from the vanishing gradient problem during the backpropagation phase. The number of filters in the first layer is set to 64. Instance normalization is used after each convolutional layer, followed by a rectified linear unit (Relu).

3.3.3 Loss functions

A loss function with multiple terms is used for more detailed optimization. We employ three terms, each operating at a different scale. At a local scale we use an ℓ_1 or ℓ_2 loss, as defined by:

$$\mathcal{L}_{\ell n}(G) = \mathbb{E}_{a_s,a_d\sim p_{A_s,A_d}}\left[\|a_d - G(a_s)\|_n\right], \tag{2}$$

where $n \in 1, 2$ is the rank of the distance measure and p_{data} denotes the probability distribution of the data samples. This loss term operates on pixel level.

A perceptual loss, which operates on patch level, is used for regional features and is defined by:

$$\mathcal{L}_{\text{VGG}}(G) = \mathbb{E}_{a_s,a_d \sim p_{A_s,A_d}} \left[\sum_{i=1}^{N} \frac{1}{M_i} \left\| \Delta F^{(i)} \right\|_1 \right],$$
$$\text{where } \Delta F^{(i)} = F^{(i)}(a_d) - F^{(i)}(G(a_s)). \tag{3}$$

Here, $F^{(i)}$ denotes the i-th layer with M_i total network elements. It minimizes the ℓ_1-distance of the network's intermediate feature representations between the predicted and ground-truth samples. The applied network is VGG16 [43], which is pre-trained with Imagenet [44] data.

For the global features, we have trained the prediction network together with a discriminator network. The network then becomes a generative adversarial network (GAN) [45], which can also be used for image-to-image translation [30] when adding conditional inputs. In this case, a least squares GAN (LSGAN) loss [46] is used, which consists of a generator loss and discriminator loss, resulting in the following specification:

$$\begin{aligned}\mathcal{L}_{\text{cLSGAN}}(D) &= \frac{1}{2}\mathbb{E}_{a_s,a_d \sim p_{A_s,A_d}} \left[(D(a_s, a_d) - 1)^2 \right] \\ &\quad + \frac{1}{2}\mathbb{E}_{a_s \sim p_{A_s}} \left[(D(a_s, G(a_s)))^2 \right], \\ \mathcal{L}_{\text{cLSGAN}}(G) &= \frac{1}{2}\mathbb{E}_{a_s \sim p_{A_s}} \left[(D(a_s, G(a_s)) - 1)^2 \right]. \end{aligned} \tag{4}$$

Unlike cross-entropy functions, the squares in Eq. (4) stronger penalize samples far from the decision boundaries, even when classified correctly, which helps to stabilize the training process [47]. For the discriminator, we have used a multi-scale Patch-GAN [30], operating at a receptive field of 70 and 140 pixels (which is the default operation setting), each with three convolutional layers. Also here, all layers are followed by a normalization and activation layer, while the first layer starts with 64 filters.

Finally, we can construct the resulting loss function as a linear combination of the aforementioned terms, where a part is minimized over G, and the last part over D, such that:

$$\begin{aligned}\mathcal{L}_{\text{total}}(G, D) &= \min_G \lambda_{\text{loc}}\mathcal{L}_{\ell n}(G) + \lambda_{\text{reg}}\mathcal{L}_{\text{VGG}}(G) \\ &\quad + \lambda_{\text{glob}}\mathcal{L}_{\text{cLSGAN}}(G) + \min_D \lambda_{\text{glob}}\mathcal{L}_{\text{cLSGAN}}(D). \end{aligned} \tag{5}$$

3.3.4 Training process for pre-training

The data are divided in a training, validation and test set, consisting of 70%, 5% and 25% of the data, respectively. The test set is carefully constructed so that all possible depths are represented. Training is done in randomized batches of 16 images. As already mentioned, data augmentation is performed on the fly. The amount of noise added to the images is uniformly distributed between zero and the specified maximum σ required to mimic the detector noise. After empirical experiments, this turned out to be the best choice. A possible reason for this choice is that the network is not able to establish proper kernel filters when only receiving very noisy images. The Adam optimizer [48] is used for minimizing the total loss function, for 300 epochs, with a learning rate of 0.0002 and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Multiple networks are trained with loss functions specified by different values for λ_{glob}, λ_{reg} and λ_{loc}. If not zero, then $\lambda_{\text{glob}} = 1$, $\lambda_{\text{reg}} = 10$ and $\lambda_{\text{loc}} = 10$. Training performance is assessed by reviewing the depth performance metrics on the validation set. The following metrics were used for model comparison on the validation set:

- Mean Relative Error: $\frac{1}{N}\sum_y \frac{y_{\text{gt}} - y_{\text{pred}}}{y_{\text{gt}}}$
- Average $\log_{10}$ Error: $\frac{1}{N}\sum_y |\log_{10} y_{\text{gt}} - \log_{10} y_{\text{pred}}|$
- Root Mean Square Error: $\sqrt{\frac{1}{N}\sum_y \left(y_{\text{gt}} - y_{\text{pred}}\right)^2}$
- Accuracy with threshold t: % of y_{pred} s.t. $\max(\frac{y_{\text{gt}}}{y_{\text{pred}}}, \frac{y_{\text{pred}}}{y_{\text{gt}}}) = \delta < t$ ($t \in [1.25^{0.25}, 1.25^{0.5}, 1.25, 1.25^2, 1.25^3]$)

where y_{pred} and y_{gt} are the predicted and ground truth depth map. N is the total number of pixels.

3.4 Depth estimation from experimental data

The shift in distributions between the experimental domain and the synthetic domain requires an extra step. In this case, we have paired experimental SEM data with available OCD data. Due to the lack of local information in the OCD data, the ground truth is only partially present, which makes that this method can be classified as a weakly-supervised learning approach.

3.4.1 Experimental datasets

We have employed a CD-SEM system to measure a focus exposure matrix (FEM) wafer just after a lithography step. On a FEM wafer, the focus and dose of the scanner is gradually changed during exposure, which results in considerable geometry variations over different locations on the wafer. The wafer contained 16-nm dense lines (32-nm pitch) and 16-nm isolated trenches (112-nm pitch). The available data consist of two measurements for 1341 unique locations on the wafer. One SEM measurement with an FOV of approximately 1 μm^3 is available, as well as one OCD measurement with an FOV of 25 μm^3. The OCD measurement contains several parameters (scalars) that are directly

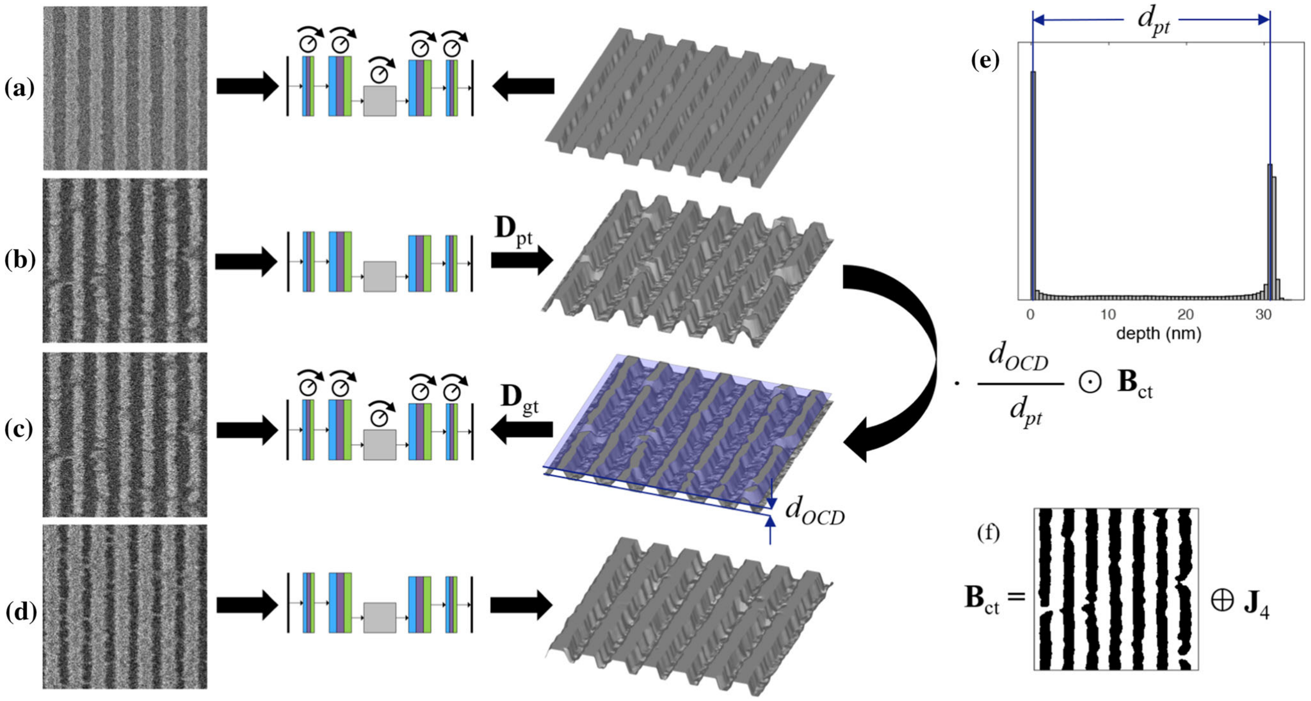

Fig. 8 Schematic overview of the training procedure. **a** The network is pre-trained on synthetic data. **b** Inference with experimental data on the pre-trained network. The resulting depth map ($\mathbf{D}_{\mathrm{pt}}$) is scaled by $d_{\mathrm{OCD}}/d_{\mathrm{pt}}$. **e** d_{pt} is determined by the peak distance of the histogram of the pre-trained depth map $\mathbf{D}_{\mathrm{pt}}$. Isolated trenches are also element-wise multiplied (operation denoted by $\odot$) by a binary matrix $\mathbf{B}_{\mathrm{ct}}$, in order to remove charging artifacts. **f** This binary matrix is obtained from the output of a contouring algorithm dilated (operation denoted by $\oplus$) with $\mathbf{J}_4$ (which is a 4×4 matrix of ones). This results in a new pixel-wise ground truth. **c** The network is fine-tuned with experimental data and pixel-wise ground truth. **d** With the final network inference on experimental data is possible

related to a (multi-) trapezoid model representing the cross-sectional profile of a line, similar to Fig. 4. One parameter of this model expresses the total depth of the line. Furthermore, we assume that global statistics of one SEM image are sufficiently averaged to correlate with OCD values.

We have constructed two datasets, one with dense lines and one with isolated trenches. The dense-line dataset contains 331 images, where the depth varies between 17 and 24 nm. The isolated-trenches dataset contains 682 images, where the depth is within 26–27 nm. Although the depth range of the isolated trenches is insufficient for testing the depth predictions, we use these data to perform other useful experiments. The total number of measurements is lower than the total number of measurements on the wafer, since cases where the OCD trapezoid model has not converged properly are omitted.

3.4.2 Pixel-wise fine-tuning

The domain adaptation step is implemented by a novel method, further referred to as pixel-wise fine-tuning. In general, fine-tuning with a single value as ground truth entails that the optimization problem of the model is under-constrained. In order to prevent the network drifting from the manifold of realistic structures, some training regularization is required. The inference on experimental data without fine-tuning the network turned out to be qualitatively correct in terms of lateral shape information, but quantitatively incorrect in terms of depth information in the axial direction. Therefore, we have decided to generate a new ground-truth by combining information from the resulting depth maps with corresponding OCD depth values. This re-enables pixel-wise training, thereby solving the under-constrained problem. This domain adaptation method is valid for this use case because the properties of a lithographic multilayer etch process imply that the structure height within the field-of-view of and OCD measurement is very constant. Alternatively, we have tried to regularize the network by fine-tuning only a subset of the layers or adding a discriminator to the loss function that was specifically trained on realistic depth maps. The results of both methods were not satisfactory because artifacts were introduced, so that it will not be treated further.

The pixel-wise ground truth is produced by scaling the depth maps ($\mathbf{D}_{\mathrm{pt}}$) obtained from inference of the experimental images on the pre-trained network. The scaling is defined by

$$\mathbf{D}_{\mathrm{gt}} = \frac{d_{\mathrm{OCD}}}{d_{\mathrm{pt}}} \cdot \mathbf{D}_{\mathrm{pt}}, \tag{6}$$

where d_{OCD} denotes the depth parameter from the OCD model and d_{pt} is the depth derived from the depth map $\mathbf{D}_{\mathrm{pt}}$. Matrix $\mathbf{D}_{\mathrm{gt}}$ is the resulting depth map. The value of d_{pt} is determined by the distance between the two peaks in the histogram of the depth map, displayed in Fig. 8. More specifically, the histogram bins have a width of 0.01, and the largest bin of the lower half and the largest bin of the upper half of the histogram are selected. These peaks represent the values of the averaged bottom-layer surface depth and the averaged depth of the LSs. This method is robust for the presence of noise in $\mathbf{D}_{\mathrm{pt}}$ and produces consistent results.

3.4.3 Artifact removal

The predicted depth maps of isolated trenches suffer from artifacts at the surface between the trenches, most likely due to charging effects present in the experimental data. These artifacts are present as small pits from the surface of the depth map and do not interfere with the border of the trench or the trench itself. We have solved this issue by adding one processing step, just prior to the pixel-wise scaling operation. The processing step entails element-wise multiplication with a dilated binary map ($\mathbf{b}_{\mathrm{ct}}$) originating from a SEM contouring algorithm, which exploits an adaptive-threshold method. This step is also depicted in Fig. 8 at step (b). It removes the artifacts while preserving the rest of the information in the depth map. With this ground-truth, the network learns to ignore charging artifacts, which results in a correct output. Since SEM contouring algorithms are available for many structures, this method can be extended to other use cases.

3.4.4 Training process for fine-tuning

The entire training process is depicted in Fig. 8. Pre-training is performed as described in the previous sections concerning synthetic data. The experimental data are separated in sets, 70% train, 5% validation and 25% test. Fine-tuning is done for 100 epochs using Adam solver, with a learning rate of 0.001. Data augmentation and detector noise are not applied. Several models are trained with different loss configurations. The same performance metrics are used as in the validation during the pre-training process.

3.5 Post-processing

Several key performance indicators that are relevant for the semiconductor industry can be inferred from the obtained depth maps. We introduce the following notations. The area at depth z is $A_z = N_z \cdot a_p$, where N_z denotes the number of pixels below (or above with dense lines) depth z within a slice at depth z of the structure, selected with a threshold operation. Parameter a_p is the area of one pixel. In this work $a_p = 1\,\mathrm{nm}^2$ for CHs and $a_p = 0.64\,\mathrm{nm}^2$ for LSs. For selecting individual structures, each unit cell is selected first, with a mask. Then the following operations are performed.

3.5.1 Semiconductor metrics for CHs

The parameters present in the model of Fig. 2 have to be retrieved for each individual contact hole. The following metrics will be used.

- TCD: $2\sqrt{A_{z_{\mathrm{top}}}/\pi}$ where $z_{\mathrm{top}} = 2\mathrm{nm}$.
- BCD: $2\sqrt{A_{z_{\mathrm{bottom}}}/\pi}$ where $z_{\mathrm{bottom}} = 0.75 z_{\max}$, where $z_{\max}$ is the deepest pixel value.
- Depth: $1/N_{z_{\mathrm{bottom}}} \sum_{ij} d_{ij} \cdot m_{ij}$. Here, d_{ij} are the individual values of the depth map $\mathbf{D}$, and $m_{ij} = 1$, where $d_{ij} > z_{\mathrm{bottom}}$, otherwise $m_{ij} = 0$.
- SWA: $180/\pi \arctan(\frac{z_{\mathrm{top}} - z_{\mathrm{bottom}}}{\mathrm{TCD-BCD}})$ degrees, when the difference TCD–BCD > 0, otherwise 90 degrees.

The critical dimension of a CH is calculated with the formula of the area of a circle. Therefore, this metric can be seen as average critical dimension.

3.5.2 Semiconductor metrics for LSs

The parameters occurring in the model of Fig. 4 representing local information will be gathered as follows.

- TCD: $A_{z_{\mathrm{top}}}/L$ where $z_{\mathrm{top}} = z_{\mathrm{ceil}} + 2\mathrm{nm}$ and L is the length of the selected structure and z_{ceil} is the location of the leftmost peak of the histogram function.
- BCD: $A_{z_{\mathrm{bottom}}}/L$ where $z_{\mathrm{bottom}} = z_{\mathrm{floor}} - 2\mathrm{nm}$, where z_{floor} is the location of the rightmost peak of the histogram function.
- Depth: Average depth difference around the line's contour, calculated with a histogram function (as described earlier) using the area around the current LS, as an input.
- SWA: $180/\pi \arctan(\frac{z_{\mathrm{bottom}} - z_{\mathrm{top}}}{\mathrm{BCD-TCD}})$ degrees, when the difference BCD–TCD > 0, otherwise 90 degrees.

Additionally, global information should be derived from the depth map to enable validation with OCD data.

- Average CD at depth z: This is $P \cdot N_z/N$ where P denotes the pitch of the pattern and N the total number of pixels in the image.
- Average depth value: This value is calculated with the histogram method as described earlier.

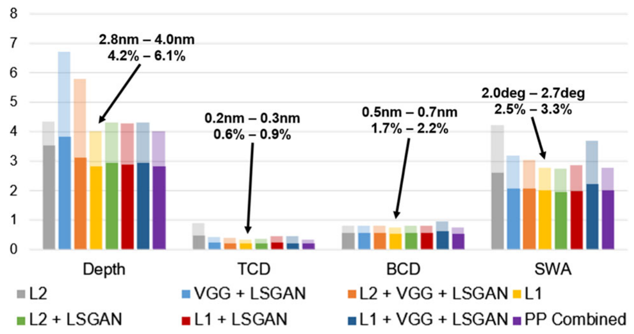

Fig. 9 Mean absolute errors of the semiconductor metrics on the CH dataset. Units are in nanometer, except for SWA, which is expressed in degrees. Several models are compared on all metrics. The darker bars represent results with a normal noise level (σ = 0.1), while the lighter bars refer to the worst case noise level (σ = 0.2). Numerical results are displayed for the best model, indicating absolute and relative errors for both bars

4 Results

In this section, we present qualitative and quantitative results. The following section elaborates on synthetic data, predominantly on the experiment with the CH dataset. The second section focuses on the experimental LS dataset.

4.1 Synthetic results

The depth estimation network is trained as explained in the previous sections. The network did not suffer from overfitting, since the performance on the validation set did not degrade at the end of the training procedure.

4.1.1 Contact holes dataset

Qualitative results of CHs are found in Figs. 1 and 12. The mean absolute errors are displayed in Fig. 9. All provided metrics are calculated with the post-processing method discussed in the previous section. It can be observed that a network with only a local ℓ_1 loss works best for all metrics. The obtained relative error of the depth is between 4.2 and 6.1% for realistic noise levels. TCD, BCD and SWA correlations of individual CHs for two different SEM images are displayed in Fig. 10. We have found that TCD and BCD always show a good correlation. Furthermore, SWA correlation is reasonable, but tends to become less accurate in images with many overhanging CHs.

In this work, we primarily focus on depth. The results of the best performing network (yellow bars in Fig. 9 with ℓ_1 loss) are displayed in Fig. 11. The depth inference by the network (indicated by *get depth*) closely follows the depth programmed in the geometry (indicated by *set depth*), which is used to generate the simulated SEM image. It can be observed that deeper holes result in less accurate predictions, since the average error grows with the depth. This is explained by the fact that when the CH becomes deeper, the change in SEM signal becomes smaller, i.e., the SEM signal scales non-linearly with the depth of the CHs. A possible physical explanation is that the total number of detected electrons is lower for deeper structures, while some noise contributions are not dependent on depth, which results in a lower SNR for deeper structures. Partially filled holes perform well (which proves applicability of this technique for defect detection) but are sometimes less correlated with

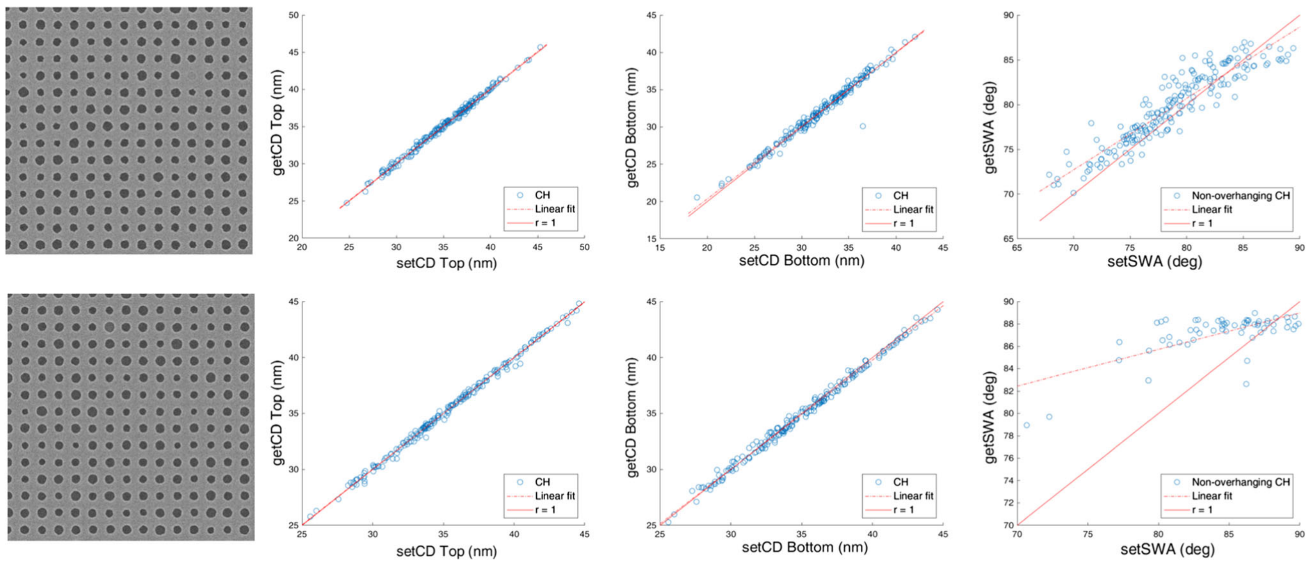

Fig. 10 Synthetic SEM images of CHs and their correlation plots on semiconductor metrics. Top row: CHs of 46 nm deep. Bottom row: CHs of 28 nm deep. From left to right: SEM image, TCD correlation, BCD correlation and SWA correlation. Measurements are done under realistic noise conditions (σ = 0.1)

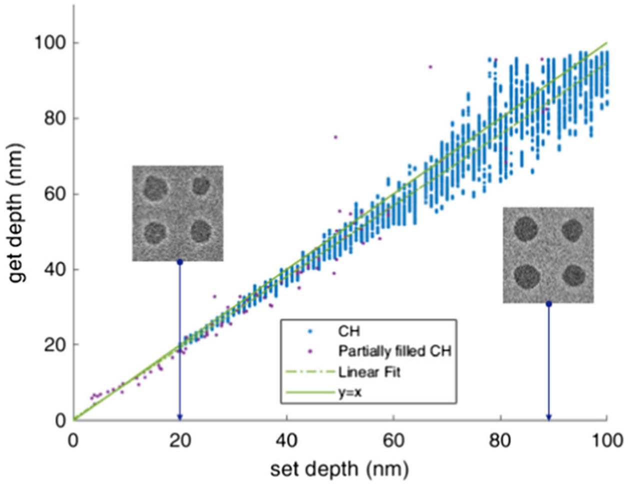

Fig. 11 Individual CH depth analysis of the test set, predicted w.r.t. ground truth, at realistic noise levels (σ = 0.1)

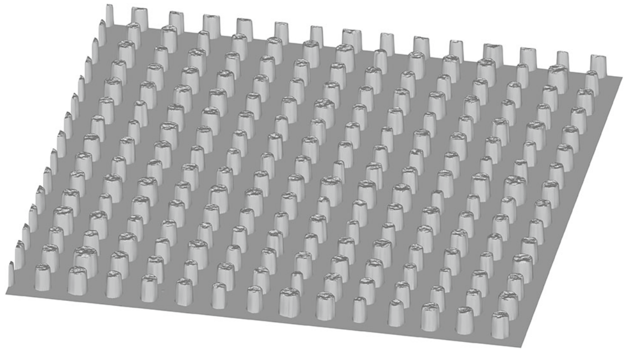

Fig. 12 Predicted map of a simulated SEM image with 70 nm deep CHs

the ground-truth depth because of the low number of training examples present in the dataset. Also equalizing effects appear, which occur from situations when the height of the partially filled hole is close to the rest of the CHs in the geometry. The first argument can be solved by creating a better balanced dataset with more partially filled CHs.

The network can also handle large field-of-view SEM images. A qualitative result of simulated data is shown in Fig. 12 and the corresponding quantitative pixel-wise absolute difference with ground truth is displayed in Fig. 13.

4.1.2 Line-spaces with roughness dataset

Global model performance on the synthetic LS dataset is summarized in Fig. 14. We observe similar behavior between the models, also here ℓ_1 performs best on all metrics except for TCD and SWA, where the model was trained with ℓ_2, LSGAN and VGG loss. It is possible to combine the metrics of different models in the post-processing to get even better predictions for SWA, as shown by the purple bars.

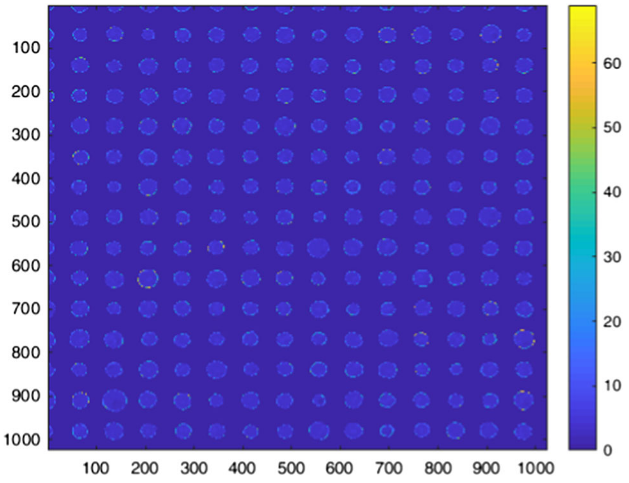

Fig. 13 Absolute difference of the predicted depth map of Fig. 12. The units of the image are pixels. The color bar with numbers indicates a scale in nanometers (color figure online)

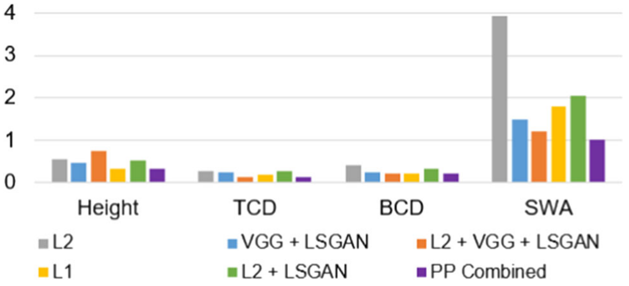

Fig. 14 Mean absolute errors of the semiconductor metrics on the synthetic LS dataset. Units are in nanometer, except for SWA, which is expressed in degrees. Four models are compared on all semiconductor metrics. The results are from data with a worst-case noise level (σ = 0.2)

4.2 Experimental results

After extensive training with synthetic data, the network was not able to give satisfactory results on experimental data. Therefore extra training steps were required to implement, which we explained in the methodology section. The results of these steps are presented in the following sections.

4.2.1 Dense lines dataset

Some examples of depth maps obtained from SEM images of dense LS patterns are displayed in Figs. 1, 8 and 17.

Figure 15 shows the performance of the model trained with ℓ_2 loss on depth estimation for individual lines. The depth inferred by the network (indicated by *get SEM depth*) closely follows the depth measured by the OCD tool (indicated by *get OCD Depth*). The average error is low, smaller than 1 nm, which means this network is able to predict depth very accurately. This is an important result because it shows

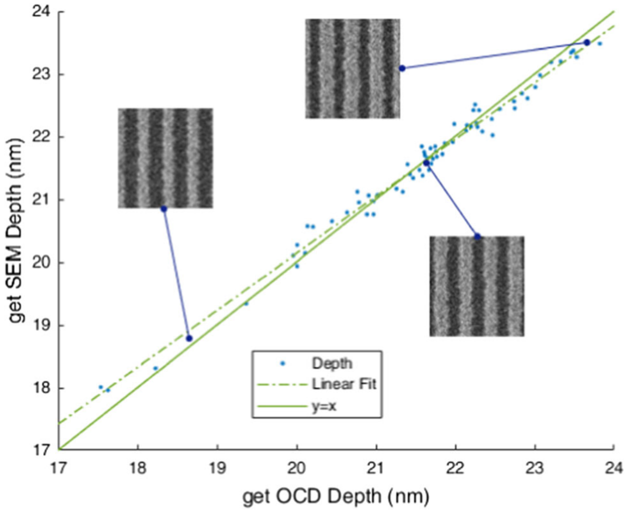

Fig. 15 OCD depth w.r.t. the calculated average depth from the predicted SEM depth map. The mean absolute error is 0.16 nm, while the mean relative error is smaller than 1%

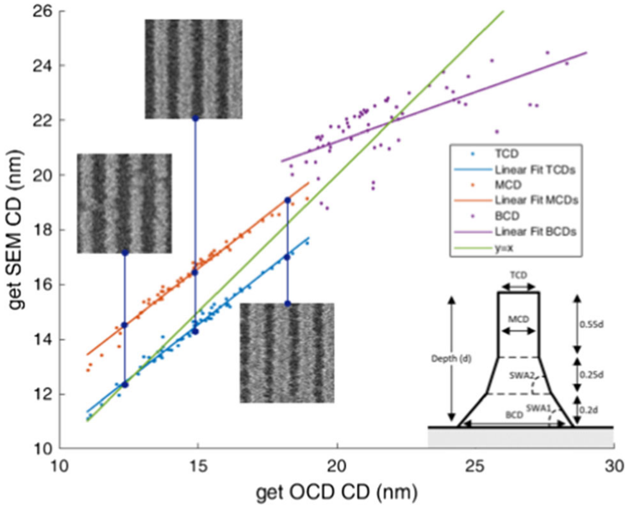

Fig. 16 OCD CDs w.r.t. the calculated average CDs from the predicted SEM depth map. Mean absolute error is 0.34, 0.44, 1.68 nm for TCD, MCD and BCD, respectively. The multi-trapezoid model used by the OCD tool is depicted in the bottom-right corner

a clear correlation between two modalities. We can also validate lateral feature information of the depth map with the OCD tool, since it additionally measures other geometric parameters. The results of average CD predictions for individual images are displayed in Fig. 16. Here, we used ℓ_1 loss for training. SEM is most sensitive for TCD, since it shows a clear correlation with the OCD data. MCD and BCD perform reasonably well. There is some offset present in the slope of the data points. This could be explained by the fact that the SEM signal is less sensitive for lower structures, than the OCD tool. Besides, the definition of MCD is not strict in the parameter model of the OCD tool.

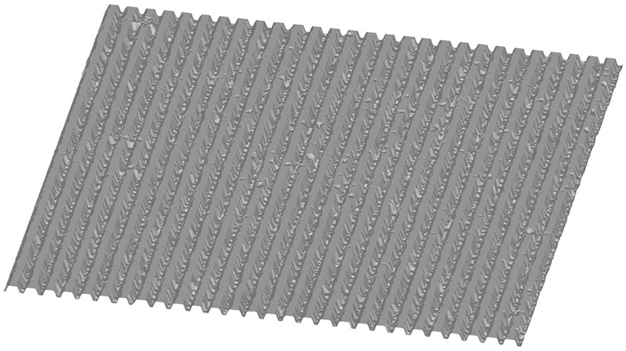

Fig. 17 Predicted map of a real SEM image with 23 nm deep lines

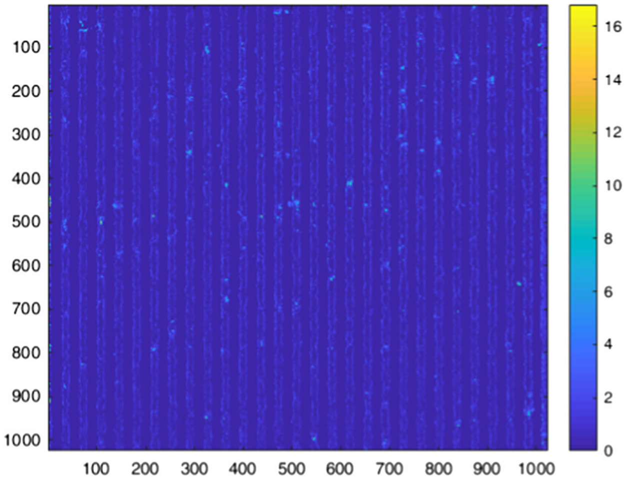

Fig. 18 Absolute difference of the predicted depth map of Fig. 17. The units of the image are pixels. The color bar with numbers indicates a scale in nanometers (color figure online)

Also with experimental data, the network is able to handle large field-of-view results. A qualitative depth map is shown in Fig. 17, and the corresponding quantitative pixel-wise absolute difference with ground truth is displayed in Fig. 18.

4.2.2 Isolated trenches dataset

Qualitative results of the depth maps before and after fine-tuning are displayed in Fig. 19. The final result shows that the charging artifacts are completely removed through better learning and modeling.

Since this dataset does not have sufficient variation in depth values, only the CD value is interesting to evaluate. The corresponding OCD model has only one CD value defined. We obtain a mean absolute error of 0.46 nm with minimal slope off-set, which indicates that the lateral information in the depth map is in accordance with both modalities. Furthermore, these depth maps can be used to measure the depth of micro-bridges inside the trenches, since the network should be able to cope well with intermediate depths values.

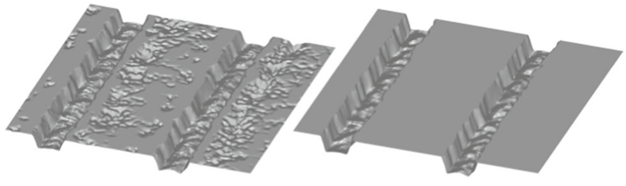

Fig. 19 Qualitative results of isolated trenches. Left: depth map prediction prior to fine-tuning. Right: depth map prediction after fine-tuning with artifact removal

5 Discussion and limitations

Although an extensive ablation study on the performance of different loss functions was performed, as well as hyperparameter tuning of the network and training process, it cannot be guaranteed that it is the optimal configuration for this use case. The most important goal of this research is to prove that the technique presented is feasible with the type of data available. Even though the results are promising, it is important to note that there are some caveats to the presented approach.

With the presented method, the measurements from the OCD tool were used as a reference, by using them to create a new ground truth. Evidently, the precision of this measurement tool is also limited. Especially because the OCD value is averaged over a much larger area of the wafer, the local accuracy cannot be guaranteed. Ideally, this method should be validated with a third metrology tool. For example, this could be implemented by comparing TEM cross sections or AFM traces with the predicted depth maps at certain points on the wafer. It would also be possible to calibrate the network with these measurements, but in the ideal case we only want to exploit it for validation, since the cost (slow, expensive, destructive, etc.) of these measurements is much higher than that of OCD metrology.

Currently, a histogram-based approach is used to match the predicted profile to the OCD measurement. This method was found empirically and showed acceptable results. However, it would be more accurate to use a Maxwell solver [49,50] for this purpose. By feeding the predicted depth map into the solver, a virtual OCD measurement can be made. This enables more accurate comparison between the modalities.

The artifact removal method for isolated trenches works well in the performed experiments. Nevertheless, it is expected that this method will degrade for certain circumstances. With specific combinations of materials and geometries, charging effects may occur more intensely, also in the deeper structures of the depth map. A straightforward solution is to incorporate the charging effects in the simulation models. However, this is not a trivial task due to the complexity of the physics involved. Alternatively, data-driven solutions, such as unsupervised domain adaptation, are interesting future research directions for this purpose.

6 Conclusions

We have shown that deep learning models are suitable as a conceptual solution for extracting 2D and 3D metrics from synthetic SEM images. The final prediction network, which is based on a image-to-image translation task, was trained with several loss functions on different scales. For depth estimation on these images, a single ℓ_1 loss turned out to be the best choice for CHs, with a mean relative error of 4.2–6.1% on depth. The ℓ_1 loss also works best for depth prediction on synthetic LSs, but for TCD and SWA a combined loss (ℓ_2 loss, perceptual loss and adversarial loss) results in the lowest error metrics. It is also possible to combine both networks (ℓ_1-based and combined-based) to obtain a slightly better performance on SWA. We also showed that the network was able to detect defected contact holes in most cases, which promises great potential for defect detection.

Furthermore, we have demonstrated that it is possible to calibrate the model in order to cope with real experimental data. We showed that it is possible to achieve an average prediction error below 1 nm after calibration with OCD data. The network can also well summarize to defects, such as microbridges, even if they are not modeled in synthetic data. This generalization power provides great potential for estimating the height of these defects. However, ideally this hypotheses should be validated first with a third metrology tool.

The result of this work makes it possible to use the three-dimensional information hidden in a SEM image. While other technologies used for this purpose have significant shortcomings in applicability or practicability, the current method may be applicable to industrial measuring equipment with limited calibration data and executed on conventional computing platforms.

Acknowledgements We thank experts from ASML Netherlands B.V., Veldhoven, NL, particularly Scott Middlebrooks and Rik Hoefnagels, and the company for offering the research project and facilities for high-performance computing.

Declarations

Conflict of interest The research leading to these results received funding from Topsector Holland High Tech under Grant Agreement No [TKI-HTSM/20.0396]. The first three authors hold a patent No [2020P00270] regarding this work.

Appendix A: Roughness details

A.1 Contact holes

For CHs, roughness is applied by connecting N equidistant points on a virtual circle, where N is set to $N = 73$. First, the distance of each point to the center of the circle is changed by a normally distributed variable ($f[m]$), where the standard deviation is uniformly distributed ($\mathcal{U}$), which is specified by

$$f[m] \sim \mathcal{N}(0, \sigma_{\mathrm{RA}}^2) \quad \text{where} \quad \sigma_{\mathrm{RA}} \sim \mathcal{U}(0.5, 1). \tag{A1}$$

The numbers $f[m]$ with $0 \le m \le N-1$ are convolved with a sampled Gaussian function, specified as:

$$g_N[x] = \frac{1}{\sigma_{\mathrm{CL}}\sqrt{2\pi}} e^{-(x[n]-\mu)^2 / 2\sigma_{\mathrm{CL}}^2}, \tag{A2}$$

with a specific correlation length σ_{CL} that is normally distributed and empirically determined as $\sigma_{\mathrm{CL}} \sim \mathcal{N}(1, 9)$. The corresponding convolution ensures smoothness around the perturbed cylindrical shape between various neighboring points, resulting in more realistic edges. This convolution is specified by

$$(f * g_N)[n] = \sum_{m=0}^{N-1} f[m] g_N[n-m]. \tag{A3}$$

After adding roughness, the top and bottom structures are connected in the z-dimension.

A.2 Line spaces

For line edge roughness (LER), the Thorsos method [51] is applied as described in [40]. This is a power spectral density-based method where the autocorrelation is approximated as

$$R(x) = \sigma^2 e^{-(|x|/l_c)^{2\alpha}}. \tag{A4}$$

For the correlation length (l_c), roughness factor (α) and standard deviation (σ), we have used 16.8 nm, 0.5 and 0.77 nm, respectively. These values are closely matching with the available experimental data.

Appendix B: Implementation details

The geometries were created with Python and Numpy and stored as text format in *.tri files. Nebula [39] was used for SEM simulations. The simulations were performed at a GPU cluster with GPU K80 video cards (memory of 24 GB). The dataset was constructed with Python and Pandas. The electron yield numbers of SEM images were stored in 8-bit unsigned integer format. The depth values in the maps were stored in 32-bit float format. The depth estimation network was implemented in Pytorch 1.3.0 with Python 3.6. Tensorboard 2.0.0 was used for visualization of the validation metrics. The neural network was trained at a GPU cluster with one V100 video card (memory of 32 GB). Post-processing was done with Python, NumPy, SciPy and OpenCV. For visualization of the data MATLAB, Matplotlib, Visio and Excel were used.

References

1. Bunday, B., Solecky, E., Vaid, A., Bello, A.F., Dai, X.: Metrology capabilities and needs for 7 nm and 5 nm logic nodes. In: Metrology, Inspection, and Process Control for Microlithography XXXI, vol. 10145, p. 101450 (2017). https://doi.org/10.1117/12.2260870
2. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), pp. 131–140 (2001)
3. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. Int. J. Robot. Res. **34**, 705–724 (2013)
4. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction (2019)
5. Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.-Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2010, pp. 275–282 (2010)
6. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, pp. 2366–2374 (2014)
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
8. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986). https://doi.org/10.1103/PhysRevLett.56.930
9. Gauvin, R.: Review of transmission electron microscopy for the characterization of materials. In: Materials Characterization and Optical Probe Techniques: A Critical Review, vol. 10291, pp. 196–225 (1997). https://doi.org/10.1117/12.279840
10. den Boef, A.J.: Optical metrology of semiconductor wafers in lithography. In: International Conference on Optics in Precision Engineering and Nanotechnology (icOPEN2013), vol. 8769, pp. 57–65 (2013). https://doi.org/10.1117/12.2021169
11. Verduin, T., Lokhorst, S.R., Hagen, C.W.: GPU accelerated Monte-Carlo simulation of SEM images for metrology. In: Metrology, Inspection, and Process Control for Microlithography XXX, vol. 9778, pp. 122–135 (2016). https://doi.org/10.1117/12.2219160
12. Csurka, G.: A comprehensive survey on domain adaptation for visual applications. In: Csurka, G. (ed.) Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and

Pattern Recognition, pp. 1–35. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58347-1_1
13. Roy, S., Meunier, J., Marian, A., Vidal, F., Brunette, I., Costantino, S.: Automatic 3D reconstruction of quasi-planar stereo scanning electron microscopy (SEM) images. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4361–4364. IEEE (2012)
14. Henao-Londoño, J.C., Riaño-Rojas, J.C., Gómez-Mendoza, J.B., Restrepo-Parra, E.: 3D stereo reconstruction of SEM images. Modern Appl. Sci. 12(12), 57 (2018). https://doi.org/10.5539/mas.v12n12p57
15. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)
16. Ito, W., Bunday, B., Harada, S., Cordes, A., Murakawa, T., Arceo, A., Yoshikawa, M., Hara, T., Arai, T., Shida, S., Yamagata, M., Matsumoto, J., Nakamura, T.: Novel three dimensional (3D) CD-SEM profile measurements. In: Metrology, Inspection, and Process Control for Microlithography XXVIII, vol. 9050, pp. 85–93 (2014). https://doi.org/10.1117/12.2047374
17. Vladár, A.E., Villarrubia, J.S., Chawla, J., Ming, B., Kline, J.R., List, S., Postek, M.T.: 10 nm three-dimensional CD-SEM metrology. In: Metrology, Inspection, and Process Control for Microlithography XXVIII, vol. 9050 (April 2014), p. 90500 (2014). https://doi.org/10.1117/12.2045977
18. Arat, K., Bolten, J., Zonnevylle, A., Kruit, P., Hagen, C.: Estimating step heights from top–down SEM images. Microsc. Microanal. **25**(4), 903–911 (2019). https://doi.org/10.1017/S143192761900062X
19. Sun, W., Zhao, P., Goto, Y., Yamamoto, T., Ninomiya, T.: Accuracy improvement of 3D-profiling for HAR features using deep learning. In: Metrology, Inspection, and Process Control for Microlithography XXXIV, vol. 11325, pp. 105–112 (2020). https://doi.org/10.1117/12.2551458
20. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision, pp. 740–756 (2016). Springer
21. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left–right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
22. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
23. Guo, X., Li, H., Yi, S., Ren, J., Wang, X.: Learning monocular depth by distilling cross-domain stereo networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 484–500 (2018)
24. Nath Kundu, J., Krishna Uppala, P., Pahuja, A., Venkatesh Babu, R.: Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2656–2665 (2018)
25. Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., Lin, L.: Single view stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 155–163 (2018)
26. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
27. Nathan Silberman, P.K. Derek Hoiem, Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012)
28. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2002–2011 (2018)
29. Chen, R., Mahmood, F., Yuille, A., Durr, N.J.: Rethinking monocular depth estimation with adversarial training. arXiv preprint arXiv:1808.07528 (2018)
30. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
31. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2018)
32. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer (2015)
33. Chaudhary, N., Savari, S.A., Yeddulapalli, S.S.: Line roughness estimation and Poisson denoising in scanning electron microscope images using deep learning. J. Micro/Nanolithogr. MEMS MOEMS **18**(02), 1 (2019). https://doi.org/10.1117/1.jmm.18.2.024001
34. Quan, T.M., Hildebrand, D.G.C., Lee, K., Thomas, L.A., Kuan, A.T., Lee, W.A., Jeong, W.: Removing imaging artifacts in electron microscopy using an asymmetrically cyclic adversarial network without paired training data. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3804–3813 (2019)
35. Krull, A., Buchholz, T.-O., Jug, F.: Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2129–2137 (2019)
36. Yan, C., Li, Z., Zhang, Y., Liu, Y., Ji, X., Zhang, Y.: Depth image denoising using nuclear norm and learning graph model. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **16**(4), 1–17 (2020)
37. Yan, C., Gong, B., Wei, Y., Gao, Y.: Deep multi-view enhancement hashing for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1445–1451 (2020)
38. Lin, L., Dou, Q., Jin, Y.-M., Zhou, G.-Q., Tang, Y.-Q., Chen, W.-L., Su, B.-A., Liu, F., Tao, C.-J., Jiang, N., et al.: Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology **291**(3), 677–686 (2019)
39. van Kessel, L., Hagen, C.W.: Nebula: Monte Carlo simulator of electron–matter interaction. SoftwareX **12**, 100605 (2020). https://doi.org/10.1016/j.softx.2020.100605
40. Mack, C.A.: Generating random rough edges, surfaces, and volumes. Appl. Opt. **52**(7), 1472–1480 (2013). https://doi.org/10.1364/AO.52.001472
41. Lee, J.H., Han, M.-K., Ko, D.W., Suh, I.H.: From big to small: multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR arXiv:1409.1556 (2015)
44. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
45. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in Neural Information Processing Systems, vol. 3(06) (2014)
46. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: On the effectiveness of least squares generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 2947–2960 (2018)
47. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of

the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)

48. Kingma, D.P., Ba, J.A.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
49. Pisarenco, M., Maubach, J.M.L., Setija, I.D., Mattheij, R.M.M.: Efficient solution of Maxwell's equations for geometries with repeating patterns by an exchange of discretization directions in the aperiodic Fourier modal method. J. Comput. Phys. **231**(24), 8209–8228 (2012). https://doi.org/10.1016/j.jcp.2012.07.049
50. van Beurden, M.C.: A spectral volume integral equation method for arbitrary bi-periodic gratings with explicit Fourier factorization. Prog. Electromagn. Res. B **36**, 133–149 (2012). https://doi.org/10.2528/PIERB11100307
51. Thorsos, E.I.: The validity of the Kirchhoff approximation for rough surface scattering using a Gaussian roughness spectrum (2004)

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